

Multi-task Learning of Dish Detection and Calorie Estimation

Takumi Ege and Keiji Yanai

Department of Informatics, The University of Electro-Communications, Tokyo
1-5-1 Chofugaoka, Chofu-shi, Tokyo 182-8585 JAPAN

ABSTRACT

In recent years, a rise in healthy eating has led to various food management applications, which have image recognition to automatically record meals. However, most image recognition functions in 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, which 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 food calories while simultaneously detecting dishes by multi-task learning of food calorie estimation and food dish detection with a single CNN. It is expected to achieve high speed and save memory by simultaneous estimation in a single network. Currently, there is no dataset of multiple-dish food photos annotated with both bounding boxes and food calories, so 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 with bounding-boxes attached and single-dish food photos with food calories. Our results show that our multi-task method achieved higher speed and a smaller network size than a sequential model of food detection and food calorie estimation.

CCS CONCEPTS

• **Computing methodologies** → *Object detection*; • **Computer systems organization** → *Real-time systems*;

KEYWORDS

food calorie estimation, food dish detection, multi-task learning



Figure 1: Examples of multiple-dish food photos.

ACM Reference Format:

Takumi Ege and Keiji Yanai. 2018. Multi-task Learning of Dish Detection and Calorie Estimation. In *CEA/MADiMa'18: Joint Workshop on Multimedia for Cooking and Eating Activities and Multimedia Assisted Dietary Management in conjunction with the 27th International Joint Conference on Artificial Intelligence IJCAI, July 15, 2018, Mässvågen, Stockholm, Sweden*. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3230519.3230594>

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 use 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 Figure 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.

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.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CEA/MADiMa'18, July 15, 2018, Mässvågen, Stockholm, Sweden
© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-6537-6/18/07...\$15.00
<https://doi.org/10.1145/3230519.3230594>

Ege et al. [2] 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, to correspond with multiple dishes, we apply object detection to multiple-dish food photos and then detect single food dishes one by one and estimate food calories. Note that the value of food calorie output 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 outputted.

A common object detection system estimates categories and bounding boxes, which identifies the position of objects, for each object in the images. Using object detection for multiple-dish food photos, it is possible to estimate bounding boxes and categories for each dish. In this case, objects of the same category are detected one by one so that multiple dishes of the same category in an image are detected one by one. With regard to object detection, it is possible to detect with high precision and high speed using CNN. In this work, we will 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.

2 RELATED WORK

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

Miyazaki et al. [4] 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 up loaded to the commercial food logging service Food-Log¹. As with our approach, their approach outputted food calorie value per serving.

One of the CNN-based researches of detection of multiple food dishes is that of Shimoda et al. [8]. In [8], 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. [9] also proposed the method which generates region proposals by CNN. In the work of Shimoda et al. [9], 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. [1] proposed the other method for food dish segmentation. In the work of Dehais et al. [1], 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. In this work, we use a CNN-based object detection system for object detection from multiple-dish food photos.

Im2Calories by Myers et al. [5] 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.

3 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 and their categories and calories from multiple-dish food photos. In this work, we use the food calorie estimator proposed by Ege et al. [2], for image-based food calorie estimation. We apply YOLOv2 [7] which is the state-of-the-art CNN-based object detector, proposed

¹<http://www.foodlog.jp/>

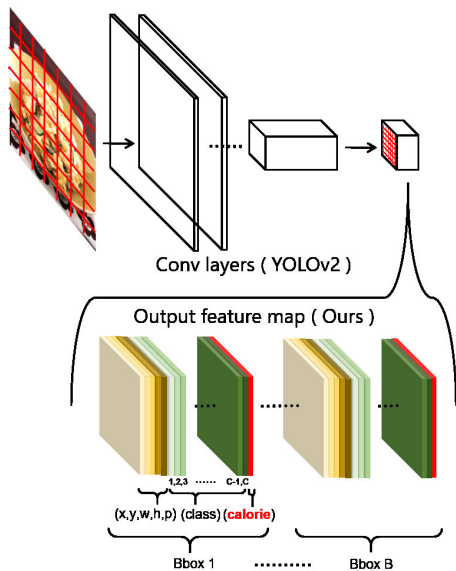


Figure 2: The architecture of YOLOv2 [7] and the output feature map of our network.

by Redmon et al. to detect dishes. YOLOv2 enabled faster and more accurate object detection by improving YOLO [6].

