DietCam: Multiview Food Recognition Using a Multikernel SVM

Hongsheng He, Fanyu Kong, and Jindong Tan

***Abstract*—Food recognition is a key component in evaluation of everyday food intakes, and its challenge is due to intraclass variation. In this paper, we present an automatic food classification method, DietCam, which specifically addresses the variation of food appearances. DietCam consists of two major components, ingredient detection and food classification. Food ingredients are detected through a combination of a deformable part-based model and a texture verification model. From the detected ingredients, food categories are classified using a multiview multikernel SVM. In the experiment, DietCam presents reliability and outperformance in recognition of food with complex ingredients on a database including 15,262 food images of 55 food types.**

***Index Terms*—Food recognition, multikernel learning, multiview classification.**

I. Introduction

Visual food recognition is a complex problem in computer vision and pattern recognition. The best results achieved so far are 84% for fast food [1] and 70% for general food [2]. Food recognition is worthy of more research effort owning to its practical signification and scientific challenges. Food recognition exposes new challenges to the current pattern recognition literature and stimulates the stemming of novel techniques for general object recognition. In addition, automatic food recognition is beneficial to healthcare-related applications, such as obesity management.

Obesity has become a severe public health problem to general population in many developed countries [3]. In the past three decades, the obesity rate in U.S. increased significantly, resulting in serious health problems, such as diabetes, strokes, heart diseases, and even cancer. Food intake assessment is one of the important methods for obesity management. Few people, however, are willing to keep track of their food intakes, owning to the inconvenience of current assessment methods and the lack of real-time feedback. Traditional food record or diary methods require manual recording of the types and portion of consumed food, and thus the accuracy essentially depends on individual estimation. The latest commercial iPhone application, Meal Snap, can assist users to record and recognize food images, whereas it still requires the user to manually recognize food types. Automatic food intake assessment that avoids the inaccuracy in manual recording and food estimation deserves more research effort for obesity management.

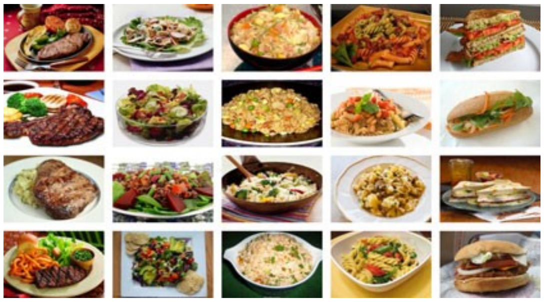


Fig. 1. Intraclass variance in different food categories. From left to right, they are steaks, salads, fried rice, pastas, and sandwiches. There are a large variety of food types and even the same food type may look different.

Object recognition has been one of the fundamental areas in pattern recognition for decades, producing prosperous results in specific object recognition, such as faces [4] and cars [5], [6]. Food recognition is challenging as compared to specific object recognition because it is essentially an intraclass recognition problem. Intraclass recognition is still unsolved, especially for objects with extreme variation [7], such as animals, furniture, flowers, and food. The appearance of food exhibits a higher degree of variance even for the same food type, as shown in Fig. 1.

General object recognition methods have been applied to food recognition. These techniques include color histogram [8], texture [9] and bag-of-feature classification [10], [11]. In our previous paper [1], we found these existing techniques could have acceptable results for regular-shaped food recognition and fast food recognition; however, the recognition performance on generally arbitrary food was not as promising. In fact, color-histogram-based classification is sensitive to lighting conditions, and the bag-of-feature method extracts the statistics of key image patches, which do not explicitly represent the necessary location information in food. Novel techniques requires to be invented for general arbitrary food recognition.

Given a sample of food, one is able to recognize its ingredients from shapes, textures and colors. With the combination of the ingredients, one could recognize the food based on the distribution of ingredients. Sometimes, persons may also have difficulties to recognize food if they do not have sufficient knowledge to distinguish food only through their appearances.

It is challenging to automate food recognition in a similar way that humans recognize food from ingredients. Different food types could have similar appearances, while the same food type

Manuscript received August 31, 2014; revised December 16, 2014; accepted March 14, 2015. Date of publication April 2, 2015; date of current version May 9, 2016. The work of H. He was supported in part by NSFC Grant 61305114.

H. He and J. Tan are with the Department of Mechanical, Aerospace and Biomedical Engineering, University of Tennessee, Knoxville, TN 37996 USA (e-mail: he@utk.edu; tan@utk.edu).

F. Kong is with the Microsoft, Inc., Seattle, WA 98109 USA (e-mail: fanyuk@microsoft.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JBHI.2015.2419251

could have different appearances. In addition, recipes, cooking methods, and chef's personal preference affect the appearances of food ingredients. The second challenge is that even though ingredients could be detected correctly, food could also have unstructured ingredient distribution. For some types of food, the ingredients are distributed randomly across a plate. The third important challenge is the occlusion in food images. Food is usually placed in certain containers such that some key elements may be covered or occluded by other ingredients. Food recognition from images in different scales is another challenge. Some types of food could not be differentiated through its own sizes in images. For example, brown rice looks similar to a baked potato without considering their relative scale.

This paper presents an ingredient-based food recognition method. Our first innovation is the improvement of the current part-based object recognition model toward texture-oriented and location-flexible to detect food ingredients. The state-of-the-art part-based detectors are not directly suitable for food ingredient detection concerning the aforementioned challenges. We modify the detector in three ways for the purpose of food ingredient detection. The ingredient detector tries to find food ingredients on a single scale in order to retain relative ingredient scales. After that, the scale invariance is achieved in a multiscale support vector machine (SVM) during food classification. The part-based model uses the shape of objects as a key property. We enrich the model with the ability to verify detection with texture models, meaning that both the shape model and texture model are used to detect food ingredients. In the existing part-based based model, the geometry locations of the parts are modeled strictly. In our model, we employ a more flexible location mechanism for food ingredients to represent food with more deformation.

