Menu-Match: Restaurant-Specific Food Logging from Images

Oscar Beijbom[1,2] Neel Joshi[1] Dan Morris[1] Scott Saponas[1] Siddharth Khullar[1]
[1]Microsoft Research [2]University of California, San Diego
obeijbom@ucsd.edu, {neel,dan,ssaponas}@microsoft.com, siddharth.khullar@gmail.com

Abstract

Logging food and calorie intake has been shown to facilitate weight management. Unfortunately, current food logging methods are time-consuming and cumbersome, which limits their effectiveness. To address this limitation, we present an automated computer vision system for logging food and calorie intake using images. We focus on the "restaurant" scenario, which is often a challenging aspect of diet management. We introduce a key insight that addresses this problem specifically: restaurant plates are often both nutritionally and visually consistent across many servings. This insight provides a path to robust calorie estimation from a single RGB photograph: using a database of known food items together with restaurant-specific classifiers, calorie estimation can be achieved through identification followed by calorie lookup. As demonstrated on a challenging Menu-Match dataset and an existing third-party dataset, our approach outperforms previous computer vision methods and a commercial calorie estimation app. Our Menu-Match dataset of realistic restaurant meals is made publicly available.

1. Introduction

Obesity has been linked to cardiovascular disease, diabetes, and cancer, and dramatically impacts both life expectancy and quality of life [16]. Furthermore, the rapid rise in the prevalence of obesity presents a critical public health concern [14]. Diet and exercise are critical to combating obesity; however, changing dietary and exercise habits are often difficult. It has been shown that exercise logging and food logging support such changes, and logging is well-correlated to increased initial weight loss and better weight maintenance [18, 7, 32].

Unfortunately, the effectiveness of logging is often limited by inconvenience. While there have been significant strides in automatic tracking of exercise and activity, such as with GPS devices and step counters, food logging is still a tedious manual process. However, given the ubiquity of smartphone cameras and the emergence of wearable cameras, taking a photo of one's meal is already a fairly easy task. Furthermore, studies have shown that the simple act of photographing your meal encourages weight loss [37]. Combining such photos with computer vision algorithms offers a compelling way to significantly reduce the barrier to food logging.

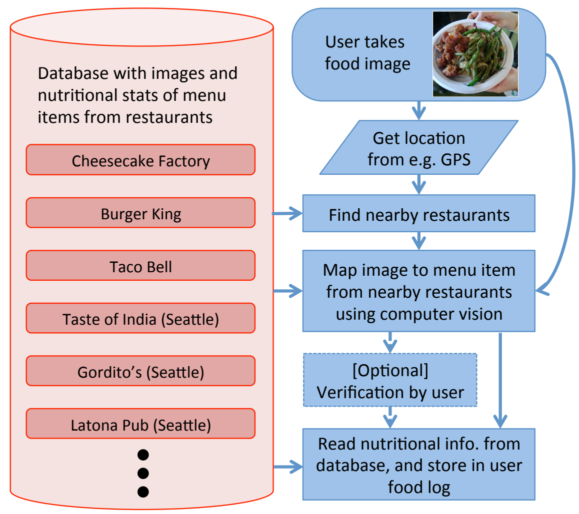

Figure 1. Overview of the proposed Menu-Match system.

In this paper, we discuss approaches to lowering the barrier to food tracking using computer vision. Our work addresses a gap in previous research: we focus on the restaurant scenario, which is typically a challenge for diet management, as it is hard to control portions and track ingredients. Also, compared to home cooking, restaurant foods are generally less healthy [36]. Based on the observation that restaurant meals are typically visually and nutritionally similar across servings, our system – Menu-Match – uses computer vision to identify the food items and estimate calories from a single food image by utilizing a database of known food items, i.e., a menu[1]. This allows the challenging problem of bottom-up calorie estimation to be mapped

[1]We use calorie estimation as a running example of a food statistic in our discussion and experiment, but this can easily be replaced by other types of nutritional information.

to an easier identification problem.

Our contributions include: (1) a new paradigm of calorie estimation and food identification from a single image, (2) an end-to-end computer vision pipeline that achieves state-of-the-art calorie estimation accuracy, and (3) a benchmark dataset for calorie estimation from realistic food images.

2. Previous Work

The benefits of food logging have been studied extensively [31, 15], and here we focus on the literature related to automated information extraction from food images.

The general problem of inferring nutritional information from a single food image is challenging for several reasons. First, there may be significant occlusions (e.g., a bread stick hidden under a side of cole slaw), resulting in missing information. Second, it is highly unlikely that visual information alone conveys all the details of food preparation (amount of oil, fat content of meats, etc.) that strongly impact nutritional content. Third, accurate volume estimation from a single image is very challenging.

