Published in final edited form as:
IEEE Trans Image Process. 2011 ; 2011: 1789–1792. doi:10.1109/ICIP.2011.6115809.

COMBINING GLOBAL AND LOCAL FEATURES FOR FOOD IDENTIFICATION IN DIETARY ASSESSMENT

Marc Bosch[†], **Fengqing Zhu**[†], **Nitin Khanna**[†], **Carol J. Boushey**[‡], and **Edward J. Delp**[†]

[†]Video and Image Processing Lab (*VIPER*), School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907, USA

[‡]Department of Foods and Nutrition, Purdue University, West Lafayette, IN 47907, USA

Abstract

Many chronic diseases, such as heart diseases, diabetes, and obesity, can be related to diet. Hence, the need to accurately measure diet becomes imperative. We are developing methods to use image analysis tools for the identification and quantification of food consumed at a meal. In this paper we describe a new approach to food identification using several features based on local and global measures and a "voting" based late decision fusion classifier to identify the food items. Experimental results on a wide variety of food items are presented.

Index Terms

Feature extraction; image analysis; object recognition; supervised learning; image texture

1. INTRODUCTION

There is a growing concern in the world about chronic diseases related to diet, including obesity, cancer, diabetes, and heart disease. Of the 10 leading causes of death in the U.S., 6 are related to diet. Dietary intake, the process of determining what someone eats during the course of a day, provides valuable insights for mounting intervention programs for prevention of many chronic diseases. The use of a mobile telephone's built-in digital camera has been shown to provide unique mechanisms for reducing a users burden and improving the accuracy and reliability of dietary assessment [1]. To provide accurate estimates of food energy and nutrient intake we are developing methods to automatically estimate the food consumed at a meal from image acquired using a mobile device. We have developed a system, known as the mobile telephone food record (mpFR), to automatically identify and quantify foods and beverages consumed based on analyzing images of food obtained with a mobile telephone [2]. This systems is shown in Figure 1. The image analysis involves three main steps: image segmentation, food identification, and volume estimation. Images captured before and after foods are eaten are used to estimate the food intake. The energy of the food consumed can then be determined. This paper concentrates on new methods to

Address all correspondence to Edward J. Delp, ace@ecn.purdue.edu or see www.tadaproject.org.

identify and classify the food items after segmentation, namely the features and the classification methods. A complete description of the mpFR is presented in [2].

In general, visual categorization of scenes is still an open problem. Many recognition methods are based on the "parts and structure" model introduced by Fischler and Elshlager [3], where the object is recognized by finding its constituent parts and measuring their geometric relationships. This model has been extended to detect and learn new categories [4]. Another approach for object classification is to represent each object as unordered collections of feature descriptors creating what is known as a *bag of features (BoF)* or a *bag of words* [5]. These methods categorize the object in isolation from the contextual information (scene). In [6] general scene understanding is shown to significantly improve the performance of object classification methods.

Food identification is a difficult problem since foods can dramatically vary in appearance [7, 8, 2]. Such variations may arise not only from changes in illumination and viewpoint, but also from non-rigid deformations, and intraclass variability in shape, texture, color and other visual properties. There has been recent efforts to address food identification. In [9] a method for food identification is described by exploiting the spatial relationship among different ingredients (such as meat and bread in a sandwich). The food items are represented by pairwise statistics between local features of the different ingredients of the food items. In [10], a multiple kernel learning method is described to integrate three sets of features namely color, texture, and SIFT descriptors. All three features are fused together forming one single feature vector by assigning different weights to combine them. Finally in [11], an online food-logging system is presented that distinguishes food images from other images, analyzes the food balance, and visualizes the log. Global and local features are used to describe food items and classify them using a Support Vector Machine (SVM).

