Leveraging Context to Support Automated Food Recognition in Restaurants

Vinay Bettadapura, Edison Thomaz, Aman Parnami, Gregory D. Abowd, Irfan Essa

School of Interactive Computing
Georgia Institute of Technology, Atlanta, Georgia, USA

http://www.vbettadapura.com/egocentric/food

Abstract

The pervasiveness of mobile cameras has resulted in a dramatic increase in food photos, which are pictures reflecting what people eat. In this paper, we study how taking pictures of what we eat in restaurants can be used for the purpose of automating food journaling. We propose to leverage the context of where the picture was taken, with additional information about the restaurant, available online, coupled with state-of-the-art computer vision techniques to recognize the food being consumed. To this end, we demonstrate image-based recognition of foods eaten in restaurants by training a classifier with images from restaurant's online menu databases. We evaluate the performance of our system in unconstrained, real-world settings with food images taken in 10 restaurants across 5 different types of food (American, Indian, Italian, Mexican and Thai).

1. Introduction

Recent studies show strong evidence that adherence to dietary self-monitoring helps people lose weight and meet dietary goals [5]. This is critically important since obesity is now a major public health concern associated with rising rates of chronic disease and early death [13].

Although numerous methods have been suggested for addressing the problem of poor adherence to nutrition journaling [1, 19, 27], a truly practical system for objective dietary monitoring has not yet been realized; the most common technique for logging eating habits today remains self-reports through paper diaries and more recently, smartphone applications. This process is tedious, time-consuming, prone to errors and leads to selective under reporting [10].

While needs for automated food journaling persist, we are seeing an ever increasing growth in people photographing what they eat. In this paper we present a system and approach for automatically recognizing foods eaten at restaurants from first-person food photos with the goal of facilitating food journaling. The methodology we employ is unique because it leverages sensor data (i.e., location) captured at the time photos are taken. Additionally, online resources such as restaurant menus and online images are used to help recognize foods once a location has been identified.

Our motivation for focusing on restaurant eating activities stems from findings from recent surveys indicating a trend towards eating out versus eating at home. In 1970, 25.9 percent of all food spending was on food away from home; by 2012, that share rose to its highest level of 43.1 percent [23]. Additionally, 8 in 10 Americans report eating at fast-food restaurants at least monthly, with almost half saying they eat fast food at least weekly [9].

Research in the computer vision community has explored the recognition of either a small sub-set of food types in controlled laboratory environments [6, 26] or food images obtained from the web [11]. However, there have been only a few validated implementations that address the challenge of food recognition from images taken "in the wild" [12]. Systems that rely on crowdsourcing, such as PlateMate [18], have shown promise but are limited in terms of cost and scalability. Additionally, privacy concerns might arise when food photographs are reviewed by untrusted human computation workers [21].

In this paper, we seek an approach that supports automatic recognition of food, leveraging the context of where the photograph was taken. Our contributions are:

- An automatic workflow where online resources are queried with contextual sensor data to find food images and additional information about the restaurant where the food picture was taken, with the intent to build classifiers for food recognition.
- An image classification approach using the SMO-MKL multi-class SVM classification framework with features extracted from test photographs.
- An in-the-wild evaluation of our approach with food images taken in 10 restaurants across 5 different types