Our second contribution is development of a multiview multikernel SVM to classify various combinations of food ingredients under occlusion. We design the SVM with multiple kernels that include a hierarchy of element kernels. The top level is the viewpoint level where we adopt a multiview scheme to address occlusion. On the viewpoint level, each view corresponds to a kernel function, and all the kernel functions from multiview are combined together according to the geometry similarities between viewpoints. Under each viewpoint kernel, we design spatial pyramid kernels to achieve scale invariance. Then under each scale, there is a linear combination of linear, quasi-linear, and nonlinear element kernels to classify food ingredient features. By employing such a hierarchy of kernel functions, we accomplish a classifier that detects and classifies food of different scales and points of view. The experiments show the effectiveness of the proposed classifier.

II. Related Work

Food intake assessment has been a popular research topic in biomedical and health related areas for years. The traditional method is food diaries and records [12], [13], where people need to record food types and estimate food volumes. It is applicable to most people since it does not require any professional knowledge. The limitation of these methods is the inaccuracy and personal biases in human estimations [14].

Researchers have developed methods to monitor food intakes inside human organs to address inaccuracy caused by human estimation. Typical methods include biological assessment, e.g., doubly labeled water [15], plasma carotene [16], and chemical analysis, e.g., tracking selected elements [17]. Tough accurate, these methods are currently merely available in lab environments.

In computer vision, food recognition is a specific case of category recognition. Martin *et al.* used a general color histogram of an image to recognize food items [8]. Wu and Yang presented interest points (SIFT)-based method to recognize fast food [10], with the accuracy under 70%. Zhu *et al.* used textures to detect food from an image and then utilize the histogram of textures to classify food items [9]. Yang *et al.* proposed ingredient-based food classification method that achieved 78% accuracy in [2]. Food ingredients were detected through texture classification, and food types were classified by calculating the pairwise statistics between food ingredients. In Yang's study, semantic texton forests (STF) were used for food ingredient classification.

Part-based based recognition is an extension of template-based recognition method. Template-based methods are usually used in rigid-shaped object recognition [5], [18], [19], exhibiting good performance for single object recognition [5], [19]. An important limitation of these methods is their inflexibility to capture the variance of object appearances. Part-based recognition introduces a geometric distribution model of different parts to represent the variance of object appearances [20].

Automatic visual food recognition is a convenient way to assess food intakes. The problems to solve are food recognition from a different viewpoint, under different lighting conditions, with different backgrounds, with partial occlusion, and from different scales. In this paper, we use a combined model of texture models and part-based model to extract food ingredients, and apply a multiview multikernel SVM to classify the food ingredients. The following sections provide the details of these two parts.

III. Ingredient Detection

Ingredients are important attributes to differentiate different types of food. The dominant ingredients are generally similar for a type of food, although food could be prepared with different cooking methods and using different condiments. Therefore, if we could find the key ingredients, the food could be classified according to the combination of these ingredients. The two components in the food classification process are detection of food ingredients and classification of ingredient combinations.

It is nontrivial to find food elements only through their colors, shapes, or textures. The food colors are not completely consistent and sensitive to ambient lighting conditions. The shape of a food element is determined by many factors, such as its natural shape, cutting patterns and perspectives. Different types of food could have the same type of textures. Thus, typical color-histogram-based, shape-template-based and texture-based classification methods are not directly applicable to food ingredient detection.

The state-of-the-art part-based detector in [20] learns models at different resolutions, and builds geometrical locations of the parts. When detecting objects, it builds feature maps of the image at different resolutions and scales. The output of the detector is determined from the fused responses of root and part filters in the predefined location structure. The part-based detector is expected to perform well in detecting food items with certain structures, such as chicken wings and sandwiches; however, part-based model is not directly suitable for food with the following attributes: 1) food with a similar shape but different scales, such as rice, a meat ball, and a baked potato, 2) food with a similar shape but only differentiable through textures, such as steak, fish, and pork steak, and 3) food with more flexible geometry distributions of ingredients, such as pizza toppings, rice, and noodles.

This section presents a novel food ingredient detector with a part-based model of textures and locations. A food ingredient detector combining part-based models and texture models is developed in this paper. Part-based models are popular for rigid-shaped object detection and classification, taking into account the shape of each part and their geometric relations. The part-based models, however, cannot be directly applied to food ingredient detection because of the shape and texture variance in food appearances. We therefore integrate texture filters into part-based detectors, where textures are classified in texture filters.

A. Texture Classification

Besides shape attributes, textures are important properties to distinguish a complex food item. In this paper, a texture filter bank, STF [21], is chosen to detect food textures. STF is an image segmentation and classification technique that generates soft labels for each pixel based on their local texture properties. This is achieved through learning from manually labeled sample images and building decision forests. The detector outputs a set of ingredient bounding boxes that are determined by thresholding the returned confidence scores. The result on texture decomposition of a plate of fried rice is shown in Fig. 2, where food materials are separated by their texture distribution.

