ChinFood1000: A Large Benchmark Dataset for Chinese Food Recognition

Zhihui Fu, Dan Chen, and Hongyu Li(✉)

ZhongAn Information Technology Service Co., Ltd, Shanghai, China
{fuzhihui, chendan, lihongyu}@zhongan.io

Abstract. In this paper, we introduce an 1000-category food dataset ChinFood1000 and propose a simple and effective baseline approach. To our best knowledge, the proposed ChinFood1000 dataset enjoys the largest number of food categories among all publicly available food dataset currently. The categories of the ChinFood1000 dataset are carefully selected to include the most popular Chinese dishes. The dataset includes 852 categories of Chinese dishes, together with 91 classes of drinks and snacks, 26 kinds of fruits and 31 kinds of other food. The images in the dataset present both large inter-class affinity and high intra-class variance. To illustrate the challenges presented by the dataset, a baseline based on a very deep CNN is proposed. In the experiments, the baseline approach is evaluated on three most widely used food datasets and achieves the best performance on all of them. The baseline approach is also applied to the ChinFood1000 dataset, with a promising accuracy reported.

Keywords: Chinese food recognition · Benchmark dataset · CNN · Chin-Food1000

1 Introduction

Food recognition is an attractive computer vision problem, which can be regarded as a basic building block, aiding tasks including but not limited to visual calorie estimation, recording and analyzing of food intaking and food recipe matching.

Different to typical objects like dogs and cars, food rarely has structured layout, which makes Chinese food classification more challenging. Firstly, the ingredients of Chinese food have various shapes, such as stripes, chops and blocks. Secondly, Chinese food have many cooking styles, leading to complex mixing ways of the ingredients, e.g. stewing, steaming and potting. In addition, Chinese food has a great variety of dishes. There exists at least eight series of traditional Chinese cuisines, each of which is composed of hundreds of dishes.

For now, there have been limited food datasets publicly available. PFID [1] introduced a small dataset of fast food, containing only 13 classes. Moreover, the 4545 still images and 606 stereo pairs of fast food in the dataset are captured in labs and restaurants. UEC100 [2] is mainly composed of Japanese dishes, with 100 categories. The images are provided together with bounding boxes to include both the food objects and the environment. The number of food categories is enlarged to 256 in the dataset UEC256 [5]. But it still focuses on the Japanese food rather than the Chinese dishes.

D.-S. Huang et al. (Eds.): ICIC 2017, Part I, LNCS 10361, pp. 273–281, 2017.
DOI: 10.1007/978-3-319-63309-1_25

Bossard et al. [6] introduced a western food dataset Food101, with high noise resulting in the unsatisfactory classification accuracy in the practical applications. VIREO Food-172 [7] is the only available Chinese food dataset until now. It collected 172 classes of Chinese food images labeled by both food category and food ingredients. However, the number of food categories is still small, which is hard to capture the broadness and variety of the Chinese food. As a first step to increase the number of categories for Chinese food, we introduce an 1000-category food dataset, namely ChinFood1000. The 1000 categories are selected to contain the most popular Chinese food. Figure 1 shows some examples of the images from the ChinFood1000 dataset. The images in the dataset demonstrate large inter-class similarity and intra-class variance. The first two rows illustrates the large inter-class similarities. For example, Braised tofu (row2 left) and Mapo tofu (row 2 right) are very visually similar, containing the same major ingredient tofu and share the red color. The last three rows demonstrate the large intr-class variance, e.g. Vermicelli with Minced Pork (row 5 left) has numerous cooking styles, different shapes and colors within the category.

