CNN-based Food Image Segmentation without Pixel-Wise Annotation

Wataru Shimoda Keiji Yanai

Department of Informatics, The University of Electro-Communications, Tokyo
1-5-1 Chofugaoka, Chofu-shi, Tokyo 182-8585 JAPAN
{shimoda-k,yanai}@mm.inf.uec.ac.jp

Abstract. We propose a CNN-based food image segmentation which requires no pixel-wise annotation. The proposed method consists of food region proposals by selective search and bounding box clustering, back propagation based saliency map estimation with the CNN model fine-tuned with the UEC-FOOD100 dataset, GrabCut guided by the estimated saliency maps and region integration by non-maximum suppression. In the experiments, the proposed method outperformed RCNN regarding food region detection as well as the PASCAL VOC detection task.

Keywords: food segmentation, convolutional neural network, deep learning, UEC-FOOD

1 Introduction

Food image recognition is one of the promising applications of visual object recognition, since it will help estimate food calories and analyze people's eating habits for health-care. Therefore, many works have been published so far [2,9,12,14,3,1,20]. However, most of the works assumed that one food image contained only one food item. They cannot handle an image which contains two or more food items such as a hamburger-and-french-fries image. To list up all food items in a given food photo and estimate calories of them, segmentation of foods is needed. Some works attempted food region segmentation [14,15,10,8].

Matsuda et al. [14] proposed to used multiple methods to detect food regions such as Felzenszwalb's deformable part model (DPM) [5], a circle detector and the JSEG region segmentation method [4]. He et al. [8] employed Local Variation [6] to segment food regions for estimating total calories of foods in a given food photo. In some works for mobile food recognition [15,10], they asked users to point rough locations of each food item in a food photo, and perform GrabCut [16] to extract food item segments.

Meanwhile, recently the effectiveness of Deep Convolutional Neural Network (DCNN) have been proved for large-scale object recognition at ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) 2012. Krizhevsky et al. [13] won ILSVRC2012 with a large margin to all the other teams who employed a conventional hand-crafted feature approach. In the DCNN approach, an input data of DCNN is a resized image, and the output is a class-label probability. That is, DCNN includes all the object recognition steps such as local feature extraction, feature coding, and learning. In general, the advantage of DCNN is that it can estimate optimal feature representations for datasets adaptively, the characteristics of which the conventional hand-crafted feature approach do not have. In the conventional approach, we extract local features such as SIFT and

SURF first, and then code them into bag-of-feature or Fisher Vector representations. Regrading food image recognition, the classification accuracy on the UEC-FOOD100 dataset [14] was improved from 59.6% [12] to 72.26% [11] by replacing Fisher Vector and liner SVM with DCNN.

By taking advantage of excellent ability of DCNN to represent objects, DCNN-based region detection and segmentation methods are proposed. RCNN [7] is the representative one of object detection, while Simonyan et al. [17] proposed a DCNN-based weakly-supervised segmentation method employing back-propagation-based saliency maps and GrabCut [16]. Both of them needs no pixel-wise annotation. The former method needs bounding box annotation, while the latter method needs even no bounding box annotation.

In this paper, we propose a new region segmentation method which combines the ideas of RCNN [7] and Simonyan et al. [17]. In RCNN, firstly, region proposals were generated by selective search [19], then extracted DCNN activation features from all the proposal, applied SVM to evaluate proposals and integrated them by non-maximum suppression to produce object bounding boxes. They fine-tuned DCNN pre-trained with ImageNet 1000 categories using the PASCAL VOC dataset having 20 categories.

Meanwhile, Simonyan et al. [17] proposed a method to generate object saliency maps by back propagation (BP) over a pre-trained DCNN, and showed it enabled semantic object segmentation by applying GrabCut [16] using saliency maps as seeds.

In this paper, we firstly obtain region proposals by selective search [19], secondly estimate saliency maps with BP-based methods over the pre-trained DCNN for each of the region proposals after aggregation of overlapped proposals, thirdly apply GrabCut using the obtained saliency maps as seeds of GrabCut, and finally apply non-maximum suppression to obtain final region results.

In the experiments, we examined food region segmentation with UEC-FOOD100 [14] and compared the proposed method and RCNN [7] regarding food detection performance in the bounding box level. In addition, we used PASCAL VOC 2007 as well. Our method outperformed RCNN by both of the dataset.