B. Flexible Part-Based Model

The distribution of food ingredients has different geometric patterns due to the intraclass deformation. Therefore, we model the deformation in a high level using the distribution of texture parts. The distribution of the texture parts is obtained by analyzing the arrangement of features from the food image. Fig. 3 shows the part location of chicken wings learned from a sample image. The location of texture parts appears to be a discriminative feature. For instance, a burger or an apple has a clustered distribution of texture parts, while pizza toppings usually have a uniform distribution of texture parts that are scattered across the whole image.

The location of the ingredients is a key parameter to distinguish different food. For example, two pieces of breads cannot form a sandwich if one pieces is not above another one. Therefore, besides the histogram of the ingredients, the feature vector is designed to encode location and spatial relation of food ingredients. We define location relation between the detected ingredients as "above," "below," and "overlapping," whereas the spatial relation "left" and "right" are not evaluated as much food is left-right symmetric. The textures in the feature vector are sorted in the primary order that the ingredients distributed from "above" to "below," and secondary order that the "overlapping" size decreases. Through the sorting, the obtained features encode the relative spatial relation of food ingredients.

The detectors that search objects at different scales lose the relative scale between different objects. For example, we found that a part-based model of a baked potato would also detect

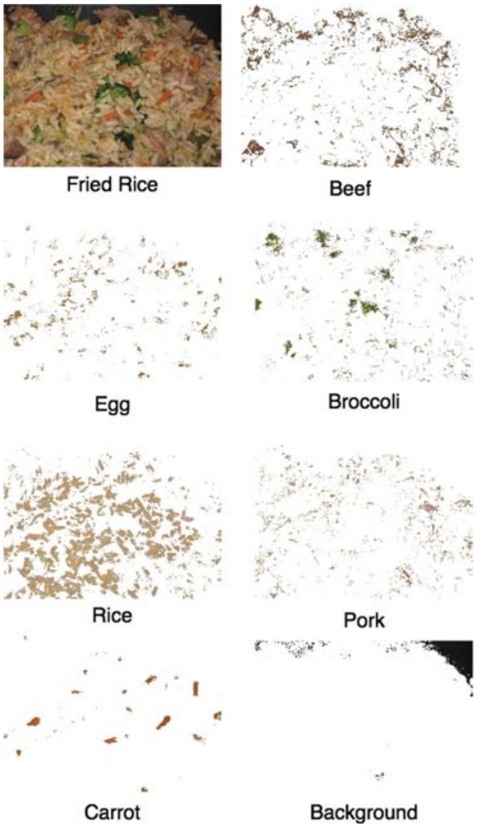


Fig. 2. Extracted food ingredient texture of a plate of fried rice.

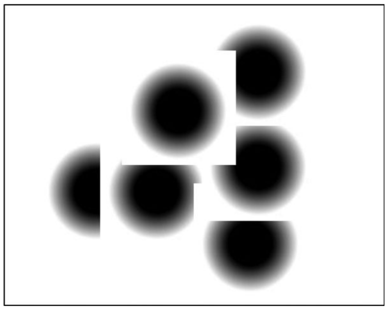


Fig. 3. Part location model of chicken wings. The black dots represent the possible part locations. A part will have a higher probability to appear at a place with darker dots. The combination of the dots represents the location relationship of different parts.

a piece of rice during the test. Thus, we restrict the part-based model such that at the ingredient detection phrase, only one scale of food images is used to retain the relative scales of different ingredients. Instead, we tackle the multiscale problem in the food classifier in the classification phrase. The food ingredients are detected in three pyramid levels, corresponding to one, four, and sixteen spatial subdivisions. By defining a reference scale in each pyramid level, we can determine the relative size of food elements from food images.

IV. Food Classification

The results of ingredient detection using the proposed part-based model include a set of bounding boxes of food ingredients and locations. With the texture-based detector, the result of the ingredient detection is a histogram of the food ingredients appearing in the image. For d food ingredients, we use a vector $\mathbf{z}$ to represent the histogram, $\mathbf{z} = (z_1, z_2, \ldots, z_d) \in \mathbb{R}^d$. The priorities of each ingredient in the feature vector are determined by their relative spatial relation as defined in Section III-B. The purpose of food recognition is to find the relative locations of these ingredients and to classify $\mathbf{z}$.

In order to classify the ingredient histogram, we use an SVM, which is one of the most successful techniques in classification problems. It could find a unique global optimal solution and it has solid mathematical derivations. SVM tries to solve the following classification problem. Given n samples $\{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}\}$ of vectors in d-dimensional space

$$\mathbf{x_i} = (x_{i_1}, x_{i_2}, \ldots, x_{i_d})^T \in \mathbb{R}^d \tag{1}$$

and the corresponding labels $y_i = \pm 1$, the classification function is in the form of

$$f(\mathbf{z}, \alpha^*, b^*) = \sum_{i=1}^{n} y_i \alpha_i^* K(\mathbf{x_i}, \mathbf{z}) + b^* \tag{2}$$

where α^* and b^* are the optimal parameters learned from the training samples $\mathbf{x}$, and $\mathbf{z}$ is the new sample to predict. The kernel function $K(\mathbf{x_i}, \mathbf{z})$ that measures the distance between two features plays an important role in classification accuracy. In the following sections, we will investigate candidate kernel functions in SVM.

