Improvement in automatic food region extraction based on saliency detection

Takuya Futagami and Noboru Hayasaka

Department of Engineering Informatics, Osaka Electro-Communication University, Neyagawa, Osaka, JAPAN

ABSTRACT
In this paper, we propose a method for extracting pixel-wise food regions on the basis of saliency detection using the multi-scale information network (MSI-Net), which includes convolutional layers at different dilated rates, and GrabCut, which can revise food regions on the basis of graph theory. Our comparative experiment, which used 241 actual food images, clarified that the proposed method significantly increased the F-measure, which was used as a comprehensive metric of the food-extraction accuracy, by 3.76% or more compared with conventional methods using saliency detection. The proposed method tended to preserve the contour structure of the food regions better than the conventional methods. In addition, the F-measure was significantly increased by 6.72% compared with the low-cost SegNet, which was trained with a publically available dataset. However, the experiment suggested that the proposed method can be further improved by revising its algorithm for determining food and background candidates. Further discussion and implications are provided herein.

ARTICLE HISTORY
Received 31 August 2021
Revised 21 February 2022
Accepted 13 March 2022

KEYWORDS
Ood image; Food region extraction; segmentation; GrabCut; Food record

Introduction

Chronic diseases, examples of which include diabetes and obesity, are major causes of worldwide mortality and morbidity and have reached epidemic proportions.\textsuperscript{[1]} To control and prevent them, food behavior and patterns should be improved, and the World Health Organization (WHO) has a coordinated strategy for healthy eating.\textsuperscript{[2]} Thus, the assessment of food intake has captured international attention as a means of promoting public health. One accurate and simple way of assessing food intake is a food diary, in which all food and beverages are manually recorded over several days. Because a food diary is traditionally done by pen and paper, issues with the burden of having to accurate recording of food and finding the motivation to do so have been a challenge.\textsuperscript{[3]} Hence, automatic assessment via food images, the main subject of which is food, has been developed to reduce user burden in recording a food diary.\textsuperscript{[4]} Image-based assessment can broaden the applicability of the food diary because it is available with smartphone applications such as FoodLog\textsuperscript{[5]} and Mobile Food Record (MFR),\textsuperscript{[6]} which support automatic energy and nutritional intake estimation. However, the accuracy of image-based assessment needs to be improved, as the energy intake estimation of the MFR had a mean error rate of 38% for 347 analyzed food occasions.\textsuperscript{[7]}

Image-based assessment extracts regions that include food items (hereafter called “food regions”) from regions that include no food items (hereafter called “background regions”) at the pixel level.\textsuperscript{[8]} Subsequently, the extracted regions are classified as different food categories.\textsuperscript{[9]} On the basis of the number of the pixels in each food region, the energy and nutritional values of the food intake are estimated. Food region extraction is mainly based on deep neural networks (DNNs),\textsuperscript{[10]} which have been actively applied and proposed owing to the advent of the fully convolutional network.
(FCN), SegNet, and saliency detection, which assumes that food regions tend to attract the human eye. The use of saliency detection is based on the characteristics of food images, in which foods tend to be salient objects because of the weak contrast and simple texture patterns of the background such as plates and dishes. Although the accuracy of food extraction has not been compared between DNNs and saliency detection in the literature, DNNs are expected to improve the accuracy due to their reliable and outstanding performance. However, a large amount of pixel-wise annotated food images is required to train DNNs due to there being diverse food and background types. Because the number of annotated food-image datasets that have been made publicly available is insufficient, such datasets need to be collected and prepared. This requirement considerably increases the introduction cost of food extraction based on DNNs. However, saliency detection requires no annotated food-image dataset, thereby decreasing the introduction cost. In the literature, eye tracking using glasses-type wearable devices or image recognition techniques is used for saliency detection in food extraction. Because wearable devices are not used in the existing image-based assessment, image recognition is used to measure salient regions in this paper.

estimated salient regions by using the Center Surround Extremas (CenSurE) detector, which obtains local extrema of intensity, to extract food regions (hereafter called the “CenSurE method”). The CenSurE method is based on the assumption that the local extrema are found in salient regions. In our previous work (Futagami et al., 2020), the Multi-Scale Information Network (MSI-Net), which is a DNN including convolutional layers at different dilated rates, was used to detect salient regions (hereafter called the “MSI method”). Although an annotated saliency dataset is required to train MSI-Net, the introduction cost can be decreased owing to the advent of large-scale and publically available saliency datasets such as SALICON. An experiment demonstrated that the MSI method increased the accuracy compared with the CenSurE method by 2.0% (Futagami et al., 2020). However, our careful analysis found that the MSI method can be further improved by overcoming its critical shortcoming outlined later.