2. VISUAL REPRESENTATION OF FOOD ITEMS

An essential step in solving any object categorization problem is to adequately represent the visual information of the object. This is commonly known as feature extraction. Our goal is to find features that describe the object so that in the classification step we can achieve maximum inter-class discrimination and maximum intra-class robustness. We explored many different features to assess the role of each type of feature for visual description of food item. Our current approach uses both global and local features. By global features we mean features that describe the entire object with a single feature vector. Global features can be of limited use when there is object occlusion, different poses, lighting and intra-class variation. In order to overcome this problem local features are used, which ideally describe the local visual characteristics around salient or invariant points. In our system, feature extraction is done after the segmentation step [12], which locates the object boundary for each food item within the image. Segmentation is beyond the scope of this paper. The segmentation method currently used in the mpFR is described in[12].

2.1. Global Features

Global features incorporate statistics of the overall distribution of the visual information in the object. Our proposed method uses two classes of global features: color and texture.

We consider three types of color features namely *color statistics, entropy statistics,* and *predominant color statistics. Color statistics* consists of estimating the 1st and 2nd moment statistics of the *R, G, B, Cb, Cr, a, b, H, S, V* channels for the entire object/segment. The segmented object is divided into rectangular blocks to measure *entropy statistics*. For each block ($N \times N$ pixels), 1st and 2nd moment statistics of the entropy in the *R, G, B* channels are estimated. The final feature is obtained by averaging the values for all the blocks. The last global color feature is inspired by the Dominant Color Descriptor of the MPEG-7 (DCD) [13]. The distribution of the salient colors in the object is estimated by selecting the four most representative colors (in RGB space) for an object. The RGB space is quantized into a 1000-cell cube where each cell represents a color value. This cube can be seen as a 1000-bin histogram. The four largest peaks of the food item color histogram are considered to be the predominant colors. The feature vectors are given by:

$$F=\{(c_1, p_1, v_1), \ldots, (c_4, p_4, v_4)\} \quad (1)$$

where c_i represents the 3-D color vector from the RGB cube, p_i is the percentage of color in the total object, and v_i is the color variance inside the region described by the predominant color.

Finally to capture texture information we use Gabor filterbank [14, 15]. We divide the segmented object into rectangular blocks and filter each block with a bank of Gabor filters at four scales and six orientations. Each block is filtered with 24 Gabor filters, and the energy of the responses is estimated. From the energy we extract mean and variance to form the texture signatures. We average the mean and variance values over all the blocks.

2.2. Local Features

Extracting local features consists of describing visual information from a patch or neighborhood ($M \times M$ pixels) around a point of interest in the food item. As a result of this, for each type of local feature we form as many feature vectors as points of interest detected in that food item. We have evaluated 8 different types of local features including high and low level descriptors: *local color, local entropy color, Tamura perceptual features, Gabor filters, SIFT descriptor, Haar wavelets, Steerable filters*, and *DAISY descriptor*.

The local color features capture color information around each point of interest by estimating 1st and 2nd moment statistics of the *R, G, B, Cb, Cr, a, b, H, S, V* channels. As in the global case, entropy color features are also extracted. The 1st and 2nd moment statistics of the entropy in the *R, G, B* channels are estimated within the neighborhood of feature points.

To capture local texture descriptions, we used again a bank of Gabor filters at four scales and six orientations. We estimated mean and variance of the energy response of the filtered signal. In addition, we use *Tamura features* to represent texture. Tamura *et al.* proposed 6 perceptual texture measures namely coarseness, contrast, directionality, line-likeness, regularity, and roughness selected by psychological experiments to better describe the human perception [16]. In our experiments we have discovered that coarseness, contrast, and directionality are the most discriminative. Coarseness provides information about the size of the texture elements. The higher the coarseness, the rougher the texture. Contrast reflects picture quality because it is influenced by the dynamic range of the gray levels, sharpness of edges and period of repeating patterns. Finally directionality is estimated by a histogram of gradient orientations giving the directions that the texture patterns follow.

The SIFT descriptor [17], introduced by Lowe in 2004, consists of estimating the gradient at each pixel in a 16×16 neighborhood around a point of interest. At each 4×4 sub-block an 8-bin histogram of the gradient orientation is formed, these 128 values are normalized to form a unit length feature vector. Normalization is done to achieve robustness against photometric variations.

