Comparison of Two Approaches for Direct Food 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 this paper, we compare CNN-based estimation and search-based estimation for image-based food calorie estimation. As the up-to-date direct food calorie estimation methods, we proposed a CNN-based calorie regression in [5], while Miyazaki et al. [9] proposed an image-search-based estimation method. The dataset used in the CNN-based direct estimation [5] contained 4877 images of 15 kinds of food classes, while the dataset used in the search-based work [9] consisted of 6522 images without any category information. In addition, in [9], hand-crafted features are used such as BoF and color histogram. The problems are that both the datasets are small and as far as we know there are no work to clearly compare CNN-based and search-based with the same dataset. In this work, we construct a calorie-annotated 68,774 food image dataset, and compare CNN-based estimation [5] and search-based estimation [9] with the same datasets. For the search-based estimation, we use CNN features instead of hand-crafted features used in [9].

Keywords: food recognition, image-based food calorie estimation, CNN

1 Introduction

In recent years, because of a rise in health thinking on eating, many mobile applications for recording everyday meals have been released so far. Some of them employ food image recognition which can estimate not only food names but also food calories. However, since these applications often require users to enter information such as food categories and size or volume, there are problems that it is troublesome and subjective evaluation. To solve these problems, automatic recognition of food photos on mobile devices is effective [10, 11, 15, 4, 6, 1]. However, in most of the cases, estimated calories are just associated with estimated food categories, or a relative size compared to the standard size of each food category which is usually indicated by a user manually. Currently, no applications which can estimate food calories automatically exist. Although most of the image recognition tasks including food category recognition have great progress of due to CNN-based image recognition methods [16, 2, 7], fully-automatic food calorie estimation from a food photo has still remained as an unsolved problem.

Regarding food calorie estimation, a lot of approaches have been proposed so far. The major approach is to estimate calories based on the estimated food

category and its size or volume [10, 11, 3, 4, 6, 12]. Since food calories strongly depend on food categories and volumes, this approach is effective and important. In this approach, since it is costly to create a food image dataset with pixel-wise annotation for segmentation, the number of the foods the calories of which can be estimated was very limited.

The other approach is to estimate calories from food photos directly without estimating food categories and volumes. Only two works adopted this approach [9, 5].

Miyazaki et al. [9] estimated the amount of food calories from food photos directly without estimating food categories and volumes. The biggest difficulty on direct calorie estimation is creating datasets which contains calorie-annotated food images. They hired dietitians to annotate calories on 6512 food photos which up-loaded to the commercial food logging service, Food-Log[1]. In their work, they adopted image-search based calorie estimation, in which they searched the calorie-annotated food photo database for the top k similar images based on conventional hand-crafted features. Since their method ignored information on food categories, their method was applicable for any kinds of foods. However, the number of food images in the database was not enough for the search-based method, and the employed image features was too simple.

On the other hand, we proposed a CNN-based direct food calorie estimation [5]. They employed multi-task CNNs for simultaneous estimation of food categories and calories from food photos. In [5], we collected calorie annotated recipe data from the online cooking recipe sites, and trained multi-task CNN that outputs food calories and food categories from a food photo that contained only one dish. Since there exists strong correlation between food categories and calories, we expected that simultaneous training of both brought performance boosting compared to independent single training. we were inspired by the work of Chen and Ngo [2] in which we proposed a multi-task CNN to estimate food categories and food ingredients at the same time, and proved that simultaneous estimation boosted estimation performance on both tasks. In [5], the recipe sites has various kind of foods, but we used only 15 food categories in the recipe dataset for multi-task learning of food calories and food categories. In addition, there is no fair comparison with previous works.

In this paper, regarding two representative methods on direct food calorie estimation, we compare CNN-based estimation [5] and search-based estimation [9] with the same datasets. To do that, we construct a calorie-annotated 68,774 food image dataset without food category. In addition, for the search-based method, we use CNN features instead of hand-crafted features used in [9].

2 Method

In this work we compare two methods. This section briefly describes the details of the two methods.

[1] http://www.foodlog.jp/

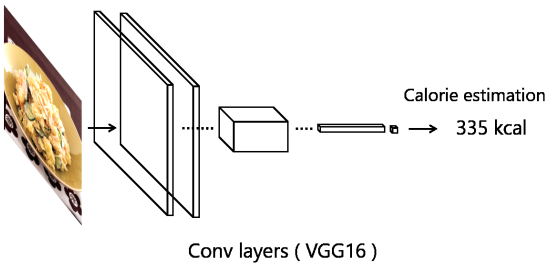


Fig. 1. The architecture for CNN-based direct food calorie estimation ([5]).

