IEICE TRANS. INF. & SYST., VOL.E102–D, NO.7 JULY 2019

PAPER Special Section on Multimedia for Cooking and Eating Activities

Simultaneous Estimation of Dish Locations and Calories with Multi-Task Learning

Takumi EGE[†a)], ***Nonmember and*** **Keiji YANAI**[†b)], ***Member***

SUMMARY In recent years, a rise in healthy eating has led to various food management applications which have image recognition function to record everyday meals automatically. However, most of the image recognition functions in the existing applications are not directly useful for multiple-dish food photos and cannot automatically estimate food calories. Meanwhile, methodologies on image recognition have advanced greatly because of the advent of Convolutional Neural Network (CNN). CNN has improved accuracies of various kinds of image recognition tasks such as classification and object detection. Therefore, we propose CNN-based food calorie estimation for multiple-dish food photos. Our method estimates dish locations and food calories simultaneously by multi-task learning of food dish detection and food calorie estimation with a single CNN. It is expected to achieve high speed and small network size by simultaneous estimation in a single network. Because currently there is no dataset of multiple-dish food photos annotated with both bounding boxes and food calories, in this work we use two types of datasets alternately for training a single CNN. For the two types of datasets, we use multiple-dish food photos annotated with bounding boxes and single-dish food photos with food calories. Our results showed that our multi-task method achieved higher accuracy, higher speed and smaller network size than a sequential model of food detection and food calorie estimation.

key words: *food calorie estimation, food dish detection, multi-task learning*

1. Introduction

In recent years, owing to the rise in healthy eating, various food photo recognition applications for recording meals have been released. However, some of them need human assistance for calorie estimation such as manual input and the help of a nutrition expert. Additionally, even if it is automatic, food categories are often limited, or images from multiple viewpoints are required. Recently, some applications have begun to estimate food categories from food photos automatically by image recognition. However, in the case of multiple-dish food photos, as shown in Fig. 1, users are required to take pictures one by one for each dish or to crop single dishes manually from images, which takes time and labor.

Meanwhile, in the research community of image recognition, the methods using CNN monopolize the highest accuracy of main tasks such as classification and object detection. Using these methods, it is possible to classify food categories and detect single dishes one by one from multiple-dish food photos.

Fig. 1 Examples of multiple-dish food photos.

In this work, we propose food calorie estimation for multiple-dish food photos using CNN. Our model is trained to perform multi-task learning of dish detection and food calorie estimation so that it detects single dishes and estimates food calories simultaneously from multiple-dish food photos.

Ege et al. [1] proposed food calorie estimation from food photos by learning of regression with CNN. They also created a calorie-annotated food photo dataset for learning of regression, which estimates food calories directly from food photos. Since this approach does not depend on food category classification, different food calories are estimated for the same food category, which potentially makes it possible to account for the intra-food category differences. However, the input of this CNN corresponds only to the single-dish food photos, and it is not possible to estimate the food calorie of individual dishes one by one from multiple-dish food photos. Therefore, in this work, for estimation of food calories of each of multiple dishes, we integrate object detection for multiple-dish food photos and food calorie estimation for single-dish food photos, and propose a new method on simultaneous dish detection and calorie estimation. Note that the output value of food calories by our network is the calorie value per serving. In this work, regardless of the quantity of food in the photo, the food calorie corresponding to the quantity of one dish is estimated.

A common object detection system estimates categories and bounding boxes, which identifies the position of objects for each of the detected objects in a given image. Using object detection for multiple-dish food photos, it is possible to estimate bounding boxes and categories for each

Manuscript received October 12, 2018.
Manuscript revised January 22, 2019.
Manuscript publicized April 25, 2019.
[†]The authors are with Department of Informatics, The University of Electro-Communications, Chofu-shi, 182–8585 Japan.
a) E-mail: ege-t@mm.inf.uec.ac.jp
b) E-mail: yanai@cs.uec.ac.jp
DOI: 10.1587/transinf.2018CEP0004

dish. By using object detection, dishes in a multiple-dish food photo are expected to be detected one by one separately. With regard to object detection, it is possible to achieve high precision and high speed using recent CNN-based methods. In this work, we use an object detection method based on CNN to detect single dishes from multiple-dish food photos. Moreover, we build a network that estimates food calories and detects multiple dishes simultaneously. Although a method on object detection estimates bounding boxes and categories in general, in this work, we detect multiple dishes and estimate food calories simultaneously by learning the food calorie estimation task in addition to object detection.