A. *Multiple Kernels*

Classifying food items from a single viewpoint would be inaccurate because of occlusion that blocks key food ingredients. To deal with partial occlusion, we develop a multiview kernel for the food classification task, by considering food appearances from more than one perspective. Given m viewpoints, $V = \{v_1, v_2, \ldots, v_m\}$, the kernel function for all the viewpoint is defined as

$$K^{\text{mv}}(\mathbf{x}, \mathbf{z}) = \sum_{\ell=1}^{m} g_\ell K^{v_\ell}(\mathbf{x}, \mathbf{z}). \tag{3}$$

The weight of each viewpoint g_ℓ is calculated through the relation between each view. Consider two views, the recognition result of the second view is related to the one in the first image, and therefore the kernel is used to model the occlusion and faulty recognition. Intuitively, if a food is occluded in the first image, it shows a small probability for that type of food to exist in the scene. The probability of existence becomes high in an image where the food is visible or partial occluded. Conversely, if the food does not exist, the detection in the first image turns into faulty recognition. The recognition accuracy is enhanced in this manner that the output of multiple views vote.

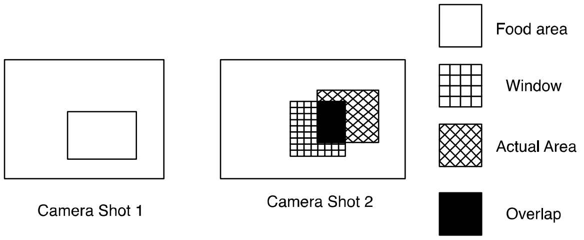


Fig. 4. Geometric similarity between two viewpoints. The rectangle area in the first camera shot is the food area, and the corresponding window area of the food area in the second camera shot is the grid window. At the same time, in the second camera shot, the detected food area is the grid actual area, which has an overlap with the corresponding window. The fraction of the overlap area in the corresponding window area represents the geometric similarity.

Starting from v_i, for $v_j \in V$, g_j is defined by the geometrical similarity $\tau(i, j)$ between viewpoints i and j

$$g_j = \tau(i, j) = \frac{\text{Area}(j)}{\text{Window}(i, j)} \tag{4}$$

where the geometrical similarity is illustrated in Fig. 4. $\text{Window}(i, j)$ denotes the projected area in the jth image from the detected food ingredients in the ith image. A set of correspondent bounding boxes can be found in the other image with the geometric projection. These bounding boxes compose a correspondent window in the current interested image. $\text{Area}(j)$ represents the actual area of detected food ingredients in the jth image. The geometrical similarity is therefore defined as the ratio of the projected and detected areas.

Under each viewpoint, multiple kernels are used to classify food from different feature channels. We use the technique in [22] to learn the optimal kernel function. If we have k element kernel functions, the optimal kernel is defined as

$$K^{v_\ell}(\mathbf{x}, \mathbf{z}) = \sum_{i=1}^{k} d_i K_i^{v_\ell}(\mathbf{x_i}, \mathbf{z}) \tag{5}$$

where d_i is the optimal weight of the ith kernel. We fuse different kinds of elementary kernels to address food variance. We consider three basic kernels commonly used in object recognition, including linear kernels of the form

$$K_{\text{linear}}(\mathbf{x}, \mathbf{z}) = \langle \mathbf{x}, \mathbf{z} \rangle \tag{6}$$

and quasi-linear kernels in the form of

$$K_{\text{quasi-linear}}(\mathbf{x}, \mathbf{z}) = \frac{1}{2}(1 - \chi^2(\mathbf{x}, \mathbf{z})) \tag{7}$$

and nonlinear RBF-χ^2 kernels of the form

$$K_{\text{nonlinear}}(\mathbf{x}, \mathbf{z}) = e^{-\gamma \chi^2(\mathbf{x}, \mathbf{z})}. \tag{8}$$

B. Multikernel SVM

Choosing a right kernel function for a specific problem is tricky when using SVMs for classification. The kernel function can be viewed as discriminative properties of a visual feature encoded in several channels. The tradeoff between the discrimination and invariance distinguishes one feature from another. This tradeoff varies from task to task so that a single kernel function cannot be optimal for all situations.

Motivated by the recent multiple kernel learning method [23], an optimal combination of kernel functions are learned in this paper, where each kernel function captures a different feature channel. Our features include the contribution from different viewpoints, distribution of food ingredients, and these features at different spatial pyramid levels.

All these features are organized in a hierarchy of kernel functions. On the top level is a linear combination of multiple viewpoints. The weight of each viewpoint is calculated from the relation between points of view. Under each viewpoint, there is another level of kernel functions, which are a linear combination of element kernel functions at multiple scales.

By organizing kernel functions in a hierarchy structure, the classifier is in the form of

$$f(\mathbf{z}, \alpha^*, b^*) = \sum_{i=1}^{n} \sum_{j=1}^{k} \sum_{\ell=1}^{m} y_i \alpha_i^* g_\ell d_j K_j^{v_\ell}(\mathbf{x_i}, \mathbf{z}) + b^*. \qquad (9)$$

The kernel weights d is learned through optimization based on the training set [22]. It should be noted that the learned kernels might not be optimal owning to the introduced multiview kernel, optimization of which is impossible without a close-form representation of the parameters.

V. Experiment

The two key components of DietCam, ingredient detection and food classification, were evaluated and studied on a food database. We compared DietCam with four methods commonly used in food recognition. The compared methods include SIFT with a nearest neighbor classifier, texture classification, color histogram with an SVM classifier, and the part-based model.

