Multimed Tools Appl
https://doi.org/10.1007/s11042-017-5346-x

Context based image analysis with application in dietary assessment and evaluation

Yu Wang[1] · Ye He[2] · Carol J. Boushey[3] · Fengqing Zhu[1] · Edward J. Delp[1]

Received: 17 November 2016 / Revised: 16 May 2017 / Accepted: 23 October 2017

Abstract Dietary assessment is essential for understanding the link between diet and health. We develop a context based image analysis system for dietary assessment to automatically segment, identify and quantify food items from images. In this paper, we describe image segmentation and object classification methods used in our system to detect and identify food items. We then use context information to refine the classification results. We define contextual dietary information as the data that is not directly produced by the visual appearance of an object in the image, but yields information about a user's diet or can be used for diet planning. We integrate contextual dietary information that a user supplies to the system either explicitly or implicitly to correct potential misclassifications. We evaluate our models using food image datasets collected during dietary assessment studies from natural eating events.

Keywords Image analysis · Image segmentation · Object classification · Context information · Dietary assessment

1 Introduction

There is a health crisis in the US related to diet that is further exacerbated by our aging population and sedentary lifestyles. Six of the ten leading causes of death in the United States, including cancer, diabetes, and heart disease, can be directly linked to diet [38]. Dietary intake, the process of determining what someone eats during the course of a day, provides

✉ Ye He
yehe@google.com

[1] Purdue University, West Lafayette, IN, USA

[2] Google Inc, Mountain View, CA, USA

[3] University of Hawaii Cancer Center, Hawaii, USA

Published online: 25 November 2017

valuable insights for mounting intervention programs for prevention of many of the above chronic diseases. Measuring accurate dietary intake is considered to be an open research problem in the nutrition and health fields. Traditional dietary assessment is comprised of written and orally reported methods that are time consuming and tedious, often requiring a nutrition professional to complete, and are not widely acceptable or feasible for everyday monitoring [29, 52].

For the past 7 years we have been investigating the use of images that a user takes of their meal before and after eating to assess their diet. We have developed a system, known as the Technology Assisted Dietary Assessment System (TADA), to acquire and process food images [8, 60]. The TADA system, and the associated mobile Food Record (mFR), allows users to acquire food images using a mobile telephone. Image processing and analysis methods are then used to determine the food type, the energy (kilocalories) and nutrients of the food [13, 59, 60]. The TADA system has been used by more than 14 scientifically controlled user studies, including free-living environments, by more than 800 users who have taken more than 60,000 food images.

The goals of image analysis are to identify the food objects in the image and segment the regions in the image corresponding to the foods (see Fig. 1). This information can then be used to estimate the food portion size used for energy and nutrient estimation [13] using the food density [23, 51]. Before we describe the work proposed in this paper, we will overview the current state of food record systems.

In recent years, mobile-based "nutrition" applications and Internet-based tools and services have become popular. Some of these applications also include the capability of taking images of foods eaten at different eating occasions and using the images as part of a food diary to assist users in manually recording their diets. These applications include *Diet Camera* [11], FoodLog [25], *uHealth* [9], *Tuingle* [55], *Argus* [2], *FoodCam* [20], *DietCam* [26] and *Im2Calories* [35]. Diet Camera and Argus are basically visual food trackers that allow users to log their diet along with various activities. DietCam, designed for automatic food intake assessment, is based on images acquired from multiple views using a mobile telephone. FoodLog provides both a mobile application and cloud service that allows users to record daily dietary intake by acquiring images of food. In many of these systems, a user must first identify the names and quantities of food items and then the nutrient values are estimated, which places a large portion of the dietary assessment on the user (or on a human analyst). The uHealth service is a image based wellness information system. However, the authors only used a single texture feature and evaluated the system using pizza images. Tuingle, advertised as the mobile food scanner, is able to automatically recognize food items

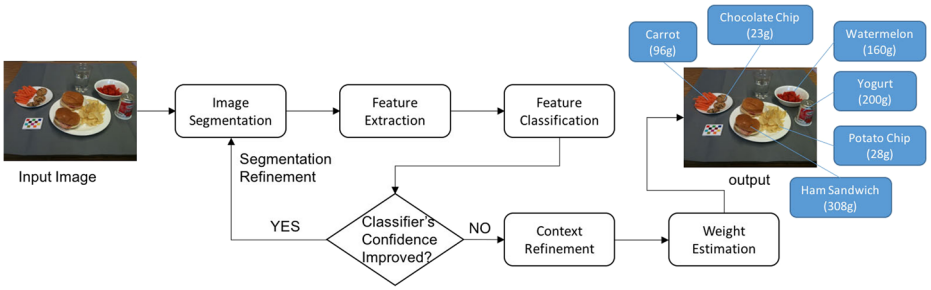


Fig. 1 Food image analysis. Given an eating occasion image, our goal is to identify all the food items in the image and estimate the food weight

and it requires users to manually input portion size to complete nutrient and calorie assessment. Im2Calories developed by Google uses deep learning techniques to recognize food but it has not been released to public. Some researchers [41] focus on improving the scalability of the food image analysis service by introducing queuing management, indexing food images, allocating cloud resources.

We have described our approach to food image analysis in [60] and our recent work in food volume estimation and food portion size estimation in [13]. In this paper we present new approaches to food image analysis based on the use of contextual information. Contextual dietary information is the data that is not directly produced by the visual appearance of a food object in the image, but yields information about a user's diet or can be used for diet planning. Examples of contextual information in dietary assessment include the time, date, and location (GPS coordinates) of a meal occasion, the dietary patterns or combinations.

Our previous work on food classification has shown that there are several issues that need to be addressed [60]. These include the inability to differentiate visually similar food items, e.g. diet coke vs. regular coke, nonfat milk vs. 2% milk, solely based on their appearance in the image. Another issue is the selection of training data for different classes. Increasing the number of food training classes could cause a drastic increase in the food classification error. Using contextual dietary information the classifier can assign different weights to the food classes that are more relevant to what are commonly eaten by the individual at similar times, dates or locations. For example, we can learn that an individual is more likely to have scrambled eggs in the morning rather than in the evening from the temporal data. GPS information is able to indicate where a person has the meal, whether at home, at work or in a restaurant. Assuming that people consume different foods at home/work compared to any meal served in a restaurant, GPS data can be treated as a priori in the classification process. Thus, the contextual information can reduce the number of classes that the classifier has to select from and hence can learn the dietary habits of the participant. In this paper we integrate contextual dietary information that a user supplies to the system either explicitly or implicitly to correct potential misclassifications. Contexutal information such as temporal data, food co-occurrence pattern along with a personalized model is the focus of this paper. We extend our earlier work on food image classification [60] and on the use of contextual information [18, 58] in which we introduced the idea of incorporating food consumption frequency and food co-occurrence pattern in assisting food classification. The main contributions of this paper include improving image segmentation and refinement methods and integrating a personalized learning model to enhance the food classification. We show that both our segmentation-to-classification pipeline with handcrafted features and a region proposal based method with deep features benefit from the contextual data.

2 Food image segmentation and classification

As we indicated in Section 1 the goal of this paper is to describe how the use of contextual information can improve food image segmentation and food classification. We first describe the segmentation and classification methods we will use for the contextual studies. The goal of segmentation is to locate and isolate food items in eating occasion images. The accuracy of food segmentation plays a crucial role in the overall performance of the system because false segmentation will cause degradation of the subsequent image analysis steps including food identification and pose size (weight) estimation. We have previously evaluated several approaches to generate initial segmentation results such as active contours [22], normalized cuts [49] and local variation [14]. We reported a comparative study in [17] to quantify

the importance of the segmentation method. Feedback from the classifier is then used to refine the initial segmentation results [18, 60]. We use local variation for the segmentation technique in this paper as it is fast and relatively good at preserving edges. Based on our previous evaluation [17], local variation also shows better results in terms of being more stable to the change of parameters.

2.1 Segmentation refinement

In this section we overview how we refine the segments generated by local variation [14]. Local variation is a graph based segmentation method, in which two regions are segmented if the difference between the two regions is large relative to the internal difference within at least one of the two regions. The degree to which the difference between regions must be larger than minimum internal difference is controlled by a threshold β [14]. β roughly controls the size of the regions in the resulting segmentation. Smaller values of β yield smaller regions and favor over-segmentation. We use $\beta = 150$ in the segmentation experiment. Since the image segmentation method is limited by a particular choice of input parameters, some food items may be under-segmented, while others may be over-segmented. We seek to overcome the segmentation problem by using classification feedback to refine the segmentation results. In our approach, the image segments are classified to a particular food label using the features extracted from that segment. The K most probable candidate classes along with their classification confidence scores are used to refine initial segmentation results. This approach is similar to our earlier work of the joint segmentation/classification described in [60]. In our earlier work segmentation refinement is achieved through adjusting a set of segmentation parameters for salient regions. In this paper we do not use salient regions, our segmentation refinement is based on image segments generated from initial segmentation. Figure 2 shows our segmentation refinement approach.

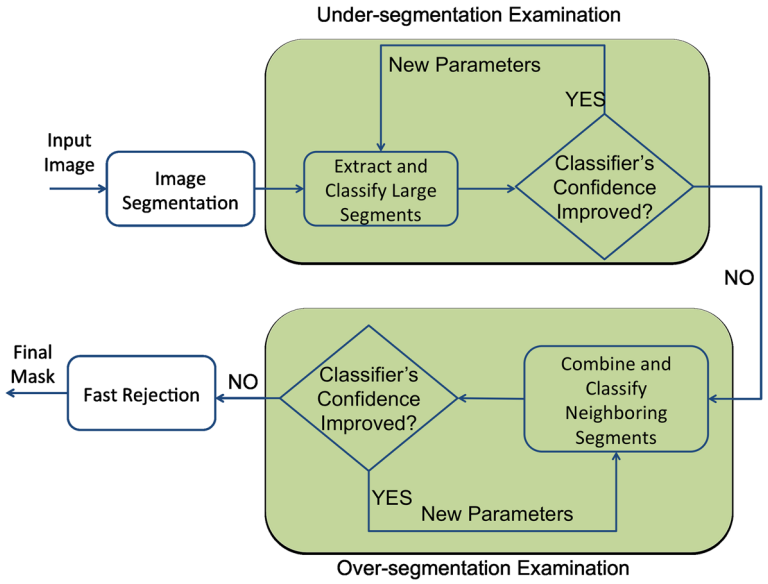


Fig. 2 Segmentation refinement

To detect under-segmentation, we first scan all the segments produced by the use of local variation in the image to filter out small segments. We define "small segments" as segments that contain less than 1/50 pixels of the original image. Each remaining segment is re-segmented and classified again. If the food classification confidence score is improved by re-segmentation, we accept the new segmentation; otherwise the original segmentation is kept as final segmentation. After under-segmentation examination, we update the label of the segments to $\{s_0, s_1, ..., s_{Q-1}\}$ and the corresponding food category label as $\{(c_{0,0}, c_{0,1}, ..., c_{0,K-1}), ..., (c_{Q-1,0}, c_{Q-1,1}, ..., c_{Q-1,K-1})\}$.