As shown in Figure 2, the network of YOLOv2 consists of all convolution layers and takes an input image, and then 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 $S \times S$ grids on the input image. In object detection, bounding boxes consisted of the center coordinates and the width and height, categories which the class probability corresponding to each category, and a probability that the target object exists in the grid are outputted. Hence, let B be the number of estimated bounding boxes for each grid and C be the number of categories, the number of channels of the output feature map is defined as $B \times (5 + C)$.

In this paper, we propose a network for multi-task learning of dish detection and food calorie estimation. We modified the network of YOLOv2 so that it can output a food calorie value on each food bounding box. To modify YOLOv2, we carry out multi-task learning of food calorie estimation as well as dish detection. In this proposed method, we add the output channels of estimated food calories to the output feature map so that our network estimates food calories in addition to bounding boxes and categories. Hence, the number of channels of our output feature map is defined as $B \times (5 + C + 1)$.

In this case, it is necessary to give an annotation of food calories to a ground-truth grid corresponding to the ground-truth of bounding boxes of food dishes for estimating food



Figure 3: Examples of food detection training images with pseudo-bounding boxes represented by red boxes. (Upper left: pilaf 375 kcal, upper right: simmered meat and potatoes 262 kcal, bottom left: spaghetti 391 kcal, and bottom right: hamburger steak 440 kcal)

calories for each estimated bounding box. That is, the annotation of the ground-truth grid corresponding to the ground-truth of bounding boxes is required for a calorie-annotated dataset, for estimating food calories for each estimated bounding box. However, no annotation of the ground-truth grid, such as bounding boxes in the calorie-annotated food photo dataset, exists currently [2]. Therefore, in this work, we give each image in the calorie-annotated food photo dataset a pseudo-bounding box as shown in Figure 3, using the following procedure. First, a calorie-annotated food image is embedded in a random position, and the embedded image region is set as a ground-truth bounding box. Further, in order to make the boundary line with the background inconspicuous, the same embedded image is inverted and embedded in the background portion.

3.2 Image-based food calorie estimation

In this work, we use image-based food calorie estimation based on regression learning with CNN [2] 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 the value per serving as in the case of [2]. Also, we use Equation (1) according to [2] 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



Figure 4: Examples of multi-label food photos in UEC Food-100 [3].

the loss function of Equation (1). We denote L_{ab} as an absolute error and L_{re} as a relative error, and L_{cal} is defined as follows:

$$L_{cal} = \lambda_{re}L_{re} + \lambda_{ab}L_{ab} \quad (1)$$

where λ are the weight on the loss function of each task. The absolute error is the absolute value of the difference between the estimated value and the ground-truth, and 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:

$$L_{ab} = |y - g| \quad (2)$$

$$L_{re} = \frac{|y - g|}{g} \quad (3)$$

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 [3] which includes multiple-dish food photos with a bounding box attached, and a calorie-annotated food photo dataset [2] which contains single-dish food photos with food calories.



Figure 5: Examples of calorie-annotated food photos of 15 food categories.

4.1 UEC Food-100

UEC Food-100 [3] 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 4 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 [2] collected from commercial cooking recipe sites on the web and the collected recipe data have food calorie information for one person. In this experiment, we used this dataset for food calorie estimation. Figure 5 shows example photos with food from 15 categories in a calorie-annotated food photo dataset.

5 EXPERIMENTS

We used both UEC Food-100 [3] and a calorie-annotated food photo dataset [2] 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 are alternately performed by switching the dataset by mini-batch. In the learning of the dish detection task, UEC Food-100 and the loss term related to the dish detection task are used. In the learning of the food calorie estimation task, a calorie-annotated food photo dataset and the loss term related to the food calorie estimation task are used.

