Automatic Chinese Food Identification and Quantity Estimation

Mei-Yun Chen*, Yung-Hsiang Yang, Chia-Ju Ho, Shih-Han Wang, Shane-Ming Liu, Eugene Chang, Che-Hua Yeh, Ming Ouhyoung
Department of CSIE & GINM
National Taiwan University

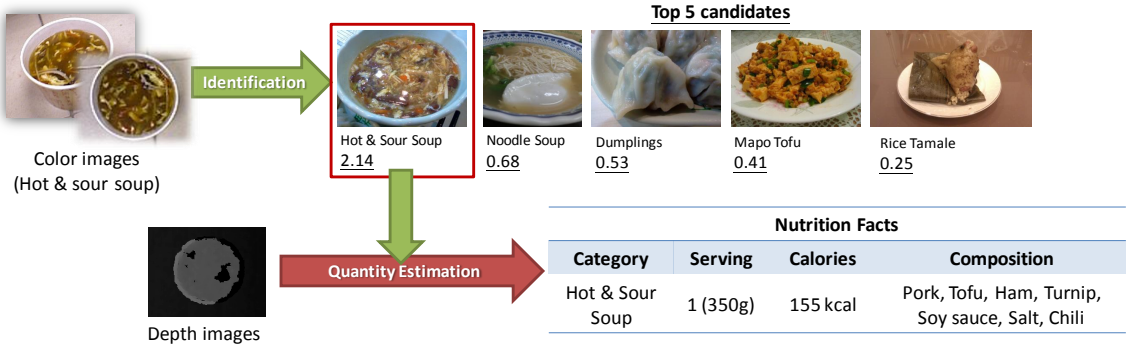

Figure 1: *The proposed system for food identification and quantity estimation. The identification function provides top 5 candidates to users as a reference, and the estimation function measures the quantity of food. By combining these two functions, the nutrition facts and calories can be calculated more precisely.*

Abstract

Computer-aided food identification and quantity estimation have caught more attention than before due to the growing concern of health and obesity. The identification problem is usually defined as an image categorization or classification problem and several researches on this topic have been proposed. In this paper, we address the issues of feature descriptors in the food identification problem and introduce a preliminary approach for the quantity estimation using depth information. Sparse coding is utilized in the SIFT and Local binary pattern feature descriptors, and these features combined with Gabor and color features are used to represent food items. A multi-label SVM classifier is trained for each feature, and these classifiers are combined with multi-class Adaboost algorithm. For evaluation, 50 major categories of worldwide food are used, and each category contains 100 photographs from different sources, such as photos taken manually or from Internet web albums. An overall accuracy of 68.3% is achieved, and success at top-N candidates achieved 80.6%, 84.8%, and 90.9% accuracy accordingly when N equals 2, 3, and 5, thus making mobile application practical. The experimental results show that the proposed methods greatly improve the performance of original SIFT and LBP feature descriptors. On the other hand, for quantity estimation using depth information, a straight forward method is proposed for certain food, while transparent food ingredients such as pure water and cooked rice are temporarily excluded.

*e-mail:meiyun@cmlab.csie.ntu.edu.tw

1 Introduction

Automatic food recognition has been paid more attention in recent years since the growing concern in dietary related health issues such as obesity. There is a rise in the demand of E-Healthcare system in recent years with the widespread of mobile devices and wireless communication networks, such as smartphones and the third generation (3G) telecommunications. With mobile devices in hands, people can record their lives with photos, videos, location information, and even vital signs, and these information can be send, stored or analyzed over cloud services. Daily diet is one key item in the system due to its strong correlation to health and chronic diseases, such as obesity, diabetes, heart disease, and cancer. Aizawa *et al.* proposed the FoodLog [Kitamura et al. 2008], a web-based system, which allows people to keep a log of their dietary intake by taking and uploading the photographs of the food they eat. The system tries to locate and analyze nutritional composition of the meals from photographs, and calculates the dietary balance according to "My Pyramid Specification"[United States Department of Agriculture 2012], which categorizes food into five groups: grains, vegetable, meat/beans, milk, and fruit.