To summarize our contributions in this work, (1) we propose food calorie estimation from multiple-dish food photos, (2) we realize the multi-task learning of dish detection and food calorie estimation with a single CNN and, (3) because there is no dataset currently with both annotated bounding boxes and food calories for each dish, we use two datasets for multi-task learning of CNN, which are multiple-dish food photos with bounding boxes and single-dish food photos with food calories. Note that this paper is based on our previous conference paper [2].

2. Related Work

Recently, various automatic food calorie estimation techniques employing image recognition have been proposed.

Miyazaki et al. [3] estimated calories from food photos directly. They adopted image-search based calorie estimation, in which they searched the calorie-annotated food photo database for the top n similar images based on conventional hand-crafted features, such as color histogram and Bag-of-Features. They hired dietitians to annotate calories on 6512 food photos which were uploaded to the commercial food logging service Food-Log[†]. As with our approach, their method estimates food calorie value per serving.

One of the CNN-based researches of detection of multiple food dishes is that of Shimoda et al. [4]. In [4], firstly, region proposals are generated by selective search. Secondly, for each region proposal, the food area is estimated by saliency maps obtained by CNN. Finally, overlapped region proposals are unified by non-maximum suppression (NMS). In practice, their method enables segmentation of the food area. It can be applied to detection because segmentation is a pixel-by-pixel classification. In addition to the above work, Shimoda et al. [5] also proposed the method which generates region proposals by CNN. In the work of Shimoda et al. [5], firstly, region proposals are generated by saliency maps obtained by CNN. Secondly, each region proposal is classified. Finally, overlapping region proposals are unified by non-maximum suppression.

Dehais et al. [6] proposed another method for food dish segmentation. In their work, firstly, the Border Map which represents rough boundary lines of a food region is obtained by CNN. Then, the boundary lines of Border Map are refined by the region growing/merging algorithm. On the other hand, in our work, we use a CNN-based object detection for multiple-dish food photos.

Im2Calories by Myers et al. [7] estimates food categories, ingredients, and the regions of each of the dishes included in a given food photo and finally outputs food calories by calculation based on the estimated volumes and the calorie density corresponding to the estimated food category. In their experiments, they faced the problem that the calorie-annotated dataset was insufficient and evaluation was not sufficiently performed.

[†]http://www.foodlog.jp/

3. Proposed Method

This section describes our network for the multi-task learning of dish detection and food calorie estimation.

3.1 Multi-Task Learning of Dish Detection and Food Calorie Estimation

We implement a network that estimates bounding boxes of food dishes and their calories simultaneously by multi-task learning of dish detection and food calorie estimation with a single CNN. In other words, our network estimates bounding boxes of dish regions, their categories, and their calories from multiple-dish food photos. In this work, we use the food calorie estimator proposed by Ege et al. [1], for image-based food calorie estimation. We apply SSD [8] which is a high-speed and highly accurate detection network to detect dishes. As shown in Fig. 2, SSD has various scale output features to gain robustness of the object scale.

As shown in Fig. 3, the network of SSD consists of only convolution layers, takes an input image, and outputs a feature map, so that the output holds position information. Consequently, each pixel on the feature map of the output corresponds to a certain region on the input image. Let S be the width and height of the output feature map, bounding boxes and categories of the object are estimated for each of $S \times S$ grids on the input image. In general, object detection method estimates object bounding boxes including coordinates and sizes with class probabilities. Hence, let B be the number of estimated bounding boxes for each grid and C be the number of categories, the total number of channels of the output feature map is defined as $B \times (4 + C)$.

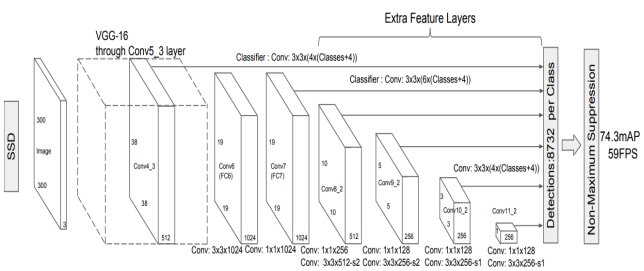

Fig. 2 The architecture of SSD (This figure is quoted from [8]).