After under-segmentation examination, for each adjacent pair of segments, if a food category label in one segment equals to a label in the other segment, and the sum of the confidence score is greater than the highest individual score, we combine these two segments with their updated K category labels corresponding to the K largest confidence scores in the descending order. This process of over-segmentation examination is done iteratively until the overall confidence score of a segment cannot be improved.

After under-segmentation and over-segmentation are examined, we may still have redundant segments, such as in the background area. We use a fast rejection step to remove these redundant segments [60]. We filter out the segments with low confidence scores from the classifier. Illustration of the complete image segmentation refinement process is shown in Fig. 3.

2.2 Feature selection

Features are used for describing the characteristics of objects. An essential step in solving the food classification problem is to select suitable features to distinguish one food from another. Some foods may have very distinctive color or very distinctive patterns, but for most food items it is the combination of these aspects that make them distinctive. Previously, we have investigated various features for food classification [18, 60]. In this paper, we overview three types of features, color, texture and local region descriptors, among which we regard color and texture as the global features. Based on the evaluation of these feature descriptors and their combinations, we select the optimal strategy for our food image analysis system.

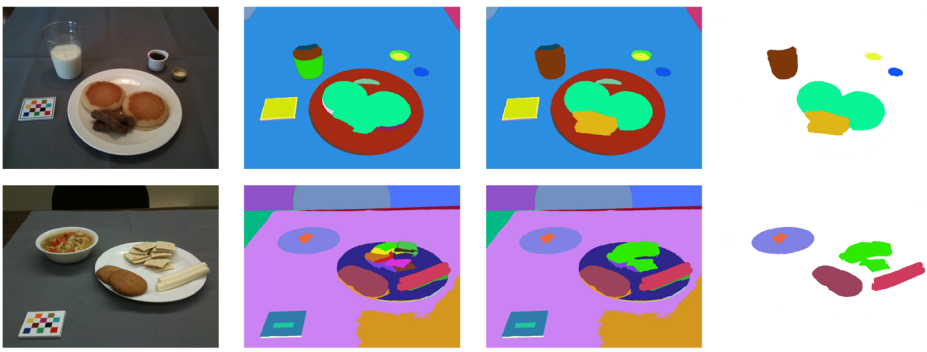


Fig. 3 Examples of food image segmentation and segmentation refinement. **a** original food images, (**b**) initial segmentation results using the local variation segmentation, (**c**) segmentation refinement using food classification confidence score, and (**d**) final image segmentation results after fast rejection

It should be noted that in this section we will describe a set of features used in our contextual experiments. Other approaches such as deep networks [35, 45] could also be used to investigate the use of contextual information.

Color features have been extensively studied in image retrieval [32]. Some foods may exist in a wide variety of colors, but many have a distinctive color. Color information is sensitive to environmental conditions, such as changes in light source and shadows. We investigated two color descriptors, namely, Dominant Color Descriptor (DCD) and Scalable Color Descriptor (SCD) [32]. DCD is a vector of D representative colors from the *CIE-Luv* color space using the generalized Lloyd algorithm for color clustering [10, 24] and their corresponding percentages. SCD is determined by quantizing the colors in the *HSV* color space uniformly into 256 bins, which includes 16 levels in H, 4 levels in S, and 4 levels in V as suggested by the MPEG-7 standard [32].

Texture, similar to color, is a very descriptive low-level feature. In general, texture describes the arrangement of basic elements of a material on a surface [3, 21]. We selected two texture descriptors for food classification: Entropy-Based Categorization and Fractal Dimension Estimation (EFD) and Gabor-Based Image Decomposition and Fractal Dimension Estimation (GFD) [60]. EFD can be seen as an attempt to characterize the variation of roughness of homogeneous parts of the texture in terms of complexity [60]. GFD is based on fractal dimension estimation [60].

Local region features are described for points of interest and/or local regions. The idea is to find points in the object which can be reliably found in other samples of the same object regardless of variations between images. An invariant local region feature describes such points of interest in the same way in different images with illumination, scale and viewpoint changes. Many local region features have been proposed to represent the characteristics of points of interest [5, 30, 36, 53]. We investigated the following two local region features for food classification: Scale Invariant Feature Transforms (SIFT) [30] and Multi-scale Dense SIFT (MDSIFT) [60]. Table 1 summarizes the features used in our experiments.

2.3 Classification

Once the food items are segmented from a eating occasion image and the features are extracted, we classify the color and texture features using K-Nearest Neighbors (KNN) [12] and the local features using the Vocabulary Tree (VT) classifier [18]. Ths classification results based on the automatic segmentation are presented in Section 5. The context integration described in this paper is independent of the type of classifier. As we indicated we are using KNN and VT but other machine learning approaches such as SVM and deep networks could also be used [35, 45].

Table 1 List of features investigated and their types and dimensions

Feature	Feature type	Dimension
DCD	Color feature	20
SCD	Color feature	256
EFD	Texture feature	120
GFD	Texture feature	120
SIFT	Local region feature	128
MDSIFT	Local region feature	384

KNN [12] is widely used. Given a query feature vector, KNN predicts the classification label based on a number of closest training vectors. The distance measure chosen for the KNN classifier depends on the feature characteristics. We use the L1 norm for the histogram based features (SCD, EFD and GFD). We use the L2 norm for the DCD feature [18].

Suppose the number of trained food classes is N. Given a query feature vector (f_q), we find the L nearest training feature vectors using KNN [18]. Given the food class c_l of each of the L nearest training feature vectors ($f_{t,l}$), we estimate the classification "confidence score" of each food class as in (1). The confidence score $\phi(c_l)$ describes the classifier's confidence that its inferred class label c_l is the correct label of the query feature vector f_q [60].

$$\phi(c_l) = \sum_{t,i=l} \exp(-d(f_q, f_{t,l})/(d_{1-NN} + \epsilon)) \tag{1}$$

where $d(f_q, f_{t,l})$ is the distance between the query feature vector and the training feature vector belonging to class c_l. d_{1-NN} is the distance between the query feature vector of the input segmented region and the nearest neighbor (1-NN). We added ϵ to denominator to avoid the case of division by zero.

Vocabulary trees (Fig. 4) have been shown to be efficient in large-scale image retrieval and object recognition [1, 39, 42, 48, 57]. A vocabulary tree is a hierarchical quantization that is based on hierarchical k-means clustering [39, 57]. K-means clustering proceeds by assigning data samples to their closest cluster centers and re-estimating the cluster center positions iteratively. To build a vocabulary tree, the training data is first partitioned into P clusters using an initial k-means clustering. Each cluster center serves as a branch for the root of the vocabulary tree. Then for each cluster, the same k-means clustering process is done again with the cluster center as the root of the cluster. The same process is done recursively to each new cluster until the maximum number of leaves H has been reached. We use the branching factor $P = 3$ and the maximum number of leaves ($H = 10,000$) [18]. Given a trained vocabulary tree, we classify a query image by assigning it to an existing food class. We first extract a set of local region feature vectors from the query image. The query

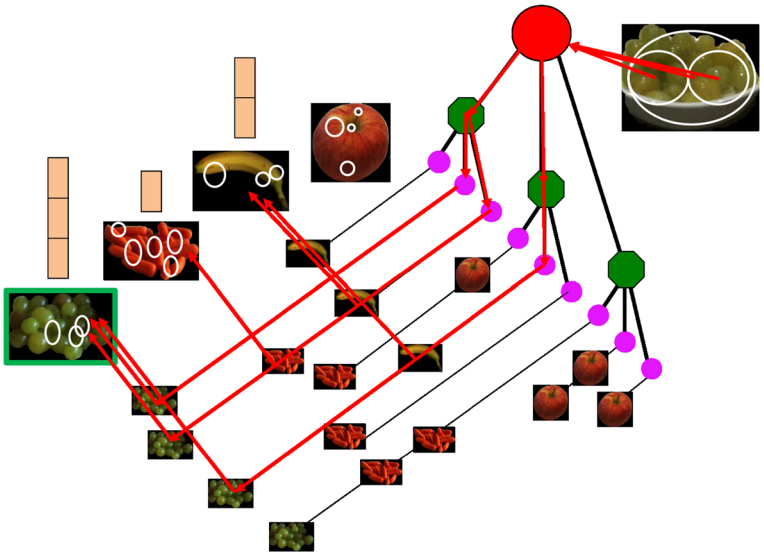


Fig. 4 The food classification process using a vocabulary tree

feature vectors are then classified by propagating them down the tree, up to the maximum level of the tree in the same manner as the dense features in [39]. At each level of the vocabulary tree, the descriptor feature vector is compared to all the children of the current root, and assigned to the branch of the closest child. The same process is done recursively until a leaf node is met. A path from the root to a leaf at the maximum level of a vocabulary tree is known as a "visual word." We associate each path down the vocabulary tree with a single integer and then use it in scoring [39].

One significant advantage of using a vocabulary tree over one step k-means clustering is the computational efficiency of adding new training data to the vocabulary tree. When a new training descriptor vector is added to the vocabulary tree, only a part of the tree needs to be modified, while in one-step k-means clustering the entire vocabulary may be modified. Finally, we classify each segment using the weighted sum of confidence scores from KNN classifiers for each global features and from the vocabulary tree for local features.