Hence, this paper aims to propose a method that can overcome the shortcoming of the MSI method. The effectiveness of the proposed method is evaluated in a comparative experiment, where the accuracy is compared with that of the conventional methods. In order to thoroughly clarify the effectiveness of the proposed method, our experiment evaluates food extraction based on the DNN, the introduction cost of which is decreased by using a publically available dataset. Since DNN-based food extraction and saliency detection have not been compared in the literature, our experiment may contribute to providing more useful and acceptable findings. The rest of this paper is structured as follows. In Section 2, the conventional methods are outlined and discussed, and in Section 3 the proposed method is described in detail. In Section 4, the comparative experiment is conducted to confirm and discuss the effectiveness of the proposed method. The present study is summarized and concluded in Section 5.

Conventional methods

This section outlines the conventional methods (i.e., CenSurE and MSI methods), which extract food regions from food images such as in Figure 1. Sections 2.1 and 2.2 verview the CenSurE and MSI methods, respectively. Section 2.3 provides an important analysis for improvement.

CenSurE method

The algorithm of the CenSurE method mainly consists of two steps. In the first step, local extrema are detected by applying the CenSurE detector to food images. Subsequently, an algorithm is applied to create a convex hull, which is the smallest polygon including all the detected local extrema. The inner and outer areas of the convex hull are determined as food and background candidates, respectively.
In the second step, GrabCut,\cite{24} which revises both food and background regions on the basis of graph theory, is applied. GrabCut creates a graph, the nodes of which represent both pixels and endpoints of food and background regions. The edges between two adjacent pixels have weights based on their color similarity (hereafter called “local similarity”). In addition, the edges between endpoints and each pixel have weights based on the probability that the pixels belong to either region (hereafter called “global similarity”). The initial food and background regions of the graph are determined on the basis of the food and background candidates. GrabCut revises food and background regions by applying a min-cut/max-flow algorithm\cite{25} so as to increase the local and global similarity. Background candidates are not changed to food regions by GrabCut (this setting hereafter called “hard-labelled”), whereas food candidates are allowed to be changed to background regions (this setting hereafter called “soft-labelled”). Figure 2 presents the process of the CenSurE method for the input image in Figure 2(a). The red points and green lines of Figure 2(b) show the detected local extrema and their convex hull, respectively. The green and pink pixels of Figure 2(c) are the inner and outer areas of the convex hull, which are determined as the food and background candidates, respectively. Figure 2(d) depicts the food regions extracted by GrabCut, initialized in Figure 2(c).

**MSI method**

The algorithm of the MSI method also mainly consists of two steps. In the first step, MSI-Net, which uses a trained VGG-16 encoder\cite{26} to extract saliency features, is applied to create a saliency map, which quantitatively evaluates how much each pixel attracts the human eye (hereafter called the “saliency value”). The reason for using MSI-Net stems from the fact that it could increase the accuracy compared with other forms of saliency detection in,\cite{20} which aimed to determine the most effective saliency detection in the food extraction task.

**Figure 1.** Examples of food images.
In the second step, all the pixels are classified as food and background regions by thresholding at a predetermined saliency threshold level \( t' \) as:

\[
  r_i = \begin{cases} \text{Food region}(s_i > t'), & \text{Food region}(s_i > t') \\ \text{Background region}(s_i \leq t'), & \text{Background region}(s_i \leq t') \end{cases}
\]  

(1)

where \( r_i, s_i \) denote the type of region and the saliency value at pixel position \( i \), respectively. Figure 3 presents process of the MSI method for the input image in Figure 2(a). The saliency map given by MSI-Net is depicted in Figure 3(a), where the strength of the white color represents the intensity of the saliency value. The extracted food regions are shown in Figure 3(b), which were obtained by applying Equation (1) to the saliency map.