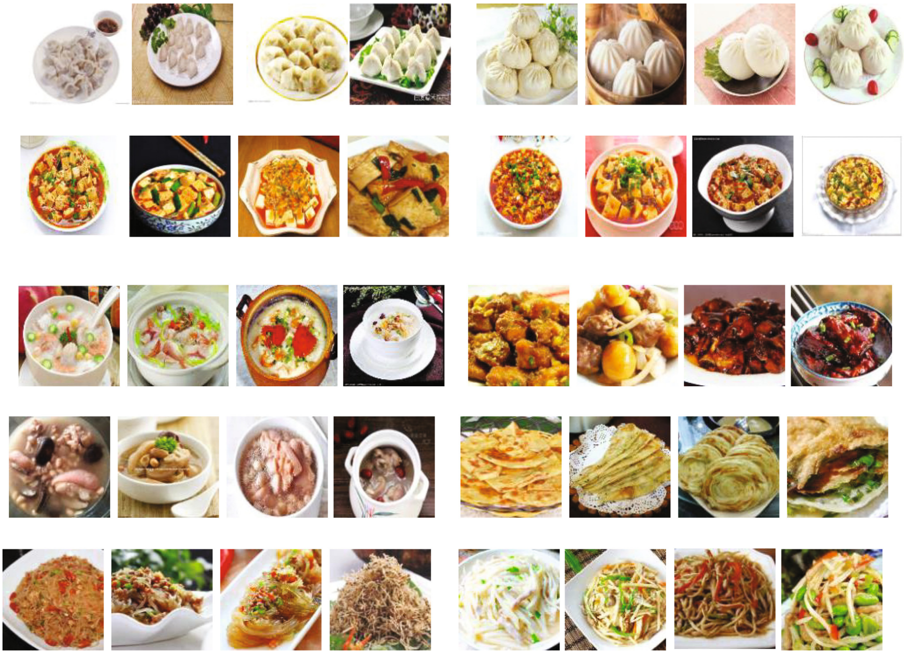


Fig. 1. Examples from ChinFood1000. Each row contains two different groups of Chinese dishes. Note the images in the dataset demonstrate large inter-class similarity and intra-class variance. The first two rows show the examples of the inter-class similarities in the dataset. For example, Braised tofu (row 2 left) and Mapo tofu (row 2 right) are very visually similar, containing the same major ingredient tofu and with close red color. The last three rows show examples of the high intra-class variances, e.g. Vermicelli with Minced Pork (row 5 left) has numerous cooking styles, different shapes and colors within the category.

Following traditional object recognition approaches, methods based on color histogram or SIFT BoW features have been widely applied to food classification [1, 3]. To further improve the robustness to the noise, Random Forest is utilized to mine discriminative parts of specific food [6]. However, the hand designed features is limited to low-level representations such as edges and textures, which restricts the classification accuracy. Liu C. et al. proposed a food recognition method [8] based on DCNN. Although the adopted network uses an efficient bottleneck structure, the accuracy is still limited.

In this paper, we propose a simple but yet effective baseline approach. Inspired by the strong modeling ability and the ease of optimization of the ResNet [9], we finetuned a deep 50-layer ResNet on ChinFood1000 as a feature learner. Upon it, a One-Vs-All logistic regressor is trained. Under the multi-stage learning framework, the number of food categories can be easily expanded and the classifiers can be trained in an online learning style. The baseline approach is evaluated on three publicly available food datasets. It is shown that the baseline approach achieves the best performances on the UEC and the Food101 datasets. It also achieves a promising accuracy on the ChinFood1000 dataset.

The rest of the paper is organized as follows: We first describe how we collect the ChinFood1000 dataset and analyse its properties in Sect. 2. The proposed baseline approach is described in Sect. 3. Experimental results are presented in Sect. 4 to show the effectiveness of the baseline approach and the challenges in the ChinFood1000 dataset.

2 The ChinFood1000 Dataset

In this paper, we introduce a 1000 categories data, the ChinFood1000 dataset, mainly composed of Chinese dishes. Some drinks, fruits and snacks are also included. To our best knowledge, for now the number of food categories is the largest among all the publicly available food datasets.

2.1 Dataset Collection

The dataset collection is divided into two steps. The names of the dishes to collect are firstly specified and selected with a naming strategy. Then images are crawled and cleaned according to the selective names.