3 Region proposal based approach with deep features

Since the interest in Convolutional Neural Networks (CNN) was rekindled by AlexNet [27] in 2012, the number of applications using deep networks has grown exponentially. CNNs have dominated many aspects of object classification and detection [28]. Besides, recent research indicates that the generic descriptors extracted from the convolutional neural networks are very effective [45]. The success of CNNs is larged attributed to big data and carefully designed models. In terms of food image analysis, some researchers [33] focused on improving the network structure by considering the food structure in the image, or "vertical food layer". However, they did not utilize any contextual information to improve classification for foods that do not have obvious structure in their appearances.

In the section, we describe a region proposal based method to identify multiple food items in an image. In contrast to the segmention-to-classification pipeline we discussed in Section 2, deep features are extracted from redundant proposed regions instead of segments. Then, support vector machine (SVM) [45] is used to classify each region as either a specific food item or background. Similar to RCNN [16], we adopted selective search [56] as the generic region proposal method. We finetuned VGG-16 [50] on Food-101 dataset [7] and used the output from the first fully connected layer as the deep features.

Before we forward propagate the region proposals through the network, we first run a fast rejection to eliminate tiny regions and regions with large aspect ratios. The fast rejection is based on the assumption that the food items in a food image usually occupies the majority of the scene.

We used the PyTorch [43] implementation of the VGG-16 [50] to obtain the 4096-dimensional features from each region proposal that was not rejected. In order to convert a region proposal to the dimension compatible with the deep network, we proposed a random 10-crop technique. Since VGG-16 requires the input image of 224×224, for any region proposal, we first resize it so that the shorter dimension of the region is 224. Then we randomly select 10 224×224 cropped regions from the resized proposal. Features are computed by propagating a mean-subtracted 224×224 RGB image through the network.

Finally, we used a SVM to classify food items and the background. For regions that are classified as a certain food class with greater than 75% confidence, we apply non-maximum suppression to select the best proposal. We combine the majority vote from the 10-crop technique with the confidence score from SVM to finalize our prediction. As improving deep features or region proposal methods is not the focus of this paper, we only show the

experimental result using our own dataset in Section 5 as a comparison to the method we discussed in Section 2. More importantly, we show that the contextual information can be integrated as easily in the region proposal based approach as in the previously discussed segmentation-to-classifiation pipeline.

4 Context refinement

Contextual information has gained more attention in image analysis and computer vision in the past few years [6, 15, 34, 37, 40, 44, 54]. Context such as semantic, spatial images and poses has been proved to be effective for natural images. Examples of contextual information in the dietary applications include the date, time and location of the eating occasion, who the subject is eating with, and personal eating habits. In this section we overview our previous approaches for integrating contextual information which the participant supplies to the system either explicitly or implicitly [58] and propose a new approach for combining temporal eating information and food co-occurrence into a personalized learning model. Note that the contextual information we investigate are independent of the classifier and can be used with other types of machine learning techniques (e.g SVM and deep networks [35, 45]).

4.1 Temporal dietary information

We explore temporal information of food images to generate the preference of different food classes based on time of an eating occasion. People usually eat different types of foods with regard to the time of a day, such as breakfast vs. dinner. We incorporate this contextual dietary information to assign a weight to different food classes.

We divide eating time into three time intervals: *12am - 11am*, *11am - 4pm*, and *4pm - 12am (midnight)*. For example, from our "free-living" dietary assessment study [58] the food consumption frequency of these three time intervals are shown in Figs. 5, 6 and 7, respectively. We can see unique food consumption patterns for different time intervals. For example, from *12am* to *11am*, "Bagel", "English Muffin" and "Pancake" are more likely to be consumed than other foods such as "Chicken Wrap," "Frozen Meal Meatloaf" and "Ham Sandwich." From *11am* to *4pm*, people tend to eat more "Ham Sandwich" and "Potato Chips" than earlier in the day. When it comes to *4pm* to 12am, there is a significant increase in the consumption of "Garlic Bread" and "Lasagna." Given a participant's food consumption frequency over time, we assign different weights to different food classes according to the eating occasion time.

4.2 Food co-occurrence patterns

A food co-occurrence pattern describes the likelihood of food combinations. It is the joint probability of food items existing together in a single eating occasion [58]. Semantic context can provide valuable information for improving classification. In this section, we describe the use of the co-occurrence of food items in order to reach a labeling agreement for all the segmented regions in an image. The goal is to detect potential misclassifications and refine the classification results that were obtained by using only visual features.

After image segmentation and food classification, an eating occasion image I is segmented into multiple regions $s_0, ..., s_q, ..., s_{Q-1}$. A segment s_q is assigned K food labels. A classification confidence score $\phi(c_{q,k})$ measuring the probability that any food label

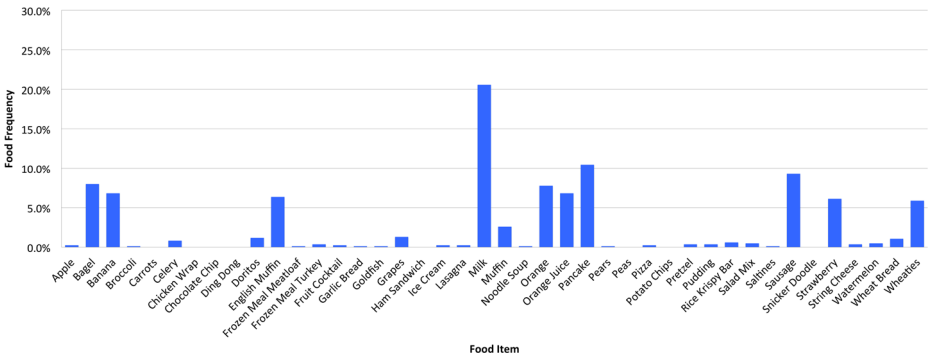


Fig. 5 Examples of food consumption preference between 12am (midnight) and 11am

matches the segment s_q based on the distance of the visual features between the segmented region and the training data. We now want to adjust the food labels to achieve maximal global contextual agreement with respect to the food co-occurrence pattern given the constraints of the segments' visual features. For example, in most cases "fries" has a higher contextual agreement with "ketchup" than with "pepper."

Graphical models provide a simple way to visualize the structure of a probabilistic model. Since the number of food segments in an eating occasion is relatively small, we construct a weighted complete digraph between all segments [19]. In our graph, each node in the weighted complete digraph represents a segment and its associated food labels from the food classification results. Therefore, the graph contains Q nodes and each contains K food labels. Obviously, one node has $Q-1$ outgoing edges and $Q-1$ incoming edges.

The food co-occurrence probability of food label $c_{j,k}$ given food label $c_{i,k'}$, denoted as $P(c_{j,k}|c_{i,k'})$ is defined as follows,

$$P(c_{j,k}|c_{i,k'}) = \max\{P(c_{j,k}|c_{i,1}), ..., P(c_{j,k}|c_{i,K-1})\} \tag{2}$$

where k and k' independently indeces K. Then the influence of segment i on the k^{th} food label of segment j, $v(c_{j,k}|s_i)$, is calculated as:

$$v(c_{j,k}|s_i) = \phi(c_{i,k'})P(c_{j,k}|c_{i,k'}) \tag{3}$$

where $\phi(c_{i,k'})$ is the classification confidence score of food label $c_{k'}$ in the i^{th} segment. Finally, the weight of the edge from node i to node j, $\mathbf{v}_{ij}$, is defined as an influence vector

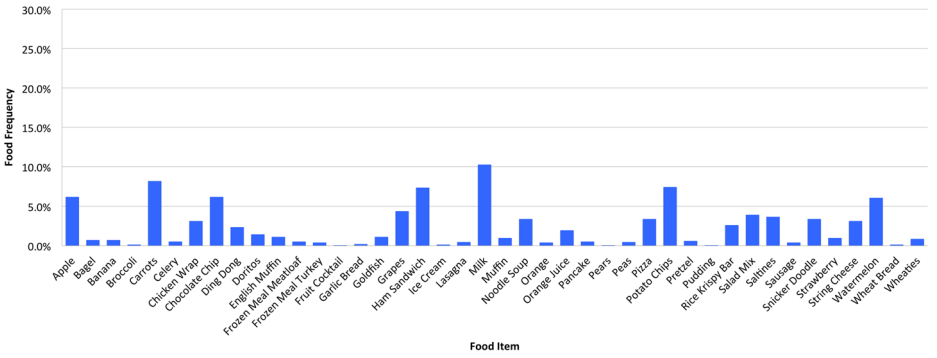


Fig. 6 Examples of food consumption preference between 11am and 4pm

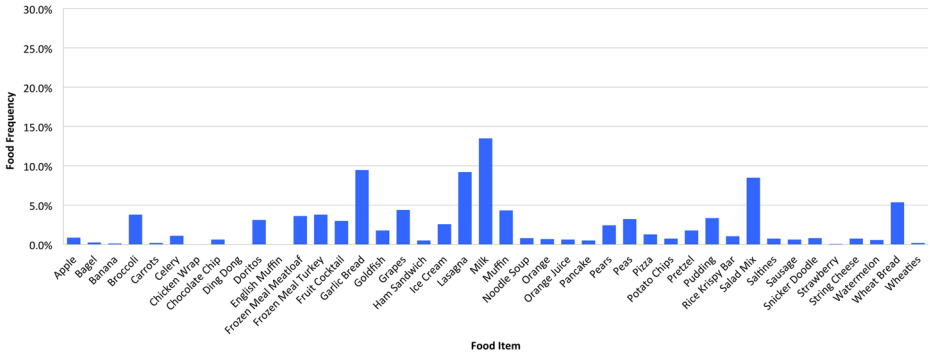


Fig. 7 Examples of food consumption preference between 4pm and 12am (midnight)

indicating how much influence the food labels in segment i have on all the food labels in segment j. An influence vector of K-dimension is computed from the food occurrence pattern as follows,

$$\mathbf{v}_{ij} = [v(c_{j,0}|s_i), ..., v(c_{j,k}|s_i), ..., v(c_{j,K-1})|s_i)] \tag{4}$$

To estimate the co-occurrence probability, we first construct a food co-occurrence matrix M_{FCO} that contains the food co-occurrence counts among food labels in the training set of the database [58]. Figure 8 shows an example of a food co-occurrence matrix. The entry (i, j) in a food co-occurrence matrix is the number of times that food λ_j is in an eating occasion image when food λ_i is in the image [58].

