Food Image Recognition with Deep Convolutional Features

Yoshiyuki KAWANO
Department of Informatics
The University of
Electro-Communications,
Tokyo
1-5-1 Chofugaoka, Chofu-shi,
Tokyo, 182-8585 Japan
kawano-y@mm.inf.uec.ac.jp

Keiji YANAI
Department of Informatics
The University of
Electro-Communications,
Tokyo
1-5-1 Chofugaoka, Chofu-shi,
Tokyo, 182-8585 Japan
yanai@mm.inf.uec.ac.jp

Abstract

In this paper, we report the feature obtained from the Deep Convolutional Neural Network boosts food recognition accuracy greatly by integrating it with conventional hand-crafted image features, Fisher Vectors with HoG and Color patches. In the experiments, we have achieved 72.26% as the top-1 accuracy and 92.00% as the top-5 accuracy for the 100-class food dataset, UEC-FOOD100, which outperforms the best classification accuracy of this dataset reported so far, 59.6%, greatly.

Author Keywords

food recognition, Deep Convolutional Neural Network, Fisher Vector

Introduction

Food image recognition is one of the promising applications of object recognition technology, since it will help estimate food calories and analyze people's eating habits for healthcare. Therefore, many works have been published so far [2, 4, 7, 9, 11]. To make food recognition more practical, increase of the number of recognizable food is crucial. In [7, 9], we created 100-class food dataset, UEC-FOOD100, and made experiments with 100-class food classification. The classification accuracy reported so far was 59.6% [7], which was not enough for practical use.

UbiComp'14, September 13–17 2014, Seattle, WA, USA.

ACM 978-1-4503-3047-3/14/09 $15.00.
http://dx.doi.org/10.1145/2638728.2641339

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. [8] 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 [8], 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 HoG first, and then code them into bag-of-feature or Fisher Vector representations.

However, DCNN is not always applicable for any kinds of datasets, because it requires a lots of training images to achieve comparable or better performance to the conventional local-feature-based methods. In our preliminary experiments on DCNN-based food recognition where we trained DCNN with the UEC-FOOD100 dataset, we failed to confirm that the DCNN-based method outperformed the conventional method. This is mainly because the amount of training data is not enough. We had only 100 images per food category, while ILSVRC dataset has 1000 images per category. In general, DCNN does not work well for a small-scale dataset, while DCNN works surprisingly well for a large-scale dataset [6]. Then, as a method to utilize DCNN for a small-scale dataset, using a pre-trained DCNN with a large-scale ILSVRC dataset as a feature vector extractor has been proposed [3]. DCNN features can be easily extracted from the output signals of the layer just before the last one of the pre-trained DCNN. Chatfield et al. made comprehensive experiments employing both DCNN features and conventional features such as SIFT and Fisher Vectors on PASCAL VOC 2007 and Caltech-101/256 which can be regarded as small-scale datasets where they have only about one hundred or less images per class [3]. They showed that the DCNN-feature was effective for a small-scale dataset, and they achieved the best performance for PASCAL VOC 2007 and Caltech-101/256 by combining DCNN features and Fisher Vectors.

Regarding food datasets, the effectiveness of the DCNN features is still unclear, because food datasets are a kind of fine-grained datasets which is different from generic datasets such as PASCAL VOC 2007 and Caltech-101/256. In food datasets, images belonging to different categories sometimes look very similar to each other. Food image recognition is regarded as the more difficult task than image recognition of generic categories. Then, in this paper, we apply DCNN features for 100-class food dataset and examine the effectiveness of DCNN features for food photos by following Chatfield et al.'s work [3].

Methods

DCNN Features

Recently, it has been proved that Deep Convolutional Neural Network (DCNN) is very effective for large-scale object recognition. However, it needs a lot of training images. In fact, one of the reasons why DCNN won the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) 2012 is that the ILSVRC dataset contains one thousand training images per category [8]. This situation does not fit food datasets most of which have only about one hundred images a food category. Then, to make the best use of DCNN for food recognition, we use the pre-trained DCNN with the ILSVRC 1000-class dataset as a feature extractor.

Following [3], we extract the network signals just before the last layer of the pre-trained DCNN as a DCNN feature vector. Since we used the same network structure proposed by Krizhevsky et al. [8], the number of elements in the last layer is the same as the number of the classes, 1000, and the number of elements in the layer just before the last one is 4096. Therefore, we obtain a 4096-dim DCNN feature vector for a food image. As implementation of DCNN, we used OverFeat [1].