In this paper, we target to identify food categories and to estimate their quantities. We observed from previous literature that the feature descriptor is the key for food identification, and the most used include SIFT, color, and texture feature descriptors; therefore, we exploited these feature descriptors for food identification. The experimental comparisons and results for these feature descriptor are described carefully in this paper. We also introduce our preliminary idea for the quantity estimation based on the depth information which has not yet been utilized in previous literatures.

The contribution of our work includes:

1. Our system can automatically identify food categories and has been implemented as an Android application (not disclosed due to the blind review). The overall accuracy for 50 categories of food achieves 68.3% by cooperating with SVM and multi-class Adaboost algorithm. Success at top-3 and top-5

candidates can reach 84.8% and 90.9% accuracy.

2. A preliminary approach is introduced for the quantity and nutrition estimation, based on the utilization of depth information.

3. Our collection of food database is provided as a research site (http://www.cmlab.csie.ntu.edu.tw/project/food/), open to researchers interested in this topic and for fair comparison, but not for commercial usage.

2 Related Works

Several works have been proposed for food recognition and identification, and the most popular method is to treat them as an image categorization or classification problem. The FoodLog system extract color, circle edge, and SIFT features from food images, and utilize support vector machine (SVM) for training and prediction [Kitamura et al. 2008; Kitamura et al. 2010]. 91.8% accuracy is achieved in food-non-food image detection (10-fold cross validation in 9000 images), and 38.2% accuracy is achieved in food balance estimator using "My Pyramid Specification" 5 categories (10-fold cross validation in 900 images). In [Joutou and Yanai 2009; Hoashi et al. 2010], color, texture, gradient, and SIFT features are extracted from food images, and a separate classifier is trained for each feature. Finally, all the classifiers are weighted combined with the multiple kernel learning method, and 61.3% and 62.5% accuracy is achieved for 50 and 85 categories of Japanese food using 9 and 17 features (5-fold cross validation in 8500 images). In [Yang et al. 2010], food items are represented with calculating pairwise statistics between local features computed over a soft pixel level segmentation of the image into eight ingredient types. In [Bosch et al. 2011; Zhu et al. 2011], local features extracted from patches and global statistic features calculated from all pixels in an image, are combined as a food item descriptor. Then food descriptors can be used to locate, segment, and identify food images. 86.1 % accuracy is achieved in a dataset of 179 hand-segmented images from 39 food categories, however, the dataset is too small (179 images).

3 Food Category Identification

3.1 Feature Extraction

3.1.1 SIFT with Sparse Coding

SIFT has been proven its robustness in feature detection and matching since proposed in [Lowe 2004]. The *bag-of-feature* (BOF) model combined with SIFT obtained a great success in image categorization. The approach constructs a set of "visual words" by quantizing descriptors extracted from image sets, and each image can be represented with the histogram of visual words instead of raw feature descriptors.

Recently, sparse coding/representation has attracted much interest and has been applied to image classification [Yang et al. 2009]. Sparse coding algorithm treats an image as a linear combination of a few basis elements from a dictionary or visual words. The learned dictionary has shown its performance over those learned by previous methods, such as by K-mean quantization [Yang et al. 2009]. Therefore, we apply sparse coding to SIFT features in the proposed framework.

First, a dictionary will be learned from a set of training images by sparse coding, and the dictionary will contain a set of SIFT basis descriptor. Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N] \in \mathbf{R}^{M \times N}$ be the SIFT descriptors extracted from the image set, where M is the dimensionality of each SIFT descriptor and N is the number of descriptors. A dictionary $D = [\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_K] \in \mathbf{R}^{M \times K}$ is trained using sparse coding formulation:

$$\begin{aligned} \min_{\mathbf{D},\mathbf{U}} \quad & \sum_{i=1}^{N} \|\mathbf{x}_i - \mathbf{u}_i \mathbf{D}\|^2 + \lambda |\mathbf{u}_i| \\ \text{subject to} \quad & \|\mathbf{d_k}\| \leq 1, \ \forall k = 1, 2, ..., K \end{aligned} \tag{1}$$

where $\mathbf{u}_i \in \mathbb{R}^K$ is the coefficient vector of $\mathbf{x}_i$ when encoded by the dictionary D. The dictionary D is an *overcomplete* basis set, i.e. $K \gg M$. The parameter λ is a trade-off between reconstruction error and sparsity. Increasing in λ leads to enlarge the sparsity of the learned coefficient vector and vice verse. The L2-norm constrain on $\mathbf{d_k}$ is to avoid trivial solutions on $\mathbf{u_i}$.

