A New Large-scale Food Image Segmentation Dataset and Its Application to Food Calorie Estimation Based on Grains of Rice

Takumi Ege Wataru Shimoda Keiji Yanai
The University of Electro-Communications, Tokyo
{ege-t,shimoda-k,yanai}@mm.inf.uec.ac.jp

ABSTRACT

To estimate food calorie accurately from food images, accurate food image segmentation is needed. So far no large-scale food image segmentation datasets which have pixel-wise labels exists. In this paper, we added segmentation masks to the food images in the existing dataset, UEC-Food100, semi-automatically. To estimate segmentation masks, we revised bounding boxes included in the UEC-Food100 dataset so that bounding boxes bounds not dish regions but only food regions. As results, by applying GrubCut, we obtained good segmentation masks, and we checked and corrected 1000 images of them if needed by hand for the testing masks.

We trained segmentation networks with the newly-created food image masks. As results, segmentation accuracy was much improved, which is expected to bring more accurate food calorie estimation. In addition, we propose a new method on food calorie estimation using grains of steamed rice which are typically contained in Japanese foods instead of a reference card. By the experiments, we show real food size can be estimated from rice images, which helps accurate food calorie estimation.

CCS CONCEPTS

• **Computing methodologies** → *Scene understanding*; *Object recognition.*

KEYWORDS

food image segmentation, food calorie estimation, food image recognition

ACM Reference Format:
Takumi Ege Wataru Shimoda Keiji Yanai. 2019. A New Large-scale Food Image Segmentation Dataset and Its Application to Food Calorie Estimation Based on Grains of Rice. In *5th International Workshop on Multimedia Assisted Dietary Management (MADiMa '19), October 21, 2019, Nice, France.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3347448.3357162

MADiMa '19, October 21, 2019, Nice, France

ACM ISBN 978-1-4503-6916-9/19/10…$15.00
https://doi.org/10.1145/3347448.3357162

1 INTRODUCTION

The food calorie is considered to be strongly dependent on the food categories and volumes, and it is useful in terms of food management if it becomes possible to estimate food categories and volumes automatically from food images. In food classification from food images, the method using CNN has already achieved highly accurate classification. Recently, there are some applications that estimate food categories automatically from food images.

However, food volume estimation that is essential for food calorie estimation still remains as a difficult problem. In most of the cases, the estimated calories are just associated with the estimated food categories, or the relative size compared to the standard size of each food category which is usually indicated by a user manually.

Therefore, in this paper, we propose estimating food calorie from a food photo considering an area of food regions. The area of food region is calculated from a real scale and a food region.

The real scale of a pixel on an image is estimated with CNN from rice grains of food images containing rice. Because of the size of rice grains is almost constant, it works as a reference object for estimating the real scale. In this paper, we construct a CNN that takes patch images of rice grains as inputs and outputs real size of the patch images. The food region is estimated by food segmentation with CNN. Note that in this method, it is assumed that the food photo is taken vertically from right above of the table surface.

In the previous works using a reference object such as a card, a user always needs to take food photos with the pre-registered reference object. In contrast to that, rice is usually included in a Japanese meal, so that the method of using rice grains as a reference object is not required to prepare that. So it is enabled to apply to the food photos taken in the past uploaded on Web.

To summarize our contributions, we propose the method to estimate food calories from a food photo considering the area of food regions. The area of food regions is calculated from an estimated real scale and food region with CNN. Firstly, for estimating the real scale of the pixel, we construct a CNN which can estimate real size of the patch images of rice grains. Secondly, the food regions are obtained by food segmentation based on CNN. Finally, the food calories considering the food area are estimated.

2 RELATED WORK

Various approaches has been proposed so far and the main approach is to estimate calories based on estimated food categories and its size or volume using the value of food calorie per unit area or volume.

Chen et al. [3] proposed an image-based food calorie estimation method that estimates food categories and volumes by depth cameras such as Kinect. Depth cameras such as Kinect are special devices, so it is thought that ordinary people are difficult to use usually.

Kong et al. [7] proposed a mobile application to estimate food calories from images multiple images,"DietCam". They carried out segmentation and food item recognition, and in addition reconstructed 3D volumes of food items and calculate food calories based estimated volumes. 3D reconstruction was performed with SIFT-based keypoint matching and homography estimation which were a standard method of 3D stereo vision.