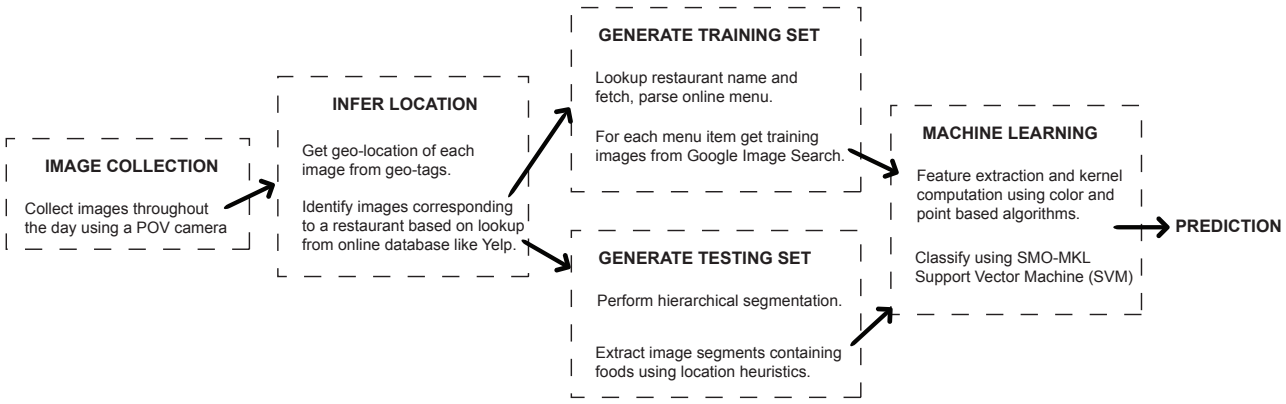

Figure 1. An overview of our automatic food recognition approach.

of cuisines (American, Indian, Italian, Mexican and Thai).

- A comparative evaluation focused on the effect of location data in food recognition results.

In this paper, we concentrate on food recognition, leveraging the additional context that is available (location, websites, etc.). Our goal in this paper is to in essence, using food and restaurants as the domain, demonstrate the value of external context, coupled with image recognition to support classification. We believe that the same method can be used for many other domains.

2. Related Work

Various sensor-based methods for automated dietary monitoring have been proposed over the years. Amft and Troster [1] explored sensors in the wrists, head and neck to automatically detect food intake gestures, chewing, and swallowing from accelerometer and acoustic sensor data. Sazonov et al. built a system for monitoring swallowing and chewing using a piezoelectric strain gauge positioned below the ear and a small microphone located over the laryngopharynx [19]. Yatani and Truong presented a wearable acoustic sensor attached to the user's neck [27] while Cheng et al. explored the use of a neckband for nutrition monitoring [7].

With the emergence of low-cost, high-resolution wearable cameras, recording individuals as they perform everyday activities such as eating has been gaining appeal [2]. In this approach, individuals wear cameras that take first-person point-of-view photographs periodically throughout the day. Although first-person point-of-view images offer a viable alternative to direct observation, one of the fundamental problems is image analysis. All captured images must be manually coded for salient content (e.g., evidence of eating activity), a process tends to be tedious and time-consuming.

Over the past decade, research in computer vision is moving towards "in the wild" approaches. Recent research has focussed on recognizing realistic actions in videos [15], unconstrained face verification and labeling [14] and objection detection and recognition in natural images [8]. Food recognition in the wild using vision-based methods is growing as a topic of interest, with Kitamura et al. [12] showing promise.

Finally, human computation lies in-between completely manual and fully-automated vision-based image analysis. PlateMate [18] crowdsources nutritional analysis from food photographs using Amazon Mechanical Turk, and Thomaz et al. investigated the use of crowdsourcing to detect [22] eating moments from first-person point-of-view images. Despite the promise of these crowdsourcing-based approaches, there are clear benefits to a fully automated method in economic terms, and possibly with regards to privacy as well.

3. Methodology

Recognizing foods from photographs is a challenging undertaking. The complexity arises from the large number of food categories, variations in their appearance and shape, the different ways in which they are served and the environmental conditions they are presented in. To offset the difficulty of this task, the methodology we propose (Figure 1) centers on the use of location information about the eating activity, and also restaurant menu databases that can be queried online. As noted, our technique is specifically aimed at eating activities in restaurants as we leverage the context of restaurant related information for classification.

Figure 2. Weakly-labeled training images obtained from Google Image search for 3 classes of food: **Left:** Basil Fried Rice; **Center:** Curry Katsu; **Right:** Lo Mein.

3.1. Image Acquisition

The first step in our approach involves the acquisition of food images. The popularity of cameras in smartphones and wearable devices like Google Glass makes it easy to capture food images in restaurants. In fact, many food photographs communities such as FoodGawker have emerged over the last several years, all centered on food photo sharing. Photographing food is also hitting major photo sharing sites like Instagram, Pinterest and Flickr, and food review sites like Yelp. These food-oriented photo activities illustrate the practicality of using manually-shot food photos for food recognition.