Similar to SIFT, the Haar wavelet features capture the distribution of gradients within the neighborhood around the point of interest. In this case the gradient response is approximated by Haar wavelet responses in horizontal and vertical directions. In [18], a novel low level descriptor known as SURF was introduced that is based on the 2D Haar wavelet responses and made efficient use of the box filter. The dominant orientation of each image patch is estimated to obtain orientation invariance. The patch is then pre-filtered with a Gaussian kernel to obtain scale invariance. We use the Haar wavelet features based on the SURF descriptors.

Steerable filters refer to randomly oriented filters synthesized using a linear combination of the basis filters [19]. We use 2D circularly symmetric Gaussian functions and obtain the 1^{st} and 2^{nd} moment statistics of the response of filtered patch with the steerable filter. We use 5 orientations and up to 5^{th} order Gaussian derivative.

We also use the DAISY descriptor, which depends on histograms of gradients like SIFT but uses a Gaussian weighting and circularly symmetrical kernel [20] to estimate these histograms. The final feature vector is formed by normalizing the histograms.

3. GLOBAL AND LOCAL FEATURE CLASSIFICATION

We categorize the food items based on the 12 global and local features vectors we described above. Each feature is independently classified. As a result of this process, we have 12 decisions of candidate classes for each food item. The final decision is made by a majority vote of the individual classifiers.

3.1. Global Features Classification

The global features are classified independently using Support Vector Machine (SVM) classifiers [21, 22]. SVM produces a model of classification space by constructing an N-dimensional hyperplane that optimally separates the training data into categories. We used the Radial Basis Function (RBF) kernels for our study.

3.2. Local Features Classification

We use the *Bag of Features* (BoF) [5, 23] approach for classifying local features. The BoF estimates the distribution of visual “words” found in the object and compares this distribution to those found in the training set. These visual “words” are formed from orderless collections of local descriptors around the points of interest. The main steps of the BoF approach are: point of interest detection, local descriptor representation, vocabulary construction, and supervised classification. Difference of Gaussian (DoG) [17] has proven to be an efficient method to detect points of interest. Thus, we adopt DoG as our point of interest detector. The DoG representation is obtained by estimating a set of sub-octave DoG filters. A DoG filter is generated by subtracting two successive Gaussian smoothed images. The points of interest are identified by looking for a space and scale extrema in the resulting structure. Each pixel in the DoG images is compared to its 8 neighbors at the same scale, plus the 9 corresponding neighbors at neighboring scales. If the pixel is a local maximum or minimum, it is selected as a candidate point. Low contrast points and responses along the edges are eliminated from the group of candidates. For each point of interest, we estimate a gradient orientation so that the features are rotation invariant. The only requirement is to have at least 20 points of interest for each object.

Once the points of interest have been identified in the food item, we extract 8 types of local features (Section 2.2) around a neighborhood of each point (image patch). For stability purposes the size of the neighborhood is proportional to the size of the food item. For each feature type we obtain the visual vocabulary by performing a hierarchical K-means clustering. This method proceeds by iterated assignments of points to their closest cluster centers and recomputation of the cluster centers. Each cluster is subdivided into further clusters until the largest distance between the cluster center and its member is smaller than a certain threshold. Finally, for each feature type, we form the final representation of the food item which is known as the signature of the object,

$$S_j=\{(t_1,m_1),\ldots,(t_i,m_i),..,(t_N,m_N)\} \quad (2)$$

where S_j represents the signature of the object for the jth feature type, t_i represents the frequency term. We use a normalized version of the frequency term by dividing it by the total number of points of interest, and m_i is the *medoid* of the ith cluster. N is the number of clusters.

Once the signatures of the food items are formed, we classify them based on the nearest neighbor criteria to assign the corresponding food category.