Dehais et al. [5] proposed the other method for food dish segmentation. In the work of Dehais et al. [5], 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. The system which requires some photos taken from multiple viewpoints needs to calibrate devices, which limits the situation where the system can be used.

Im2Calories by Myers et al. [9] 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 experiment, they faced the problem that the calorie-annotated dataset is insufficient and evaluation is not sufficiently performed.

Pouladzadhe et al. [11] proposed a food calorie estimation system which needed two dish images taken from the top and the side and used a thumb of a user as a reference object. Their method to estimate volumes were calculated by multiplying the size of food items estimated from the top-view image by the height estimated from the side-view image, which was relatively a straight-forward way. In our method, the real size estimation is performed using rice grains with the fixed size as reference objects.

Okamoto et al. [10] proposed an image-based calorie estimation system which estimates food calories automatically by simply taking a meal photo from the top with a pre-registered reference object. Firstly, both regions of food and reference object are estimated from meal photo by GrabCut [15], Secondly, the real area of food is calculated by comparing regions of food and that of the reference object, Finally, the calorie of the food is estimated based on the real area of the food. Contrastly, our method treats rice grains as reference objects and estimate real size from rice grains with CNN.

3 DATASET: UECFOODPIX

A lot of food image datasets have been published so far. In food categorization tasks, some of them such as Food-101 [1], UECFood-100 [8] and UECFood-256 [6] are used for standard benchmarks of food image recognition tasks. Among them, UECFood-100/256 are the only datasets which have bounding boxes of food areas of all the images in the dataset. However, the bounding boxes in UECFood-100/256 are originally annotated for food classification of multiple dishes, so that the food category is limited to 100 kinds of foods included in UECFood-100/256 categories, and the other foods than 100 categories are ignored. For example, if an image contains bread toast and butter, a bounding box is annotated with only the area of the bread toast and no bounding box is provided to the area of the butter, since butter is not contained in the categories of the UECFood dataset. In addition, bounding boxes in the dataset do not consider instances. If an image contains many pieces of sushi, only one large bounding box is usually given to a set of sushi.

Currently, there are no large-scale food image dataset with segmentation masks. Only the UNIMIB2016 dataset provides food region information as polygons [4] which are equivalent to segmentation masks. However, its scale is not so large (1027 multiple-dish images with 73 food categories), and the food images in UNIMIB2016 are biased and not generic since all the food images were taken at the same canteen.

3.1 Construction

Therefore, we have decided to construct a food image dataset annotated with instance-based bounding boxes and segmentation masks by expanding the conventional UECFood-100. Firstly, we annotated new instance-based bounding boxes to 10,000 images included in UECFood-100 manually [1] Secondly, we annotated segmentation masks to 9000 images automatically by GrabCut [15], and 1000 images manually for the evaluation. As shown in the Figure 1, the new bounding boxes are applied to all of the food categories and instances so that they enclose foods rather than dishes.

The number of bounding boxes in the new dataset becomes more than twice that of UECFood-100. To save annotation cost, we annotated only bounding boxes on each of the food instances and did not assign food categories manually. Instead, the food category of each new bounding box is annotated automatically based on the overlap ratio a_o between a new bounding box $B_{uecfoodseg}$ and one $B_{uecfood100}$ contained in the original UECFood-100 dataset by Eq.(1). The new bounding boxes have the food category of conventional one with overlap ratio a_0 which exceed 0.5, and that has no overlapped conventional one are assigned to "Other foods" category.

$$overlap = \frac{area(B_{uecfoodseg} \cap B_{uecfood100})}{area(\min\{B_{uecfoodseg}, B_{uecfood100}\})} \quad (1)$$

Since it costs too much to create all segmentation masks by hand, we annotate new segmentation masks automatically by GrabCut [15] using new bounding boxes. The segmentation mask for each bounding box was generated one by one by using GrabCut. Figure 2 shows some segmentation masks generated automatically from the newly annotated instance-based bounding boxes.