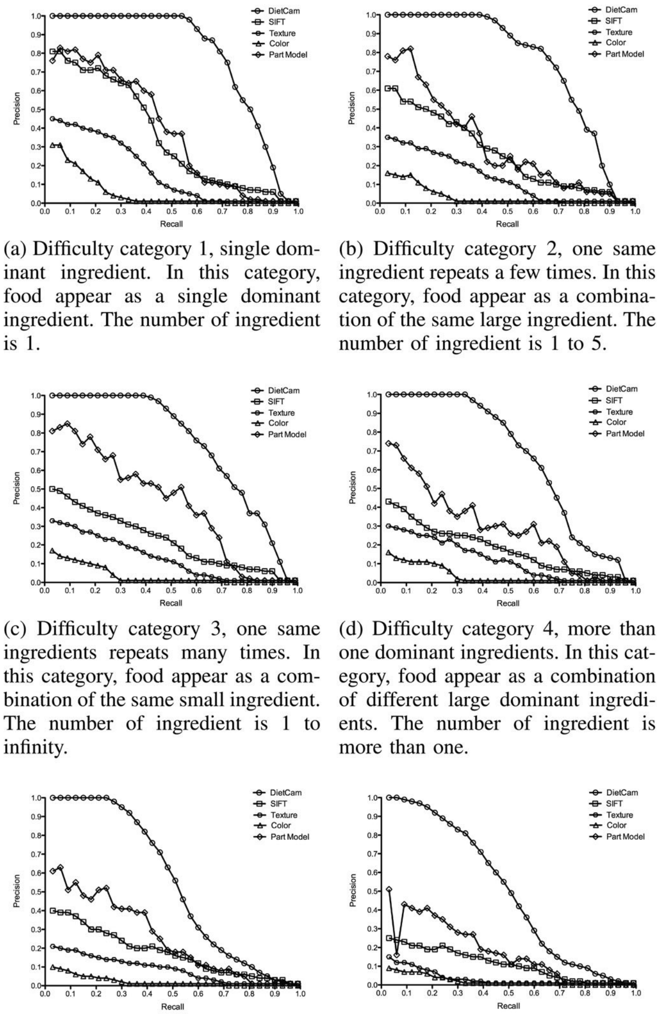
A. Dataset

The community has developed many image databases for general object recognition; however, there are few complete food image databases available. PFID [24] is one published image dataset for research purpose, containing 4545 images of fast food.

We developed a food image database consisting of 15 262 images for both training and testing purposes. Training of the SVM consists of two parts, training of the food ingredient detector and training of food classifiers. We collect a list of 55 popular American food categories, ranging from drinks, pies, to sandwiches and Hoppin' Johns. For each food category, food images were obtained through keyword search using Google Image search, and they were manually validated. A small amount of multiview images were collected at restaurants, and merged with the PFID database. The database has been split into 50% for training and validation, and 50% for testing. The distribution of images, food objects, and food ingredients are approximately equal across the data sets. There are totally 55 food categories and 15 262 food images in the database, with average 277 images for each class. The statistics of the database is given in Table I.

TABLE I
Food Database Statistics

	Training/Validation/Test		
Food Types	Image	Food	Ingredient
Hoppin' John	62/61/123	103/107/211	1056/978/2011
Buttermilk biscuits	75/75/150	128/121/234	359/366/762
Whole lobster	67/67/134	110/113/239	130/143/242
Shrimp and hushpuppies	78/77/155	540/532/1009	540/521/112
Barbecue ribs steak	65/65/130	89/81/198	108/121/231
Krispy Kreme	65/64/129	76/73/172	556/567/1098
Tacos	73/73/146	98/101/210	435/441/987
Lime pie	66/66/132	69/70/154	219/217/445
Philly steak sandwich	65/65/130	80/76/156	530/578/1231
Pork barbecue sandwich	73/72/145	101/99/200	880/760/1770
Lowcountry boil	72/72/144	73/76/161	354/334/720
Huckleberry pie	74/74/148	81/82/160	145/140/243
Clam chowder	65/64/129	76/69/155	204/198/345
Burger	80/80/160	123/119/154	549/567/989
Eggs Benedict	65/65/130	100/101/189	193/191/432
Pastrami on rye sandwich	63/63/126	108/87/199	121/129/231
Pancakes with syrup	65/65/130	68/67/123	68/76/123
Bagel	64/64/128	98/88/192	98/93/228
Soft pretzel	64/64/128	72/70/157	72/73/156
funnel cake	67/67/134	69/67/140	135/125/267
Snow cone	65/64/129	67/68/142	67/70/140
Smoked salmon	61/61/122	99/92/207	109/103/234
Persimmon pudding	76/76/152	97/98/211	99/98/254
Corn dog	79/79/158	86/91/178	89/108/202
French fry	61/60/121	98/77/187	109/112/215
Chicken wings	71/70/141	78/73/143	78/84/156
Drink	72/71/143	113/121/257	132/145/278
Chili dog	65/64/129	81/79/175	342/356/723
Spam musubi	72/71/143	231/240/460	681/665/1428
Fluffernutter sandwich	65/65/130	77/78/155	357/360/750
Cookie	73/73/146	78/76/156	82/90/150
BLT sandwich	69/69/138	80/85/166	459/450/924
Baked beans	65/65/130	68/71/133	680/657/1428
Pumpkin pie	60/60/120	64/66/143	134/135/266
Fajitas	58/57/115	59/58/122	335/340/680
Succotash	63/62/125	69/69/151	379/411/760
Cornbread	61/61/122	66/65/136	79/90/165
Barbecue chicken pizza	71/71/142	74/76/137	657/660/1328
Chicken fried steak	73/73/146	77/76/170	157/140/325
Burrito	75/75/150	79/78/155	349/350/766
Pecan pie	72/71/143	78/77/145	180/195/354
Catfish	74/73/147	90/91/185	213/231/454
Mashed potato	76/76/152	89/88/190	95/106/210
Meatloaf	76/76/152	87/85/185	95/113/207
Green bean casserole	75/75/150	79/79/157	130/143/257
French's fried onions	76/76/152	109/106/178	457/435/956
Sopaipillas	65/65/130	93/90/194	250/260/468
Cheesecake	66/65/131	75/70/159	88/99/533
Turkey sandwich	65/65/130	97/87/206	211/231/466
Salad	78/78/156	83/80/183	355/320/779
Fried rice	78/78/156	81/80/166	320/313/620
Pasta	79/79/158	83/86/178	449/446/993
Noodles	79/79/158	80/90/195	414/430/820
Steaks with broccoli	67/66/133	156/145/323	177/198/330
Sushi	75/75/150	83/85/177	166/210/340



(a) Difficulty category 1, single dominant ingredient. In this category, food appear as a single dominant ingredient. The number of ingredient is 1.