We firstly specify the category names of the food in the dataset. The name selection strategy is shown in Fig. 2. The collected food in the dataset should be non-duplicated, and should contain the most ordinary dishes in China. Several lists of food names are obtained from Chinese cooking websites ‘Xinshipu’[1], ‘CNDZYS’[2] and diet website ‘39Jiankang’[3]. In addition to food names, we also crawled the food ingredients. As a

[1] http://www.xinshipu.com/.

[2] http://www.cndzys.com/.

[3] http://fitness39.net/.

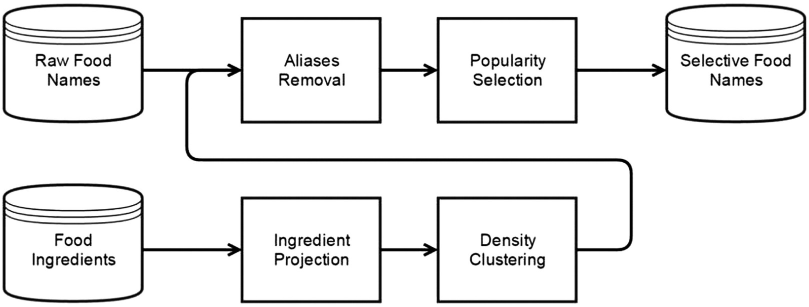


Fig. 2. The strategy for food name selection. In the projected ingredient space, density clusters are utilized to find name aliases. After aliases removal, popular food names are filtered out to get the selective food names.

result, each food in the lists corresponds a food ingredient vector. To obtain the most popular food in China, all lists are intersected to form the raw food names. However, a food name may have numerous aliases. For example, dumpling and Jiaozi are the same food. Fortunately, we observe that aliases usually share very similar major ingredients. Donate S as ingredient matrix, each row S_i is an ingredient vector contains 0s and 1s, where 1 indicates the ingredient is used in the food and vice verse. We aim to find an ingredient space that projects the ingredients to the directions with the most variations:

$$T = W\hat{S} \tag{1}$$

Where $\hat{S}$ is the centering matrix of S,W is the left singular vectors of $\hat{S}$ To find density peaks, density clustering such as DBSCAN is performed on T. After that, the density peaks are manually looped to remove the food aliases. In addition, it is assumed that the length of the names for popular food tends to be short according to the Chinese food traditions. Any name longer than a specific threshold is discarded to obtain the selective results. In practice, the threshold is configured as 7 Chinese characters. Finally, we manually add some popular food and obtain the names of food in the dataset. As a result, the food category names include 852 Chinese dishes, 91 classes of drinks and snacks, 26 kinds of fruits adn 31 kinds of other food.

With the selective food names as keyword, the relative images were down- loaded from the Baidu image search engine[4]. The search results are truncated and only the most relevant images are preserved.

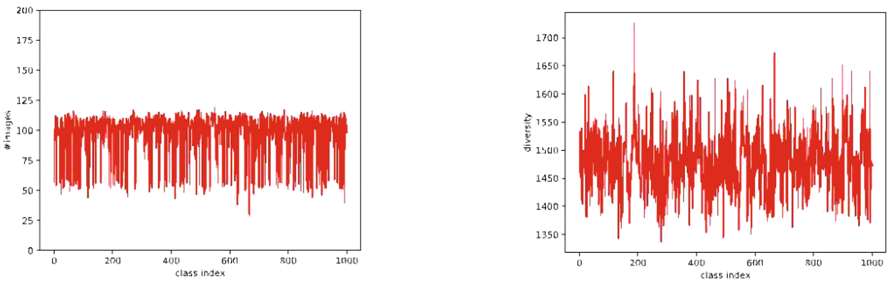
In addition, since a certain image may appear in the search results of different food categories due to the close major ingredients, we perform filtering [12] on the crawled images by re- moving replicative pictures. During filtering, a table is constructed to remove the replicative images, where the keys are the MD5 hash of the image files and the value of each key is the occurrence in the dataset. Finally, we double check the images to avoid the occurrence of the irrelevant images.