Although DCNN [11,9] has been applied to food image classification problem so far, no work tackled food image segmentation problems with DCNN-based methods. As long as we know, this is the first work to apply a DCNN-based segmentation method to food image segmentation task.

2 Proposed Method

The proposed method on DCNN-based region detection consists of the following steps as shown in Fig. 1:

(1) Apply selective search and obtain 2000 bounding box proposals at most.
(2) Group them and select bounding boxes.
(3) Perform back propagation over the pre-trained DCNN regarding all the selected bounding boxes.
(4) Obtain saliency maps by averaging BP outputs within each group.
(5) Extract segments based on the saliency maps with GrabCut.
(6) Apply non-maximum suppression (NMS) to obtain final region results.

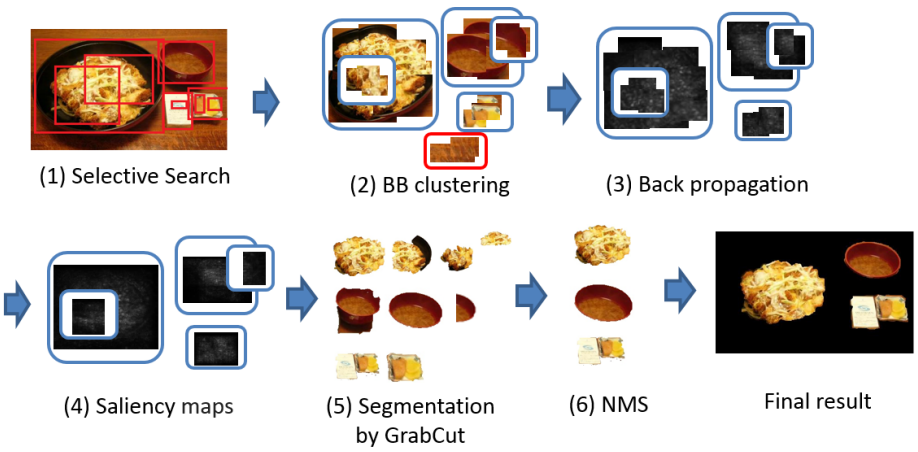


Fig. 1. The processing flow of the propose method.

2.1 Selective Search

In the work by Simonyan et al. [17], they applied their method to a whole image. This brings acceptable results for images containing only one prominent object, while it it difficult to handle images containing many objects. Especially, in case that a target image includes multiple same-class objects, Simonyan et al.'s method sometimes extracts multiple objects as one large object region and fails to extract individual object regions, since they employed GrabCut which is a generic region segmentation method.

Then, first, we apply selective search [19] to obtain food region candidates where we perform estimation of saliency maps and region segmentation, which is inspired by RCNN [7]. We obtain 2000 region proposals represented by bounding boxes at most from the selective search implementation. [1]

2.2 Bounding Box Grouping

2000 bounding boxes (BB) are too many to perform estimation of BP-based saliency maps and GrabCut within each of them. Therefore, we perform bounding box clustering to reduce the number of bounding boxes. We group the bounding boxes based on the ratio of intersection over union (IOU) into 20 BB groups at most, and we removed the groups the number of the members of which is less than 15 BBs. The rest BB groups are regarded as food region candidates. Note that BB groups sometimes contain other BB groups inside them, as shown in Fig. 1(2), because we cluster BBs according to the ratio of intersection over union (IOU).

2.3 Saliency Maps by Back Propagation over Trained DCNN

According to Simonyan et al. [17], we estimate food saliency maps which represents rough position of target objects employing back propagation (BP) over the trained

[1] Downloaded from http://koen.me/research/selectivesearch/

DCNN. In general, BP is used for training of DCNNs, which propagates errors between estimated values and ground truth values in an output layer from an output layer to an input layer in the backward direction. In case of training, the weights of DCNNs are modified so that total errors are reduced. Reducing errors is equivalent to increasing the output scores of given classes. If propagating errors to an input image, we can obtain a map indicating which pixels need to be changed to increase the scores of given classes. Such pixels are expected to correspond to the object location in the images. This is the explanation why BP can be used for object region estimation. The advantage of this method is that it does not need neither pixel-wise annotation or bounding box annotation as training data. The only thing needed is a trained DCNN with labeled images.

We estimate saliency maps of each of the selected bounding boxes (Fig. 1 (3)) and unify saliency maps within each BB group. In the experiments, we fine-tuned AlexNet [13] with the UEC-FOOD100 dataset [14] and used it to estimate food categories and saliency maps.