Figure 8 illustrates the structure and content of a food co-occurrence matrix. As we can see, some food items have a high probability of existing together in the same image, e.g. "Wheaties" with "Milk," "Garlic Bread" with "Lasagna;" while some food items rarely appear together in the same image, e.g. "Carrots" with "Celery." Note that the co-occurrence matrix is trained on training data, where we have perfect segmentation and food labels from our dietary studies. The matrix is only updated when we receive a participant's confirmation from the review process of the TADA system where the participant can confirm, change or add food labels [58].

So far we have found the influence vector from s_i to s_j. Following the same approach, we find the influence vectors from all other segments to s_j. The next step is to find the total influence on s_j from all other nodes in the graph. We propose to use the maximal influence vector $\mathbf{w}_j$:

$$\mathbf{w}_j = [\max(\{\mathbf{v}_{ij}(0), i \neq j\}), ..., \max(\{\mathbf{v}_{ij}(K-1), i \neq j\})]$$

where $\{\mathbf{v}_{ij}(k), i \neq j\}$ is the set containing the k^{th} element of each of the influence vectors that point to node j. We choose the largest influence given to each food label of node j as the final influence vector $\mathbf{w}_j$. We finally update the food classification confidence score of each food label of node j as follows:

$$\phi'(c_{j,k}) = \phi(c_{j,k})(w_j(k) + \epsilon) \tag{5}$$

where $\epsilon > 0$ and $w_j(k)$ is the k^{th} element of $\mathbf{w}_j$. The term ϵ is used to avoid setting $\phi'(c_{j,k})$ to zero when $w_j(k) = 0$. The max influence of one segment on another is constrained by the max confidence score of the food label. For each adjacent pair of segments, we only accept the new segmentation if a food category label in one segment equals to a label in the other segment, and the sum of the confidence score is greater than the highest individual score.

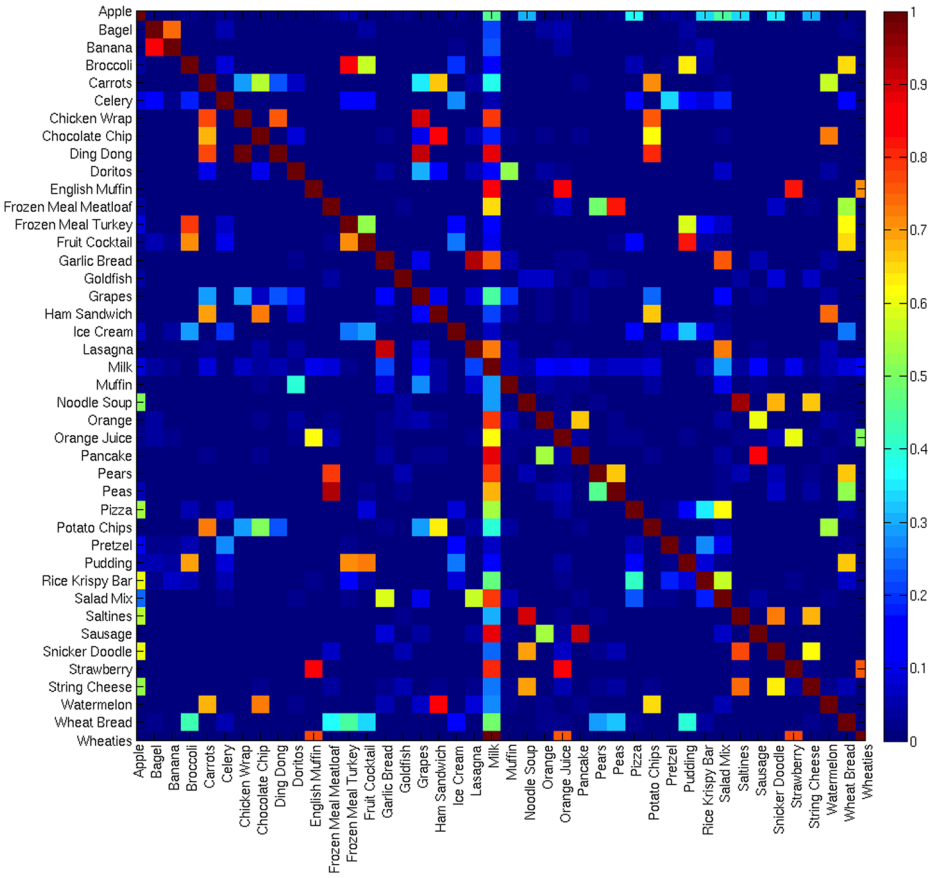


Fig. 8 An example of food co-occurrence patterns

After the confidence scores for segment i is updated to $\{\phi'(c_{i,0}), ..., \phi'(c_{i,K-1})\}$, we update the order of food labels accordingly, with the Top 1 food label being the one associated with the largest updated confidence score, and the Top M food label being the one associated with the M^{th} largest updated confidence score.

4.3 Personalized learning model

The goal of a personalized learning model is to improve food classification by using dietary preferences (Fig. 9). For example, if it learns that a person prefers diet coke and he/she never drinks regular coke from his/her dietary history, the personalized learning model will adjust the prediction of different coke products if a classifier initially assigns similar confidence scores to those classes.

Figure 10 illustrates a list of food consumption frequency patterns for various participants in our free-living study. For example in Fig. 10, most of the participants shown here drink milk quite frequently but participant 36 rarely drinks milk. We can also tell that the favorite fruit of participant 3 is "Grapes;" the favorite drink for participant 35 is "Orange Juice;" and the favorite food for participant 36 is "Lasagna." The figure shows the differences in

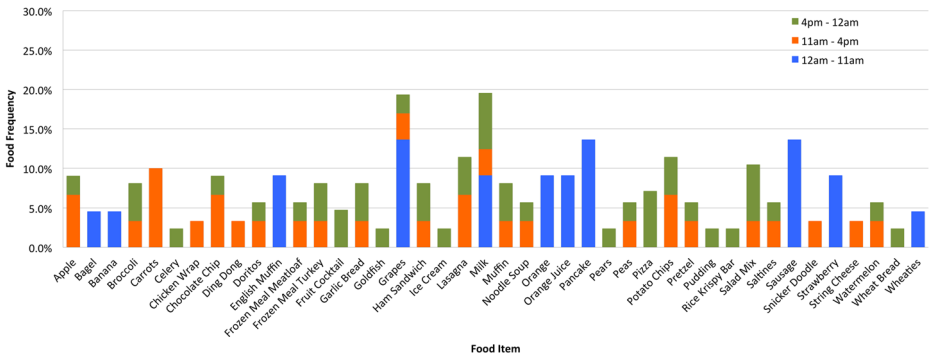


Fig. 9 An example of food consumption frequency by a participant with the integrated time information integrated. The colors indicated different eating time intervals

individual eating habits using food consumption frequency for this subset of foods. The food consumption frequency of food item λ_i for a participant S_j is:

$$F(S_j, \lambda_i) = \frac{\gamma_i(S_j, \lambda_i)}{\sum_k \gamma_k(S_j, \lambda_i)} \text{ for } k = 1, ..., K \tag{6}$$

where γ_i is the food consumption counts of a participant and K is the number of food classes.

The personalized learning model takes into account both temporal dietary information and food co-occurrence patterns. Figure 11 shows how we propose to do context-based classification refinement. Given a set of labeled segments $\{(c_{0,0}, ..., c_{0,K-1}), ..., (c_{q,0}, ..., c_{q,K-1})\}$ with associated confidence scores $\{(\phi(c_{0,0}), ..., \phi(c_{0,K-1})), ..., (\phi(c_{q,0}), ..., \phi(c_{q,K-1}))\}$, the food co-occurrence pattern generates an updated confidence scores for each segment in the image,

$$\{(\phi'(c_{0,0}), ..., \phi'(c_{0,K-1})), ..., (\phi'(c_{q,0}), ..., \phi'(c_{q,K-1}))\}.$$

In the temporal information block of Fig. 11, we use recursive Bayesian estimation to incrementally learn a participant's dietary pattern [4, 46, 58]. We model whether a participant, S_j eats a particular food, λ_i in time internal, V, as a Bernoulli trial,

$$W = \begin{cases} 1, X \\ 0, 1 - X \end{cases} .$$

where $W = 1$, X represents S_j eats λ_i in V with a possibility, X, and X is assumed to follow a Gaussian-like distribution with the support from 0 to 1. As discussed in Section 4.1, we used three time intervals *12am - 11am*, *11am - 4pm*, and *4pm - 12am (midnight)*.

We would like to estimate the probability, P_{λ_i}, that a participant, S_j, will eat a particular food, λ_i on the next day given the history [58]. Let $p_{\lambda_i}(x^n)$ be the probability density function (PDF) representing S_j eats λ_i, in the time interval V on the n^{th} day, and z^n be the observation whether S_j eats λ_i in V on the n^{th} day [58].