3.2 Benchmarks

For the benchmark of our dataset, we perform food detection and food segmentation with our new dataset, UECFoodPix, of 10,000 food images. In this experiments, 9000 and 1000 images are used for training and evaluation, respectively. Note that the bounding boxes and segmentation masks of the evaluation dataset are annotated manually.

For food detection, we use YOLOv2 proposed by Redmon et al. [13] to detect dishes. YOLOv2 improves YOLO [12] based on CNN previously proposed, it enables high-speed and high-accuracy

[1] In fact, UECFood-100 has 12,740 food images. The annotation work is still on the way. We will release it after all annotation is done.

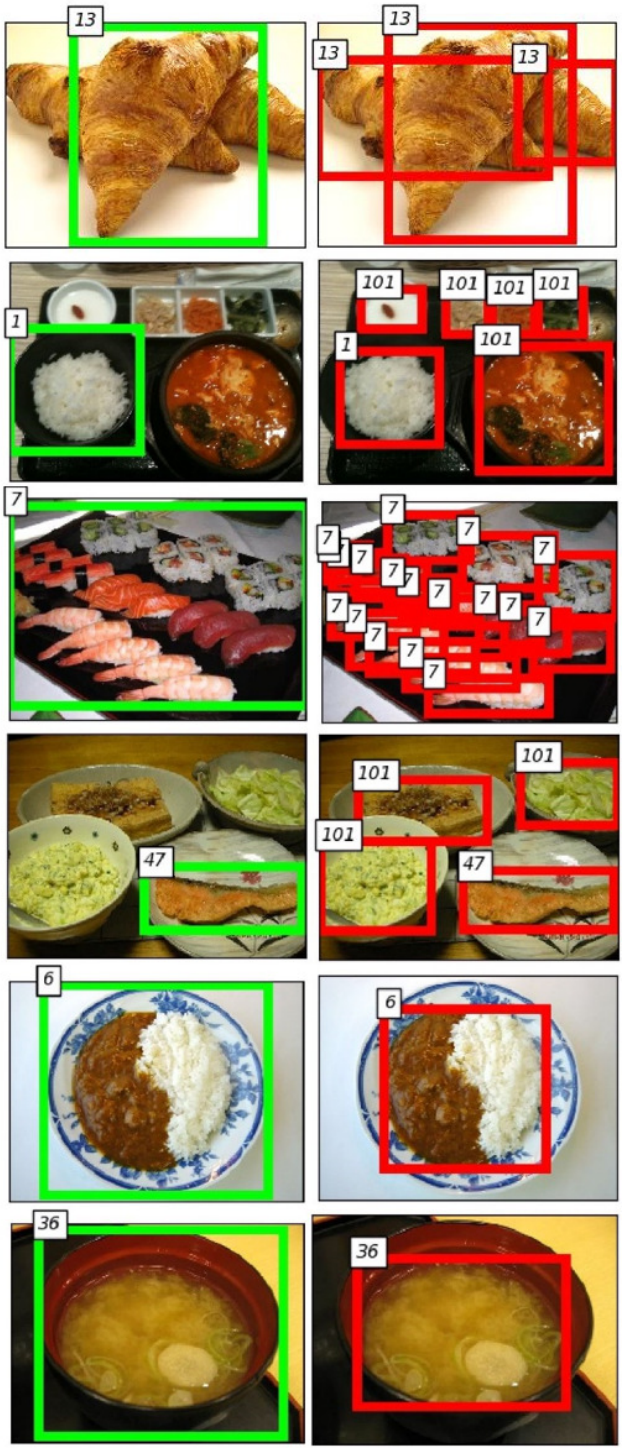

Figure 1: Bounding boxes, left is UECFood-100 [8], right is UECFoodPix (Ours).

object detection. As a result, mean Average Prediction (mAP) of food 101 classes is 60.4 %.