Consequently, there is no work that we are aware of that attempts to estimate nutritional statistics (e.g., calories) from a single image of a realistic meal. One line of work relaxes the single-image assumption and utilizes auxiliary hardware such as calibration targets [38], multiple images [22, 12], laser scanners [29], or structured light [10]. This reduces the usability of the proposed systems and often assumes unrealistic arrangements of the food items on a plate. Other work relaxes the goal of estimating nutritional statistics from realistic images and focuses on various aspects of the computer vision challenges. Yang *et al*. [35] propose a novel feature descriptor but evaluate only on the highly controlled Pittsburgh food dataset [9], which does not represent real-world food images. Bosch *et al*. [4] use both local and global features in a voting scheme, Anthimopoulos *et al*. [1] propose a system based on a bag-of-features model, and Hoashi *et al*. [17] use multiple kernel learning for food classification, however none attempt to estimate calories. Noronha *et al*. [26] bypass computer vision altogether and investigate the feasibility of crowd-sourced assessment of nutritional information. They demonstrate results very similar to those supplied by a dietitian at the cost of significant human input.

The work of Kitamura *et al*. [21] is the closest to our approach. Their work utilizes a nutritional table with five categories: grain, vegetable, meat/fish/beans, fruit, and milk. User-supplied images are mapped to these categories, and serving sizes are supplied by the user. This work is limited by the granularity of the nutritional table: coarse nutritional information carries large standard deviations of serving counts, preventing accurate calorie estimation. In addition, it relies on user-supplied volume estimates.

3. Restaurant-Specific Recognition

Any vision-based method for calorie estimation must have access to a database of nutritional information for a number of food items, and it is in the granularity of this database where methods differ. On one extreme, databases could contain fundamental nutritional building blocks such as oils, fats, proteins, and minerals. Such a database could be made very short and accurate (e.g., one gram of olive oil contains 8.8 calories). However, mapping visual information to this database is hard and inevitably inaccurate. Instead, most methods resort to a coarser database containing food categories such as grain, fish, and fruit [21] or atomic food items such as hamburger, orange, apple, steak, and sandwich [22, 35]. This level of granularity is easier to resolve visually, but the database entries carry large standard deviations (e.g., one hamburger may differ radically from another in calories, even independent of size).

In this work, we consider an alternative problem formulation, where the database contains atomic items *as they are served at specific restaurants*. Equipped with such a database, a meal can be directly classified as, for example, "the cheeseburger at Joe's at Solo Grill in Toronto", and accurate nutritional statistics can be read from the database. Given that in many restaurants any given food item is fairly nutritionally consistent from plate to plate, such identification offers a compelling path to accurate nutritional information. This is, to our knowledge, the first paper to consider this approach.

Restaurant-specific food recognition has the potential to resolve the problems listed above for the "restaurant" scenario. First, as we no longer need to identify every item, but rather do a holistic assessment of the plate, occlusions cause fewer problems. Second, by considering the meal as a whole entity, ingredients and preparation details are encoded into the database. Third, volume estimation is no longer needed.

While a large database of menu items, nutritional information, and sample images does not yet exist, we argue that it is feasible to create. Restaurants' menus and nutritional information are commonly available online though sites such as Yelp, Foursquare, or restaurants' websites. We believe that the associated sample images could be collected either as a top-down database (e.g., a company deploying this system could bootstrap the database by collecting data for restaurants in major cities), or as a bottom-up database (populated by leveraging prevalent social media use in restaurants, e.g., Yelp, Twitter, Foursquare, and Instagram, with direct contributions from participating restaurants).

Given such a database for a large number of restaurants, we propose an application in which location information commonly available on mobile devices (e.g., GPS) restricts the search for a particular image to a small set of nearby

Figure 2. A sample of the food images from our Menu-Match dataset. Top row: Asian restaurant, middle row: Italian restaurant, bottom row: soup restaurant.

restaurants, which greatly simplifies recognition and offers a plausible path to robust, accurate mapping of images to nutritional information.

4. Menu-Match

The proposed method is detailed in this section.

4.1. Assumptions

While we assume the existence of a detailed database of food items, we allow for several food items in a single image, e.g., a side of bread next to soup, or a serving of curry along with a bowl of rice and naan. Again, our assumption implies that these items are specific to the restaurant where they are served, allowing identification to accurately predict nutritional content.

While we assume that the nutritional content and general visual appearance is consistent for the same menu item at the same restaurant, we make *no assumptions* about the consistency of the spatial arrangement of that meal. On the contrary, our computer vision framework is specifically designed to be invariant to spatial arrangement. Foods that fall outside the scope of this paper include meals where serving sizes and ingredients vary by customer, such as salad bars. Home cooking also largely falls outside the scope of the proposed method, although a user-specific database of home-cooked or pre-made meals could also leverage our techniques.