**Analysis for improvement**

Figure 4 depicts examples of extracted food regions that were obtained by applying the MSI method to the food images in Figure 1. The figures indicate that the MSI method tended to erroneously determine the actual background as a food region. This erroneous determination is due to the fact that the contour structure of the food items could not be preserved because the MSI method does not increase the local similarity as GrabCut does. On the basis of this analysis, the algorithm of the proposed method needs to be revised in order to preserve the contour structure so as to improve the accuracy.
Proposed method

This section describes the algorithm of the proposed method, which is designed to overcome the shortcoming of the MSI method discussed in Section 2.3. Figure 5 depicts the flowchart of the proposed method, which mainly consists of two steps. Each step is outlined in the following subsections.
Specifying product and background candidates

In the first step, food and background candidates are created under the assumption that actual food regions tend to have higher saliency values. The proposed method also uses MSI-Net since the saliency map provided by this model increased the prediction accuracy for food regions (but not equal to the extraction accuracy) by 2.0% compared with that by other types of state-of-the-art saliency detection using DNNs in.\textsuperscript{[20]} In addition, our preliminary comparative experiment found that MSI-Net could markedly increase the food extraction accuracy of the proposed method compared with other saliency models tested in that work. Each pixel of the saliency map given by MSI-Net ($s_i$) is simply thresholded at level $t'$ as:

$$c_i = \begin{cases} 
\text{Food candidate}(& s_i > t'), \\
\text{Background candidate}(& s_i \leq t')
\end{cases}$$

(2)

where $c_i$ returns the type of the candidates at pixel position $i$. Subsequently, the food candidates are further classified into more and less certain food candidates on the basis of the aforementioned assumption by maximizing inter-class variance like the Otsu algorithm\textsuperscript{[27]} as:

$$c_i = \begin{cases} 
\text{More certain food candidate}(& s_i > t^*), \\
\text{Less certain food candidate}(& s_i \leq t^*)
\end{cases}$$

(3)

$$t^* = \arg\max_{t} N_{h,t}N_{l,t} \left( \mu_{h,t} - \mu_{l,t} \right)^2,$$

(4)

where $\mu_{h,t}$, $\mu_{l,t}$ denote the mean saliency value of pixels, the saliency values of which are higher than $t$ or not, respectively. In addition, $N_{h,t}$, $N_{l,t}$ also denote pixel numbers.

Applying Grabcut

In the second step, the food regions are extracted by applying GrabCut, which initialized by the candidates determined in the first step. The reason for using GrabCut stems from the fact that it is expected to be effective in the preservation of contour structures by increasing the local similarity as
discussed in Section 2.1. In fact, GrabCut has been used in contour extraction tasks in several works. More certain food candidates are set to be hard-labeled on the basis of the assumption, whereas less certain ones and background candidates are set to be soft-labeled.

Figure 6 presents the process of the proposed method for the input image in Figure 2(a). Figure 6(a) describes food and background candidates based on the saliency map in Figure 3(a). The green and blue pixels represent the more and less certain food candidates, respectively, whereas the pink pixels represent the background candidates. Figure 6(b) depicts the food regions extracted by applying GrabCut, which was initialized by the candidates. In addition, Figure 7 also demonstrates the food
regions extracted by the proposed method. These figures suggest that the proposed method tended to overcome the shortcoming of the MSI method since it could preserve the contour structure better compared with the MSI method (see Figures 3(b) and 7).

**Experiment**

This section quantitatively evaluates the accuracy of the proposed method compared with other methods. Sections 4.1 and 4.2 provide experimental conditions and results, respectively. Section 4.3 discusses the effectiveness of the proposed method on the basis of a detailed analysis.

**Experimental condition**

The experiment used 241 food images, which were captured to record food intake and were also used in (Futagami et al., 2020). All the images were resized uniformly so that longer sides of the width or height were set to 320 pixels. Ground truth images, which correctly distinguished food and background regions, were manually prepared. Figure 8 depicts a food image and its ground truth image, the white and black pixels of which represent the food and background regions, respectively. As metrics of the accuracy, precision, recall, and F-measure were used. They are defined as:

\[
\text{Precision} = \frac{1}{M} \sum_{m=0}^{M-1} \frac{\text{TP}_m}{(\text{TP}_m + \text{FP}_m)}, \tag{5}
\]

\[
\text{Recall} = \frac{1}{M} \sum_{m=0}^{M-1} \frac{\text{TP}_m}{(\text{TP}_m + \text{FN}_m)}, \tag{6}
\]

\[
\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{7}
\]

where \(M\) denotes the total number of images included in the evaluation dataset. In addition, \(\text{TP}_m\) denotes the number of pixels correctly determined as a food region for the \(m\)-th image. Furthermore, \(\text{FP}_m, \text{FN}_m\) also denote the number of pixels erroneously determined as food and background regions, respectively. Because the precision and recall are generally in a trade-off, the F-measure, which is the harmonic mean of them, was used as a comprehensive evaluation metric.