2.1 CNN-based food calorie estimation

As a direct calorie regression method, we estimate food calories from a food photos by CNN according to [5]. They collected calorie annotated recipe data from the online cooking recipe sites, and trained multi-task CNN that outputs food calories and food categories directly from a food photo that contained only one dish. According to [5], we train the network shown in Figure 1 by the recipe dataset for food calorie estimation. This network is a single-task CNN that outputs food calories only. Initially, each layer is pre-trained by the ImageNet 1000-class dataset. The architecture of this network is based on VGG16 [14]. As shown in Figure 1, only the output layer (fc8) is replaced by a single unit which outputs food calories.

According to [5], we use a loss function as shown below for food calorie estimation task.

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

L_{ab} and L_{re} are absolute error and relative error. 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 ground-truth. Generally, in the regression problem, a mean square error is used as the loss function, however in [5] we used this loss function. This loss function improves performance. where λ_{re} and λ_{ab} are the weight on the loss function, and it is usually determined so that all loss terms converge to the same value. In this work, λ_{re} and λ_{ab} are determined as follows. Firstly, the weights of the loss terms are set to 1 and train once. In the training, the values of the losses for each iteration are preserved. Finally, the reciprocal of the average value of the loss in all iterations is used as the weight for the loss term of each task. In this experiments, we fixed λ_{re} to 1.

Let y_i as the estimated value of an image x_i and g_i as the ground-truth, L_{ab} and L_{re} are defined as following:

$$L_{ab} = |y_i - g_i| \tag{2}$$

$$L_{re} = \frac{|y_i - g_i|}{g_i} \tag{3}$$

2.2 Search-based food calorie estimation

In [9], Miyazaki et al. adopted image-search based calorie estimation, in which they searched the calorie-annotated food photo database for the top k similar images based on conventional hand-crafted features such as SURF-based BoF and color histograms and estimated food calories by averaging the food calories of the top k food photos.

As an image-search based calorie estimation method, we follow this search-based method in [9]. However we use CNN features instead of conventional features such as SURF-based BoF and color histograms. In this experiments, we use VGG16 [14] which is pre-trained with the ImageNet 1000-class dataset for a feature extractor. We extract activation signals of fully connected layers (fc layers) of the VGG16 network as CNN features. Both fc6 layer and fc7 layer of VGG16 [14] are 4096-dim, so we obtain a 4096-dim feature vector for each food image. Initially, we extract CNN features for each training image and create a database of CNN features. Then, for each test image, we search the database for the top k similar images based on CNN features. Finally, we obtain a food calorie by calculating their average value of top k similar images.

3 Dataset

The datasets used in [9, 5] are small. Then, we did not perform fair comparison of CNN-base [5] and search-base [9] with 15 categories dataset used in [5]. Therefore in this work we use following two datasets for comparison between CNN-base [5] and search-base [9].

3.1 15 categories dataset

In this work, we used calorie-annotated recipe data in [5] for food calorie estimation. It costs too much to create calorie-annotated food image dataset by hand. In [5], we focused on using commercial cooking recipe sites on the web and collected recipe data which has food calorie information for one person. Then we manually collected data on 15 categories, and created a total of 4877 images dataset. In this experiment, we used this dataset for food calorie estimation by both CNN-based method and search-based method.

3.2 All recipe dataset

In [5], we used only 15 food categories in the recipe dataset for multi-task learning of food calories and food categories. In this work, in order to correspond to every category, we used all recipe data. Then, we excluded photos of multiple dishes, and photos lower than 256×256. For excluding photos with more than one dishes, we used Faster R-CNN [13] trained by UEC FOOD-100 [8] which is the food image dataset annotated bounding boxes for each image. Faster R-CNN is the basis of the latest research on object detection using CNN, and achieves high-speed and highly accurate detection. In the end, we created 68,774 food images dataset.

4 Experiments

In this experiments we compared CNN-based method and search-based method for 15 categories dataset and all recipe dataset respectively. In CNN-based method, for the test, 10 models obtained at the 100 iteration intervals from the last 1000 iterations in training were used, and the average value of the estimated values obtained from each model was taken as the final estimated value. In search-based method, we searched the database for the top k similar images based on CNN features ($k = 1, 5, 10$). Also, we used fc6 and fc7 features of VGG16 [14] as CNN features in this experiments.

4.1 Calorie estimation with 15 categories dataset

In this experiment, we used 15 categories dataset for food calorie estimation.

In CNN-based method, according to [5], we used 70% of the dataset for training of single-task CNN in Figure 1 and multi-task CNN, and the rest for testing. For optimization of the CNN, we used Momentum SGD, the momentum value 0.9. Then we used 0.001 of the learning rate for 50,000 iterations, and then 0.0001 for 20,000 iterations with size of mini-batch 8.