3.2. Geo-Localizing Images

The second step involves associating food photos with longitude and latitude coordinates. If the camera that is being used supports image geo-tagging, then the process of localizing images is greatly simplified. Commodity smartphones and cameras like the Contour and SenseCam come with built-in GPS capabilities. If the geo-tag is not available, image localization techniques can be used [28]. Once location is obtained for all captured images, the APIs of Yelp and Google Places are valuable for matching the images' geo-tags coincide with the geo-tag of a restaurant.

3.3. Weakly Supervised Learning

Being able to localize images to a restaurant greatly constrains the problem of food classification in the wild. A strong assumption can be made that the food present in the images must be from one of the items on the restaurant's menu. This key observation makes it possible to build a weakly supervised classification framework for food classification. The subsequent sections describe in detail the gathering of weakly-labeled training data, preparing the test data and classification using the SMO-MKL multi-class SVM classification framework [25].

3.3.1 Gathering Training Data

We start with collecting images localized to a particular restaurant R. Once we know R, we can use the web as a knowledge-base and search for R's menu. This task is greatly simplified thanks to online data-sources like Yelp, Google Places, Allmenus.com and Openmenu.com, which provides comprehensive databases of restaurant menus.

Let the menu for R be denoted by M_R and let the items on the menu be m_i. For each $m_i \in M_R$, the top 50 images of m_i are downloaded using search engines like Google Image search. This comprises the weakly-labeled training data. Three examples are shown in Figure 2. From the images, it is possible to see that there is a high degree of intra-class variability in terms of color and presentation of food. As is the case with any state-of-the-art object recognition system, our approach relies on the fact that given sufficient number of images for each class, it should be possible to learn common patterns and statistical similarities from the images.

3.3.2 Preparing Testing Data

The test images, localized to restaurant R, are segmented using hierarchical segmentation and the segments are extracted from parts of the image where we expect the food to be present [3]. The final set of segmented images forms our test data. An example is shown in Figure 3.

3.3.3 Feature Descriptors

Choosing the right combination of feature detectors, descriptors and classification backend is key to achieving good accuracy in any object recognition or image categorization task. While salient point detectors and corresponding region descriptors can robustly detect regions which are invariant to translation, rotation and scale [16, 17], illumination changes can still cause performance to drop. This is a cause of concern when dealing with food images, since images taken at restaurants are typically indoors and under

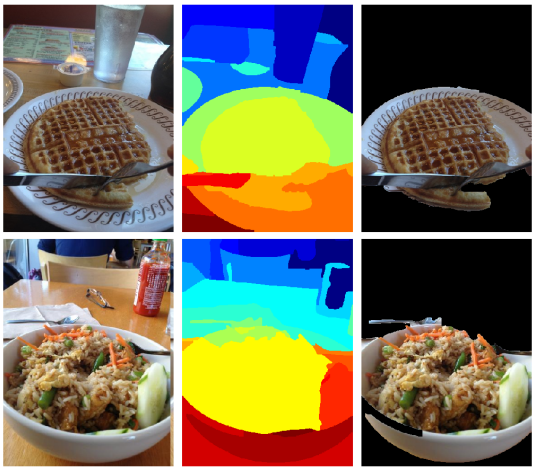

Figure 3. Extracting segments using hierarchical segmentation. The final segmented image is shown on the right.

varying lighting conditions. Recent work by van de Sande et al. [24] systematically studies the invariance properties and distinctiveness of color descriptors. The results of this study guided the choice of the descriptors in our approach. For the classification back-end, we use Multiple Kernel Learning (MKL), which in recent years, has given robust performance on object categorization tasks [4, 20, 25].

For feature extraction from the training and test data, a Harris-Laplace point detector is used since it has shown good performance for category recognition tasks [29] and is scale-invariant. However the choice of feature descriptor is more complicated. As seen in Figure 2, there is a high degree of intra-class variability in terms of color and lighting. Based on the recent work by van de Sande et al. [24] that studies the invariance properties and distinctiveness of various color descriptors on light intensity and color changes, we pick the following six descriptors, 2 color-based and 4 SIFT-based (Scale-Invariant Feature Transform [16]):

Color Moment Invariants: Generalized color moments M_{pq}^{abc} (of order $p+q$ and degree $a+b+c$) have been defined as $M_{pq}^{abc} = \int\int x^p y^q [I_R(x,y)]^a [I_G(x,y)]^b [I_B(x,y)]^c dxdy$. Color moment invariants are those combinations of generalized color moments that allow for normalization against photometric changes and are invariant to changes and shifts in light intensity and color.