To perform BP, both forward pass and backward pass computation are needed. Forward pass computation is equivalent to classification by DCNN. We provide a region cropped within each selected BB to DCNN in the forwarding direction, and obtain softmax scores of all the categories. Then, we select the top five categories, and provide the vector where only the elements corresponding to the top five categories are 1 and the rest elements are 0 into the backward pass. Note that the size of an input image is fixed to 227×227 in case of using AlexNet. We resize (shrink or enlarge) cropped regions to fit to the fixed size.

To estimate object saliency maps, two other methods than the BP-based method proposed by Simonyan et al. [17] exists. One is deconvolution (deconv) proposed by Zeiler et al. [21], the other is guided back propagation (guided BP) proposed by Springerberg et al. [18]. Basic ideas of the three method are the same. Only the ways to back propagation through ReLUs (rectified linear units) are different. Refer the further detail to [18]. Originally, guided BP and deconv were proposed as visualizing methods of inside of a CNN which was regarded as a black box for analysis and understanding of it. Guided BP can emphasis edges of objects, which is good for visualizing trained filters inside a DCNN. Fig. 2 shows saliency maps, and GrabCut results obtained by the three methods.

After obtaining saliency maps of BBs, we average them within each BB group and obtain saliency maps of BB groups as shown in Fig. 1(4). The pixels with higher values are expected to correspond to objects.

2.4 Segmentation by GrabCut

In this step, we apply GrabCut [16] to each BB group region to extract whole object regions, because BP can estimate only most discriminative parts of objects. To use GrabCut, both foreground and background color models are needed. In the similar way as Simonyan et al. [17], the foreground model are estimated from the pixels with the top 3the lower 40the foreground and the background regions in the thresholded images shown in Fig. 2. Because we apply GrabCut to each BB group independently, we obtain several regions for one objects as shown in Fig.1 (5).

To integrate overlapped regions, we apply non-maximum supression (NMS), and we obtain non-overlapped regions as shown in Fig. 1 (6). Finally, we estimate rectangular regions bounding obtained segmented regions, and provide them to the trained CNN to obtained labels for each of the segmented regions. In addition, in the experiments, we use the extracted bounding boxes for evaluation.

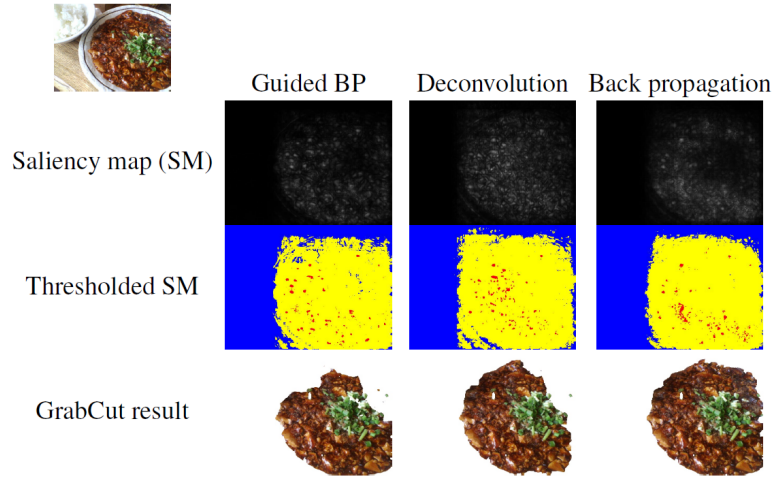


Fig. 2. Saliency maps, thresholded saliency maps and GrabCut results generated by three kinds of BP-variant methods: Guided BP, Deconvolution, Back Propagation.

3 Experiments

In the experiments, we used the UEC-FOOD100 dataset [14] and the PASCAL VOC 2007 detection dataset, both of which have bounding box information as well as class labels.

3.1 Food Detection Evaluation

The UEC-FOOD100 dataset [14] contains one hundred kinds of food photos. The total number of the food photos is 12740 including 1174 multiple-food photos. In the experiment, we used 1174 multiple-food photos including 3045 food items for testing, while we used the rest 11566 photos for fine-tuning a DCNN pre-trained with the ImageNet 1000 dataset.