![Figure 8. Food image and its ground truth image.](image)
As methods to compare against the proposed method, the CenSurE method, MSI method, low-cost SegNet, and interactive GrabCut were. The CenSurE method (Section 2.1), MSI method (Section 2.2), and proposed method (Section 3) are based on saliency detection. MSI-Net was trained by SALICON, which has 10,000 images and their ground truths. In addition, other parameters of these methods were determined experimentally.

The low-cost SegNet is based on the DNN, the introduction cost of which was decreased by using a publically available dataset as discussed in Section 1. The reason for using SegNet\textsuperscript{[12]} is based on the fact that it could increase the accuracy higher compared with FCN\textsuperscript{[11]} ENet,\textsuperscript{[30]} EdaNet,\textsuperscript{[31]} and ResSeg\textsuperscript{[32]} in.\textsuperscript{[33]} In addition, SegNet can increase the accuracy for various images and has been extensively applied in several works.\textsuperscript{[34–36]} The low-cost SegNet was trained by using 1017 annotated food images included in UNIMIB2016,\textsuperscript{[37]} which is the only publically available annotated food image dataset.

Interactive GrabCut is initialized by using bounding boxes, which are manually prepared to include each food item. This method was used to confirm the best performance in terms food extraction done using GrabCut such as with the CenSurE method and proposed method. The bounding boxes were created on the basis of the food regions in the ground truth images. Figure 9 demonstrates the process of interactive GrabCut for the input image in Figure 2(a). The green rectangles of Figure 9(a) are the bounding boxes of the food items. The extracted food regions are shown in Figure 9(b), which were obtained by applying GrabCut, which was initialized by the candidates as shown in Figure 9(a).

**Experimental result**

Table 1 summarizes the accuracy provided by each method. The effectiveness of the proposed method is discussed on the basis of this table and its analysis using a Wilcoxon signed-rank test, which is categorized as a non-parametric paired comparison test that detects significant differences at a significance level of 5%.

Compared with the CenSurE and MSI methods, the proposed method significantly increased the F-measure by 3.76% or more. In addition, the precision was non-significantly increased by 1.21% or more. Although the recall was decreased by 1.31% compared with the MSI method, no significant

| Metrics     | CenSurE | MSI  | SegNet | IG     | Proposed |
|-------------|---------|------|--------|--------|----------|
| Precision [%]| 66.51   | 61.22| 68.78  | 85.58  | 67.72    |
| Recall [%]  | 75.89   | 86.55| 68.73  | 88.82  | 85.24    |
| F-measure [%]| 70.89  | 71.71| 68.75  | 87.16  | 75.47    |
difference was found. Compared with the low-cost SegNet, the proposed method significantly increased the F-measure and recall by 6.72% and 16.51%, respectively. Although the precision was decreased by 1.06%, no significant difference was found. These comparisons indicated that the proposed method was effective since the F-measure was significantly increased by 3.76% or more compared with the CenSurE method, MSI method, and low-cost SegNet without significantly decreasing either precision nor recall.

Compared with interactive GrabCut, the proposed method significantly decreased the precision, recall, and F-measure by 17.86%, 3.58%, and 11.69%, respectively. Although the proposed method is effective in that it does not require any manual operation to initialize GrabCut, the comparison suggests that the accuracy can be further improved by revising the algorithm to determine the food and background candidates.

Figure 10 depicts examples of the food regions extracted by each method. The first and second examples demonstrate that the proposed method increased the accuracy compared with the MSI method due to its effect on the preservation of the contour structure as discussed in Section 3.2. However, the accuracy was decreased in the third example because the pixels in the actual food regions tended to be erroneously determined as background regions. This erroneous determination is due to the fact that the proposed method misclassified pixels in the actual food regions as background candidates. Due to this misclassification, GrabCut erroneously determined the pixels around the misclassified ones as background regions to increase their local similarity. This is the main reason why the recall of the proposed method decreased compared with that of the MSI method in Table 1. Compared with the low-cost SegNet and interactive GrabCut, the proposed method could not extract each food item separately, whereas it could detect all the food items.