[4] Baidu image search: http://image.baidu.com.

2.2 Dataset Properties

In this section, three different properties of the ChinFood101 dataset are analysed. Images Per Class and Noise shows that the images of each class are fairly balanced and the labeling is accurate. Meanwhile we define Diversity to show the intra-class variances of the dataset.

Images Per Class. We plot the distribution of the number of images for each category in Fig. 3. The amount of images of each class is fairly balanced. Interestingly, the number of images of a more popular category tends to be larger and vice verse. A possible reason is that the images of popular food is more likely to be taken and uploaded to the Internet. Thus the images are more accessible by the Internet image search engines.



(a) Number of images per class

(b) Diversity of images per class.

Fig. 3. Dataset properties of ChinFood1000. Left: the number of images per class; Right: the diversity of images per class. The diversity is measured by the Lossless JPG file size of the average images in each class.

Noise. To estimate the degree of noise of the dataset, we randomly pick 5% images from the dataset and check validity of the corresponding labelings. The correctness of our labeling is about 86%, which proves that the noise is acceptable.

Diversity. Inspired by ImageNet [13], the average image of each category is computed. Lossless JPG encoding is performed on the average image. Intuitively, the average of diverse images tends to be blurrier and vice verse. We compute the average images of each category, the file sizes are depicted in Fig. 3. The classes in ChinFood1000 provides a large variance in intra-class diversity.

3 The Baseline Approach

We adopt a multi-stage learning framework for the baseline approach. There has been several papers [4, 8] that adopts DCNN as the feature learner for food recognition. However, to our best knowledge, ultra deep networks have not been used in the food classification. In this paper, a Residual Convolutional Neural Network model [9] is utilized for feature learning. The network has 3 bottleneck units with 256d output, 4 with 512d output, 6 with 1024d output and 3 with 2048d output. Every bottleneck unit has a 1 × 1 conv layer followed by a 3 × 3 conv layer and a 1 × 1 conv layer. Identity mapping is used to ease the difficulty of optimization.

The network is initialized by pretrained parameters on ImageNet. We fine- tune the network parameters via SGD to minimize the softmax log loss. The batch size is fixed as 10. The learning rate is initialized as 1e5 for the last fc layer and 1e4 for all the other layers. The learning rate is decreased by a factor of 0.1 after each epoch. The network is finetuned for 5000 iterations.

On the next stage, the output of the average pooling layer is used as the features. We apply One-Vs-All Logistic Regression on the features to classify the food. Under the multi-stage learning framework, it is relatively easy to expand the number of categories of food. Moreover for a light-weight expansion, it is very efficient since there is no need to retrain the feature learner. As a result, The baseline approach is simple but effective.

4 Experiments

4.1 Experiment Setup

The experiments are all performed on Linux servers equipped with Nvidia GPU GTX1080. To show the effectiveness of the baseline approach, we evaluate it on three publicly available food benchmarks: UEC100 [2], UEC256 [5] and Food101 [6]. We then evaluate the baseline approach upon the proposed ChinFood1000 dataset to show the challenges provided by the dataset. We split the Chin-Food1000 dataset into training/testing data. Totally 20% of images are used for testing. The training and testing splits are fixed for the convenience of bench- marking. Two measures, top-1 and top-5 accuracies, are adopted to estimate the performance of the baseline approach.

4.2 Baseline Results and Discussion

In this section, we firstly evaluate the proposed baseline approach. The base- line approach is evaluated on the following datasets: UEC100 [2], UEC256 [5] and Food101 [6]. We compare the baseline approach to both deep CNN based methods and the traditional methods. Specifically, the following approaches are utilized for comparison:

DeepFood DeepFood [8] is a state-of-the-art CNN based food classification algorithm. The authors adopt a finetuned GooLeNet [10] and report a great margin of accuracy gain over traditional methods based on hand designed features.