4.2. Dataset

To evaluate our method, we collected a dataset of actual meal images from three local restaurants. The images were captured by five photographers using a mixture of six different models of smartphones and one point-and-shoot camera. The photographers captured *one* image of each meal after that meal was ordered by the customers, with instructions to capture images from arbitrary angles and at varying distances in a manner similar to what a user of an envisioned Menu-Match system would do. We stress the evaluation of our approach on multiple restaurants; GPS will often be unable to restrict the search to a single restaurant, so our system needs to work on multiple menus concurrently.

Sample images are shown in Figure 2: images in the top row are from an Asian restaurant, the middle row from an Italian restaurant, and the bottom row from a soup restaurant. The Asian restaurant offers a buffet-style setup where customers select 1-3 toppings that are served with a fixed serving size with brown or white rice. The Italian restaurant offers a variety of pizza, lasagna, and pasta, served with sides of breadsticks or salad. The soup restaurant offers ten soups with a side of one of five breads.

The dataset contains a total of 646 images, with 1386 tagged food items across 41 categories. In addition, calorie counts for all food items were provided by a dietitian that works with the restaurants. The calorie counts are close to ground truth since the dietitian had access to ingredients and recipes. Our Menu-Match dataset is thus unique in that it contains both accurate nutritional information and realistic food images. By contrast, several other computer vision food datasets [9, 38] only contain highly controlled photographs and well-separated food items. Certain categories from the '50 foods' dataset of Chen *et al*. [10] contain realistic food images, but there is no nutritional metadata. The dataset of Kitamura *et al*. [21], which contains approximate nutritional metadata and realistic photographs, is unfortunately not publicly available. Our dataset is publically availible[2].

4.3. Recognition Framework

We employ an image recognition framework based on the *bag of visual words* approach [30, 20]. In the first step, five types of base features are extracted from the images: color [19], histogram of oriented gradients (HOG) [11], scale-invariant feature transforms (SIFT) [24], local binary patterns (LBP) [27], and filter responses from the MR8 filter bank [33]. These base features are encoded with locality-constrained linear encoding (LLC) [34], using a dictionary with 1024 words learned via k-means clustering. The encoded base features are then pooled using max-pooling [34] in a rotation-invariant pooling scheme [2]. The pooling procedure, which is shown in Figure 3, is done at six scales as follows. Let d be the largest dimension of the image (i.e. the maximum across image width and height). The width of the square pooling regions is given by 6 values between $\log_{10}(d)$ and $\log_{10}(d/5)$, equally spaced on a $\log_{10}$ scale. Encoded base features are then pooled in centered squares of these widths and concatenated. For example, if the image is 500 pixels wide and 400 pixels tall, d will be 500, and the

[2] `http://research.microsoft.com/menumatch/data/`

Figure 3. Comparison of pooling methods. Left: rotationally invariant pooling scheme [2] used in this paper. Each square represent a pooling region, with the largest square covering the whole image. Right: spatial pyramid pooling scheme of [23] with 2 levels. The image is pooled across the 16 small regions, the 4 larger regions, and the whole image for a total of 21 regions.

pooling regions will be: $[100, 134, 190, 263, 362, 500]$. After pooling, the image is represented by five feature descriptors (one for each feature type), each with $6 * 1024 = 6144$ dimensions. For all experiments below, we adopted a 10-fold cross-validation procedure, where the data was split randomly into 10 sets so that 9 were used for training and 1 for testing.

Implementation Details: Our implementation largely follow that of a publicly available vision library[3]. The only pre-processing common to all base features was that images were rescaled so that the largest dimension is 500 pixels. Neither color nor contrast correction was performed. All base features are extracted at a 4 by 4 grid across the image plane. Each base feature was encoded using LLC [34] and pooled using the rotationally invariant pooling scheme. Color base features were extracted by first mapping each pixel to an integer color code using the discriminatory color encoding scheme of Kahn *et al*. [19]. These color codes were then mean-pooled across patches of 6, 8, 10, 12, 14, and 16 pixels to create several base features at each location in the image. The SIFT base feature was extracted at patch sizes of 8, 16, and 24 pixels at each location in the image. The other base features – HOG, LBP and MR8 – were extracted on a single scale using the standard procedure as given in the original publications and implemented by the aforementioned library. For the dictionary learning, 1000 descriptors were extracted at random locations from each image. For the LLC encoding, 3 nearest neighbors were used.