4. MODEL EVALUATION AND DISCUSSION

As we mentioned earlier, our food identification system is preceded by a segmentation step [12]. In order to evaluate our features and classifiers we need to isolate the error caused by the feature extraction and classification steps from the error caused by segmentation step. To test the entire system we have created a database containing food images obtained under controlled conditions from nutritional studies conducted at Purdue University [24]. Many of these images have been hand segmented creating one of the largest groundtruth datasets of food images with more than 1000 hand segmented images. For these experiments we considered images from two user studies. In the first user study (dataset 1) there were 19 food items served in different meals, and we collected 63 images. On average there were approximately 20 images per food item. In the second user study (dataset 2) there were 28 different foods in 116 total images. We tested our food identification system for each of these two datasets. In a third experiment, we combined them for a total of 39 foods from 179 images. We divided the dataset into training and testing, for each category approximately half of the images were training data and half testing data. The minimum number of training image was 10. We evaluated our system using the groundtruth segmentation masks for feature extraction. We repeated the experiments 10 times, each time randomizing the training and testing sets.

In the results shown in Table 1 we have reported the performance of each type of feature individually and the overall classifier performance after the majority vote decision was made. Table 1 shows the results of correct classification rate averaged for all the 19 categories available in the first dataset (3^{rd} column), for all the 28 categories corresponding to the second dataset (4^{th} column), and the combination of both datasets (5^{th} column). Some food items are inherently difficult to classify due to their similarity in the feature space. Examples of such errors are *canned pineapple* labeled as *canned pear*, *ketchup* considered as *catalina dressing*, *garlic bread* classified as *white bread toasts*, or *chocolate cake* as *brownie* as shown in Figure 2.

Overall we believe that combining global and local features allows us to capture more discriminative visual information for each food item. Having highly informative features allows reliable classification with simple decision rules. As expected, color contributes the most in describing foods, however low level descriptors like SIFT, DAISY or SURF have also proven to be very efficient. We also observed that low level descriptors, and in general local features, are sensitive to the point detection criteria, *i.e.*, DoG, Laplacian of Gaussian, Harris points, Hessian-Laplace regions, because they do not always detect salient and discriminative regions.

As the size and number of categories grows, we will consider a hierarchical feature extraction and classification system. The first step will be able to separate groups of foods using a subset of features. As we go down the hierarchy, using groups of food specific features will give finer discrimination. We will also incorporate contextual information in our system to correct misidentified food items based on the likelihood of food combinations.

5. CONCLUSIONS

Measuring accurate dietary intake is an open research problem in the nutrition and health fields [2]. We are developing methods to automatically locate and identify foods in a meal image. This paper presented a food identification system that combines global and local features for a more accurate visual description of food items. We have shown that by applying late decision fusion-based rules to each individual feature channel we can increase the correct classification rate by more than 7%.

Acknowledgments

This work was sponsored by grants from the National Institutes of Health under grants NIDDK 1R01DK073711-01A1 and NCI 1U01CA130784-01.