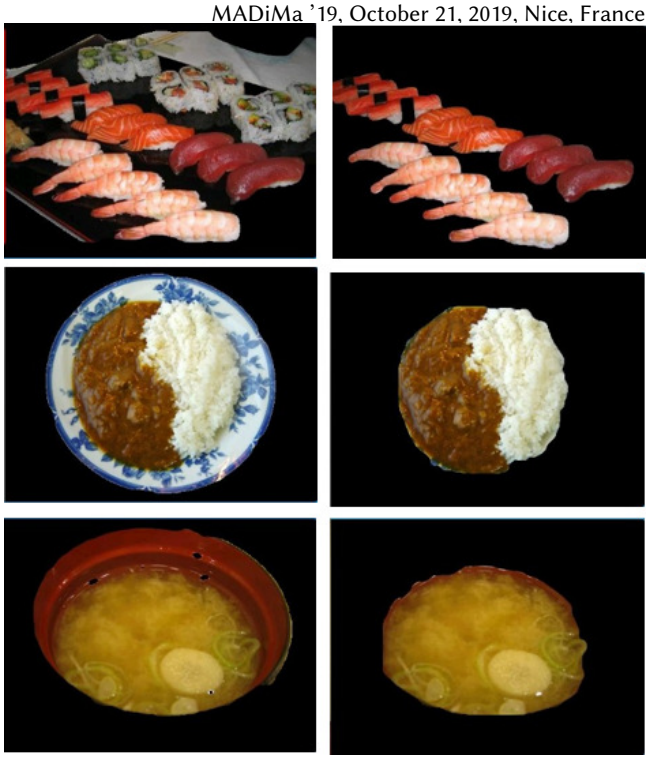

Figure 2: Segmentation masks with GrabCut [15], left is based on bounding boxes of UECFood-100 [8], right is based on that of UECFoodPix (Ours).

For food segmentation, we use DeeplabV3+ proposed by Chen et al. [2] as a baseline method. DeeplabV3+ has a deep Xception model as a backbone network and a decoder network for refinement of outline of segmentation outputs. The model is known as a current state-of-the-art semantic segmentation architecture. We obtained 41.9 % by Deeplabv3+ on mean Intersection over Union (IoU) for the 101 class food segmentation, which includes "Other foods" and a background class.

4 METHOD

Our proposed method performs food detection, food segmentation and real scale estimation for rice grain images. Finally, estimates food calories considering food area, according to the following processing.

(1) Take a food photo with rice.
(2) Food detection.
(3) Food segmentation for each of the estimated bounding boxes.
(4) Estimate real size from rice images.
(5) Calculate real size of food area from both estimated real scale and segmentation masks.
(6) Estimate food calories based on estimated food area and category-dependent calorie density.

In the proposed method, it is assumed that the food photo is taken from directly above the dish vertically to the table surface. The detail of each procesing step is described below.

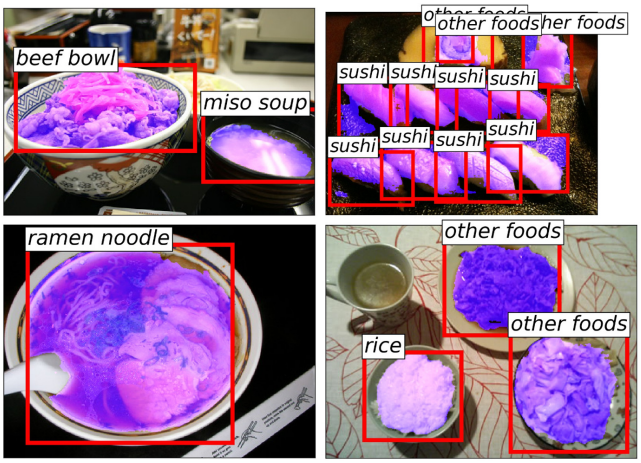

Figure 3: The results of food detection and segmentation.

4.1 Food detection and segmentation

In food detection, we use YOLOv2 [13] trained with our new bounding boxes of 9000 food images of our dataset.

In food segmentation, we use U-Net proposed by Ronneberger et al. [14] For training of U-Net, we use a binary mask which represents food foreground or background ignoring the food categories, since the detection network, YOLOv2, can estimate food categories with bounding boxes and a segmentation network does not need to estimate food categories. As a result of the evaluation with 1000 food images annotated segmentation masks manually, mean IoU is 84.1 %.

The results of food detection and estimated segmentation masks from each of the estimated bounding boxes are shown in Figure 3.