4.4. Semi-Automated Food Item Identification

Given location information and the Menu-Match database proposed in this work, a system could be designed where the user navigates the available menu options to find the correct meal. Indeed, a well-designed user interface could go a long way toward facilitating food logging *without using any computer vision* by incorporating quick search functions, auto-complete, etc. However, we have identified three reasons why it may be beneficial to incorporate computer vision into such a system. First, as shown by Zepada and Deal [37], the simple act of photographing your meal does in itself encourage weight loss. Second, new input device designs, such as smart- watches or glasses, may be smaller and less suitible for text input. Third, recent research has shown that computer vision can be incorporated in hybrid user interfaces to reduce manual effort [6]. Such hybrid interfaces could reduce the barrier to food logging compared to a fully manual interface. This is particularly important in situations where several restaurants are clustered together (e.g., a food court) and the list of nearby restaurants is large or if the menus for the nearby restaurants are large. Here we investigate to what degree computer vision can sort the food items to facilitate rapid verification by the user.

The efficacy of such sorting was evaluated on our Menu-Match dataset. A one-versus-rest linear Support Vector Machine (SVM) was trained for each of the 41 food items and each of the 5 feature types separately [13, 5]. These $5 * 41 = 205$ classifiers were then applied to the training set yielding a new 205-dimensional *joint* feature vector of concatenated decision values for each training image. Finally, a one-versus-rest linear SVM was trained for each of the 41 food items using this joint representation. This method of merging feature types is commonly referred to as late fusion [20, 19], and, as shown in Figures 6 & 7, significantly boosts performance over any individual feature type. Cross validation was used to determine the appropriate SVM regularization. Note that since there are often several food items in an image and we do *not* leverage any spatial information (e.g., bounding boxes from the labeling process), an image will often have multiple labels. This was handled during training by using any image with multiple labels as a positive sample for all of its corresponding labels. The resulting classifier takes a new image and assigns a classification score to each food item on a selected set of menus. A sorted list of food items can then be displayed to the user who verifies which items are present on the plate.

4.5. Fully Automated Estimation Of Food Statistics

In the previous section, we described how to create a sorted list of food items for verification by the user. However, a method that automatically estimates calorie content, or other nutritional statistics, is more appealing from a user perspective. We have developed a method based on regression that directly estimates calories from a meal image. Specifically, we concatenated the five feature descriptors detailed above to a $6144 * 5 = 30720$-dimensional feature vector for each meal image. This feature representation was used together with the total calorie count for each meal image to learn a mapping directly from this feature space to calories using Support Vector Regression [13]. The calorie count for a new image can then be estimated directly by re-

[3] github.com/adikhosla/feature-extraction

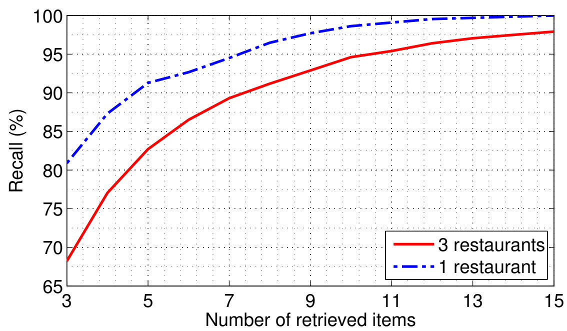

Figure 4. Average recall as a function of retrieval set size for two localization scenarios.

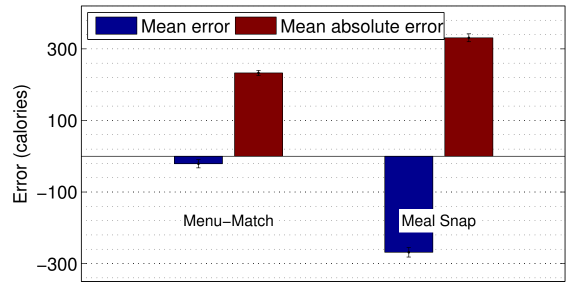

Figure 5. Mean errors (bias) and mean absolute error (average error magnitude) of the proposed method and Meal Snap on our three restaurant datasets. Error bars correspond to standard error.

gressing from feature space to calories. Note that a deployment of this method assumes, as previously, that location information is availible for a new image, so that a regressor for the nearby restaurant(s) can be utilized.

5. Results

5.1. Semi-Automated Food Item Identification

The efficacy of our semi-automated approach is shown in Figure 4, where average recall is shown as a function of the number of retrieved items for two localization scenarios. One where the localization information (e.g., from GPS) has narrowed down the search to three restaurants (in this case our whole test dataset) and one where it has narrowed down the search to a single restaurant. Note that the mean and max number of food items per plate in our Menu-Match dataset is 2.1 and 4 respectively, so the recall for less than 4 items will always be less than 100%. The results show average recall rates of 83% or 92% for a list of 5 food items if these are drawn from the menus from all three restaurants or from a single restaurant, respectively. This sorting strategy can thus facilitate rapid selection of correct food items by the user. Quantitative results are shown in Figure 9.