Discussion

This section presents a more detailed discussion to clarify and generalize the effectiveness of the proposed method thoroughly. The food images consisted of those that had a single food item and those that had multiple items. Examples of single and multiple food images are provided in Figures 11(a) and (b), respectively. They suggest that the food regions of the single food images tended to be located at the image-center area compared with the multiple food images. This trend can be confirmed in Figure 12, which compares the probability that food regions are located in each pixel position between 91 single images and 150 multiple food images included in the evaluation dataset. The probability was obtained from the resized ground truth images, the resolution of which was uniformly resized to a square size, as:
\[
p_i = \frac{1}{M} \sum_{m=0}^{M-1} g_i^m,
\]

where \( g_i^m \) denotes the resized ground truth image, which takes values of 0 or 1 for the background or food regions, respectively. Because the trend of the probability is considerably different between single and multiple food images, the accuracy for each type of the image was checked.

Tables 2(a) and (b) depict the accuracy for single and multiple food images, respectively. Compared with the CenSurE method, MSI method, and low-cost SegNet, the proposed method significantly increased the F-measure by 2.63% or more and 2.65% or more for the single and multiple food images, respectively. For single food images, the recall was significantly increased by 2.11% or more, whereas the precision was non-significantly decreased compared with the CenSurE method. However, for multiple food images, the precision and recall were significantly decreased compared with the low-cost SegNet and the MSI method, respectively. Thus, although...
the proposed method was effective for both types of food images, it is suggested that it may be less suitable for multiple food images compared with single food images. This suggestion is supported by the fact that MSI-Net, which the proposed method uses, was learned to give higher saliency values at image-center areas (hereafter called the “center prior”) as discussed in.\textsuperscript{[21]} The center prior was not effective for the multiple food images because they tended to have food regions at boundary areas as also observed in (Futagami et al., 2020).\textsuperscript{[20]} Hence, the accuracy for multiple food images may be further improved by excluding this prior.

The proposed method significantly increased the extraction accuracy compared with the CenSurE and MSI methods. However, we should note that it requires a greater number of the parameters to be adjusted compared with the conventional methods because of the combination of different types of techniques. This requirement increases the number of the images to estimate the optimal parameters for the proposed method. To build a more robust food-extraction method on the basis of a smaller number of images, it may be better to decrease the number of parameters.

### Conclusion

This paper proposed a method for food extraction using MSI-Net, which includes convolutional layers at different dilated rates, and GrabCut, which revises food regions on the basis of graph theory. The algorithm of the proposed method is structured to overcome the shortcoming of the MSI method. Our experiment, which used 241 food images, demonstrated that the proposed method significantly increased the F-measure, which was used as a comprehensive metric of the food-extraction accuracy, by 3.76\% or more compared with the conventional methods. In addition, the F-measure was significantly increased by 6.72\% compared with the low-cost SegNet, which was trained by a publically available food-image dataset. Our detailed discussion also clarified that the F-measure was significantly increased by 2.63\% or more for single and multiple food images compared with the conventional methods and the low-cost SegNet. However, the proposed method was less suitable for multiple food images since the precision and recall were significantly decreased compared with the low-cost SegNet and the conventional methods, respectively. These results may be due to the fact that MSI-Net gives higher saliency values at image-center areas. In addition, the proposed method was not effective in any metrics compared with the interactive GrabCut, which requires manual operation to initialize GrabCut. Although the proposed method can extract food regions automatically, it was found that it could be further improved by revising the algorithm.

### Disclosure statement

No potential conflict of interest was reported by the author(s).
References

[1] Lega, I. C.; Lipscombe, L. L. Diabetes, Obesity, and Cancer–pathophysiology and Clinical Implications. *Endocrine Rev.* 2020, 41(3), 33–52. DOI: 10.1210/endrev/bnz014.

[2] World Health Organization. *Global Action Plan for the Prevention and Control of Noncommunicable Diseases 2013–2020*. World Health Organization, 2013.