The following equations describe the posteriori update step in the recursive Bayesian network,

$$\begin{aligned} p_{\lambda_i}(x^n|z^{1:n}, V) &= \frac{p_{\lambda_i}(z^n|x^n, V)\, p_{\lambda_i}(x^n|z^{1:n-1}, V)}{p_{\lambda_i}(z^n|z^{1:n-1}, V)} \\ &= \frac{\text{likelihood} \times \text{prior}}{\text{normalization term}} . \end{aligned} \tag{7}$$

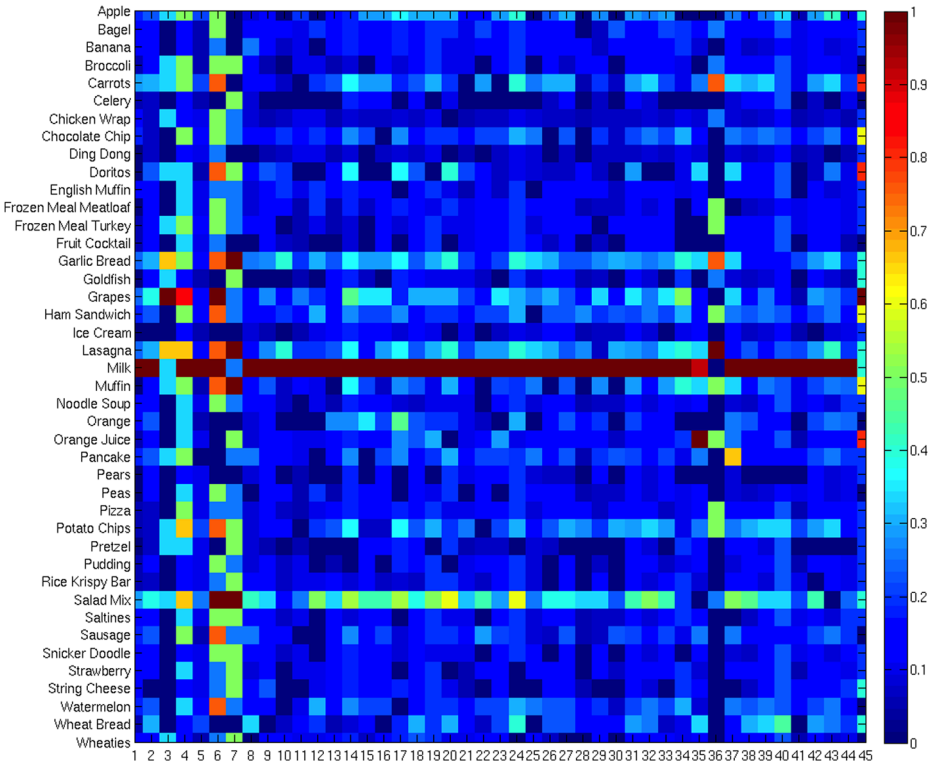


Fig. 10 An example of food consumption frequency for various participants. **Horizontal axis** shows the IDs of participants and **vertical axis** represents various food items

Initially, $p_{\lambda_i}(x^1|V)$ is assumed to have a Gaussian-like distribution centered at 0.5 with unit variance. If the participant eats λ_i in V on the n^{th} day, $p_{\lambda_i}(z^n|x^n, V)$ becomes the Gaussian-like distribution centered at 1 with unit variance, otherwise the distribution centers at 0. $p_{\lambda_i}(x^n|z^{1:n}, V)$ is used to predict $p_{\lambda_i}(x^{n+1}|z^{1:n}, V)$ and the PDF is computed by multiplying the likelihood and prior followed by normalization between 0 and 1. On the $n+1^{\text{th}}$ day, the optimal estimate of $P_{\lambda_i}(V)$ is computed as $P_{\lambda_i}(V) = \arg\max_x p_{\lambda_i}(x^n|z^{1:n}, V)$ (Fig. 9).

For all the foods in the training dataset, we have a set of probabilities, $\{P^{n+1}_{\lambda_0}(V), \ldots, P^{n+1}_{\lambda_{K-1}}(V)\}$ where N is the total number of food categories. We further define the context-based confidence scores (CCS) to be:

$$\begin{aligned}\Psi^{n+1}(V) &= \left[\psi^{n+1}_{\lambda_0}(V), \ldots, \psi^{n+1}_{\lambda_{K-1}}(V)\right]^{\mathrm{T}} \\ &= \left[\omega P^{n+1}_{\lambda_0}(V), \ldots, \omega P^{n+1}_{\lambda_{K-1}}(V)\right]^{\mathrm{T}} \quad . \end{aligned} \tag{8}$$

where ω controls the trust weight we assigned to the context-based decisions. For each image segment, the CCS associated with the Top M food labels, $(\psi_{i,0}, ..., \psi_{i,M-1})$, can

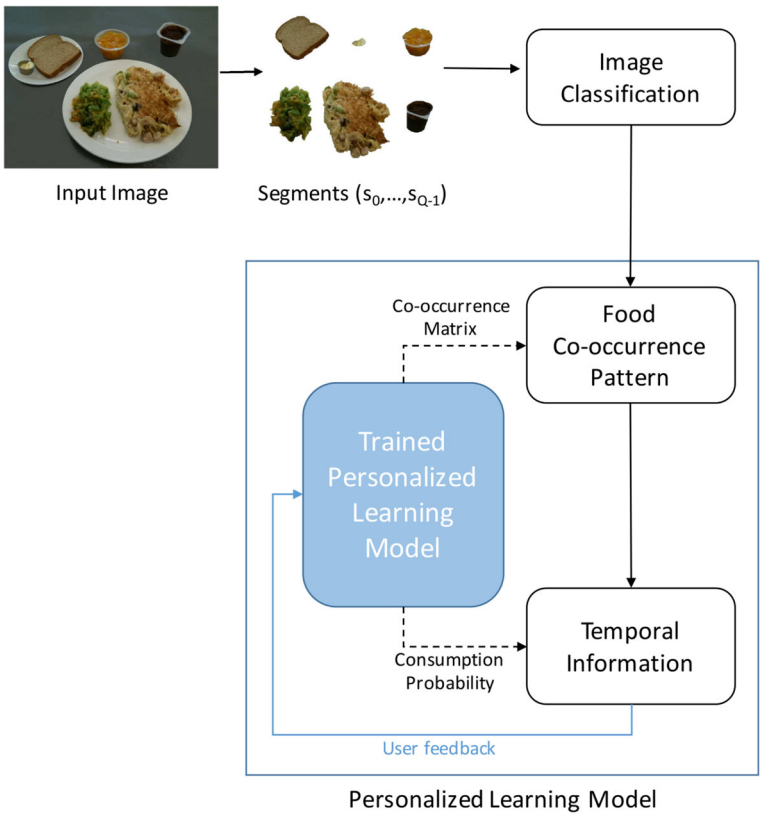


Fig. 11 The use of contextual dietary information in food classification refinement

be obtained from $\Psi^{n+1}(V)$. The final confidence scores are calculated using a strategy of majority vote,

$$\begin{aligned}(\phi''(c_{i,0}), ..., \phi''(c_{i,M-1})) &= (\phi'(c_{i,0}), ..., \phi'(c_{i,M-1})) \\ &\quad + (\psi_{i,1}(V), ..., \psi_{i,M-1}(V)) \\ &= (\phi'(c_{i,0}), ..., \phi'(c_{i,M-1})) \\ &\quad + (\omega P_{i,1}(V), ..., \omega P_{i,M-1}(V)) \end{aligned} \tag{9}$$

ω, also in (8), is set to be $1/h$ of the maximum automatic analysis based confidence score. In our experiments, we observed best results when h was set to 4-5. The food labels are updated again according to the new confidence scores to generate the final food labels $\{(c''_{0,0}, ..., c''_{1,M-1}), ..., (c''_{q,0}, ..., c''_{q,M-1})\}$.

5 Experimental results

5.1 Experiment setup and datasets

We evaluate our system using a free-living dietary study where we provided some foods to the participants and were flexible relative to their preferences regarding how and when foods

were eaten [58]. In addition, we encouraged the participants to select their own favorite foods if not provided [47]. They used the TADA mobile application to record their eating occasions and all the images were uploaded to our back-end server and then analyzed using the techniques we described in the previous sections.

This study consists of 45 participants. Each participant was asked to acquire eating occasion images at each eating occasion for a 7-day period. In total 1,453 eating occasion images were collected in the study and 42 commonly eaten food items were analyzed. Moreover, the dataset contains rich contextual information, such as participant feedback, temporal data, GPS location and nutrient information.

We used the free-living dataset as it is for evaluating features and classification. To evaluate the personalized learning model on a day-to-day basis, we selected participants in the free-living study with similar food consumption patterns to construct three datasets. We measure the similarity using Euclidean distance between each food consumption pattern and used *K-means* for clustering. For example, one of the datasets contains 119 food images from participant 14, 17, 20 and 32. As illustrated in Fig. 10, participant 14, 17, 20 and 32 all show relatively high consumption frequency of milk, mixed salad and lasagna. Each dataset features a different food consumption pattern and contains approximately 120 images. We labeled them as *Dataset 1, 2 and 3* corresponding to *User 1, 2 and 3*. Milk, lasagna, mixed salad and garlic bread are the most frequently-consumed foods in *Dataset 1*, while *Dataset 2* does not have any frequently-consumed foods except milk. *Dataset 3* represents a significant dietary pattern change within a month. The first three weeks in *Dataset 3* have similar food consumption style as *Dataset 1*. However, the eating pattern of the last week was selected to be noticeably different.

In addition, we use a subset of the 1453 images for evaluating the performance of the region proposal method with deep features. Due to the limitation of the bounding box groundtruth we have in the free-living dataset, the subset consists of 60 images of 24 food classes and we denote it as *Dataset 4*. On average, each image contains 4 instances of the 24 food classes. The bounding box groundtruth of the 24 food classes from the rest of the free-living dataset are used for training.

5.2 Feature and classification

As we discussed in Section 2, we classify food items based on the automatic segmentation result and the features extracted from each segment. The food identification accuracy is defined as: $accuracy = TP/(TP + FP/M + FN)$
where TP indicates True Positives (correctly detected food segments); FP indicates False Positives (incorrectly detected food segments or misidentified foods); FN indicates False Negatives (food not detected). Finally, M refers to the identification accuracy order. If one food classification label is generated for each image segment, then $M = 1$. In our implementation, after image segmentation and food classification, each image segment is assigned 4 food categories (or classes) with the 4 largest class labeling confidence scores ($M = 4$).

The classification accuracy of a single feature is discussed in [18]. From the results, we see that color features achieve better classification accuracy compared to texture features. DCD outperformed all other features. This suggests that in general we can represent the color content of food items using only a few colors, which is consistent with color representation of general objects [31]. As to local features, the *MDSIFT* feature achieves better food classification results than the *SIFT* feature. We combine the confidence scores of food labels that belong to the same food class and choose M food classes with Top M

Table 2 Food classification accuracy from feature combination

Features	Top 1 accuracy	Top 4 accuracy
DCD + MDSIFT	60.9%	83.27%
DCD + MDSIFT + SCD	62.9%	85.1%
DCD + MDSIFT + SCD + SIFT	64.5%	84.2%
DCD + MDSIFT + SCD + SIFT + EFD	63.5%	83.4%
DCD + MDSIFT + SCD + SIFT + EFD + GFD	62.9%	82.8%

confidence scores. Each feature is assigned with a weight from 0 to 1 based on our training experiment and the final confidence score is obtained by a weighted sum model. The food classification accuracy of Top 1 and Top 4 most probable food classes is shown in Table 2.