4.2 Real scale estimation from rice grain images

Our proposed method estimates real scale from rice grains images with CNN to estimate food area. As shown in Figure 4, our model of real scale estimation is based on VGG16 [17], which takes a rice grain image as an input and outputs the real size of the length of one side of the input image. The input image is a patch image of rice images that length of one side is 224×224 in this work. Since the real scale estimation is a regression problem, we use the mean square error as a loss function.

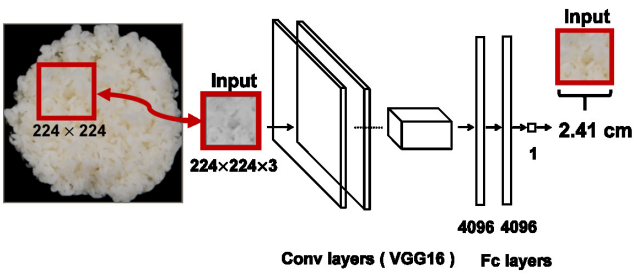

Figure 4: Real scale estimation network.

4.3 Calorie estimation considering food area

Our method estimates food calories based on the real size of food regions calculated from the real scale and the food regions. The real size of food regions F_r can be obtained by the following equation:

$$F_r = S_r * F_p \tag{2}$$

where F_p represents the number of pixels of the region of the target food, and S_r represents the real area size of region of a pixel.

After calculating the real size of the regions of target food, we estimate its food calorie value following the method proposed by Okamoto et al. [10]. They convert the 2D top-view size of the food to the calorie value according to the category-dependent quadratic curves estimated based on the training data annotated with real food calories. In this work, we use the quadratic curve of each food category trained by Okamoto et al. [10] as they are.

5 EXPERIMENTS

We perform two experiments of the real scale estimation from rice grains images and the food calorie estimation base on food area. In the real scale estimation from rice grain image, we use real scale annotated rice image dataset. In the food calorie estimation based on food area, 9000 and 1000 images of our UECFoodPix dataset are used for training and testing, respectively. Currently, since there are no dataset annotated food calories considering food volume, the evaluation of the estimated food calories are not performed.

5.1 Real scale estimation from rice grain image

In this experiment, we estimate the real scale from rice grain images. To do that, we construct a rice image dataset consisting of 360 images annotated with real size.

5.1.1 Rice image dataset. Because these rice images need to be annotated real scale, they were taken in our laboratory and the real scale was given based on the diameter of the bowl of rice.

Rice of Japonica commercially available was used for the dataset and we prepared three patterns of rice cooked with different amounts of water such as 180 ml, 200 ml and 220 ml for 150 g of rice. Then we took 60 photos for each pattern of rice with two kinds of camera respectively such as COOLPIX AW120 and iPhone8 Plus. For collecting various rice images with various real scales, the types of dishes and the distance between the rice and the camera were changed when photos were taken. Finally, a total of 360 rice images with the real scale were collected.

After taking rice photos, the real scale per pixel was given based on the diameter of the bowl of rice for each image. Furthermore, in order to remove background information, rice segmentation masks were created manually.

5.1.2 Real scale estimation. In this experiment, we estimate the real scale from rice grain image using our rice image dataset. We divide rice image dataset into six sets based on the camera type and the amount of water for steaming rice. We use five sets for training and one remaining set for evaluation. Thus, the training images are 300 images and test images are 60 images.

In training of a network, 224×224 patch images are cropped from random positions of the rice images. In addition, the augmentation such as flipping, resizing and rotating of rice images are performed. In the case of resizing a rice image with the ratio of n, the real scale of the rice image is multiplied by $1/n$.

In the evaluation, input patch images are cropped by 5×5 grid sampling from a rice image. The final output is an average value of outputs of the network for 25 input patch images. Furthermore, in both of the training and evaluation, input patch images which has a background area of 1 % or more are removed based on annotated rice segmentation masks.

Our network architecture of real scale estimation is based on VGG16 [17] and the initial weights of the network are the pre-trained model of ImageNet classification tasks except for the output layer. For optimization of the CNN, we used SGD with the momentum value 0.9 and the size of mini-batch was 16. We used 10^{-5} of the learning rate for 3,000 iterations.