After the training phase, the coding phase encodes the SIFT descriptor set X of each image with the coefficient vector obtained by optimizing Eq 1. Since different numbers of SIFT features are extracted from different images, a histogram or pooling operation is required for the encoded coefficient vectors, and the most common way is to compute histogram. In previous literature, other pooling mechanisms, such as max pooling were proposed for representing the image with the encoded coefficient vectors. Multi-scale max pooling, which pools the coefficient vectors at different grid levels of image, outperforms in most image classification and object detection tasks [Yang et al. 2009]. In our experiment; however, the best performance is achieved by computing histogram. The results will be shown and discussed in Section 5.

3.1.2 Local Binary Patterns with Multi-resolution Sparse Coding

As an efficient non-parametric representation of the local composition of an image, one of the proposed methods is Local Binary Patterns (LBP). LBP is designed for texture description [Ojala et al. 2002]. In this work, we propose to apply sparse coding to improve the LBP feature descriptor. First, we construct 3 levels of pyramid with different image sizes, e.g. 1/1, 1/4, 1/16 for each training image. We extract LBP 59-bins histograms from 16×16 patches with step size of 8 pixels in each level of image pyramids as described in [Ojala et al. 2002]. All the LBP histograms are used to learn a 59×2048 dictionary . After the training, the LBP histogram of each image patch can be encoded to a 2048-dimension vector with sparse coding using the dictionary. Comparisons of different settings in LBP are listed in Section 5.

3.1.3 Color Histograms

Color plays an important role in food identification since the uniqueness of each ingredient's color. In the proposed framework, we divide an image into 4×4 blocks and extract a 96-bin RGB color histogram for each, where each channel of RGB is quantized to 32-bin from 256-bin. Finally, a 1536-dimension color vector is formed for each image.

3.1.4 Gabor Texture

Gabor filter is able to capture the properties of spatial localization, orientation information, and spatial frequency information. Moreover, the frequency and orientation representations of Gabor filter are similar to those of human visual system, so it has been widely used in texture representation and recognition.

Each image is divided into 4×4 blocks, and each block is convolved with Gabor filter. 6 orientations and 5 scales Gabor filters are used here, and the mean and variance of the Gabor magnitudes are calculated for each block. Then the values are concatenated to

Figure 2: *Examples from the collected food categories (from left up to right down): Gongbao Chicken, Mapo Tofu, Dumplings, Curry Rice, Arepas, Chasiu, Steamed Sandwich, Omurice, Peking Duck, Braised Pork. Note that some images are from Google Image Search.*

form the Gabor feature vector. Although LBP operation and Gabor filter are able to capture texture appearance in images, there are still some difference in their capabilities so that they cannot be replaced by each other completely. We will show the experimental results in Section 5.

3.2 Multi-class Classification

After extracting feature vectors, we separately train a SVM classifier for each feature. Each classifier is a multi-class model, which contains 50 labels corresponding to the 50 kinds of food. To fuse the SVM classifiers, we adopt the Multi-class AdaBoost algorithm [Zhu et al. 2009] - Stagewise Additive Modeling using a Multi-class Exponential loss function (SAMME) for multi-class classification.

4 Quantity Estimation

After identifying the food category, we would like to measure the food quantity, which is the most important factor for calorie estimations. Previous literatures show an approach that measures the quantity using single food image with edge detector. The results are really rough since it's challenging to estimate the size of a bowl and its depth from a single image; therefore, it's necessary to acquire more information beyond color images for the quantity estimation. In this paper, we propose a preliminary approach to estimate the food quantity based on exploiting the depth information.

The depth information acquisition was an expensive task before recent introduction of commodity depth cameras, such as Microsoft Kinect. Besides depth cameras, the depth information can also be obtained from stereo images incorporating with stereo matching techniques [Scharstein and Szeliski 2002]. Thanks to the handheld devices and tablet computers with stereo camera, stereo images can be easily acquired nowadays.