(b) Difficulty category 2, one same ingredient repeats a few times. In this category, food appear as a combination of the same large ingredient. The number of ingredient is 1 to 5.

(c) Difficulty category 3, one same ingredients repeats many times. In this category, food appear as a combination of the same small ingredient. The number of ingredient is 1 to infinity.

(d) Difficulty category 4, more than one dominant ingredients. In this category, food appear as a combination of different large dominant ingredients. The number of ingredient is more than one.

(e) Difficulty category 5, dominant ingredients with many small ingredients. In this category, food appear as a combination of a small number of large dominant ingredients and many small ingredients. The number of ingredient is more than one.

(f) Difficulty category 6, a large number of small ingredients repeat many times. In this category, food appear as a combination of a large number of different small ingredient. The number of ingredient is more than one.

Fig. 5. Precision-recall curve for models trained on food categories with difficulties for comparison of recognition results of DietCam and other three baseline methods.

B. *Ingredient Detection*

The goal of ingredient detection is to predict bounding boxes of each food ingredient in the image. A predicted bounding box is considered correct if it overlaps more than 50% with a ground-truth bounding box. Otherwise, it is considered a false detection. We used precision-recall curves across all the images in the dataset to evaluate the performance of ingredient detection. Every group has a precision-recall curve shown from Fig. 5(a) to (f).

Experiment that compares the performance and the other three implemented methods was conducted on six groups of food in the dataset according to the difficulty definitions in Table II. The difficulty indices are determined in accordance with their ingredient composites and the number of ingredients. The regular-shaped and rigid-shaped food has the smallest difficulty index, while those like Hoppin' John, fried rice and noodles have the highest index. Table II shows a complete list of the food difficulty categories (DC).

TABLE II
DEFINITIONS OF DC

DC	Ingredient Composition	Examples
1	Single dominant ingredient	lobster, steak, snow cone, drink
2	One ingredient repeats a few times	pancake, corn bread
3	One ingredient repeats many times	shrimp, biscuits, French fry
4	More than one dominant ingredient	steak dish, fish dish
5	Dominant plus small ingredients	pie, burger, pizza, sushi
6	Small ingredients repeat many times	fried rice, noodles, boils, fajitas

We implemented ingredient detection based on SIFT features for performance comparison. SIFT has been widely used in general object recognition because of its tolerance to transformation and illumination variance. In the experiment, we decided the type of every food ingredient based on the frequencies of ingredient features that were extracted and classified using a nearest neighbor classifier.

Another method used for comparison was the texture-based ingredient recognition. We used the same texture detection method in DietCam for comparison, yet without the fusion of the part-based detection model. The textures of an image were found through convolving the image with a texton bank, followed by classification using decision forests. Based on the responses of texture filters, the image were segmented and each segmented piece was classified using a multiclass SVM.

For the color-based ingredient recognition, we employed an RGB 3-D histogram with four quantization levels per color band. Each pixel in the image was mapped to its closet cell in the histogram to produce a 64 dimensional histogram of the image, followed by the classification using a multiclass SVM. The performance of color-based ingredient recognition was inferior to other methods mainly because a color histogram is inadequate to represent different food appearances.

The part-based model in [20] was implemented for ingredient detection. Food images with different resolutions were employed for the training of part-based models and location models. In the experiment, accuracy of food ingredient detection using the part-based model presented was low for food with similar shapes, or different textures, or random location distribution.

Fig. 5(a) shows the result of food ingredient detection using DietCam and the other three methods for food in DC 1. The food in this category consists of one dominant ingredient, albeit an image could contain more than one food item. Since there is only one ingredient for one food type, it is expected to have the best result among all the DC. DietCam achieved the best performance in terms of precision and recall. The recognition precision of all the methods dropped as the the number of food items in the images increased. It is still difficult to detect all the ingredients even if there is one food item in the image.

Fig. 5(b) shows the results of ingredient detection for food in DC 2. In this category, there could be more than one appearance of the same dominant ingredient. Compared with food in DC 1, DietCam detected more irrelevant ingredients when recall was larger than 0.4. The general performance of the competing methods downgraded as the number of ingredients increased.

Fig. 5(c) shows the results of ingredient detection for food in DC 3. In this category, a food item consists of one kind of ingredient, yet allowing more than one appearance of this ingredient. The difference between food in DCs 2 and 3 is the size of the ingredients. In DC 3, the size is much smaller than those in DC 2. Typical examples are chopped onions, chopped green peppers, and french fries. DietCam showed a similar result in this category compared with that in DC 2 thanks to the deliberately designed feature extraction model. It was difficult for the competing methods to detect smaller ingredients effectively.