Deep-UEC Deep-UEC [4] is another DCNN based food classification method. OverFeat [11] is adopted as the feature learner. An One-Vs-All linear learner is used for classification.

Food101 Food101 [6] introduces a method to mine discriminative parts via Random Forests (RF), enabling mining parts simultaneously for all categories and sharing knowledge among them.

The comparison results are shown in Table 1. The proposed baseline approach achieves the best accuracies on all the three datasets. The pro- posed approach and the state-of-the-art DCNN based food classifier DeepFood both adopt a deep CNN based framework. The difference is that DeepFood utilizes a wider but shallower network while the proposed approach uses a thinner and deeper Residual Neural Network, which indicates that for food recognition, a well trained deeper network can model the food appearances better.

Table 1. Comparison with state-of-the-art methods on UEC100, UEC256 and Food101 datasets. The result of the baseline approach on the proposed Chin-Food1000 dataset is also reported.

Dataset	Method	Top-1	Top-5
UEC100	DeepFood [8]	76.3	94.6
	Deep-UEC [4]	72.2	92.0
	Proposed baseline	**80.6**	**95.9**
UEC256	DeepFood [8]	63.8	87.2
	Proposed baseline	**71.2**	**91.0**
Food101	DeepFood [8]	77.4	93.7
	Food101 [6]	50.76	NA
	Proposed baseline	**78.5**	**94.1**
ChinFood1000	Proposed baseline	44.1	68.4

Compared to the DeepFood method, the proposed baseline approach has a larger accuracy gain on the UEC datasets than the Food101 dataset, i.e. on average 5.85% top-1 accuracy gain vs. 1.1% top-1 accuracy gain. The UEC dataset is mainly composed of Japanese food, which is close to the Chinese dishes, implying that the proposed baseline is more good at modeling data like Chinese food.

In addition, we report the top-1 and top-5 accuracies of the baseline approach on the 1000-category ChinFood1000 dataset. The accuracies are promising but is much lower than the UEC and Food101 datasets, indicating that the Chin-Food1000 dataset is still challenging for the current food classification methods due to the great number of categories, the high inter-class similarities and the large intra-class variances.

We calculated the per class F1 score from the baseline approach. The best 10 categories and the worst 10 categories are respectively reported in Fig. 4. Note that for the 10 worst categories, the F1 scores are all 0, where the food is mainly composed of Chinese dishes. For instance, mushroom fried rice and rice are very similar classes. For the best 10 categories, a major composition is fruits, which enjoys more consistent visual appearances than typical Chinese dishes.

The proposed baseline approach is implemented via Caffe [14], and deployed on a Linux server with 1G memory, Intel Intel(R) Xeon(R) CPU E5-2650 and no CUDA acceleration. The runtime of the prediction for one image is around 500 ms.