5.2. Fully Automated Estimation of Food Statistics

The result of our calorie estimation method is shown in Figure 5. Our method achieves a calorie estimation error (i.e., bias) of -21.0 ± 11.6 and an absolute error of 232 ± 7.2 (mean $\pm$ standard error). For comparison, we ran the test images through Meal Snap, a commercial app for food logging from images. Meal Snap returns a calorie range per image, and following [26] we report the mean of this range. As shown in Figure 5, Meal Snap achieves a calorie estimation error of -268.5 ± 13.3 and an absolute error of 330.9 ± 11.0 (mean $\pm$ standard error). These errors are significantly higher than those achieved by Menu-Match and indicate the efficacy of the proposed approach[4].

To the best of our knowledge, the performance of Menu-Match is state of the art. To put our performance in some context, we consider Noronha *et al.*'s reported mean calorie estimation errors on their dataset of 18 food images. They report a calorie estimation error of (mean $\pm$ standard error): -174 ± 76, 30 ± 42, and 51 ± 52 for three expert dietitians, -211 ± 122 for Meal Snap, and 36 ± 66 for their proposed system, PlateMate, that is based on crowdsourcing [26]. The mean absolute errors were 233 ± 67, 119 ± 28, 151 ± 39, 316 ± 106, and 192 ± 50 respectively. Menu-Match thus had lower bias than the experts and PlateMate. The absolute errors of Menu-Match were on par with PlateMate and the first expert, but larger than the other two experts.

Furthermore, our method compares favorably to amateur self-reports, where error rates can exceed 400 calories per day [8, 28]. Our approach also lacks the systematic bias towards underestimating calorie counts commonly seen in self-reports, particularly among vulnerable users [28].

We were unfortunately not able to acquire Noronha *et al.*'s [26] dataset to perform direct comparisons to our approach, nor were we able to compare our calorie estimation results against those of other computer vision methods, such as Chen *et al.* [10] or Kitamura *et al.* [21] as they have not made their code publicly available nor published results that detail calorie estimation errors.

5.3. Analysis of Proposed Methodology

In this section, we provide a more detailed analysis of the proposed computer vision method.

5.3.1 Relative importance of visual features

The importance of fusing several features is established in Figure 6 where the *joint* feature representation significantly outperforms the performance of any single feature type. The color [19] feature is the strongest of any individual feature, followed by the mr8 [33] texture-based feature.

[4] Meal Snap (`mealsnap.com`) does not allow upload of previously taken images. Instead, images were displayed one-by-one on a high-resolution computer screen and re-photographed using the image capture function of the Meal Snap App running on an iPhone 4s. To ensure a fair comparison with Meal Snap, these re-photographed images were then downloaded from the iPhone and used by our method for the calorie estimation experiment.

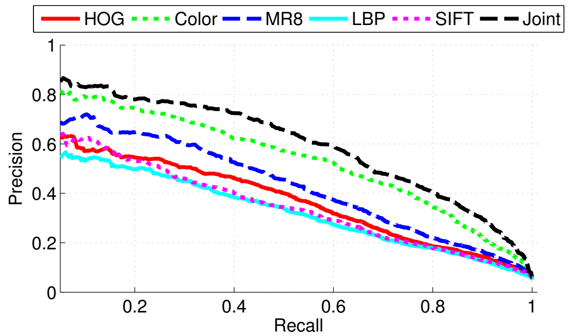

Figure 6. Precision vs. recall curve for our method applied to our dataset indicating the superior performance of the *joint* representation that fuses the other feature types.

The gradient-based HOG [11] and SIFT [24] that are widely used for object recognition are weaker, supporting the intuition that texture and color are the most useful features to describe food images.

5.3.2 Rotationally invariant pooling

The rotationally invariant pooling method increased the mean average precision (mAP [25]) for the *joint* feature from 38.3% to 51.2% compared to the traditional spatial pyramid pooling [23] (Figure 7). This may be because food images are commonly captured top-down as opposed to "regular" photographs commonly captured sideways. When pictures are taken top-down there is no ordering imposed by gravity (e.g., sky in the top of the images and grass in the bottom). Instead, a reasonable assumption is that food tends to be in the middle of the photograph, hence the efficacy of the deployed pooling scheme.

5.3.3 Evaluation of required training set size

A successful deployment of the proposed method requires training images from all included restaurants. As this is a core challenge in deploying our method, we evaluated the sensitivity of the proposed recognition framework to training set size. The results are shown in Figure 8. As expected, the results indicate a steady increase in accuracy with training set size. However, the increase slows as more training images are added, and with 500 training images, the recall is already above 83%. Since our dataset contains around 40 classes, this experiment indicates that 10 images per class is a reasonable target value.