We show the average of the relative error representing the ratio between the estimated values and the ground-truth, and the absolute error representing the differences between both. In addition, we show the correlation coefficient between the estimated value and ground-truth and the ratio of the estimated value within the relative error of 5 %, 10 %, and 20 %. Note that in the evaluation, the estimated real scale for 224 pixels is used.

Table 1 shows the result of real scale estimation. Figure 5 shows the relation between the estimated real scale values and ground-truth values. The average of the relative error and the absolute error for the estimated real scale of 224 pixels were 0.145 cm and 5.548%, respectively. In addition, the average correlation coefficient between the estimated value and ground-truth was 0.946 which showed a very high correlation.

In all combinations of training data and evaluation data, the relative error was less than 10%, the correlation coefficient was higher than 0.9, and most of the estimated values were included within the relative error of 20 %. From these results, we have confirmed that the effectiveness of the proposed model to estimates the real scale directly from the rice grain images without taking into account the variations in size and orientation of rice grains.

5.2 Food calorie estimation considering food area

In this experiment, the food calories considering food area is estimated. Firstly, the food bounding boxes are obtained by the food detection with YOLOv2 [13], then the food regions are estimated by food segmentation with U-Net [14] from each bounding boxes. Secondly, the real scale is estimated from obtained rice images. Finally, the food area is calculated by the combination of estimated real scales and food regions, and food calories are estimated based on the food area with a method proposed by Okamoto et al. [10]. Figure 6 shows an example of the result of estimated food areas and food calories.

6 CONCLUSIONS

In this work, we estimate food calories considering food area from a food photo. To do that, the food area is calculated by a combination of the estimated and food regions and real scales.

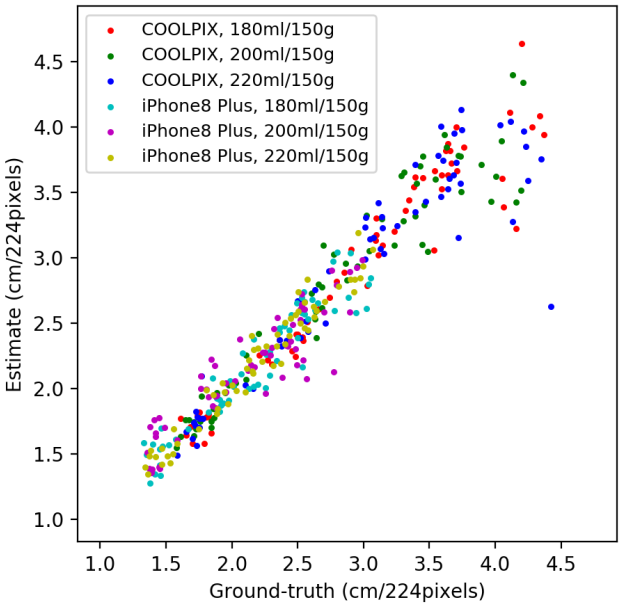

Figure 5: The correlation between the estimated real scale values and ground-truth values in the real scale estimation.

Figure 6: Example of the food calorie estimation considering the food area for a multiple-dish food photo. The tags on each red frame shows the estimated food areas and food calories.

In the real scale estimation, the average of the relative error and the absolute error for the estimated real scale of 224 pixels were 0.145 cm and 5.548 %, respectively. In addition, the average correlation coefficient between the estimated value and ground-truth was 0.946 that shows a very high correlation.

Furthermore, for food detection and food segmentation, we construct newly food image dataset with bounding boxes and segmentation masks by expanding the conventional UECFood-100 [8].

Table 1: The real scale estimation from rice grains images(the estimated values of 224 pixels is used for the evaluation).

evaluation data	abs.err.(cm)	rel.err.(%)	correlation	≤ 5 % rel.err.(%)	≤ 10 % rel.err.(%)	≤ 20 % rel.err.(%)
COOLPIX, 180ml/150g	0.152	4.822	0.963	61.667	88.333	98.333
COOLPIX, 200ml/150g	0.169	5.513	0.959	55.000	85.000	100.000
COOLPIX, 220ml/150g	0.194	5.906	0.920	55.000	86.667	96.667
iPhone8 Plus, 180ml/150g	0.123	5.706	0.949	51.667	85.000	100.000
iPhone8 Plus, 200ml/150g	0.145	7.305	0.910	56.667	66.667	91.667
iPnone8 Plus, 220ml/150g	0.086	4.037	0.976	71.667	95.000	100.000
Average	0.145	5.548	0.946	58.611	84.444	97.778

As future work, we plan to evaluate the food calorie estimation considering food areas. For the evaluation, we prepare multiple-dish food photos annotated food calories considering food volumes.