[3] Doulah, A.; McCorry, M. A.; Higgins, J. A.; Sazonov, E. A Systematic Review of Technology-driven Methodologies for Estimation of Energy Intake. *IEEE Access.* 2019, 7(1), 49653–49668. DOI: 10.1109/ACCESS.2019.2910308.

[4] Bruno, V., and Silva Resende, C. J. 2017. A Survey on Automated Food Monitoring and Dietary Management Systems. *J. Health Med. Inform.* 8(3), 1–7.

[5] Aizawa, K. Image Recognition-based Tool for Food Recording and Analysis: FoodLog. In *Connected Health in Smart Cities*; El Saddik, A.; Hossain, M. S.; Kantarci, B.; Eds.; Switzerland: Springer Cham, 2019; pp 1–9.

[6] Ahmad, Z.; Bosch, M.; Khanna, N.; Kerr, D. A.; Boushey, C. J.; Zhu, F., and Delp, E. J. (2016). A Mobile Food Record for Integrated Dietary Assessment. Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management, Netherlands, 53–62.

[7] Fang, S.; Shao, Z.; Kerr, D. A.; Boushey, C. J.; Zhu, F. An End-to-end Image-based Automatic Food Energy Estimation Technique Based on Learned Energy Distribution Images: Protocol and Methodology. *Nutrients.* 2019, 11(4), 877. DOI: 10.3390/nu11040877.

[8] Hassannejad, H.; Matrella, G.; Ciampolini, P.; De Munari, I.; Mordonini, M.; Cagnoni, S. Automatic Diet Monitoring: A Review of Computer Vision and Wearable Sensor-based Methods. *Int. J. Food Sci. Nutr.* 2017, 68(6), 656–670. DOI: 10.1080/09637486.2017.1283683.

[9] Höchsmann, C.; Martin, C. K. Review of the Validity and Feasibility of Image-assisted Methods for Dietary Assessment. *Int. J. Obes.* 2020, 44(12), 2358–2371. DOI: 10.1038/s41366-020-00693-2.

[10] Dhruv, P., and Naskar, S. (2020). 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, India, 367–381.

[11] Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 2017, 39(4), 640–651. DOI: 10.1109/TPAMI.2016.2572683.

[12] Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder–decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 2017, 39(12), 2481–2495. DOI: 10.1109/TPAMI.2016.2644615.

[13] Ullah, I.; Jian, M.; Hussain, S.; Guo, J.; Yu, H.; Wang, X.; Yin, Y. A Brief Survey of Visual Salience Detection. *Multimedia Tools Appl.* 2020, 79(45), 34605–34645. DOI: 10.1007/s11042-020-08849-y.

[14] Chen, H. C.; Jian, W.; Sun, X.; Li, Z.; Li, Y.; Fernstrom, J. D.; Sun, M.; Baranowski, T.; Sun, M. Saliency-aware Food Image Segmentation for Personal Dietary Assessment Using a Wearable Computer. *Meas. Sci. Technol.* 2015, 26(2), 025702. DOI: 10.1088/0957-0233/26/2/025702.

[15] Wang, Y.; Zhu, F.; Boushey, C. J., and Delp, E. J. (2017). Weakly Supervised Food Image Segmentation Using Class Activation Maps. Proceedings of the 2017 IEEE International Conference on Image Processing, China, 1277–1281.

[16] Ege, T.; Shimoda, W., and Yanai, K. (2019). A New Large-scale Food Image Segmentation Dataset and Its Application to Food Calorie Estimation Based on Grains of Rice. Proceedings of the 5th International Workshop on Multimedia Assisted Dietary Management, France, 82–87.

[17] Futagami, T.; Kitada, A., and Hayasaka, H. (2020). Food Region Extraction by Applying Salience Detection Model. Proceedings of the 64th Annual Conference of the Institute of Systems, Control and Information Engineers, Web conference, 57–62 (in Japanese).

[18] Sugiyama, H.; Morikawa, C.; Aizawa, K. Segmentation of Food Images by Local Extrema and GrabCut. *J. Inst. Image Inform. Television Eng.* 2012, 66(5), J179–J181. in Japanese. DOI: 10.3169/ijie.66.J179.

[19] Agrawal, M.; Konolige, K., and Blas, M. R. (2008). Censure: Center Surround Extremas for Realtime Feature Detection and Matching. Proceedings of the 10th European Conference on Computer Vision, France, 102–115 (in Japanese).