Hue Histograms: Based on the observation that the certainty of hue is inversely proportional to the saturation, each hue sample in the hue histogram is weighted by its saturation. This helps overcome the (known) instability of hue near the gray axis in HSV space. The descriptors obtained are invariant to changes and shifts in light intensity.

C-SIFT: The descriptors are built using the C-invariant (normalized opponent color space). C-SIFT is invariant to changes in light intensity.

OpponentSIFT: All the channels in the opponent color space are described using SIFT descriptors. They are invariant to changes and shifts in light intensity.

RGB-SIFT: SIFT descriptors are computed for every RGB channel independently. The resulting descriptors are invariant to changes and shifts in light intensity and color.

SIFT: The original SIFT descriptor proposed by Lowe [16]. It is invariant to changes and shifts in light intensity.

3.3.4 Classification Using SMO-MKL

For a given restaurant R, 100,000 interest points are detected in the training data and for each of the 6 descriptors, visual codebooks are built using k-means clustering with k = 1000. Using these codebooks, bag-of-words (BoW) histograms are built for the training images. Similarly, interest points are detected in the test images and BoW are built for the 6 descriptors (using the visual codebooks generated with the training data).

For each of the 6 sets of BoW features, extended Gaussians kernels of the following form are computed:

$$K(H_i, H_j) = \exp(-\frac{1}{A} D(H_i, H_j)) \quad (1)$$

where $H_i = \{h_{in}\}$ and $H_j = \{h_{jn}\}$ are the BoW histograms (scaled between 0 to 1 such that they lie within a unit hypersphere) and $D(H_i, H_j)$ is the χ^2 distance defined as

$$D(H_i, H_j) = \frac{1}{2} \sum_{n=1}^{V} \frac{(h_{in} - h_{jn})^2}{h_{in} + h_{jn}} \quad (2)$$

where V is the vocabulary size (1000, in our case). The parameter A is the mean value of the distances between all the training examples [29]. Given the set of these N base kernels $\{K_k\}$ (in our case N = 6), linear MKL aims to learn a linear combination of the base kernels: $K = \sum_{k=1}^{N} \alpha_i K_i$

But the standard MKL formulation subject to l_1 regularization leads to a dual that is not differentiable. Hence the Sequential Minimal Optimization (SMO) algorithm cannot be applied and more expensive alternatives have to be pursued. Recently, Vishwanathan et al. showed that it is possible to use the SMO algorithm if the focus is on training p-norm MKL, with $p > 1$ [25]. They also show that the SMO-MKL algorithm is robust and significantly faster than the state-of-the-art p-norm MKL solvers. In our experiments, we train and test using the SMO-MKL SVM.

4. Study & Evaluation

We perform two sets of experiments to evaluate our approach. In the first set of experiments, we compare the feature extraction and classification techniques used in this paper, with the state-of-the-art food recognition algorithms on

the PFID benchmark data-set [6]. This validates our proposed approach. In the second set of experiments, we measure the performance of the proposed approach for "in-the-wild" food recognition.

4.1. Comparative Evaluations

We study the performance of the 6 feature descriptors and SMO-MKL classification on the PFID food data-set. The PFID dataset is a collection of 61 categories of fast food images acquired under lab conditions. Each category contains 3 different instances of food with 6 images from 6 view-points in each instance. In order to compare our results with the previous published results on PFID [6, 26], we follow the same protocol used by them, i.e. a 3-fold cross-validation is performed with 12 images from one instance being used for training while the other 6 images from the remaining instance are used for testing. The results of our experiments are shown in Figure 4. MKL gives the best performance and improves the state-of- the-art [26] by more than 20%. It is interesting to note that the SIFT descriptor used in our approach achieves 34.9% accuracy whereas the SIFT descriptor used in the PFID baseline [6] achieves 9.2% accuracy. The reason for this difference is that the authors of the PFID baseline use LIB-SVM for classification with its default parameters. However, by switching to the χ^2 kernel (and ensuring that the data is scaled) and by tuning the SVM parameters (through a grid-search over the space of C and γ), we can get a significant boost in performance with just SIFT features alone.