Conventional Features

As conventional features, we extract RootHoG patches and color patches, and code them into Fisher Vector (FV) representation with Spatial Pyramid with three levels (1x1+3x1+2x2). Fisher Vector is known as a state-of-the-art coding method [10].

RootHoG is an element-wise square root of the L1 normalized HOG, which is inspired by "RootSIFT" [1]. The HOG we use consists of 2×2 blocks (totally four blocks). We extract gradient histogram regarding eight orientations from each block. The total dimension of a HOG Patch feature is 32. After extraction of HOG patches, we convert each of them into a "RootHOG".

As color patches, we extract mean and variance values of RGB value of pixels from each of 2×2 blocks. Totally, we extract 24-dim Color Patch features.

After extracting RootHoG patches and color patches, we apply PCA and code them into Fisher Vectors (FV) with the GMM consisting of 64 Gaussians. As results, we obtain a 32768-dim RootHOG FV and a 24576-dim Color FV for each image. This setting is almost the same as [7] except for the number of spatial pyramid levels.

Classifiers

We use one-vs-rest linear classifiers for 100-class food classification. For integrating both DCNN features and conventional features, we adopt late fusion with no weighting. For lower-dimensional DCNN features, we use a standard linear SVM, while for higher-dimensional FV features, we use an online learning method, AROW [5]. As their implementations, we use LIBLINEAR [2] and AROWPP [3].

Experiments

As a food dataset for the experiments, we use the UEC-FOOD100 dataset [7, 9] which is an open 100-class food image dataset [4]. Part of the food categories in the UEC-FOOD100 dataset is shown in Fig. 1. It includes more than 100 images for each category and bounding box information which indicates food location within each food photo. We extract features from the regions inside the given bounding boxes following [7]. We evaluate the classification accuracy within the top N candidates employing 5-fold cross validation.

Figure 2 shows classification accuracy within the top-N candidates with each of single features, RootHOG FV, Color FV and DCNN, the combination of RootHoG and Color FV, and the combination of all the three features.

[1] http://cilvr.nyu.edu/doku.php?id=software:overfeat:start

[2] http://www.csie.ntu.edu.tw/∼cjlin/liblinear/

[3] https://code.google.com/p/arowpp/

[4] http://foodcam.mobi/dataset/

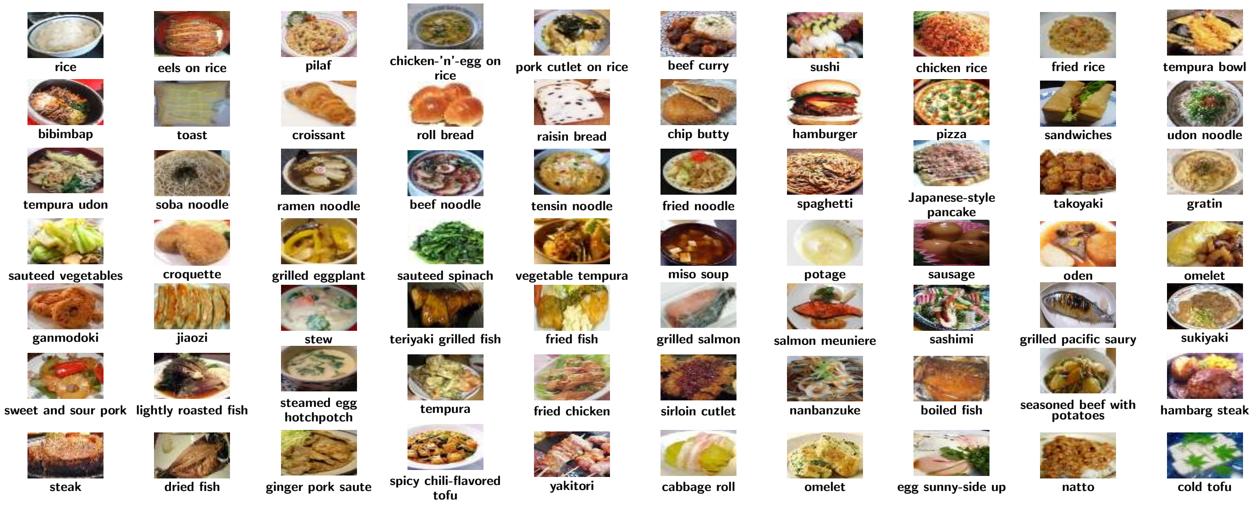

Figure 1: 70 kinds of foods in the UEC-FOOD100 dataset.