For evaluation, we use mean average precision. We count it as a correct result only if the ratio of intersection over union (IOU) exceeds 50% between the detected bounding box and the ground truth bounding box. Note that we evaluated results regarding not segmentation but only bounding boxes, since UEC-FOOD has no pixel-wise annotation.

Fig.3 shows some examples of the detected BB and food regions. The red letters with yellow backgrounds represent food IDs and corresponding output scores from the DCNN. Most of the food items were correctly detected. In the top row, "[93] kinpira-style salad" was correctly detected, although it was not annotated in the ground truth data. In the bottom row, "[24] beef noodle" was detected as only half of the ground truth region due to failure of GrabCut.

Next, we compared three kinds of BP-variant methods which are used for estimating saliency maps. Tab.1 shows mean average precisions by three methods regarding estimated bounding boxes. Although the results by BP were better than the results by the other methods, the difference was not so large.

We compared our results with the results by RCNN. For RCNN as well as the proposed method, we used the same DCNN fine-tuned with the single food images of

(1) original (2) R-CNN results (3) ours results (4) ground truth

Fig. 3. The results of food region segmentation for UEC-FOOD100. (1) original food photo, (2) detected BB, (3) estimated food segments, (4) ground truth BB. ([] represents food ID: [01] rice, [05] pork cutlet, [17] humberger, [24] beef noodle, [36] miso soup, [39] oden, [93] kinpira-style salad, [94] rice ball, [98] french fries.)

UEC-FOOD 100. Tab.2 shows the results. Unexpectedly, the mean AP by RCNN was much lower than the proposed method. Fig.4 shows some example results. Compared to the bounding boxes estimated by the proposed methods, RCNN detected too small bounding boxes which cannot be counted as correct bounding boxes.

3.2 Evaluation on Pascal VOC 2007 Detection Task

For more fair comparison with RCNN [7], we also applied our method to Pascal VOC 2007 detection dataset. We used the pre-trained model on PASCAL VOC 2007 included in the RCNN package [2]. In the same way as UEC-FOOD, we compare both performance in mean average precision. The results are shown in Fig.3. Our method outperformed RCNN by 4.5 points.

4 Conclusions

In this paper, we proposed a DCNN-based food image segmentation which requires no pixel-wise annotation. The proposed method consists of food region proposals by

[2] Downloaded from https://github.com/rbgirshick/rcnn

Table 1. Mean average precision over all the 100 categories, 52 categories (more than 10 items of which are included in the test data), and 11 categories (more than 50 items of which are included in the test data).

UEC-FOOD100 mAP	100class (all)	53class (#item $\geq$ 10)	11class (#item $\geq$ 50)
guided back propagation (GBP)	**50.7**	52.5	51.4
deconvolution (deconv)	48.0	54.1	**55.4**
back propagation (BP)	49.9	**55.3**	**55.4**

Table 2. The results by RCNN and the proposed methods.

UEC-FOOD100 mAP	100class (all)	53class (#item $\geq$ 10)	11class (#item $\geq$ 50)
R-CNN	26.0	21.8	25.7
proposed method	**49.9**	55.3	**55.4**

Table 3. The results for the PASCAL VOC 2007 detection dataset.

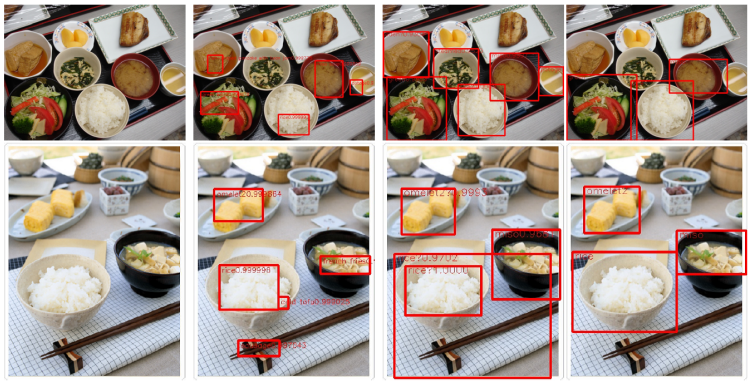
	aero	bike	bird	boat	btl	bus	car	cat	chair	cow	dtable	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
R-CNN	64.2	69.7	50.0	41.9	32.0	62.6	71.0	60.7	32.7	58.5	46.5	56.1	60.6	66.8	54.2	31.5	52.8	48.9	57.9	64.7	**54.2**
proposed	81.5	70.2	65.2	39.7	37.8	63.9	83.2	67.8	27.0	65.3	39.5	63.6	63.2	73.2	61.2	37.3	63.5	39.8	70.0	60.8	**58.7**

selective search and bounding box clustering, back propagation based saliency map estimation with the DCNN fine-tuned with the UEC-FOOD100 dataset, GrabCut guided by the estimated saliency maps and region integration by non-maximum supression. In the experiments, the proposed method outperformed RCNN regarding food region detection as well as the PASCAL VOC detection task.