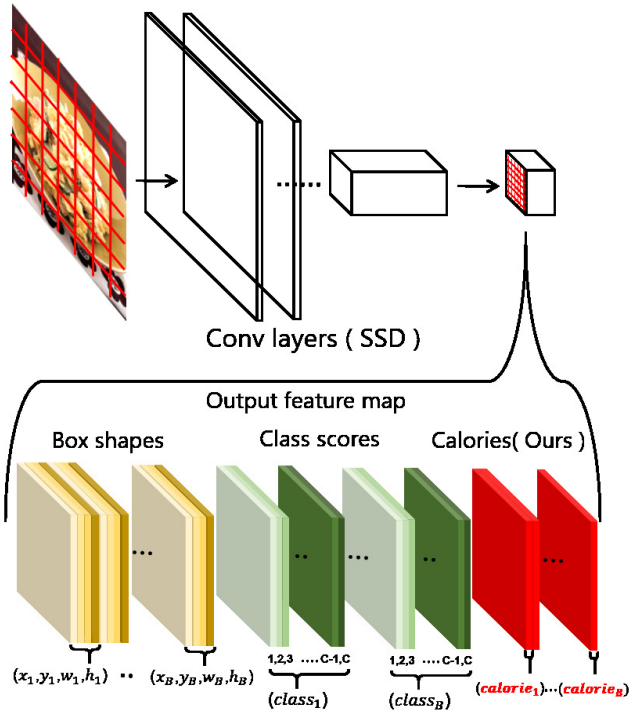

Fig. 3 The output feature map of our network.

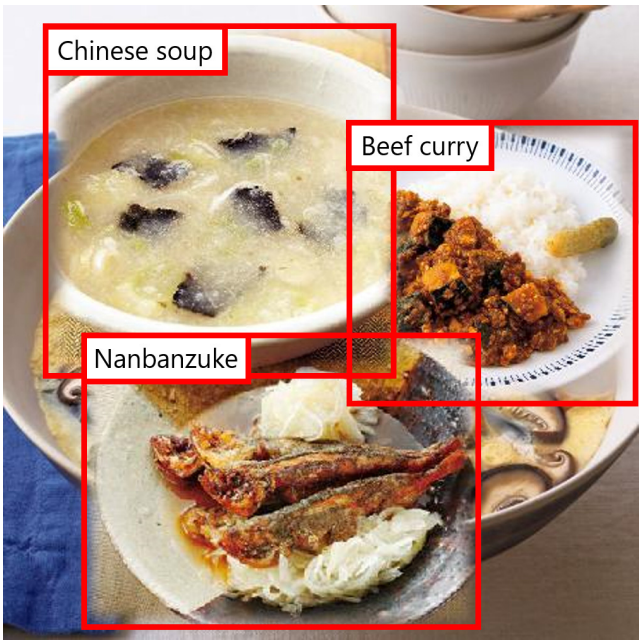

Fig. 4 Example of calorie-annotated multiple-dish food photos with pseudo-bounding boxes represented by red boxes (Chinese soup: 64 kcal, beef curry: 647 kcal, nanbanzuke: 290 kcal).

In this paper, we propose a network for multi-task learning of dish detection and food calorie estimation. We modify the network of SSD based on VGG16 [9] so that it can output a food calorie value on each food bounding box. To modify SSD, we carry out multi-task learning of food calorie estimation as well as dish detection. In this proposed method, we add the output layer of estimated food calories as well as bounding boxes and categories so that our network estimates food calories in addition to bounding boxes and categories. Hence, the total number of channels of our output feature map is defined as $B \times (4 + C + 1)$.

To estimate food calories for each of the estimated bounding boxes, a food image dataset annotated with both bounding boxes and the calorie values of the foods in each of the bounding boxes is required. However, such datasets do not exist at present [1]. Therefore, in this work, we create calorie-annotated multiple-dish food photos with pseudo-bounding boxes as shown in Fig. 4 using the following procedure. Firstly, a food image is prepared as a background image. Secondly, some random size calorie-annotated food images are embedded on the random positions. For smoothing of the boundary, an alpha blending process is used to embed calorie-annotated food images into a background. Then the embedded image region is set as a ground-truth bounding box.

3.2 Image-Based Food Calorie Estimation

In this work, we use image-based food calorie estimation based on regression learning with CNN [1] to detect dishes and estimate food calories simultaneously. The network proposed by Ege et al. was limited to an input image with a single-dish, and the estimated value of food calories corresponds to the amount for one person regardless of the amount of food in the food image. On the other hand, our network additionally supports multiple-dish food photos, and the value of the food calorie output is calorie value per serving as in the case of [1]. Also, we use Eq. (1) according to [1] as a loss function of the food calorie estimation task.