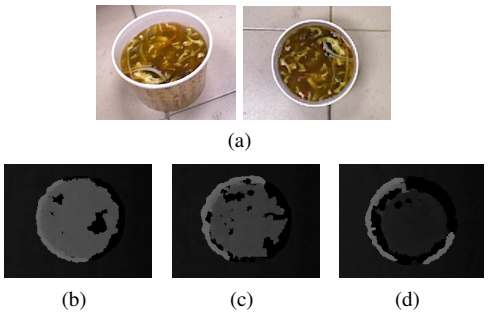

Figure 3: *(a) The color image of food "hot & sour soup" (a bowl full) (b)-(d) The depth images from the top view with different quantities: a bowl full, 1/2 bowl full, and empty.*

We take the food "hot & sour soup" as an example to show our preliminary idea. We used a depth camera to acquire the color and depth information of the noodle soup. The foreground region was segmented, and therefore the bowl containing the food can be identified in the depth image. Figure 3 shows depth images of one bowl of hot & sour soup with different quantities. Two main parameters can be obtained for the quantity estimation. The first one is the area of the bowl, which is calculated according to the diameter of the boundary as shown in Figure 3. The second one is the depth value of the contained food, which can be multiplied by the area to give the volume of a bowl. From the Figure 3 (b)-(d), we can observe that the depth value of centric region varies while the quantity of contained food changes. From our experiments, we estimated there are 7.5 centimeter displacement between one bowl full and empty. After obtaining the two parameters, we can estimate the quantities of food and measure the calories.

Unfortunately, there are still some limitations in this method. The depth of pure water and cooked rice could not be estimated precisely while taking the depth cameras based on infrared structured-light as the capture device, such as Kinect. The reason is that reflection or refraction properties of infrared will be affected by some materials, such as water. This problem can be avoided by using different depth estimation techniques, such as stereo matching from one pair of images.

5 Experiments and Results

We implemented the proposed system as a service and can support handheld devices, such as Android/iOS phones and pads. The system extracts features and identifies the category of the food shown in the images taken and uploaded by users. The service is now running on a small cluster of 4 servers which support a maximum value of 34 threads, and it takes around 12 seconds for the total response time.

We collected 50 categories of Chinese foods (major ones) for evaluation. There are 100 images (all ready-to-eat) for each category, and Figure 2 shows some examples of the collected food images. In our experiments, we adopted 5-fold cross validation to evaluate the performance of the proposed framework. The dataset is randomly partitioned into 5 sets: one set is retained as the validation data, and 4 sets are used as the training data. For the proposed framework, three of the four sets are used to train the SVM classifiers using LIBSVM library [Chang and Lin 2001], and the single set is used to determined the weighting for each SVM classifier with the multi-class Adaboost algorithm.

Figure 4: *Four examples of the collected photos for food – Popcorn*

Table 2 shows the classification accuracy of each feature and all features combined with SVM (all feature vectors are concatenated into one for SVM training and testing) and multi-class Adaboost. The proposed work achieves 68.3% average accuracy , and success at top-N achieved 80.6%, 84.8%, and 90.9% accuracy when N equals 2, 3, and 5. Table 1 shows the best and worst 5 categories in terms of precision and recall rate.

Table 1: *Top and Bottom Categories in Precision and Recall Rate*

Highest Precision		Lowest Precision		Highest Recall		Lowest Recall	
Food	**Precision**	**Food**	**Precision**	**Food**	**Recall**	**Food**	**Recall**
Popcorn	98.4%	Buns	50.7%	Corn	96.7%	Arepas	36.7%
Corn	85.5%	Donuts	51.6%	Pop corn	88.3%	Fish&Chips	41.7%
St.Stuffed Bun	83.3%	Lobster	54.2%	Fried Rice	88.3%	Croissants	43.3%
Hot&Sour Soup	82.2%	Crab	56.1%	Sashimi	88.3%	Lobster	45.0%
Shumai	80.7%	Braised Pork	56.3%	Chocolate	86.7%	Curry Rice	46.7%