[20] Futagami, T.; Hayasaka, N., and Onoye, T. (2020). Performance Comparison of Salience Detection Methods for Food Region Extraction. Proceedings of the 4th International Conference on Graphics and Signal Processing, Japan, 1–4.

[21] Kroner, A.; Senden, M.; Driessens, K.; Goebel, R. Contextual Encoder–decoder Network for Visual Saliency Prediction. *Neural Netw.* 2020, 129, 261–270. Doi:10.1016/j.neunet.2020.05.004.

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

[23] Sklansky, J. Measuring Concavity on a Rectangular Mosaic. *IEEE Trans. Comput.* 1972, 21(12), 1355–1364. DOI: 10.1109/T-C.1972.223507.
[24] Rother, C.; Kolmogorov, V.; Blas, A. GrabCut: Interactive Foreground Extraction Using Iterated Graph Cuts. ACM Trans. Graph. 2004, 23(3), 309–314. DOI: 10.1145/1015706.1015720.

[25] BoyKov, Y.; Kolmogorov, V. An Experimental Comparison of Min-cut/max-flow Algorithms for Energy Minimization in Vision. IEEE Trans. Pattern Anal. Mach. Intell. 2004, 26(9), 1124–1137. DOI: 10.1109/TPAMI.2004.60.

[26] Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-scale Image Recognition. arXiv preprint. 2014, arXiv, 1409.1556.

[27] Goh, T. Y.; Basah, S. N.; Yazid, H.; Safar, M. J. A.; Saad, F. S. A. Performance Analysis of Image Thresholding: Otsu Technique. Measurement. 2018, 114(9), 298–307. DOI: 10.1016/j.measurement.2017.09.052.

[28] Jaisakthi, S. M.; Mirunalini, P.; Aravindan, C. Automated Skin Lesion Segmentation of Dermoscopic Images Using GrabCut and K-means Algorithms. IET Comput. Vis. 2018, 12(8), 1088–1095. DOI: 10.1049/iet-cvi.2018.5289.

[29] Silva, R. H. L.; Machado, A. M. C. Automatic Measurement of Pressure Ulcers Using Support Vector Machines and GrabCut. Comput. Method Program Biomed. 2021, 200, 105867. Doi:10.1016/j.cmpb.2020.105867.

[30] Paszke, A.; Chaurasia, A.; Kim, S.; Culurciello, E. Enet: A Deep Neural Network Architecture for Real-time Semantic Segmentation. arXiv preprint. 2016, arXiv, 1606.02147.

[31] Lo, S. Y.; Hang, H. M.; Chan, S. W.; Lin, J. J. Efficient Dense Modules of Asymmetric Convolution for Real-time Semantic Segmentation. arXiv preprint. 2018, arXiv, 1809.06323.

[32] Pohlen, T.; Hermans, A.; Mathias, M., and Leibe, B. (2017). Full-resolution Residual Networks for Semantic Segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, United States, 4151–4160.

[33] Aslan, S.; Ciocca, G.; Mazzini, D.; Schettini, R. Benchmarking Algorithms for Food Localization and Semantic Segmentation. Int. J. Mach. Learn. Cybern. 2020, 11(12), 2827–2847. DOI: 10.1007/s13042-020-01153-z.

[34] Ciocca, G.; Mazzini, D., and Schettini, R. (2019). Evaluating CNN-based Semantic Food Segmentation across Illuminants. Proceedings of the 7th International Workshop on Computational Color Imaging, Japan, 247–259.

[35] Pinzon-Arenas, J. O.; Jimenez-Moreno, R.; Pachon-Suescun, C. G. ResSeg: Residual Encoder-decoder Convolutional Neural Network for Food Segmentation. Int. J. Electr. Comput. Eng. 2020, 10(2), 1017–1026.

[36] Shimoda, W., and Yanai, K. (2015). CNN-based Food Image Segmentation without Pixel-wise Annotation. Proceedings of the 20th International Conference on Image Analysis and Processing, Italy, 449–457.

[37] Ciocca, G.; Napoletano, P.; Schettini, R. Food Recognition: A New Dataset, Experiments, and Results. IEEE J. Biomed. Health Inform. 2016, 21(3), 588–598. DOI: 10.1109/JBHI.2016.2636441.