Table 2 summaries the food classification results after combining multiple features. Based on the performance of feature combinations and complexity consideration, we choose three features, namely, DCD, MDSIFT and SCD, in our food classification system. The Top 1 and Top 4 food classification accuracy for each food item using the combination of these three features is shown in Fig. 12. When tested on Intel Xeon X5550 CPU, it usually takes 30-70s for one image to complete segmentation and classification depending on the number of food items in the image.

We fully understand that automatic identification of food items in an image is not an easy problem and we will not be able to recognize every type of food. The way of packaging or the way the food is served will present problems for automatic recognition. Also some food items are inherently difficult to identify due to their visual similarity in the feature space. In some cases, even if a food is undetected or not correctly identified, it may not make much difference with respect to the energy or nutrients consumed. For example, if we fail to detect water in an eating occasion image, it will have little impact on the estimate of the energy or nutrients consumed in the meal due to the low energy content of water. Similarly, if our system identifies a "brownie" as "chocolate cake," there is no significant difference in energy or nutrients consumed.

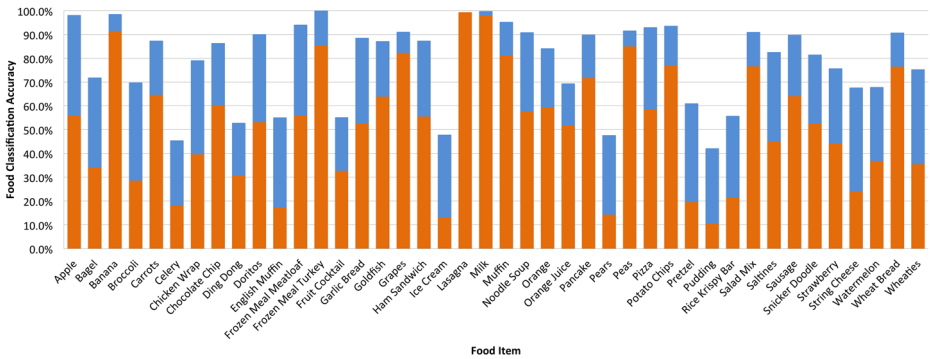


Fig. 12 Top 1 and Top 4 food classification accuracy for each food item with three features fused together, DCD, MDSIFT and SCD. The top of the orange bar: Top 1 classification accuracy; the top of the blue bar: Top 4 classification accuracy

5.3 Region proposal based approach with deep features

To evaluate the performance of the region proposal based method, we selected 24 food classes with relatively sufficient bounding box groundtruth as mentioned in Section 5.1. We start with 20 bounding boxes on average for each class and we augment the training set by applying minor shifting to the original groundtruth. We end up getting more than 100 bounding boxes for each class. We validate the SVM model using a 5-fold cross-validation. For each bounding box groundtruth, we first resize it so that the shorter dimension of the region is 224 while keeping the aspect ratio. Then the deep features are extracted from the 224×224 center-cropped region. In our experiment, we chose the SVM model with the Radial Basis Function kernel where $C = 1000$ and $\gamma = 0.001$. *Dataset 4* was used to evaluate the proposed method. For each detected region, we consider it correct if it has more than 80% overlap with the groundtruth. We report 52.4% detection and classification accuracy.

5.4 Contextual refinement

We conducted two experiments to validate the personalized learning model. First, we tested on the same 1453 eating images used in the feature and classification experiment by assuming the food assumption frequency and co-occurrence pattern are known. If the food co-occurrence pattern is given, the Top 1 and Top 4 food identification accuracy increased to 65.3% and 85.9% compared to 62.9% and 85.1% without contextual information. The accuracy is further improved to **71.4%** and **88.3%** with both food assumption frequency and co-occurrence pattern. The food identification accuracy for each food item is shown in Fig. 13. Comparing food classification accuracy obtained after contextual refinement (Fig. 13) and before contextual refinement (Fig. 12), we can see that most of the food items in our dataset achieve a better classification accuracy. Similarly, we tested the region proposal based method on *Dataset 4*. We achieved 57.5% detection and classification accuracy with contextual information by assuming the food assumption frequency and co-occurrence pattern are given compared to 52.4% without contextual information.

Next, we would like to examine how the personalized learning model behaves day by day over a month. As we mentioned in Section 5.1, we have created three datasets from

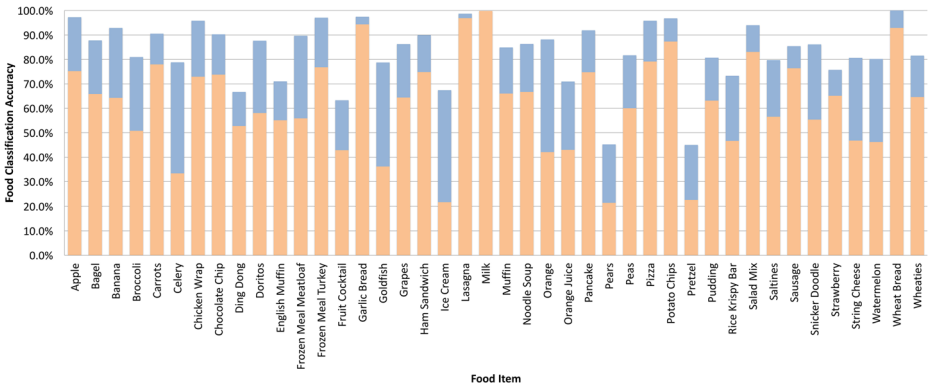


Fig. 13 Food identification accuracy for each food item after integrating contextual dietary information

the free-living study, each of which features a different food consumption pattern. In the following experiment, we use the method discussed in Section 2 as it shows better result than the region proposal based method in the previous evaluation.

Figure 14a shows how the recursive Bayesian network updates the prediction probabilities for three example food items in *Dataset 1* from 12 am to 11am. On Day 1, every food has the same prediction. In the end, the prediction of milk, orange juice and muffin converges to 0.51, 0.19 and 0.06 respectively. Figure 14b compares the prediction of milk among all three datasets from 11am to 4pm. It is clear that *User 1* consumes milk during lunch time more frequently than other *Users*.

Note we use the food label with the highest confidence score (top 1) from the classifier. As we show below the classification accuracy from the highest confidence score is in the range of 50-65%. In the TADA system we report the top 4 food labels and have a classification accuracy of 80-85% [18].

Figure 15 demonstrates the food classification accuracy improvement. The blue lines in Fig. 15a, c and e indicate the average daily food classification accuracy with temporal context, Θ_{context}, while the red lines indicate the one without, Θ_{auto}. The accuracy improvement is illustrated in Fig. 15b and it is defined as, $improvement = (\Theta_{\text{context}} - \Theta_{\text{auto}})/\Theta_{\text{auto}}$.

As shown in Fig. 15b, the accuracy improvement drops from Day 10 to Day 20 as the baseline (classification accuracy without context) increases from 47% to 57%. This implies that the proposed method is more effective when the automatic image analysis does not work well. The 80% accuracy rate achieved with temporal context on Day 25 in Fig. 15a demonstrates the effectiveness of the proposed method when the automatic image analysis result is poor (36%). In *Dataset 2*, the classification accuracy without context is always above 55% (see the red line in Fig. 15c). The drop in the first few days shown in Fig. 15d implies the undergoing learning process. Nevertheless, Fig. 15b and d both illustrate an ascent trend of accuracy improvement.

We selected images of the last 7 days to have a noticeably different food consumption pattern compared to the first 23 days in *Dataset 3*. We would like to verify the behavior of our training model under circumstance where a participant may change their eating style. We witnessed a huge drop in Fig. 15f followed by the re-learning state. The accuracy improvement is the minimum on Day 24 after the first week, because the context-based prediction puts more confidence on the specific food, which *Dataset 3* no longer contains

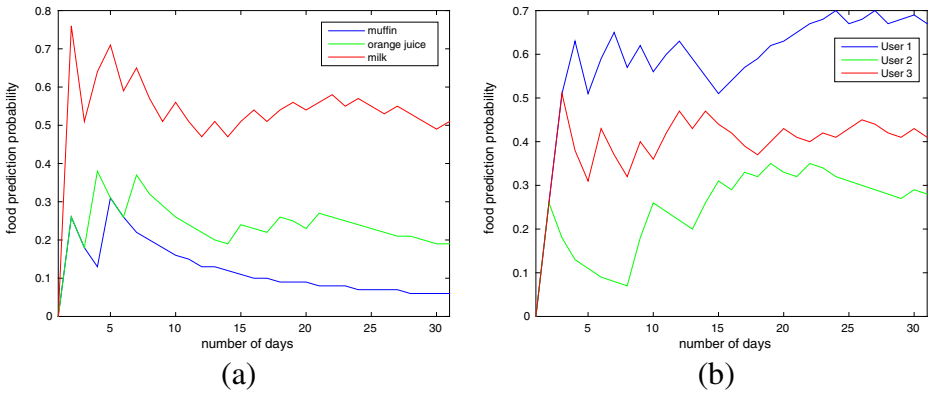


Fig. 14 **a** Food occurrence prediction of three food items, **b** Prediction of milk among three datasets