References

1. Boushey C, Kerr D, Wright J, Lutes K, Ebert D, Delp E. Use of technology in children's dietary assessment. European Journal of Clinical Nutrition. 2009:S50–S57. [PubMed: 19190645]
2. Zhu F, Bosch M, Woo I, Kim S, Boushey C, Ebert D, Delp E. The use of mobile devices in aiding dietary assessment and evaluation. IEEE Journal of Selected Topics in Signal Processing. Aug; 2010 4(4):756–766. [PubMed: 20862266]
3. Fischler M, Elschlager R. The representation and matching of pictorial structures. IEEE Transactions on Computers. 1973; 1(22):67–92.
4. Fergus, R.; Perona, P.; Zisserman, A. Sparse object category model for efficient learning and exhaustive recognition. Proceedins of the International Conference on Computer Vision and Pattern Recognition (CVPR); San Diego, CA. 2005. p. 380-387.
5. Csurka, G.; Willamowski, DCFLJ.; Bray, C. Visual categorization with bags of keypoints. Statistical Learning in Computer Vision, 2004 International Workshop on; Prague, CZ. 2004.
6. Oliva A, Torralba A. The role of context in object recognition. Trends in Cognitive Sciences. 2007; 12(11):520–527. [PubMed: 18024143]
7. Mariappan, A.; Bosch, M.; Zhu, F.; Boushey, CJ.; Kerr, DA.; Ebert, DS.; Delp, EJ. Personal dietary assessment using mobile devices. Proceedings of the IS&T/SPIE Conference on Computational Imaging VII; San Jose, CA. Jan. 2009
8. Zhu, F.; Bosch, M.; Boushey, C.; Delp, E. An image analysis system for dietary assessment and evaluation. Proceedings of the International Conference on Image Processing, ICIP; Hong-Kong, China. Sep. 2010
9. Yang, S.; Chen, M.; Pomerleau, D.; Sukhankar, R. Food recognition using statistics of pairwise local features. Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR); San Francisco, CA. Jun. 2010
10. Joutou, T.; Yanai, K. A food image recognition system with multiple kernel learning. Proceedings of International Conference on Image Processing (ICIP); Beijing, China. Oct. 2009
11. Kitamura, K.; Yamasaki, T.; Aizawa, K. Foodlog: Capture, analysis and retrieval of personal food images via web. Proceedings of ACM multimedia 2009 workshop on Multimedia for cooking and eating activities; Cairo, Egypt. Nov. 2009
12. Zhu, F.; Bosch, M.; Schap, T.; Khanna, N.; Ebert, D.; Boushey, C.; Delp, E. Segmentation assisted food classification for dietary assessment. Proceedings of the IS&T/SPIE Conference on Computational Imaging IX; San Francisco, CA. Jan. 2011
13. Manjunath, B.; Salembier, P.; Sikora, T. Introduction to MPEG-7: Multimedia Content Description Interface. Wiley & Sons; USA: 2002.
14. Jain A, Farrokhnia. Unsupervised texture segmentation using gabor filters. Pattern Recognition. 1991; 24(12):1676–1186.
15. Bovik A, Clark M, Geisler W. Multichannel texture analysis using localized spatial filters'. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI). 1990; 12:55–73.
16. Tamura H, Mori S, Yamawaki T. Textural features corresponding to visual perception. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI). 1978; 8:530–534.
17. Lowe D. Distinctive image features from scale-invariant key- points. International Journal on Computer Vision. 2004; 2(60):91–110.
18. Bay H, Ess A, Tuytelaars T, Gool LV. SURF: Speeded up robust features. Computer Vision and Image Understanding (CVIU). 2008; 110(3):346–359.
19. Freeman W, Adelson Y. The design and use of steerable filters. IEEE Transactions on Systems Man Cybernetics. 1978:460–473.
20. Tola E, Lepetit V, Fua P. DAISY: An efficient dense descriptor applied to wide-baseline stereo. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI). 2005; 32(5):815–830.
21. Vapnik, V. The nature of statistical learning theory. Springer-Verlag; New York, NY: 1995.
22. Chang, CC.; Lin, CJ. LIBSVM: a library for support vector machines. 2001. software available at http://www.csie.ntu.edu.tw/cjlin/libsvm

23. Zhang J, Marszalek M, Lazebnik S, Schmid C. Local features and kernels for classification of texture and object categories: a comprehensive study. International Journal of Computer Vision. 2007; 2(73):213–238.
24. Bosch, M.; Schap, T.; Khanna, N.; Zhu, F.; Boushey, CJ.; Delp, EJ. Integrated databases system for mobile dietary assessment and analysis. Proceedings of the 1st IEEE International Workshop on Multi-media Services and Technologies for E-health in conjunction with the International Conference on Multimedia and Expo; Barcelona, Spain. Jul. 2011