5.3.4 Generalization to other food datasets

The generality of the proposed recognition framework is evaluated on the dataset of Chen *et al.* [10]; this is to the best of our knowledge the only publicly available food dataset containing realistic food images of the type that a user might capture. We ran our recognition pipeline with *no* changes to parameters or settings on this dataset (Table 1). Our method achieved 77.4% accuracy, significantly outperforming the

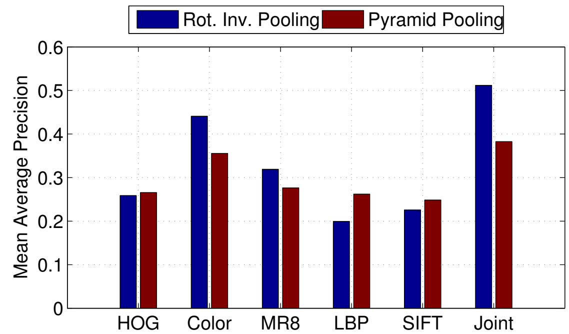

Figure 7. Mean Average Precision [25] of the rotationally invariant pooling scheme [2] as well as the spatial pyramid pooling [23].

proposed method of Chen *et al.*, which achieved 68.3%. Similarly, if allowed to suggest 5 candidate labels per photograph (this dataset only contains a single label per image), our method achieved 96.2% compared to 90.9% by Chen *et al.* This experiment strengthens our belief that the proposed feature extraction and recognition pipeline is generally applicable and has state of the art performance.

The superior performance of our method can be attributed to several differences in the algorithms. First, we use a unified framework where all base features are encoded with LLC and pooled in a rotationally invariant manner. By contrast, Chen *et al.* use a custom procedure for each feature type. For SIFT they use sparse coding, and mean pooling across the whole images plane. In contrast we use the more recent LLC encoding, which outperforms several previous methods including sparse coding, even for the same dictionary size [34] (Chen *et al.* also use a dictionary with 1024 words). The improved pooling and encoding scheme may explain why our SIFT descriptor is significantly stronger. Second, Chen *et al.* use a simple 96-bin RGB color histogram, while we adopt the method of Kahn *et al.* [19]. The color encoding of Kahn *et al.* is designed to be invariant to variations in light, but sensitive to semantically important color differences, which makes it a more suitable than RGB color histogram for classification tasks. Third, Chen *et al.* do not utilize HOG [11] and texton histograms [33], which both achieve high accuracy, but instead utilize a less discriminative descriptor based on Gabor filters. Chen *et al.* introduce a custom LBP [27] descriptor that achieves stronger result on their dataset than our generic LBP descriptor. However, due to advances mentioned above, our method achieves a higher final accuracy.

6. Conclusion and Future Work

We have presented a method for restaurant-specific food classification. This work addresses an important portion of the food-tracking space and adds to the ecosystem of methods that facilitate easier and more widespread diet logging and improve obesity management and public health. There

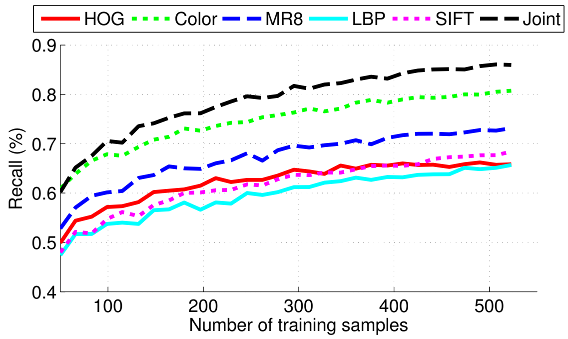

Figure 8. Average recall rate at 5 retrieved food items per image for different number of training images.

Feature	Chen [10]	Our Implementation
SIFT	53.0%	57.7%
LBP	45.9%	43.6%
Color	40.3%	45.0%
Gabor	26.5%	N.A.
HOG	N.A.	52.4%
MR8	N.A.	50.0%
Joint	68.3%	77.4%
Rank 5	90.9%	96.2%

Table 1. Results on the dataset of Chen *et al*. Note that there are differences in the implementation of the features in the two methods. The last row indicates how often the target food is in the list of 5 of the most likely items returned by the classifier.

are some situations where our approach may have limited success, such when meals do not come in discrete serving sizes, such as buffets or salad bars. It would also take additional effort to accurately handle home-cooked food, as this would require a custom menu and more consistent control over portions than typically exists in a home scenario. Lastly, takeout or delivery food is a challenge in that the location information where the food is consumed is not relevant; thus our approach would require a user to specify the restaurant manually.

Future work includes focusing on incorporating our methods into a deployed application, to evaluate user experience questions and larger, broader datasets. We also believe that these estimates can be significantly improved by incorporating user-specific customization (i.e., learning over time that a user tends to order certain items) as priors in the inference model, which can only be evaluated in a deployment context. Further, cost-sensitive learning may be utilized to directly minimize calorie estimation errors during training [3]. Other opportunities include using convolutional neural networks and the further development of images features for food recognition.