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

	rel. err. (%)	abs. err. (kcal)	$\leq 20\%$ err. (%)	$\leq 40\%$ err. (%)
Single-dish (single-task) [2]	30.2	105.7	43	76
Multiple-dish (ours)	36.1	121.7	34	64

5.1 Food calorie estimation from single-dish food photos

In this experiment, we used test data in the calorie-annotated food calorie photos[2]. Following Ege et al. [2], 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 showed 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.

We used SGD as an optimization, a momentum of 0.9, and a mini-batch of 8. We used 10^{-5} of learning rate for 40,000 iterations and then used 10^{-6} for 20,000 iterations.

In this experiment, the test images are single-dish food photos; therefore, as a final output, we used an estimated bounding box with the highest probability that the target object existed in the grid. Table 1 shows the results of food calorie estimation from single-dish food photos.

In comparison with the food calorie estimation [2] that only estimates calorie content using VGG16 [10], the accuracy of our method was lower for all of the evaluation values.

Figure 6 shows an example of dish detection and food calorie estimation.


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 extracted a bounding box of a food dish by YOLOv2 [7], and obtained a cropped image corresponding to the bounding box. Then, we put the cropped image in the image-based food calorie estimation network [2] to estimate the number of food calories in the food.

The execution speed of our network with an input image with a size of 224×224 and mini-batch of 1 is approximately 22.3 ms on a GTX 1080 Ti. Additionally, the size of our network that detects dishes and estimates food calories is 181 MB.

Figure 7 shows the results of dish detection from multiple-dish food photos. We used food photos of calorie-annotated dish cards² as test data. The calorie-annotated dish cards included 131 real-size dish cards, and each dish card included

Table 2: Comparison of execution speed and model size. The sequential model is a two stage process of YOLOv2 [7] and image-based food calorie estimation [2]

	speed (msec)	model size (MB)
Sequential model	49.5 (22.3+27.2)	840 (181+659)
Multiple-dish (ours)	22.3	181



Estimated value	412 kcal	722 kcal	25 kcal	375 kcal
	Fried noodle	Curry	Miso soup	Hamburg steak
Ground truth	491 kcal	937 kcal	47 kcal	461 kcal
	Spaghetti	Curry	Miso soup	Hamburg steak
Error	-79 kcal	-215 kcal	-22 kcal	-86 kcal

Figure 6: Examples of food calorie estimation from single-dish food photos. The blue frame is the estimated bounding box.

relevant information such as food ingredients, recipes, and food calories.

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 is no dataset of multiple-dish food photos annotated with bounding boxes and food calories. We used UEC Food-100 [3] for object detection and calorie-annotated food photos [2] for food calorie estimation.

As future work, we plan to construct calorie-annotated multiple-dish food photos. As one of the methods, it is considered to create newly by learning CNN by using food images with bounding box and food images with food calorie.

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

REFERENCES

- [1] 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*.
- [2] T. Ege and K. Yanai. 2017. Simultaneous Estimation of Food Categories and Calories with Multi-task CNN. In *Proc. of IAPR International Conference on Machine Vision Applications (MVA)*.
- [3] 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.
- [4] T. Miyazaki, G. Chaminda, D. Silva, and K. Aizawa. 2011. Image-based Calorie Content Estimation for Dietary Assessment. In *Proc. of IEEE ISM Workshop on Multimedia for Cooking and Eating Activities*. 363–368.

²<http://www.gun-yosha.com/book/balanceguide.html>



Figure 7: Examples of dish detection and food calorie estimation from multiple-dish food photos. The blue frames are estimated bounding boxes. (ES: estimated value, GT: ground-truth)

- [5] 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.
- [6] 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*.
- [7] J. Redmon and A. Farhadi. 2017. YOLO9000: Better, Faster, Stronger. In *Proc. of IEEE Computer Vision and Pattern Recognition*.
- [8] 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*.
- [9] W. Shimoda and K. Yanai. 2016. Foodness Proposal for Multiple Food Detection by Training of Single Food Images. In *Proc. of ACM MM Workshop on Multimedia Assisted Dietary Management*.
- [10] K. Simonyan and A. Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. In *arXiv preprint arXiv:1409.1556*.