Generally, in a regression problem, a mean square error is used as the loss function; however, in this paper, we use the loss function of Eq. (1). We denote L_{ab} as an absolute error and L_{re} as a relative error, and L_{cal} is defined as follows:

$$L_{cal} = \lambda_{re} L_{re} + \lambda_{ab} L_{ab}, \tag{1}$$

where λ_{re} and λ_{ab} are the weight on the loss functions. The absolute error is the absolute value of the difference between the estimated value and the ground-truth, while the relative error is the ratio of the absolute error to the ground-truth. Let y be the estimated value of an image x and g be the ground-truth, L_{ab} and L_{re} are defined as follows:

$$L_{ab} = |y - g| \tag{2}$$

$$L_{re} = \frac{|y - g|}{g} \tag{3}$$

We integrate these calorie loss functions, L_{cal}, with the loss function of SSD [8], L_{SSD}, for end-to-end training of the proposed network.

$$L = L_{SSD} + L_{cal}, \tag{4}$$

where L is the total loss to be optimized.

4. Dataset

Currently, there is no multiple-dish food photo dataset with bounding boxes for object detection and food calorie estimation. Therefore, we use two types of datasets for learning dish detection and food calorie estimation with a single CNN. For the two types of datasets, we use UEC Food-100 [10] which includes multiple-dish food photos attached with bounding boxes, and a calorie-annotated food photo dataset [1] which contains single-dish food photos with food calories. Note that all the food categories of the calorie-annotated dataset are included in the categories of UEC Food-100.

Fig. 5 Examples of multi-label food photos in UEC Food-100 [10].

Fig. 6 Examples of calorie-annotated food photos of 15 food categories.

4.1 UEC Food-100

UEC Food-100 [10] is a Japanese food photo dataset with 100 food categories including multiple-dish food photos. This dataset includes more than 100 single-dish food photos for each category, with a total of 11566 single-dish food photos. This dataset includes 1174 multiple-dish food photos. All 12740 images in the dataset are annotated with bounding boxes. Figure 5 shows examples of multi-label images in UEC Food-100.

4.2 Calorie-Annotated Food Photo Dataset

In this work, we use calorie-annotated recipe data [1] collected from commercial cooking recipe sites on the web and the collected recipe data have food calorie information for one person. In this work, we extend this dataset from 15 food categories to 50 food categories and create a total of 7687 images dataset. Figure 6 shows example photos with food from 15 categories in a calorie-annotated food photo dataset.

5. Experiments

We used both UEC Food-100 [10] and a calorie-annotated food photo dataset for multi-task learning of dish detection and food calorie estimation with a single CNN. The learning of the dish detection task and learning of the food calorie estimation task were alternately performed by switching the dataset by mini-batch. For the learning of the dish detection task, UEC Food-100 and the loss terms related to the dish detection task are used. On the other hand, for the learning of the food calorie estimation task, a calorie-annotated food photo dataset of both single-dish and multiple-dish with pseudo-bounding boxes and the loss terms related to the food calorie estimation task are used.

We used SGD as an optimizater with a momentum of 0.9 and a mini-batch size of 32. We used 10^{-3} of learning rate for 40,000 iterations and then used 10^{-4} for 10,000 iterations. We adopted non-maximum suppression (NMS) and fixed the Intersection-over-Union (IoU) threshold for NMS at 0.5.

The weights of the loss term of Eq. (1), λ_{re} and λ_{ab}, were determined in the following way. Firstly, all the weights of the loss terms are set to 1, and the model is trained once. In the training, the values of the losses for each iteration are preserved. Finally, the inverse of the average of the loss values in all iterations is used as the weights for each of the loss terms so that all the losses were equally reduced. In the experiments, we set λ_{re} and λ_{ab} as 1.12 and 0.00275, respectively. Note that because the average of L_{SSD} was close to 1.0, we did not put a weight to the loss term of SSD.