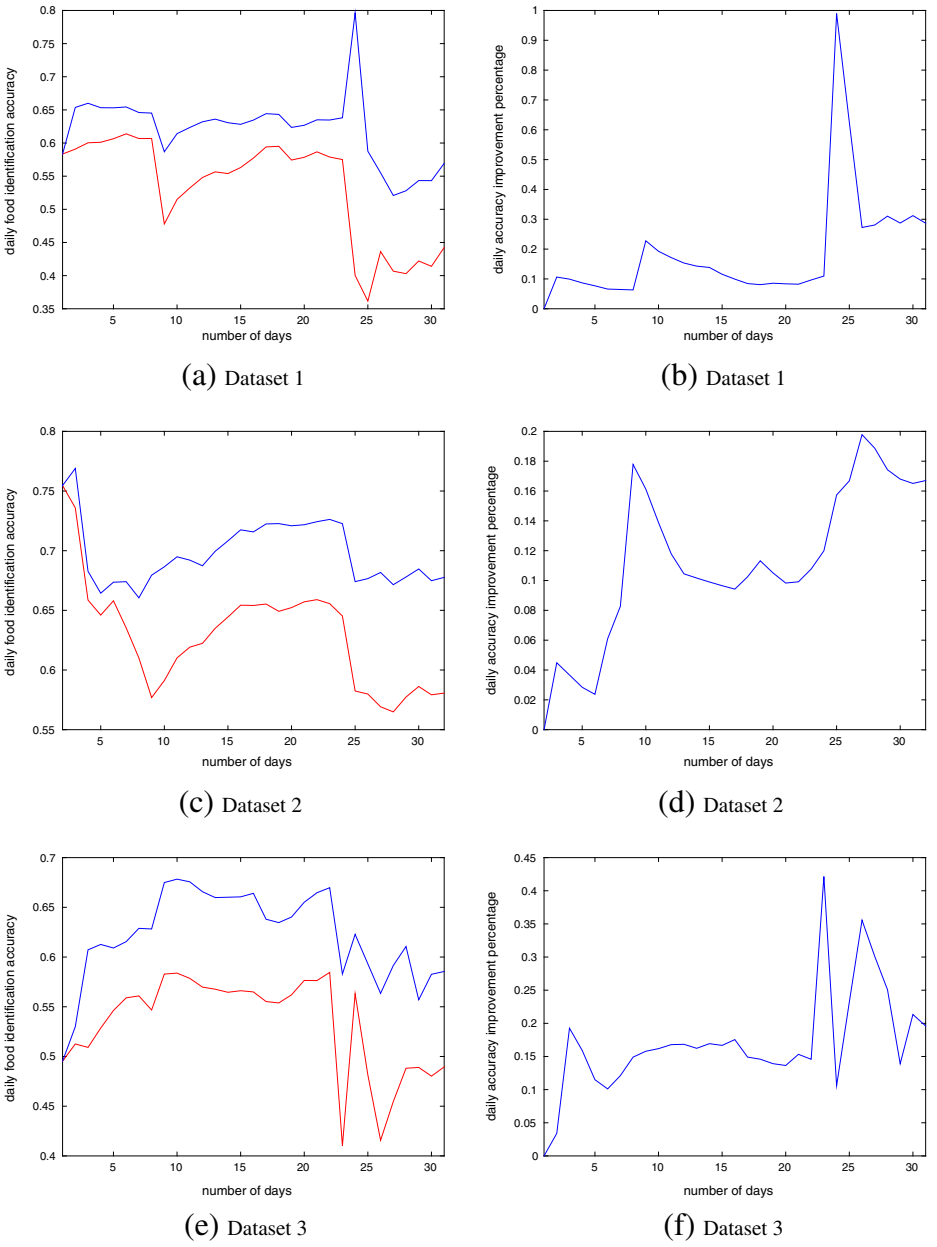


Fig. 15 Learning curves for one month. Daily classification rates with(blue) and without(red) temporal context are illustrated in (**a**),(**c**) and (**e**). Corresponding accuracy improvements are shown in (**b**),(**d**) and (**f**)

after the user 3 changes eating habit. For example, milk is not consumed on Day 24. Due to the dietary change in *Dataset 3*, the increasing trend of classification accuracy is not as obvious. Table 3 compares the average daily classification accuracy with and without

Table 3 Food Classification with Context Information

Statistics	User ID	With context	Without context
Average daily classification accuracy(%)	user1	62.90	53.23
	user2	69.81	62.90
	user3	62.12	53.28
Average daily accuracy improvement(%)	user1	18.69	
	user2	10.94	
	user3	17.05	

contextual information for each user. The average daily accuracy improvement is calculated as:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\Theta_{context}^{(i)}-\Theta_{auto}^{(i)}}{\Theta_{auto}^{(i)}} \tag{10}$$

Due to our dataset selection, the classification accuracy using automatic image analysis alone in *Dataset 2* is significantly higher than other datasets. The lower accuracy in *Dataset 1* and *Dataset 3* reflects the variation in the subset of the total 1453 testing images. Thus, the accuracy improvement for *Dataset 2* is expected to be lower (10.94%). The fact that *Dataset 2* has less frequently-consumed foods also contributed to the lower accuracy improvement. When a person has a more consistent eating pattern, such as User 1, the classification accuracy gain using temporal contextual information is higher (18.69%). On average, the proposed method of utilizing temporal context shows 15.56% improvement. In the end, three datasets obtain roughly 65% accuracy with contextual information, which slightly lower than 71.4% we reported in the first experiment. This is because the classifier gradually learns the personalized eating patterns throughout 30 days, therefore the accuracy improvement in the earlier days is expected to be relatively lower than the one of the later days.

6 Conclusions

Dietary assessment is a comprehensive evaluation of a person's food intake which may suggest risk factors for diet-related chronic diseases and help to prevent them. In this paper, we explored methods for use of contextual information dietary assessment.

After the classifying an object with visual features, we use contextual information to refine the classification decision. The temporal dietary information is used to predict the likelihood of eating certain foods based on the time of a day and improve the precision of food classification results. The food co-occurrence pattern is investigated to reach a labeling agreement for all the segmented regions of the input image. Both temporal information and food co-occurrence pattern are integrated in the personalized learning model to accommodate different users' preferences. Experimental results using a free-living dietary study show that the contextual refinement step is able to improve the classification accuracy of segmented regions in the images by 15.56%. The increase in food identification accuracy through incorporating contextual dietary information indicates that our contextual models are promising and further investigation is warranted. In the future, dietary patterns can also be incorporated as "side" information for the classifier, if an individual has a repeated

dietary behavior, the classifier can dynamically select the food classes in the training dataset that match the eating pattern.

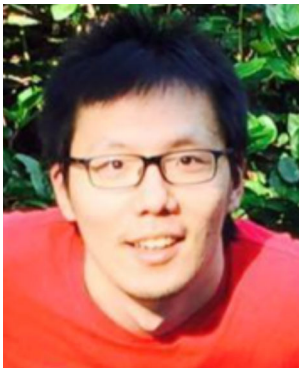
Acknowledgements This work was sponsored by the US National Institutes of Health under grant NIH/NCI 1U01CA130784-01 and NIH/NIDDK 2R56DK073711-04,1R01-DK073711-01A1. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the US National Institutes of Health.

References

1. Agarwal S, Snavely N, Simon I, Seitz SM, Szeliski R (2009) Building rome in a day. In: Proceedings of the IEEE international conference on computer vision. Elsevier, Kyoto, pp 72–79
2. Argus. http://www.azumio.com/s/argus/index.html
3. Amadasun M, King R (1989) Textural features corresponding to textural properties. IEEE Trans Syst Man Cybern 19(5):1264–1274
4. Arulampalam S, Maskell S, Gordon N, Clapp T (2002) A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. IEEE Trans Signal Process 50(2):174–188
5. Bay H, Ess A, Tuytelaars T, Gool LV (2008) Speeded-up robust features (SURF). Journal of Computer Vision and Image Understanding 110(3):346–359
6. Biederman I, Mezzanotte R, Rabinowitz J (1982) Scene perception: detecting and judging objects undergoing relational violations. Cogn Psychol 14(2):143–177
7. Bossard L, Guillaumin M, Van Gool L (2014) Food-101 – mining discriminative components with random forests. European Conference on Computer Vision 8694:446–461
8. Boushey CJ, Kerr DA, Wright J, Lutes KD, Ebert DS, Delp EJ (2009) Use of technology in children's dietary assessment. Eur J Clin Nutr 63:S50–S57
9. Choi T, Chin S (2013) An intelligent wellness keeper for food nutrition with graphical icons. International Journal of Multimedia and Ubiquitous Engineering 8:207–214
10. Deng Y, Manjunath BS, Kenney C, Moore MS, Shin H (2001) An efficient color representation for image retrieval. IEEE Trans Image Process 10:140–147
11. Diet Camera. http://www.dietcamera.com/
12. Duda R, Hart P (1973) Pattern classification and scene analysis. Wiley, Hoboken
13. Fang S, Liu C, Zhu F, Delp E, Boushey C (2015) Single-view food portion estimation based on geometric models. In: Proceedings of the IEEE international symposium on multimedia. Elsevier, Miami, pp 385–390
14. Felzenszwalb P, Huttenlocher D (1998) Image segmentation using local variation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE, Santa Barbara, pp 98–104
15. Galleguillos C, Belongie S (2010) Context based object categorization: a critical survey. Comput Vis Image Underst 114:712–722
16. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE, Columbus, pp 580–587
17. He Y, Khanna N, Boushey C, Delp E (2013) Image segmentation for image-based dietary assessment: A comparative study. In: Proceedings of the IEEE international symposium on signals, circuits and systems. Springer, Iasi, pp 1–4
18. He Y, Xu C, Khanna N, Boushey C, Delp E (2014) Analysis of food images: Features and classification. In: Proceedings of the IEEE international conference on image processing. IEEE, Paris, pp 2744–2748
19. Jordan MI, Ghahramani Z, Jaakkola TS, Saul LK (1999) An introduction to variational methods for graphical models. Mach Learn 37(2):183–233
20. Joutou T, Yanai K (2009) A food image recognition system with multiple kernel learning. In: Proceedings of the IEEE international conference on image processing. Springer, Cairo, pp 285–288
21. Julesz B (1981) Textons, the elements of texture perception and their iteractions. Nature 290:91–97
22. Kass M, Witkin A, Terzopoulos D (1988) Snakes: active contour models. International journal Of Computer Vision 1(4):321–331
23. Kelkar S, Stella S, Okos M (2010) X-Ray micro computed tomography (CT): a novel method to measure density of porous food. In: Proceedings of the IFT annual meeting and food expo. ACM, Chicago
24. Kenney C, Deng Y, Manjunath BS, Hewer G (2001) Peer group image enhancement. IEEE Trans Image Process 10:326–334