Fig. 5(d) shows the results of ingredient detection for food in DC 4. In this category, more than one type of dominant ingredients present in the image. The result deteriorated with an increasing number of irrelevant ingredients predicted by DietCam. The performance deterioration was caused by the increased number of ingredients. The performance of the implemented competing methods dropped in a similar pattern.

Fig. 5(e) shows the results of ingredient detection for food in DC 5. In this food category, a food item may contain dominant ingredients together with small other ingredients. The precision of DietCam declined sharply because of the increased number of small ingredients in food images. As we have discussed for Fig. 5(c), the competing methods could not detect those small ingredients effectively. The relative performance decrease of the competing methods, due to the inclusion of small ingredients, was not as severe as that caused by the increase of dominant ingredients.

Fig. 5(f) shows the results of ingredient detection for food in DC 6. The food in this category is most difficult to detect, consisting of small ingredients in irregular shapes. The recognition accuracy of all the compared methods went downhill because of the complexity and deformation of food items. In some cases, the methods would completely fail to detect food ingredients and positions.

The ingredient detection is the foundation to classify food types. From Fig. 5(a) to (f), we can see that the precision decreases when the number of ingredients increases in the images. The influence of small ingredients is greater than that of dominant ingredients. In the following section, we will investigate the performance of the multiview classifier built on the intermediate results of ingredient detection.

C. Food Classification

We implemented three competing methods to compare with DietCam in food recognition: SIFT-based nearest neighbor classifier, texture-based SVM, and DietCam using a single-view kernel SVM. We trained and tested the classifiers with entire images, rather than with segmented food ingredients in bounding boxes. The recognition accuracy is presented in Fig. 6.

The recognition precision of DietCam was around 90% for general food items, and 85% for food items in DCs 5 and 6. The recognition precision using SIFT, which is commonly using in fast food classification, was 60% for food in DC 1. As food composition became complex, e.g., for food in DCs 4–6, the accuracy of SIFT classification dropped down to 30%. The texture-based classification presented comparable yet inferior results to SIFT-based classification. The low accuracy was due to the fact that food textures themselves are usually indistinctive and volatile without considering the relation among each other. Without the multiview kernels, the single-view kernel only achieved accuracy of 60% to 80%. With multiview kernels, the accuracy increased by 10% for food in DCs 1 and 2, and about 20% for food in DCs 3–6.

Image samples in the database were given in Fig. 7. The food that was correctly recognized shows clear parts, unique textures, and distinguishable colors. The food that was not correctly

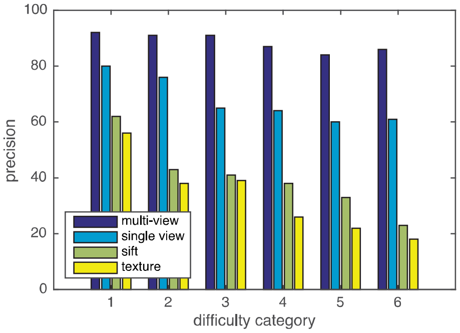
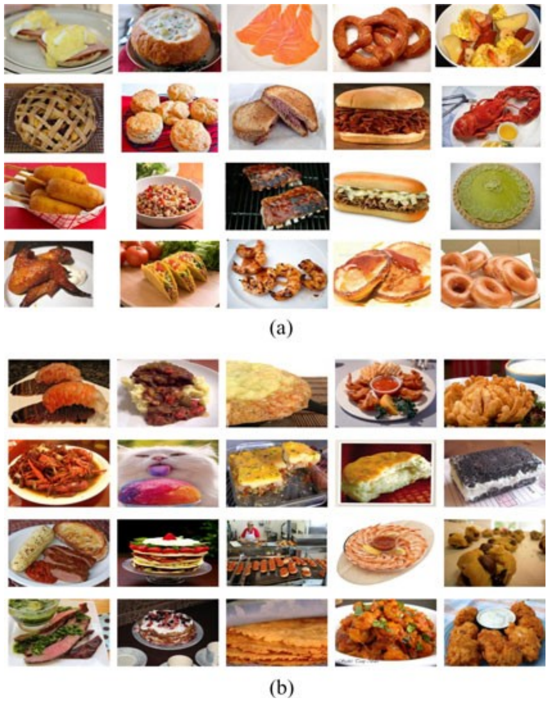


Fig. 6. Food recognition accuracy comparison between DieCam, and SIFT, texture, and single view classifier. The accuracy statistics are obtained under food types from difficulty 1 to 6. Each bar corresponds to each difficulty, and the sum of the blue bar and red bar is the accuracy of DietCam.


(a)

(b)

Fig. 7. Samples of recognized and unrecognized food images. (a) Recognized food images. (b) Unrecognized food images.

recognized exists in every DC. The typical cases of unrecognized food are: key ingredients are covered by sources, creams or decoration ingredient; food are prepared in irregular shapes; food is cooked with special recipes; food in the image is bitten or incomplete.

D. Discussion

Though promising results and increased performance have been achieved compared with other food recognition methods, improvements are required toward a successful field application of automatic food recognition. The first improvement is to fulfill a complete food image database. Currently, we have a database of 55 popular food classes, yet most food categories are not covered, especially for food with small and vague ingredients. The second possible improvement is toward real-time performance. At present, the classification method needs many computing resources, mostly for the ingredient detection part. Another problem we met was that some kinds of food and ingredients were not naturally separable visually from images. For these types of food, human beings use other information and experience to recognize them. For example, mayonnaise usually appears with salads or sandwiches, while yogurt does not. With the context, people could guess whether the white cream is mayonnaise or yogurt. Context-aware models could be adopted in this case to improve overall performance.