Table 1 The results of food calorie estimation from single-dish food photos.

	rel. err.(%)	abs. err.(kcal)	20% err.(%) ≤	40% err.(%) ≤
Single-task [1]	30.2	105.7	43	76
Multi-task [1]	27.9	94.1	48	**80**
Detection+Calorie estimation (ours)	**26.6**	**89.4**	**51**	79

Table 2 Comparison of execution speed and model size. The sequential model is a two stage process of SSD [8] and image-based food calorie estimation [1].

	speed (msec)	model size (MB)
Sequential model	104.8 (77.6+27.2)	796 (136.9+659)
Multiple-dish (ours)	73.5	137.4

5.1 Food Calorie Estimation from Single-Dish Food Photos

In this experiment, we used 70% of calorie-annotated single-dish food photos and 10k multiple-dish food photos with pseudo-bounding boxes for the training of the food calorie estimation task, and 30% of calorie-annotated single-dish food photos for the evaluation.

Note that considering the number of test images per food category, we used food photos included in the 15 food categories of Fig. 6 for evaluation. In addition, the test images are single-dish food photos; therefore, as a final output, we used an estimated bounding box with the highest scores for the class probability that is the presence of each object category.

Following Ege et al. [1], we used several evaluation values, including an absolute error, a relative error and a ratio of the estimated value within the relative errors of 20% and 40%. We evaluated the absolute error representing the differences between estimated values and the ground-truth, and the relative error representing the ratio between the absolute error and the ground-truth.

Table 1 shows the results of food calorie estimation for single-dish food photos. In comparison with the food calorie estimation [1] that estimates food calories and food categories simultaneously using multi-task learned VGG16 [9], our proposed method outperformed the previous method except for 40% error ratio.

In addition we showed the execution speed and model size of our network in Table 2. We prepared the following sequential model for comparison. Firstly, we extract a bounding box of a food dish by SSD [8], and crop the region corresponding to the bounding box. Then, we input the cropped image into the image-based food calorie estimation network [1] in order to estimate the value of food calories in the image.

The execution speed of our network with an input image with a size of 300×300 and mini-batch of 1 is approximately 73.5 ms on a GTX 1080 Ti. Additionally, the size of our network that detects dishes and estimates food calories is 137.4 MB, while the size of VGG16-based calorie estimation network is 659MB and the size of original SSD is 136.9 MB.

Fig. 7 Examples of calorie-annotated multiple-dish food photos.

5.2 Food Calorie Estimation from Multiple-Dish Food Photos

Unfortunately, no standard dataset on calorie-annotated multiple-dish food photos exists. For evaluating the performance of the proposed method for multiple-dish food photos, we prepared calorie-annotated multiple-dish food photos by using calorie-annotated life-sized food cards[†]. Note that since the volumes of foods in the food cards are normalized to a serving for one person, calorie estimation of them is easier than real food photos in general.

Firstly, 30 single-dish food cards included in the 15 food categories of Fig. 6 are selected. Secondly, the 30 single-dish food photos and 50 multiple-dish food photos composed of 25 two-dish photos and 25 three-dish photos are taken by a camera. The food cards of multiple-dish food photos are randomly selected from the 30 single-dish food cards. Finally, bounding boxes and total food calories are annotated for each food photos. The average value of total food calorie of single-dish and multiple-dish are 400kcal and 966kcal, respectively. Figure 7 shows examples of calorie-annotated multiple-dish food photos.

[†]http://www.gun-yosha.com/book/{ryori.html, balanceguide.html, gaishokucard.html}

Table 3 The results of both food detection and food calorie estimation from multiple-dish food photos.

	rel. err.(%)	abs. err.(kcal)	≤ 20% err.(%)	≤ 40% err.(%)	precision (%)	recall (%)
Single-dish (top1)	28.2	106.0	46.7	76.7	100.0	100.0
Single-dish (threshold)	32.7	110.4	43.3	73.3	93.8	100.0
Multiple-dish (threshold)	29.7	292.3	40.0	70.0	85.7	86.4

5.3 Calorie-Annotated Food Photo Dataset

In this experiment, we used the model used in Sect. 5.1 and newly created single-dish and multiple-dish food photos for evaluation of performance of both food detection and food calorie estimation. To prevent redundant bounding boxes from being detected, we used only the estimated bounding boxes the class probability of which are higher than a fixed threshold value. We empirically fixed the threshold value as 0.3.

For evaluation of food detection, we used both of the precision and the recall of bounding box detection. We defined true positive as IoU ≥ 0.5, which is the ratio of an intersection of the estimated and ground-truth bounding boxes to their union. If there is more than one true positive bounding box for a ground-truth, one is considered as true positive and the others are false positive. For evaluation of food calorie estimation for multiple-dish food photos, we calculate total food calories from the sum of the food calories of each of the estimated bounding boxes and use the same evaluation metrics as Sect. 5.1.