25. Kitamura K, Yamasaki T, Aizawa K (2009) Foodlog: capture, analysis and retrieval of personal food images via web. In: Proceedings of the ACM multimedia workshop on Multimedia for cooking and eating activities. MIT Press, Beijing, pp 23–30
26. Kong F, He H, Raynor HA, Tan J (2015) Dietcam: multi-view regular shape food recognition with a camera phone. Pervasive Mob Comput 19:108–121
27. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Proceedings of advances in neural information processing systems, pp 1097–1105
28. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436–444
29. Livingstone MBE, Robson PJ, Wallace JMW (2004) Issues in dietary intake assessment of children and adolescents. Br J Nutr 92:S213–S222
30. Lowe D (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 2(60):91–110
31. Ma WY, Deng Y, Manjunath B (1997) Tools for texture- and color-based search of images. In: Proceedings of the SPIE human vision and electronic imaging II 3016, San Jose, pp 496–507
32. Manjunath B, Ohm JR, Vasudevan V, Yamada A (2001) Color and texture descriptors. IEEE Trans Circuits Syst Video Technol 11(6):703–715
33. Martinel N, Foresti GL, Micheloni C (2016) Wide-slice residual networks for food recognition. arXiv:1612.06543
34. McFee B, Galleguillos C, Lanckriet G (2011) Contextual object localization with multiple kernel nearest-neighbor. IEEE Trans Image Process 20(2):570–585
35. Meyers A, Johnston N, Rathod V, Korattikara A, Gorban A, Silberman N, Guadarrama S, Papandreou G, Huang J, Murphy KP (2015) Im2calories: Towards an automated mobile vision food diary. In: Proceedings of the IEEE international conference on computer vision. MIT Press, Santiago, pp 1233–1241
36. Mikolajczyk K, Schmid C (2004) Scale and affine invariant interest point detectors. Int J Comput Vis 1(60):63–86
37. Murphy K, Torralba A, Freeman W (2003) Using the forest to see the trees: a graphical model relating features, objects and scenes. Adv Neural Inf Proces Syst 16:1499–1506
38. National Vital Statistics System U.S. (2009) Quickstats: age-adjusted death rates for the 10 leading causes of death. Morb Mortal Wkly Rep 58(46):1303
39. Nister D, Stewenius H (2006) Scalable recognition with a vocabulary tree. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Elsevier, Washington, pp 2161–2168
40. Oliva A, Torralba A (2007) The role of context in object recognition. Trends Cogn Sci 11(12):520–527
41. Peddi SVB, Kuhad P, Yassine A, Pouladzadeh P, Shirmohammadi S, Shirehjini AAN (2017) An intelligent cloud-based data processing broker for mobile e-health multimedia applications. Futur Gener Comput Syst 66:71–86
42. Philbin J, Chum O, Isard M, Sivic J, Zisserman A (2007) Object retrieval with large vocabularies and fast spatial matching. In: Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE, Minneapolis, pp 1–8
43. Pytorch. http://www.pytorch.org/. Tensors and Dynamic neural networks in Python with strong GPU acceleration
44. Rabinovich A, Vedaldi A, Galleguillos C, Wiewiora E, Belongie S (2007) Objects in context. In: Proceedings of the IEEE international conference on computer vision. IEEE, Rio de Janeiro, pp 1–8
45. Razavian A, Azizpour H, Sullivan J, Carlsson S (2014) CNN features off-the-shelf: an astounding baseline for recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. Cambridge University Press, Columbus, pp 806–813
46. Sarkka S (2013) Bayesian filtering and smoothing. Cambridge University Press, Cambridge
47. Schap T, Zhu F, Delp E, Boushey C (2014) Merging dietary assessment with the adolescent lifestyle. J Hum Nutr Diet 27(s1):82–88
48. Schindler G, Brown M, Szeliski R (2007) City-scale location recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE, Minneapolis, pp 1–7
49. Shi J, Malik J (2000) Normalized cuts and image segmentation. IEEE Trans Pattern Anal Mach Intell 22(8):888–905
50. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
51. Stella S, Kelkar S, Okos M (2010) Predicting and 3D laser scanning for determination of apparent density of porous food. In: Proceedings of the IFT annual meeting and food expo. Elsevier, Chicago
52. Thompson FE, Subar AF, Loria CM, Reedy JL, Baranowski T (2010) Need for technological innovation in dietary assessment. J Am Diet Assoc 110(1):48–51
53. Tola E, Lepetit V, Fua P (2010) DAISY: an efficient dense descriptor applied to wide baseline stereo. IEEE Trans Pattern Anal Mach Intell 32(5):815–830

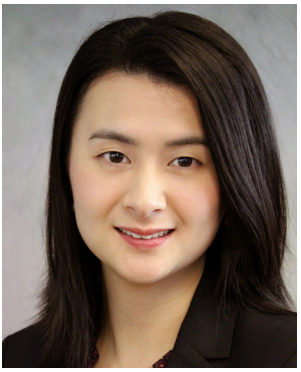
54. Torralba A, Murphy KP, Freeman WT, Rubin MA (2003) Context-based vision system for place and object recognition. In: Proceedings of the IEEE international conference on computer vision, Nice, pp 273–280
55. Tuingle. http://tuingle.com/
56. Uijlings JR, van de Sande KE, Gevers T, Smeulders AW (2013) Selective search for object recognition. Int J Comput Vis 104(2):154–171
57. Wang X, Yang M, Cour T, Zhu S, Yu K, Han TX (2011) Contextual weighting for vocabulary tree based image retrieval. In: Proceedings of the IEEE international conference on computer vision, pp 209–216
58. Wang Y, He Y, Zhu F, Boushey C, Delp E (2015) The use of temporal information in food image analysis. In: Murino V, Puppo E, Sona D, Cristani M, Sansone C (eds) New Trends in image analysis and processing – ICIAP 2015 workshops, lecture notes in computer science, vol 9281. Springer International, Berlin, pp 317–325
59. Zhu F, Bosch M, Woo I, Kim S, Boushey C, Ebert D, Delp E (2010) The use of mobile devices in aiding dietary assessment and evaluation. IEEE J Sel Top Sign Proces 4(4):756–766
60. Zhu F, Bosch M, Khanna N, Boushey C, Delp E (2015) Multiple hypotheses image segmentation and classification with application to dietary assessment. IEEE journal of Biomedical and Health Informatics 19(1):377–388



Yu Wang received B.S from Nanjing University of Aeronautics and Astronautics, China in 2012. Currently, he is a PhD candidate in Electrical and Computer Engineering at Purdue University, USA. His research interest includes image processing, computer vision and machine learning.

Ye He is a Software Engineer at Google Inc. Dr. He received her Ph.D. in Electrical and Computer Engineering from Purdue University in 2014. Her research interests include Machine Learning, Image processing and analysis, and Computer Vision. Her graduate study was partially supported by a Ross Fellowship.

Carol J. Boushey is an Associate Research Professor in the Epidemiology Program at the University of Hawaii Cancer Center. She directs the Nutrition Support Shared Resource (NSSR) for the Cancer Center. She also holds an adjunct professor position at Purdue University in West Lafayette, Indiana. Dr. Boushey serves on the Board of Editors of the Journal of the Academy of Nutrition and Dietetics and Nutrition Today. She is the co-editor for the fourth edition of the Elsevier publication, Nutrition in the Treatment and Prevention of Disease, to be released in the 2017. Her research focuses on dietary assessment and examining the relationship of dietary intakes or dietary patterns as an exposure for health or risk for disease. Her research has appeared in book chapters and journals, such as the Journal of the Academy of Nutrition and Dietetics (JAND), Pediatrics, the Journal of Nutrition, the American Journal of Clinical Nutrition, and JAMA. She is a member of the JAND 'statistical team' which has published papers and book chapters to guide practitioners, students, and scientists to conduct successful research and report findings. She received the B.Sc. degree from the University of Washington, Seattle, WA, USA, and the Masters of Public Health from the University of Hawaii at Manoa, Honolulu, HI, USA, and the Ph.D. degree from the University of Washington through the interdisciplinary nutrition program and the epidemiology program.



Fengqing Zhu is an Assistant Professor of Electrical and Computer Engineering at Purdue University, West Lafayette, IN. Dr. Zhu received her Ph.D. in Electrical and Computer Engineering from Purdue University in 2011. Prior to joining Purdue in 2015, she was a Staff Researcher at Huawei Technologies (USA), where she received a Huawei Certification of Recognition for Core Technology Contribution in 2012. Her research interests include Image processing and analysis, video compression, computer vision and computational photography. Dr. Zhu was selected to participate in the National Institutes of Health (NIH) mHealth Summer Institute in 2011. She was a Student Intern at the Sharp Laboratories of America in the summer of 2007 and attended the McKinsey Insight Engineering & Science Program in the summer of 2008. Her graduate study was partially supported by a Charles C. Chappelle Graduate Fellowship and a Motorola Foundation Fellowship.

Edward J. Delp (S'70-M'79-SM'86-F'97) was born in Cincinnati, OH, USA. He received the B.S.E.E. (cum laude) and the M.S. degrees from the University of Cincinnati and the Ph.D. degree from Purdue University, West Lafayette, IN, USA. In May 2002, he received an Honorary Doctor of Technology from the Tampere University of Technology, Tampere, Finland. From 1980 to 1984, he was with the Department of Electrical and Computer Engineering, The University of Michigan, Ann Arbor, MI, USA. Since August 1984, he has been with the School of Electrical and Computer Engineering and the School of Biomedical Engineering, Purdue University. He is currently The Charles William Harrison Distinguished Professor of Electrical and Computer Engineering and Professor of Biomedical Engineering. His research interests include image and video compression, multimedia security, medical imaging, multimedia systems, communication, and information theory. Dr. Delp is a Fellow of the SPIE, a Fellow of the Society for Imaging Science and Technology (IS&T), and a Fellow of the American Institute of Medical and Biological Engineering. In 2000, he was selected a Distinguished Lecturer of the IEEE Signal Processing Society. He received the Honeywell Award in 1990, the D. D. Ewing Award in 1992 and the Wilfred Hesselberth Award in 2004 all for excellence in teaching. In 2001, he received the Raymond C. Bowman Award for fostering education in imaging science from the Society for Imaging Science and Technology (IS&T). In 2004, he received the Technical Achievement Award from the IEEE Signal Processing Society for his work in image and video compression and multimedia security. In 2008, he received the Society Award from IEEE Signal Processing Society (SPS). This is the highest award given by SPS and it cited his work in multimedia security and image and video compression. In 2009, he received the Purdue College of Engineering Faculty Excellence Award for Research.