In search-based method, initially, we created CNN features database by 70% of the dataset, then the rest of the dataset was used for testing.

In addition, in this experiment, we estimated food calories by CNN-based classification and treated this as a baseline method. Initially, we calculated an average value of food calories for each food category using 70% of the dataset. Then for each test image, we estimated food category by CNN, and regarded the average calorie values over the estimated category as the estimated food calorie. In baseline, we finetuned VGG16 [14] for 15-class food classification. Then we used 0.001 of the learning rate for 20,000 iterations with size of mini-batch 8.

Table 1 shows the result of food calorie estimation with 15 categories dataset. 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 estimated value and ground-truth and the ratio of the estimated value within the relative error of 20% and 40%.

Figure 2 (b) shows the relation between the ground truth values and the estimated calorie values by search-based method, while Figure 2 (c) shows the relation between that by the CNN-based method.

Table 1 indicates the performance improve by single-task CNN. Compared with search-base method (fc6, $k = 15$), in single-task CNN, 18.3% and 8.8 kcal were reduced on the relative error and the absolute error, and 0.069 and 8.1% were increased on the correlation coefficient and the ratio of the estimated calories within 40% error. However, single-task CNN is not much different from the baseline.

Table 1. Comparison of CNN-based method and search-based method with 15 categories dataset. In search-based method, we used fc6 and fc7 feature vectors and the top k similar images. The feature vector of fc6+fc7 means that fc6 and fc7 are concatenated. Multi-task CNN [5] is simultaneous learning of food categories and calories.

	rel. err.(%)	abs. err.(kcal)	correlation	20% err.(%) ⩒	40% err.(%) ⩒
Baseline	32.4	93.6	0.784	50.0	76.8
fc6 (4096-d), k=5	47.9	117.2	0.673	43.1	68.8
fc6 (4096-d), k=10	47.4	111.9	0.699	45.4	70.1
fc6 (4096-d), k=15	47.4	110.4	0.707	45.6	70.5
fc7 (4096-d), k=5	52.5	119.2	0.657	42.8	69.3
fc7 (4096-d), k=10	52.3	116.5	0.675	44.2	69.7
fc7 (4096-d), k=15	54.0	116.3	0.672	44.0	69.2
fc6+fc7 (8192-d), k=5	48.2	117.4	0.673	43.0	69.4
fc6+fc7 (8192-d), k=10	47.4	112.2	0.698	45.3	70.1
fc6+fc7 (8192-d), k=15	47.6	110.7	0.706	45.5	71.1
Single-task CNN [5]	29.1	101.6	0.776	45.5	78.6
Multi-task CNN [5]	27.2	96.2	0.805	48.3	80.2

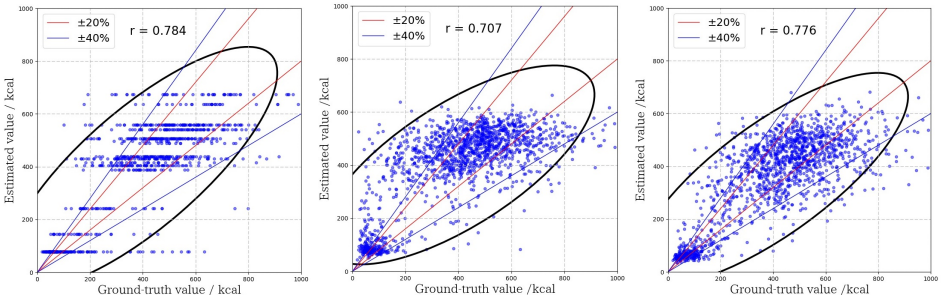
4.2 Calorie estimation with all recipe dataset

In this experiment, we used all recipe dataset. In CNN-based method, we trained single-task CNN in Figure 1 by 80% of the dataset. In this case, since we cannot use food category, we use only single-task without multi-task. We used 0.001 of the learning rate for 150,000 iterations, and then used 0.0001 for 50,000 iterations. In search-based method, initially, we created CNN features database by 80% of the dataset, then the rest of the dataset was used for testing.

Table 2 shows the result of food calorie estimation with all recipe dataset. In Table 2, similar to 15 categories dataset, it was confirmed that the CNN-based regression is superior that the search-based methods. However, their difference is not so significant.