Table 3 shows the results of both food detection and food calorie estimation for single-dish and multiple-dish food photos. Single-dish (top1) is the result in which we used the criteria of selecting only the bounding box with the highest class probability in the same way as the experiments in Sect. 5.1. The absolute error of multiple-dish photos shown in the table was calculated as the average value of the differences between the total estimated food calories and the ground-truth calories within each multiple-dish photo. The estimated total food calories are strongly influenced by detection errors including false-positive, false-negative and partially detected regions. This is the main reason why the absolute error of multiple-dish was much larger than single-dish, but the relative error of multiple-dish is smaller than single-dish.

6. Conclusions

In this work, we proposed food calorie estimation from multiple-dish food photos by multi-task learning of dish detection and food calorie estimation with a single CNN. Currently, there exists no dataset of multiple-dish food photos annotated with bounding boxes and food calories. We used UEC Food-100 [10] for the learning of food detection task and calorie-annotated food photos [1] for the learning of food calorie estimation task. In addition, for evaluating performance for multiple-dish food photos, we prepared newly calorie-annotated multiple-dish food photos.

As future work, we plan to construct large-scale calorie-annotated multiple-dish food photo dataset. In addition, we plan to realize food calorie estimation taking account of food volume because the current method estimates the standard calorie value per serving and cannot estimate calorie values for a large or small serving of dishes.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 15H05915, 17H01745, 17H05972, 17H06026, and 17H06100.

References

[1] T. Ege and K. Yanai, "Image-based food calorie estimation using recipe information," IEICE Trans. Inf. & Syst., vol.E101-D, no.5, pp.1333–1341, 2018.
[2] T. Ege and K. Yanai, "Multi-task learning of dish detection and calorie estimation," Proc. ACM MM Workshop on Multimedia Assisted Dietary Management, pp.53–58, 2018.
[3] T. Miyazaki, G. Chaminda, D. Silva, and K. Aizawa, "Image-based calorie content estimation for dietary assessment," Proc. IEEE ISM Workshop on Multimedia for Cooking and Eating Activities, pp.363–368, 2011.
[4] W. Shimoda and K. Yanai, "CNN-based food image segmentation without pixel-wise annotation," Proc. IAPR International Conference on Image Analysis and Processing, vol.9281, pp.449–457, 2015.
[5] W. Shimoda and K. Yanai, "Foodness proposal for multiple food detection by training of single food images," Proc. ACM MM Workshop on Multimedia Assisted Dietary Management, pp.13–21, 2016.
[6] J. Dehais, M. Anthimopoulos, and S. Mougiakakou, "Food image segmentation for dietary assessment," Proc. ACM MM Workshop on Multimedia Assisted Dietary Management, pp.23–28, 2016.
[7] A. Myers, N. Johnston, V. Rathod, A. Korattikara, A. Gorban, N. Silberman, S. Guadarrama, G. Papandreou, J. Huang, and P.K. Murphy, "Im2calories: towards an automated mobile vision food diary," Proc. IEEE International Conference on Computer Vision, pp.1233–1241, 2015.
[8] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, and A.C. Berg, "SSD: Single shot multibox detector," Proc. European Conference on Computer Vision, vol.9905, pp.21–37, 2016.
[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," Proc. International Conference on Representation Learning, 2015.
[10] Y. Matsuda, H. Hajime, and K. Yanai, "Recognition of multiple-food images by detecting candidate regions," Proc. IEEE International Conference on Multimedia and Expo, pp.25–30, 2012.

Takumi Ege is a master course student at Department of Informatics, the University of Electro-Communications, Tokyo, Japan. He is working on food image recognition.

Keiji Yanai is a professor at Department of Informatics, the University of Electro-Communications, Tokyo, Japan. He received B.Eng., M.Eng. and D.Eng. degrees from the University of Tokyo in 1995, 1997 and 2003, respectively. From 1997 to 2006 he was a research associate and till 2015 he was an associate professor at Department of Computer Science, the University of Electro-Communications, Tokyo. From November, 2003 to September, 2004, he was a visiting scholar at Department of Computer Science, University of Arizona, USA. His recent research interests include object recognition, deep learning, and Web multimedia mining.