In addition, in order to construct a system that can estimate the food calories considering food volumes even when there is no rice, we are considering combining the method [10, 16] which employed segmentation and a reference object, or the depth information obtained from the camera of iPhone.

Note that we plan to release the new pixel-wise annotated food image segmentation dataset, "UECFoodPix", at http://foodcam.mobi/dataset/.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number 17J10261, 15H05915, 17H01745, 17H06100 and 19H04929.

REFERENCES

[1] L. Bossard, M. Guillaumin, and L. Van Gool. 2014. Food-101 – Mining Discriminative Components with Random Forests. In *Proc. of European Conference on Computer Vision.*

[2] L-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proc. of European Conference on Computer Vision.*

[3] M. Chen, Y. Yang, C. Ho, S. Wang, E. Liu, E. Chang, C. Yeh, and M. Ouhyoung. 2012. Automatic Chinese Food Identification and Quantity Estimation. In *Proc. of SIGGRAPH Asia Technical Briefs.* 1–4.

[4] G. Ciocca, P. Napoletano, and R. Schettini. 2017. Food recognition: a new dataset, experiments and results. *IEEE Journal of Biomedical and Health Informatics* 21, 3 (2017), 588–598.

[5] J. Dehais, M. Anthimopoulos, and S. Mougiakakou. 2016. Food Image Segmentation for Dietary Assessment. In *Proc. of ACM MM Workshop on Multimedia Assisted Dietary Management.*

[6] Y. Kawano and K. Yanai. 2014. Automatic Expansion of a Food Image Dataset Leveraging Existing Categories with Domain Adaptation. In *Proc. of ECCV Workshop on Transferring and Adapting Source Knowledge in Computer Vision (TASK-CV).*

[7] F. Kong and J. Tan. 2012. DietCam: Automatic dietary assessment with mobile camera phones. *Pervasive Mob. Comput.* 8, 1 (2012), 147–163.

[8] Y. Matsuda, H. Hajime, and K. Yanai. 2012. Recognition of Multiple-Food Images by Detecting Candidate Regions. In *Proc. of IEEE International Conference on Multimedia and Expo.* 25–30.

[9] A. Myers, N. Johnston, V. Rathod, A. Korattikara, A. Gorban, N. Silberman, S. Guadarrama, G. Papandreou, J. Huang, and P. K. Murphy. 2015. Im2Calories: towards an automated mobile vision food diary. In *Proc. of IEEE International Conference on Computer Vision.* 1233–1241.

[10] K. Okamoto and K. Yanai. 2016. An Automatic Calorie Estimation System of Food Images on a Smartphone. In *Proc. of ACM MM Workshop on Multimedia Assisted Dietary Management.*

[11] P. Pouladzadeh, S. Shirmohammadi, and R. Almaghrabi. 2014. Measuring calorie and nutrition from food image. In *IEEE Transactions on Instrumentation and Measurement.* 1947–1956.

[12] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. 2016. You Only Look Once: Unified, Real-Time Object Detection. In *Proc. of IEEE Computer Vision and Pattern Recognition.*

[13] J. Redmon and A. Farhadi. 2017. YOLO9000: Better, Faster, Stronger. In *Proc. of IEEE Computer Vision and Pattern Recognition.*

[14] O. Ronneberger, P. Fischer, and T. Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. *Springer* (2015), 234–241.

[15] C. Rother, V. Kolmogorov, and A. Blake. 2004. "GrabCut": Interactive Foreground Extraction Using Iterated Graph Cuts. *ACM Trans. Graph.* 23, 3 (2004), 309–314.

[16] W. Shimoda and K. Yanai. 2015. CNN-Based Food Image Segmentation Without Pixel-Wise Annotation. In *Proc. of IAPR International Conference on Image Analysis and Processing.*

[17] K. Simonyan and A Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. In *arXiv preprint arXiv:1409.1556.*