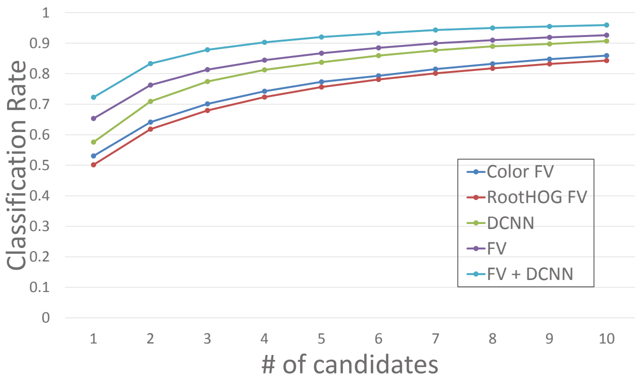

Figure 2: Classification accuracy within the top N candidate on UEC-FOOD100 with DCNN, RootHoG-FV, Color-FV and their combinations.

Among the three single features, DCNN, RootHoG-FV, and Color-FV, the DCNN feature achieved the best performance, 57.87%, in the top-1 accuracy, while RootHoG-FV and Color-FV achieved 50.14% and 53.04%, respectively. Although the combination of both FVs achieved 65.32% which was better than single DCNN features, the total dimension of the FV combination was 57,344, which 14 times as larger as the dimension of DCNN features.

The combination of all the three features achieved 72.26% in the top-1 accuracy and 92.00% in the top-5 accuracy, which were the best performance for the UEC-FOOD100 dataset, while the previous best was 59.5% [7]. This indicates that DCNN features has different characteristics from the conventional local features and Fisher Vectors, and integration of them is important to achieve better performance rather than use of single ones. This is a very promising result for practical use of food image recognition technology.

Conclusions

In this work, we proposed introducing DCNN features which are extracted from the pre-trained DCNN with the ILSVRC 1000-class dataset into food photo recognition. In the experimental results, we have achieved the best classification accuracy, 72.26%, for the UEC-FOOD100 dataset, which proved that that DCNN features can boosted the classification performance by integrating it with the conventional features.

For future work, we will implement the proposed framework on mobile devices. To do that, it is needed to reduce the amount of the pre-trained DCNN parameters which consist of about 60 million floating values.

References

[1] Arandjelovic, R., and Zisserman, A. Three things everyone should know to improve object retrieval. In *Proc. of IEEE Computer Vision and Pattern Recognition* (2012), 2911–2918.

[2] Bosch, M., Zhu, F., Khanna, N., Boushey, C. J., and Delp, E. J. Combining global and local features for food identification in dietary assessment. In *Proc. of IEEE International Conference on Image Processing* (2011).

[3] Chatfield, K., Simonyan, K., Vedaldi, A., and Zisserman, A. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531* (2014).

[4] Chen, M., Yang, Y., Ho, C., Wang, S., Liu, S., Chang, E., Yeh, C., and Ouhyoung, M. Automatic chinese food identification and quantity estimation. In *SIGGRAPH Asia 2012 Technical Briefs* (2012).

[5] Crammer, K., Kulesza, A., and Dredze, M. Adaptive regularization of weight vectors. In *Advances in Neural Information Processing Systems* (2009), 414–422.

[6] Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531* (2013).

[7] Kawano, Y., and Yanai, K. Foodcam: A real-time food recognition system on a smartphone. *Multimedia Tools and Applications* (2014), 1–25.

[8] Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (2012).

[9] Matsuda, Y., Hoashi, H., and Yanai, K. Recognition of multiple-food images by detecting candidate regions. In *Proc. of IEEE International Conference on Multimedia and Expo* (2012), 1554–1564.

[10] Perronnin, F., Sánchez, J., and Mensink, T. Improving the fisher kernel for large-scale image classification. In *Proc. of European Conference on Computer Vision* (2010).

[11] Yang, S., Chen, M., Pomerleau, D., and Sukthankar, R. Food recognition using statistics of pairwise local features. In *Proc. of IEEE Computer Vision and Pattern Recognition* (2010).