4.2. Food Recognition in Restaurants

To study the performance and the practicality of our approach, experiments were conducted on images collected from restaurants across 5 different cuisines: American, Indian, Italian, Mexican and Thai. To discount for user and location bias, 3 different individuals collected images on different days from 10 different restaurants (2 per cuisines). The data collection was done in two phases. In the first phase, the food images were captured using smartphone cameras. In total, 300 "in-the-wild" food images (5 cuisines $\times$ 6 dishes/cuisine $\times$ 10 images/dish) were obtained. In the second phase, data collection was repeated using a Google Glass and an additional 300 images were captured. These 600 "in-the-wild" images, form our test data-set. A sample of these test images is shown in Figure 5.

Using the geo-location information, the menu for each restaurant was automatically retrieved. For our experiments, we restricted the training to 15 dishes from each cuisine (selected based on online popularity). For each of the 15 dishes on the menu, 50 training images were downloaded using Google Image search. Thus, a total of 3,750 weakly-labeled training images were downloaded (5 cuisines $\times$ 15 menu-items/cuisine $\times$ 50 training-images/menu-item).

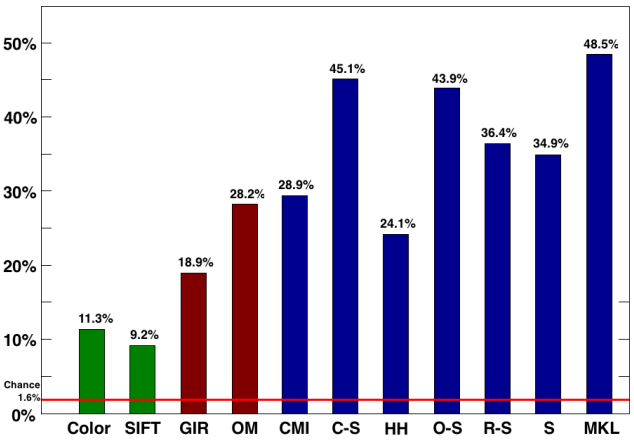

Figure 4. Performance of the 6 feature descriptors and SMO-MKL on the PFID data-set. The first two results (shown in green) are the baseline for PFID published by [6]. The next two (shown in red) are the results obtained by using Global Ingredient Representation (GIR) and Orientation and Midpoint Category (OM) [26]. The rest of the results (in blue) are one ones obtained using the 6 feature descriptors and MKL (CMI: Color Moment Invariant, C-S: C-SIFT, HH: Hue-Histogram, O-S: OpponentSIFT, R-S: RGB-SIFT, S: SIFT and MKL: Multiple Kernel Learning). MKL gives the best performance on this data-set.

Figure 5. Sample (12 out of 600) of the "in-the-wild" images used in testing.

Next, we perform interest point detection, feature extraction, codebook building for BoW representation, kernel pre-computation and finally classification using SMO-MKL. The results are summarized in Table 1 and the individual confusion matrices are shown in Figure 6. We achieve good classification accuracy with American, Indian and Italian cuisines. However, for the Mexican and Thai

	CMI	C-S	HH	O-S	R-S	S	MKL
American	45.8%	51.7%	43.3%	43.3%	37.5%	29.2%	**67.5%**
Indian	44.2%	74.2%	55.0%	59.2%	69.2%	65.0%	**80.8%**
Italian	33.3%	52.5%	67.5%	**74.2%**	66.7%	49.2%	67.5%
Mexican	36.7%	35.8%	20.8%	37.5%	24.2%	33.3%	**43.3%**
Thai	27.5%	36.7%	25.0%	33.3%	**50.8%**	30.8%	**50.8%**

Table 1. Classification results showing the performance of the various feature descriptors on the 5 cuisines. The columns are: CMI: Color Moment Invariant, C-S: C-SIFT, HH: Hue-Histogram, O-S: OpponentSIFT, R-S: RGB-SIFT, S: SIFT and MKL: Multiple Kernel Learning.

cuisines, the accuracy is limited. It could be due to the fact that there is a low degree of visible variability between food types belonging to the same cuisines. For example, in the confusion matrix for Thai, we can see that Basil Fried Rice is confused with Mandarin Fried Rice and Pad Thai Noodles is confused with Lo Mein. It could be very hard, even for humans, to distinguish between such classes by looking at their images.