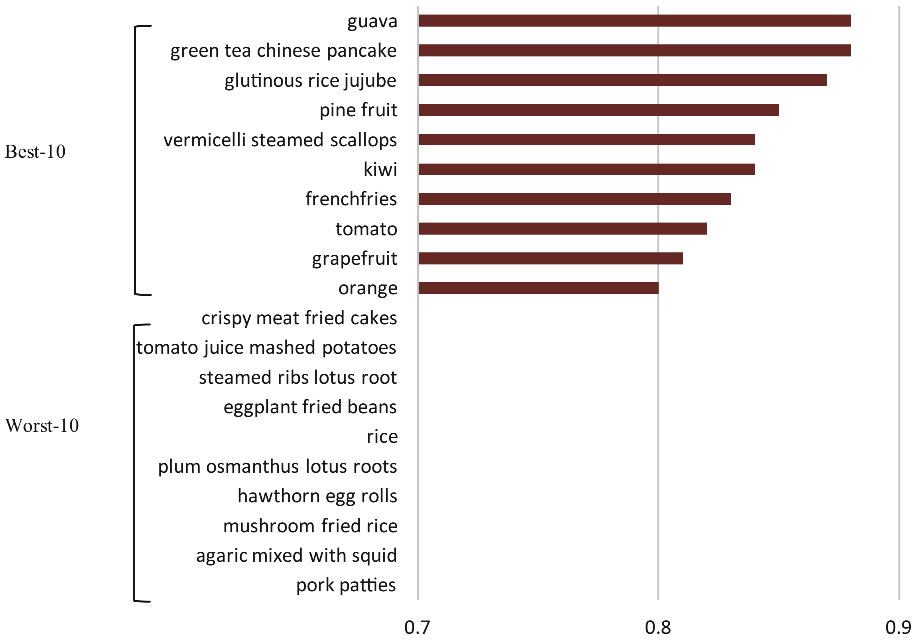


Fig. 4. The best 10 categories and the worst 10 categories by the baseline approach. The best and worst predictions are evaluated by the F1 score. For the 10 worst categories, the F1 scores are 0. The worst categories are mainly composed by Chinese dishes while the best categories are mainly fruits, which enjoy more consistent visual appearances than typical Chinese dishes.

5 Conclusion and Future Work

In this paper, we introduce the 1000-category food dataset ChinFood1000. The dataset is very challenging, with large inter-class similarities and intra-class variance. Since the categories in the dataset are carefully designed to include the most popular Chinese dishes, fast food, snacks, drinks and fruits, the dataset is meaningful and practical for real applications. We further propose a simple and effective baseline. With the power of a very deep network for food images, the baseline achieves the currently best performance on UEC100, UEC256 and Food101 datasets. We study the ChinFood1000 dataset with the baseline approach. The promising results indicate that the ChinFood1000 dataset is still challenging for the current food classification methods. Our future work will improve the classification accuracies with the help of the hierarchical structures.

References

1. Chen, M., Dhingra, K., Wu, W., Yang, L., Sukthankar, R., Yang, J.: PFID: Pittsburgh fast-food image dataset. In: ICIP (2009)
2. Matsuda, Y., Hoashi, H., Yanai, K.: Recognition of multiple-food images by detecting candidate regions. In: ICME (2012)
3. Kawano, Y., Yanai, K.: FoodCam: a real-time food recognition system on a smartphone. Multimedia Tools Appl. **74**, 5263–5287 (2014)
4. Kawano, Y., Yanai, K.: Food image recognition with deep convolutional features. In: CEA (2014)
5. Kawano, Y., Yanai, K.: Automatic expansion of a food image dataset leveraging existing categories with domain adaptation. In: Agapito, L., Bronstein, M.M., Rother, C. (eds.) ECCV 2014. LNCS, vol. 8927, pp. 3–17. Springer, Cham (2015). doi:10.1007/978-3-319-16199-0_1
6. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101 – mining discriminative components with random forests. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8694, pp. 446–461. Springer, Cham (2014). doi:10.1007/978-3-319-10599-4_29
7. Chen, J-J., Ngo, C-W.: Deep-based ingredient recognition for cooking recipe retrival. In: ACM Multimedia (2016)
8. Liu, C., Cao, Y., Luo, Y., Chen, G., Vokkarane, V., Ma, Y.: DeepFood: deep learning-based food image recognition for computer-aided dietary assessment. In: Chang, C.K., Chiari, L., Cao, Y., Jin, H., Mokhtari, M., Aloulou, H. (eds.) ICOST 2016. LNCS, vol. 9677, pp. 37–48. Springer, Cham (2016). doi:10.1007/978-3-319-39601-9_4
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
10. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: CVPR (2015)
11. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: OverFeat: integrated recognition, localization and detection using convolutional networks. In: CVPR (2015)
12. Krause, J., Sapp, B., Howard, A., Zhou, H., Toshev, A., Duerig, T., Philbin, J., Fei-Fei, L.: The unreasonable effectiveness of noisy data for fine-grained recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 301–320. Springer, Cham (2016). doi:10.1007/978-3-319-46487-9_19
13. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: CVPR (2009)
14. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093 (2014)