References

[1] M. Anthimopoulos, L. Gianola, L. Scarnato, P. Diem, and S. Mougiakakou. A food recognition system for diabetic patients based on an optimized bag-of-features model. *J. Biomed. Health Inform.*, 2014.

[2] O. Beijbom, P. J. Edmunds, D. I. Kline, B. G. Mitchell, and D. Kriegman. Automated annotation of coral reef survey images. In *CVPR*, 2012.

[3] O. Beijbom, M. Saberian, D. Kriegman, and N. Vasconcelos. Guess-averse loss functions for cost-sensitive multiclass boosting. In *ICML*, 2014.

[4] M. Bosch, F. Zhu, N. Khanna, C. J. Boushey, and E. J. Delp. Combining global and local features for food identification in dietary assessment. In *ICIP*, 2011.

[5] S. Branson, O. Beijbom, and S. Belongie. Efficient large-scale structured learning. In *CVPR*, 2013.

[6] S. Branson, C. Wah, F. Schroff, B. Babenko, P. Welinder, P. Perona, and S. Belongie. Visual recognition with humans in the loop. In *ECCV*, 2010.

[7] L. Burke, J. Wang, and M. Sevick. Self-monitoring in weight loss: A systematic review of the literature. *J Am Diet Assoc*, 2011.

[8] C. M. Champagne, G. A. Bray, A. A. Kurtz, J. B. R. Montiero, E. Tucker, J. Volaufova, and J. P. Delany. Energy intake and energy expenditure: a controlled study comparing dietitians and non-dietitians. *J Am Diet Assoc*, 2002.

[9] M. Chen, K. Dhingra, W. Wu, L. Yang, R. Sukthankar, and J. Yang. Pfid: Pittsburgh fast-food image dataset. In *ICIP*, 2009.

[10] M.-Y. Chen, Y.-H. Yang, C.-J. Ho, S.-H. Wang, S.-M. Liu, E. Chang, C.-H. Yeh, and M. Ouhyoung. Automatic chinese food identification and quantity estimation. In *SIGGRAPH Technical Briefs*, 2012.

[11] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[12] J. Dehais, S. Shevchik, P. Diem, and S. G. Mougiakakou. Food volume computation for self dietary assessment applications. In *BIBE*, 2013.

[13] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *JMLR*, 2008.

[14] K. Flegal, M. Carroll, C. Ogden, and L. Curtin. Prevalence and trends in obesity among us adults, 1999-2008. *JAMA*, 2010.

[15] K. Glanz, S. Murphy, J. Moylan, D. Evensen, and J. D. Curb. Improving dietary self-monitoring and adherence with hand-held computers: A pilot study. *Am J Health Promot*, 2006.

[16] D. Haslam and W. James. Obesity. *Lancet*, 2005.

[17] H. Hoashi, T. Joutou, and K. Yanai. Image recognition of 85 food categories by feature fusion. In *ISM*, 2010.

[18] J. Hollis, C. Gullion, and V. Stevens. Weight loss during the intensive intervention phase of the weight-loss maintenance trial. *Amer J Prev Med*, 2008.

[19] R. Khan, J. Van de Weijer, F. S. Khan, D. Muselet, C. Ducottet, and C. Barat. Discriminative color descriptors. *CVPR*, 2013.