VI. Conclusion

Food intake assessment is a building block of many treatments to public health problems, especially for obesity control. In this paper, we presented an automatic food recognition method named DietCam. To address the variance problem of food appearances, we developed a new food ingredient detector and a multiview multikernel-based SVM to classify food items. Based on the experiment on the developed food database of 15 262 food images, DietCam presented promising performance as compared with commonly used food classification methods. The proposed method has the potential to be implemented on mobile devices such as smart phones for convenient daily use.

References

[1] F. Kong and J. Tan, "Dietcam: Regular shape food recognition with a camera phone," in *Proc. Int. Conf. Body Sens. Netw.*, 2011, pp. 127–132.

[2] S. Yang, M. Chen, D. Pomerleau, and R. Sukthankar, "Food recognition using statistics of pairwise local features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2010, pp. 2249–2256.

[3] E. Finkelstein, I. Fiebelkorn, and G. Wang, "National medical spending attributable to overweight and obesity: How much, and who's paying?," *Health Affairs Web Exclusive*, vol. 5, no. 14, pp. 219–226, 2003.

[4] W. Zhao, R. Chellappa, P. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Comput. Surveys*, vol. 35, no. 4, pp. 399–458, Jan. 2003.

[5] M. Weber, M. Welling, and P. Perona, "Unsupervised learning of models for recognition," in *Proc. Eur. Conf. Comput. Vis.*, Jan. 2000, pp. 18–32.

[6] H. He, Z. Shao, and J. Tan, "Recognition of car makes and models from a single traffic-camera image," *IEEE Trans. Intell. Transp. Syst.*, to be published.

[7] R. Szeliski, *Computer Vision: Algorithms and Applications*. New York, NY, USA: Springer, 2010.

[8] C. Martin, S. Kaya, and B. Gunturk, "Quantification of food intake using food image analysis," in *Proc. IEEE Annu. Int. Conf. Eng. Med. Biology Soc.*, 2009, pp. 6869–6872.

[9] F. Zhu, M. Bosch, I. Woo, S. Kim, C. Boushey, D. Ebert, and E. Delp, "The use of mobile devices in aiding dietary assessment and evaluation," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 4, pp. 756–766, Aug. 2010.

[10] W. Wu and J. Yang, "Fast food recognition from videos of eating for calorie estimation," in *Proc. IEEE Int. Conf. Multimedia Expo.*, Jan. 2009, pp. 1210–1213.

[11] M. Anthimopoulos, L. Gianola, L. Scarnato, P. Diem, and S. Mougiakkou, "A food recognition system for diabetic patients based on an optimized bag of features model," *IEEE J. Biomed. Health Informat.*, vol. 18, no. 4, pp. 1261–1271, Jul. 2014.

[12] M. A. Murtaugh, K. Ma, T. Greene, D. Redwood, S. Edwards, J. Johnson, L. Tom-Orme, A. P. Lanier, J. A. Henderson, and M. L. Slattery, "Validation of a dietary history questionnaire for american indian and alaska native people," *Ethnicity Disease*, vol. 20, no. 4, pp. 429–36, Feb. 2011.

[13] N. D. Wright, A. E. Groisman-Perelstein, J. Wylie-Rosett, N. Vernon, P. M. Diamantis, and C. R. Isasi, "A lifestyle assessment and intervention tool for pediatric weight management: The habits questionnaire," *J. Human Nutrition Dietetics*, vol. 24, no. 1, pp. 96–100, Feb. 2011.

[14] A. F. Smith, S. D. Baxter, J. W. Hardin, C. H. Guinn, and J. A. Royer, "Relation of children's dietary reporting accuracy to cognitive ability," *Amer. J. Epidemiology*, vol. 173, no. 1, pp. 103–9, Jan. 2011.

[15] A. E. Dutman, A. Stafleu, A. Kruizinga, H. A. Brants, K. R. Westerterp, C. Kistemaker, W. J. Meuling, and R. A. Goldbohm, "Validation of an FFQ and options for data processing using the doubly labelled water method in children," *Public Health Nutrition*, vol. 14, pp. 1–8, Aug. 2010.

[16] M. Aubertin-Leheudre, A. Koskela, A. Samaletdin, and H. Adlercreutz, "Plasma alkylresorcinol metabolites as potential biomarkers of whole-grain wheat and rye cereal fibre intakes in women," *British J. Nutrition*, vol. 103, no. 3, pp. 339–43, Feb. 2010.

[17] P. B. Ryan, K. A. Scanlon, and D. L. MacIntosh, "Analysis of dietary intake of selected metals in the nhexas-maryland investigation," *Environ. Health Perspectives*, vol. 109, no. 2, pp. 121–128, Feb. 2001.

[18] P. Viola and M. Jones, "Robust real-time face detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jul. 2001, pp. 747–747.

[19] H. Schneiderman and T. Kanade, "A statistical method for 3d object detection applied to faces and cars," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2000, pp. 746–751.

[20] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[21] J. Shotton, M. Johnson, and R. Cipolla, "Semantic texton forests for image categorization and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2008, pp. 1–8.

[22] M. Varma and D. Ray, "Learning the discriminative power-invariance trade-off," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

[23] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman, "Multiple kernels for object detection," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, 2009, pp. 606–613.

[24] M. Chen, K. Dhingra, W. Wu, and L. Yang, "PFID: Pittsburgh fast-food image dataset," in *Proc. IEEE Int. Conf. Image Process.*, Jan. 2009, pp. 289–292.

Authors' photographs and biographies not available at the time of publication.