From Table 1, we can see that there is no single descriptor that works well across all the 5 cuisines. This could be due to the high-degree of variation in the training data. However, combining the descriptors using MKL yields the best performance in 4 out of the 5 cases.

4.3. Recognition Without Location Prior

Our approach is based on the hypothesis that knowing the location (through geo-tags) helps us in narrowing down the number of food categories which in turn boosts recognition rates. In order to test this hypothesis, we disregard the location information and train our SMO-MKL classifier on all of the training data (3,750 images). With this setup, accuracy across our 600 test images is 15.67%. On the other hand, the overall average accuracy across the 5 cuisines (from Figure 6) is 63.33%. We can see that the average performance increased by 47.66% when location prior was included. This provides validation that knowing the location of eating activities helps in food recognition, and that it is better to build several smaller restaurant/cuisine specific classifiers rather than one all-category food classifier.

5. Discussion

In this section we discuss several important points pertaining to the generalizability of our approach, implementation issues, and practical considerations.

Generalizability The automatic food identification approach that we propose is focused on eating activities in restaurants. Although this might seem limiting, eating out has been growing in popularity and 43.1% of food spending was reported to having been spent in foods away from home in 2012 [9, 23]. Moreover, we feel that eating and food information gathered in restaurants is more valuable for dietary self-monitoring than food information obtained at home, since individuals are more likely to know food types and the composition of food items prepared in their own homes.

We designed our study and evaluation with the goal of maximizing the external validity of our results. We evaluated our approach by having three individuals collect images from the most popular restaurant types by cuisine in the US on different days and using two different devices (smartphones and Google Glass). We feel confident that our methodology will scale in the future, especially since it leverages sensor data, online resources and practices around food imagery that will become increasingly more prevalent in years to come.

One important aspect of the approach is that it depends on weakly-labeled training images obtained from the web. The high-degree of intra-class variability for the same food across different restaurants has a negative effect on performance. A promising alternative is to train on (automatically acquired) food images taken at the same restaurant as the one where the test images were taken. While getting this kind of data seems difficult, it may soon be possible. A recently launched service by Yelp (among others), allows users to upload photos of their food. With such crowd-sourced imagery available for a given restaurant, it may soon be possible to train specialized classifiers for that restaurant. In our future work, we plan to test this hypothesis and improve the recognition accuracies.

Location Error We not only identify the cuisine that the individual is eating, but also identify the specific dish that is being consumed. Our approach hinges on identifying the restaurant the individual is at, and retrieving the menu of said restaurant. Although latitude and longitude can be reliably obtained with GPS sensors in mobile and wearable devices today, there might be times when the association between location data and the exact restaurant the person is visiting is erroneous (e.g. person is inside a shopping mall, or when two or three restaurants are in close proximity to each other). Although this might seem like a limitation of our method, it is usually not of practical concern since restaurants that are physically close are typically significantly different in their offerings. Thus, it is often enough to identify the general physical area the individual is at (as