For future work, we plan to implement the proposed method on mobile devices as a real-time food region recognition system for estimating more accurate food calorie intake.

References

1. Bosch, M., Zhu, F., Khanna, N., Boushey, C.J., Delp, E.J.: Combining global and local features for food identification in dietary assessment. In: Proc. of IEEE International Conference on Image Processing (2011)
2. Bossard, L., Guillaumin, M., Gool, L.V.: Food-101 - mining discriminative components with random forests. In: Proc. of European Conference on Computer Vision (2014)
3. Chen, M., Yang, Y., Ho, C., Wang, S., Liu, S., Chang, E., Yeh, C., Ouhyoung, M.: Automatic chinese food identification and quantity estimation. In: SIGGRAPH Asia (2012)
4. Deng, Y., Manjunath, B.S.: Unsupervised segmentation of color-texture regions in images and video. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(8), 800–810 (2001)
5. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(9), 1627–1645 (2010)
6. Felzenszwalb, P.F., Huttenlocher, D.P.: Image segmentation using local variation. In: Proc. of IEEE Computer Vision and Pattern Recognition. pp. 98–104 (1998)
7. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proc. of IEEE Computer Vision and Pattern Recognition. pp. 580–587 (2014)



(1) original (2) R-CNN results (3) ours results (4) ground truth

Fig. 4. Examples of the detection results by R-CNN and the proposed method.

8. He, Y., Xu, C., Khanna, N., Boushey, C.J., Delp, E.J.: Food image analysis: Segmentation, identification and weight estimation. In: Proc. of IEEE International Conference on Multimedia and Expo. pp. 1–6 (2013)
9. Kagaya, H., Aizawa, K., Ogawa, M.: Food detection and recognition using convolutional neural network. In: Proc. of ACM International Conference Multimedia. pp. 1085–1088 (2014)
10. Kawano, Y., Yanai, K.: Real-time mobile food recognition system. In: Proc. of IEEE CVPR International Workshop on Mobile Vision (IWMV) (2013)
11. Kawano, Y., Yanai, K.: Food image recognition with deep convolutional features. In: Proc. of ACM UbiComp Workshop on Workshop on Smart Technology for Cooking and Eating Activities (CEA) (2014)
12. Kawano, Y., Yanai, K.: Foodcam: A real-time food recognition system on a smartphone. Multimedia Tools and Applications pp. 1–25 (2014)
13. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (2012)
14. Matsuda, Y., Hoashi, H., Yanai, K.: Recognition of multiple-food images by detecting candidate regions. In: Proc. of IEEE International Conference on Multimedia and Expo. pp. 1554–1564 (2012)
15. Morikawa, C., Sugiyama, H., Aizawa, K.: Food region segmentation in meal images using touch points. In: Proc. of ACM MM WS on Multimedia for Cooking and Eating Activities (CEA). pp. 7–12 (2012)
16. Rother, C., Kolmogorov, V., Blake, A.: Grabcut: Interactive foreground extraction using iterated graph cuts. ACM Transactions on Graphics (TOG) 23(3), 309–314 (2004)
17. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. In: Proc. of International Conference on Learning Represenation Workshop Track (2014), http://arxiv.org/abs/1312.6034
18. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: The all convolutional net. In: Proc. of International Conference on Learning Represenation Workshop Track (2015), http://arxiv.org/abs/1412.6806
19. Uijlings, J.R.R., van de Sande, K.E.A., Gevers, T., Smeulders, A.W.M.: Selective search for object recognition. International Journal of Computer Vision 104(2), 154–171 (2013)
20. Yang, S., Chen, M., Pomerleau, D., Sukthankar, R.: Food recognition using statistics of pairwise local features. In: Proc. of IEEE Computer Vision and Pattern Recognition (2010)
21. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Proc. of European Conference on Computer Vision, pp. 818–833 (2014)