5 Discussion

Compared with search-base method, the CNN-based method showed improvement in performance. However, in calorie estimation with 15 categories dataset, single-task CNN is not much different from the baseline. Compared with the performance of single-task CNN with 15 categories dataset, that with all recipe dataset was significantly lower. Because of these facts, it seems that it is effective to consider food category for food calorie estimation, because there are strong correlation between food calories and food categories. In addition, for calorie estimation, we think it is necessary to recognize food ingredients and sizes explicitly as well as food categories. In order to realize highly accurate food calorie



(a) Baseline (Classification). **(b) Search-based method (fc6 feature vector, $k = 15$).** **(c) CNN-based method (Single-task CNN).**

Fig. 2. The relation between the ground-truth values and the estimated calorie values.

Table 2. Comparison of CNN-based method and search-based method with all recipe dataset. In search-based method, we used fc6 and fc7 feature vectors and the top k similar images. The feature vector of fc6+fc7 means that fc6 and fc7 are concatenated.

	rel. err.(%)	abs. err.(kcal)	correlation	20% err.(%) ↑	40% err.(%) ↑
fc6 (4096-d), k=15	122.5	141.6	0.353	25.1	47.8
fc7 (4096-d), k=15	128.4	144.7	0.329	24.2	46.8
fc6+fc7 (8192-d), k=15	122.3	141.6	0.350	25.1	47.6
Single-task CNN [5]	60.0	132.0	0.436	23.5	48.0

estimation, it is considered to be urgent to create high-quality datasets. It is biggest issue how to build a large-scale calorie-annotated food photo dataset.

6 Conclusion

In this paper, we estimated food calories directly from a food photo. In this experiments we compared CNN-based method and search-based method for 15 categories dataset and all recipe dataset respectively. Then, compared with search-base method, CNN-based method showed improvement in performance.

As future work, we plan to combine object detection to calorie estimation, because it is necessary to recognize more detailed information such as food ingredients and multiple objects. In parallel we prepare a calorie-annotated food image dataset for highly accurate food calorie estimation.

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

References

1. V. Bettadapura, E. Thomaz, A. Parnami, D. G. Abowd, and A. Essa. Leveraging context to support automated food recognition in restaurant. In *Proc. of the 2015 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2015.
2. J. Chen and C. W. Ngo. Deep-based ingredient recognition for cooking recipe retrival. In *Proc. of ACM International Conference Multimedia*, 2016.
3. M. Chen, Y. Yang, C. Ho, S. Wang, E. Liu, E. Chang, C. Yeh, and M. Ouhyoung. Automatic chinese food identification and quantity estimation. In *Proc. of SIGGRAPH Asia Technical Briefs*, page 29, 2012.
4. J. Dehais, M. Anthimopoulos, and S. Mougiakakou. Gocarb: A smartphone application for automatic assessment of carbohydrate intake. In *Proc. of ACM MM Workshop on Multimedia Assisted Dietary Management*, 2016.
5. T. Ege and K. Yanai. Simultaneous estimation of food categories and calories with multi-task cnn. In *Proc. of IAPR International Conference on Machine Vision Applications(MVA)*, 2017.
6. F. Kong and J. Tan. Dietcam: Automatic dietary assessment with mobile camera phones. In *Proc. of Pervasive and Mobile Computin*, pages 147–163, 2012.
7. N. Martinel, G. L. Foresti, and C. Micheloni. Wide-slice residual networks for food recognition. In *arXiv preprint arXiv:1612.06543*, 2016.
8. Y. Matsuda, H. Hajime, and K. Yanai. Recognition of multiple-food images by detecting candidate regions. In *Proc. of IEEE International Conference on Multimedia and Expo*, 2012.
9. T. Miyazaki, G. Chaminda, D. Silva, and K. Aizawa. Image- based calorie content estimation for dietary assessment. In *Proc. of IEEE ISM Workshop on Multimedia for Cooking and Eating Activities*, 2011.
10. 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. In *Proc. of IEEE International Conference on Computer Vision*, 2015.
11. K. Okamoto and K. Yanai. An automatic calorie estimation system of food images on a smartphone. In *Proc. of ACM MM Workshop on Multimedia Assisted Dietary Management*, 2016.
12. P. Pouladzadeh, S. Shirmohammadi, and R. Almaghrabi. Measuring calorie and nutrition from food image. In *IEEE Transactions on Instrumentation and Measurement*, pages 1947–1956, 2014.
13. S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.
14. K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *arXiv preprint arXiv:1409.1556*, 2014.
15. R. Tanno, K. Okamoto, and K. Yanai. Deepfoodcam: A dcnn-based real-time mobile food recognition system. In *Proc. of ACM MM Workshop on Multimedia Assisted Dietary Management*, 2016.
16. K. Yanai and Y. Kawano. Food image recognition using deep convolutional network with pre-training and fine-tuning. In *Proc. of IEEE International Conference on Multimedia and Expo*, 2015.