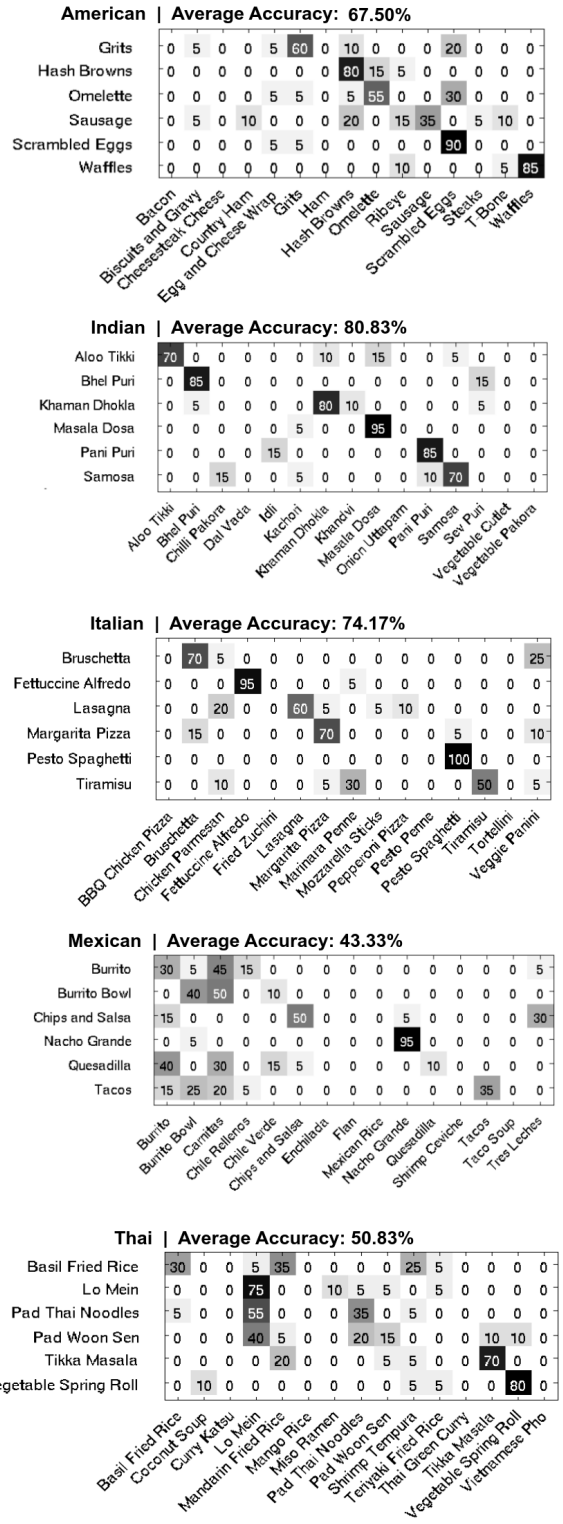

Figure 6. Confusion matrices for the best performing features of Table 1 (for each of the 5 cuisines). Darker colors show better recognition. The 6 food classes in the rows are the ones used for testing and the 15 food classes in the columns are the ones used for training. The overall average accuracy is 63.33%

opposed to the exact restaurant) and retrieve the menu of all restaurants and their respective food photos.

Semi-Automating Food Journaling Dietary self-monitoring is effective when individuals are actively engaged and become aware of their eating behaviors. This, in turn, can lead to reflection and modifications in food habits. Our approach to food recognition is designed to facilitate dietary self-monitoring. Engagement is achieved by having individuals take a picture of their food; the tedious and time-consuming task of obtaining details about the food consumed is automated.

6. Conclusion

Although numerous solutions have been suggested for addressing the problem of poor adherence to nutrition journaling, a truly practical system for dietary self-monitoring remains an open research question. In this paper, we present a method for automatically recognizing foods eaten in restaurants leveraging location sensor data and online databases.

The contributions of this work are (1) an automatic workflow where online resources are queried with contextual sensor data (e.g., location) to assist in the recognition of food in photographs.; (2) image classification using the SMO-MKL multi-class SVM classification framework with features extracted using color and point-based algorithms; (3) an in-the-wild evaluation of our approach with food images taken in 10 restaurants across 5 different types of food (American, Indian, Italian, Mexican and Thai); and (4) a comparative evaluation focused on the effect of location data in food recognition results.

Acknowledgements Funding and sponsorship was provided by the U.S. Army Research Office (ARO) and Defense Advanced Research Projects Agency (DARPA) under Contract No. W911NF-11-C-0088 and the Intel Science and Technology Center for Pervasive Computing (ISTC-PC). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of our sponsors.