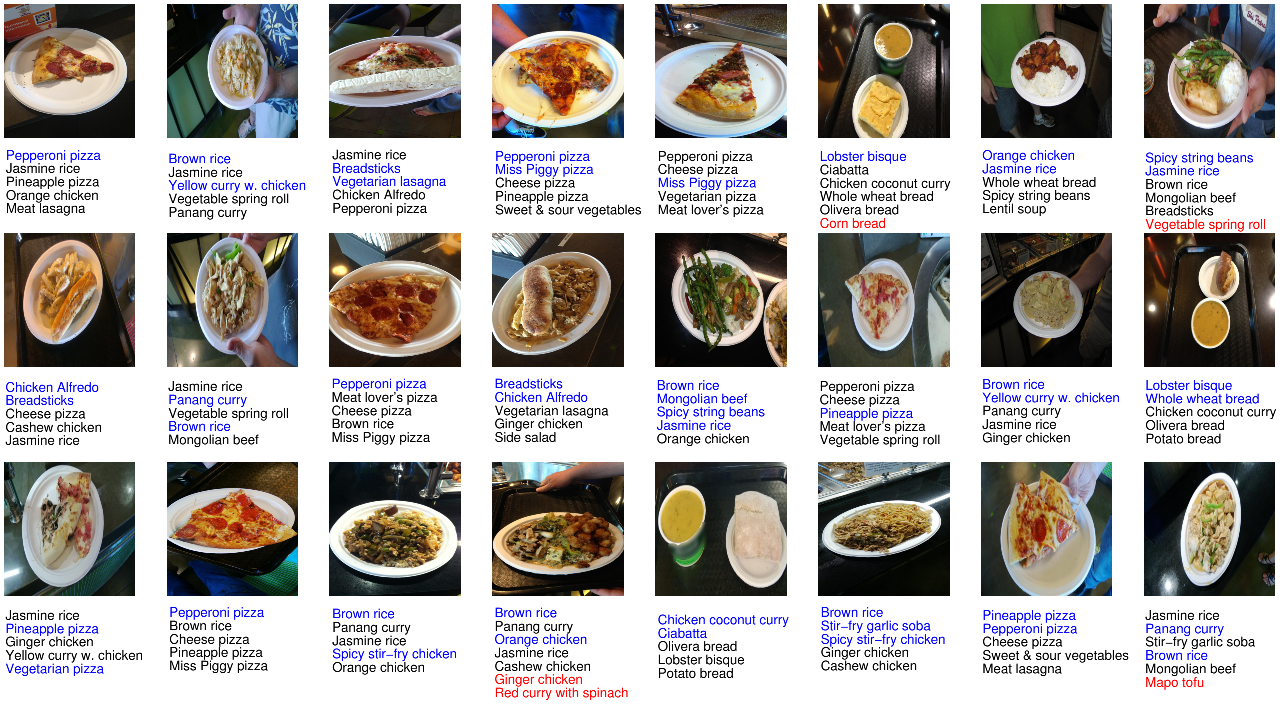

Figure 9. Qualitative results of our Menu-Match tagging system showing 24 randomly selected images. Under the images are the top 5 retrieved from the system as described in Section 4.4, shown in the retrieved order. Correctly retrieved food items (*i.e.* food items that are on the plate) are written in blue, and incorrectly retrieved food items (*i.e.* retrieved food-items that were *not* on the plate) are written in black. Food items on the plate that were not among the five retrieved candidates are written in red.

[20] A. Khosla, J. Xiao, A. Torralba, and A. Oliva. Memorability of image regions. In *NIPS*, 2012.

[21] K. Kitamura, T. Yamasaki, and K. Aizawa. Foodlog: capture, analysis and retrieval of personal food images via web. In *ACM workshop*, 2009.

[22] F. Kong and J. Tan. Dietcam: Automatic dietary assessment with mobile camera phones. *PMC*, 2012.

[23] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[24] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999.

[25] H. Müller, W. Müller, D. M. Squire, S. Marchand-Maillet, and T. Pun. Performance evaluation in content-based image retrieval: overview and proposals. *Pattern Recogn Lett*, 2001.

[26] J. Noronha, E. Hysen, H. Zhang, and K. Z. Gajos. Platemate: crowdsourcing nutritional analysis from food photographs. In *UIST*, 2012.

[27] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *PAMI*, 2002.

[28] D. A. Schoeller, L. G. Bandini, and W. H. Dietz. Inaccuracies in self-reported intake identified by comparison with the doubly labelled water method. *Can J Physiol Pharm*, 1990.

[29] J. Shang, E. Pepin, E. Johnson, D. Hazel, A. Teredesai, A. Kristal, and A. Mamishev. Dietary intake assessment using integrated sensors and software. In *IS&T/SPIE Electronic Imaging*, 2012.

[30] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *CVPR*, 2003.

[31] J. M. Spahn, R. S. Reeves, K. S. Keim, I. Laquatra, M. Kellogg, B. Jortberg, and N. A. Clark. State of the evidence regarding behavior change theories and strategies in nutrition counseling to facilitate health and food behavior change. *J Am Diet Assoc*, 2010.

[32] L. Svetkey and et al. Comparison of strategies for sustaining weight loss: The weight loss maintenance randomized controlled trial. *JAMA*, 2008.

[33] M. Varma and A. Zisserman. A statistical approach to texture classification from single images. *IJCV*, 2005.

[34] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010.

[35] S. Yang, M. Chen, D. Pomerleau, and R. Sukthankar. Food recognition using statistics of pairwise local features. In *CVPR*, 2010.

[36] L. R. Young and M. Nestle. The contribution of expanding portion sizes to the us obesity epidemic. *Amer J Public Health*, 2002.

[37] L. Zepeda and D. Deal. Think before you eat: photographic food diaries as intervention tools to change dietary decision making and attitudes. *Int J Consum Stud*, 2008.

[38] F. Zhu, M. Bosch, I. Woo, S. Kim, C. J. Boushey, D. S. Ebert, and E. J. Delp. The use of mobile devices in aiding dietary assessment and evaluation. *STSP*, 2010.