Table 2: *Accuracy of Features*

Features	**Accuracy**
SIFT (SC+Histogram)	53.0%
LBP (SC+Histogram)	45.9%
Color	40.3%
Gabor	26.6%
All (SVM)	62.7%
All (Multi-class Adaboost, Top-1 accuracy)	**68.3%**
All (Multi-class Adaboost, Top-3 accuracy)	**84.8%**
All (Multi-class Adaboost, Top-5 accuracy)	**90.9%**

Table 3 shows the results of different settings in SIFT feature descriptor. Bag-of-feature, sparse coding with multi-scale max pooling, and sparse coding with the histogram are evaluated, and the dictionary size is set to 1024 for all. The results show that sparse coding actually improves the performance. The best performance, 53.0%, is obtained by the histogram, not the multi-scale max pooling, which achieved the best performance in previous image and object categorization problems. The reason may come from the non-rigid properties in foods, which can be deformed dramatically in appearance for the same food.

Table 3: *Comparisons of SIFT*

Features	**Accuracy**
SIFT (bag-of-feature)	40.2%
SIFT (SC+Multi-scale Max Pooling)	43.4%
SIFT (SC+Histogram, **Our**)	**53.0%**

Table 4 shows the results of different settings in LBP feature descriptor. The best performance is obtained by cooperating LBP with sparse coding and the histogram statistics of 2048-dimension.

Table 4: *Comparisons of LBP*

Features	**Accuracy**
LBP (Original)	36.2%
LBP (SC+Histogram, 1024-Dim)	39.9%
LBP (SC+Histogram, 2048-Dim, **Our**)	**45.9%**

6 Future Work

As a future work, we would like to complete our preliminary idea for quantity estimation. Then, to improve the performance in response time, we plan to run the service in cloud computing environment, and so the reduction of food identification time is expected. We also target to combine the food identification system with location information. Since we can take food photographs with our smartphones, the location information, for instance, from GPS, can be used to predict the food category more precisely. Current system utilizes the non-linear SVM, which requires high computational complexity while the training data grows. We will evaluate the performance of feature descriptors in the linear SVM model.

References

BOSCH, M., ZHU, F., KHANNA, N., BOUSHEY, C., AND DELP, E. 2011. Combining global and local features for food identification in dietary assessment. In IEEE ICIP, 2011, 1789 –1792.

CHANG, C.-C., AND LIN, C.-J. 2001. LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

HOASHI, H., JOUTOU, T., AND YANAI, K. 2010. Image recognition of 85 food categories by feature fusion. In IEEE ISM, 2010, 296–301.

JOUTOU, T., AND YANAI, K. 2009. A food image recognition system with multiple kernel learning. In IEEE ICIP, 2009, 285–288.

KITAMURA, K., YAMASAKI, T., AND AIZAWA, K. 2008. Food log by analyzing food images. In ACM MM, 2008, ACM, 999–1000.

KITAMURA, K., DE SILVA, C., YAMASAKI, T., AND AIZAWA, K. 2010. Image processing based approach to food balance analysis for personal food logging. In IEEE ICME, 2010, 625 –630.

LOWE, D. G. 2004. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision 60, 2 (nov), 91–110.

OJALA, T., PIETIKÄINEN, M., AND MÄENPÄÄ, T. 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE PAMI, 2002 24, 7 (jul), 971 –987.

SCHARSTEIN, D., AND SZELISKI, R. 2002. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Int. J. Comput. Vision 47, 1-3 (Apr.), 7–42.

UNITED STATES DEPARTMENT OF AGRICULTURE, 2012. USDA ChooseMyPlate.gov. Website. http://www.choosemyplate.gov/.

YANG, J., YU, K., GONG, Y., AND HUANG, T. 2009. Linear spatial pyramid matching using sparse coding for image classification. In IEEE CVPR, 2009, 1794 –1801.

YANG, S., CHEN, M., POMERLEAU, D., AND SUKTHANKAR, R. 2010. Food recognition using statistics of pairwise local features. In IEEE CVPR, 2010, 2249 –2256.

ZHU, J., ROSSET, S., ZOU, H., AND HASTIE, T. 2009. Multi-class adaboost. Statistics and Its Interface 2, 3, 349–360.

ZHU, F., BOSCH, M., KHANNA, N., BOUSHEY, C., AND DELP, E. 2011. Multilevel segmentation for food classification in dietary assessment. In IEEE ISPA, 2011, 337 –342.