References

[1] O. Amft and G. Tröster. Recognition of dietary activity events using on-body sensors. *Artificial Intelligence in Medicine*, 42(2):121–136, Feb. 2008. 1, 2

[2] L. Arab, D. Estrin, D. H. Kim, J. Burke, and J. Goldman. Feasibility testing of an automated image-capture method to aid dietary recall. *European Journal of Clinical Nutrition*, 65(10):1156–1162, May 2011. 2

[3] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *PAMI*, 33(5):898–916, 2011. 3

[4] F. Bach, G. Lanckriet, and M. Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *ICML*, 2004. 4

[5] L. E. Burke, J. Wang, and M. A. Sevick. Self-Monitoring in Weight Loss: A Systematic Review of the Literature. *YJADA*, 111(1):92–102, Jan. 2011. 1

[6] M. Chen, K. Dhingra, W. Wu, L. Yang, R. Sukthankar, and J. Yang. Pfid: Pittsburgh fast-food image dataset. In *ICIP*, 2009. 1, 5

[7] J. Cheng, B. Zhou, K. Kunze, C. C. Rheinländer, S. Wille, N. Wehn, J. Weppner, and P. Lukowicz. Activity recognition and nutrition monitoring in every day situations with a textile capacitive neckband. In *the 2013 ACM conference*, page 155, New York, New York, USA, 2013. ACM Press. 2

[8] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010. 2

[9] GALLUP. Fast food still major part of u.s. diet, Aug. 2013. 1, 6

[10] A. Goris, M. Westerterp-Plantenga, and K. Westerterp. Undereating and underrecording of habitual food intake in obese men: selective underreporting of fat intake. *The American journal of clinical nutrition*, 71(1):130–134, 2000. 1

[11] H. Hoashi, T. Joutou, and K. Yanai. Image recognition of 85 food categories by feature fusion. In *Multimedia (ISM), 2010 IEEE International Symposium on*, pages 296–301, 2010. 1

[12] K. Kitamura, C. de Silva, T. Yamasaki, and K. Aizawa. Image processing based approach to food balance analysis for personal food logging. *IEEE International Conference on Multimedia. Proceedings*, pages 625–630, July 2010. 1, 2

[13] F. KM, G. BI, W. DF, and G. MH. Excess deaths associated with underweight, overweight, and obesity. *JAMA*, 293(15):1861–1867, 2005. 1

[14] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar. Attribute and simile classifiers for face verification. In *ICCV*, 2009. 2

[15] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos "in the wild". In *CVPR*, 2009. 2

[16] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 3, 4

[17] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Gool. A comparison of affine region detectors. *IJCV*, 65(1):43–72, 2005. 3

[18] J. Noronha, E. Hysen, H. Zhang, and K. Z. Gajos. Platemate: crowdsourcing nutritional analysis from food photographs. *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 1–12, 2011. 1, 2

[19] E. Sazonov, S. Schuckers, P. Lopez-Meyer, O. Makeyev, N. Sazonova, E. L. Melanson, and M. Neuman. Non-invasive monitoring of chewing and swallowing for objective quantification of ingestive behavior. *Physiological Measurement*, 29(5):525–541, Apr. 2008. 1, 2

[20] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf. Large scale multiple kernel learning. *The Journal of Machine Learning Research*, 7:1531–1565, 2006. 4

[21] E. Thomaz, A. Parnami, J. Bidwell, I. A. Essa, and G. D. Abowd. Technological approaches for addressing privacy concerns when recognizing eating behaviors with wearable cameras. *UbiComp*, pages 739–748, 2013. 1

[22] E. Thomaz, A. Parnami, I. A. Essa, and G. D. Abowd. Feasibility of identifying eating moments from first-person images leveraging human computation. *SenseCam*, pages 26–33, 2013. 2

[23] USDA. Food consumption and demand, Nov. 2013. 1, 6

[24] K. Van De Sande, T. Gevers, and C. Snoek. Evaluating color descriptors for object and scene recognition. *PAMI*, 32(9):1582–1596, 2010. 4

[25] S. Vishwanathan, Z. Sun, N. Theera-Ampornpunt, and M. Varma. Multiple kernel learning and the smo algorithm. *NIPS*, 2010. 3, 4

[26] S. Yang, M. Chen, D. Pomerleau, and R. Sukthankar. Food recognition using statistics of pairwise local features. In *CVPR*, 2010. 1, 5

[27] K. Yatani and K. N. Truong. BodyScope: a wearable acoustic sensor for activity recognition. *UbiComp '12: Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 341–350, 2012. 1, 2

[28] A. Zamir and M. Shah. Accurate image localization based on google maps street view. *ECCV*, 2010. 3

[29] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *IJCV*, 73(2):213–238, 2007. 4