NIH-PA Author Manuscript
NIH-PA Author Manuscript

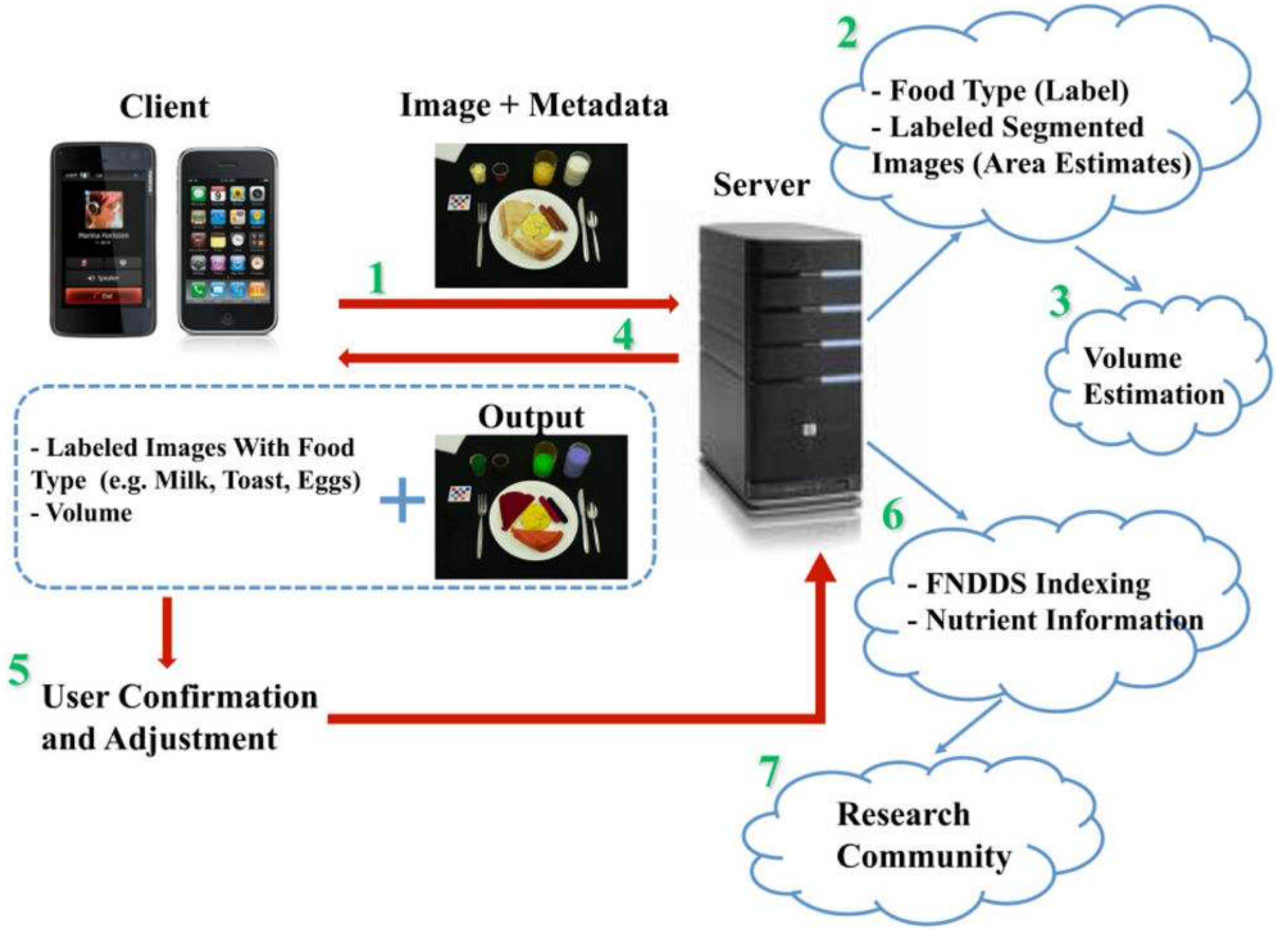

Fig. 1.
Overall System for Dietary Assessment.

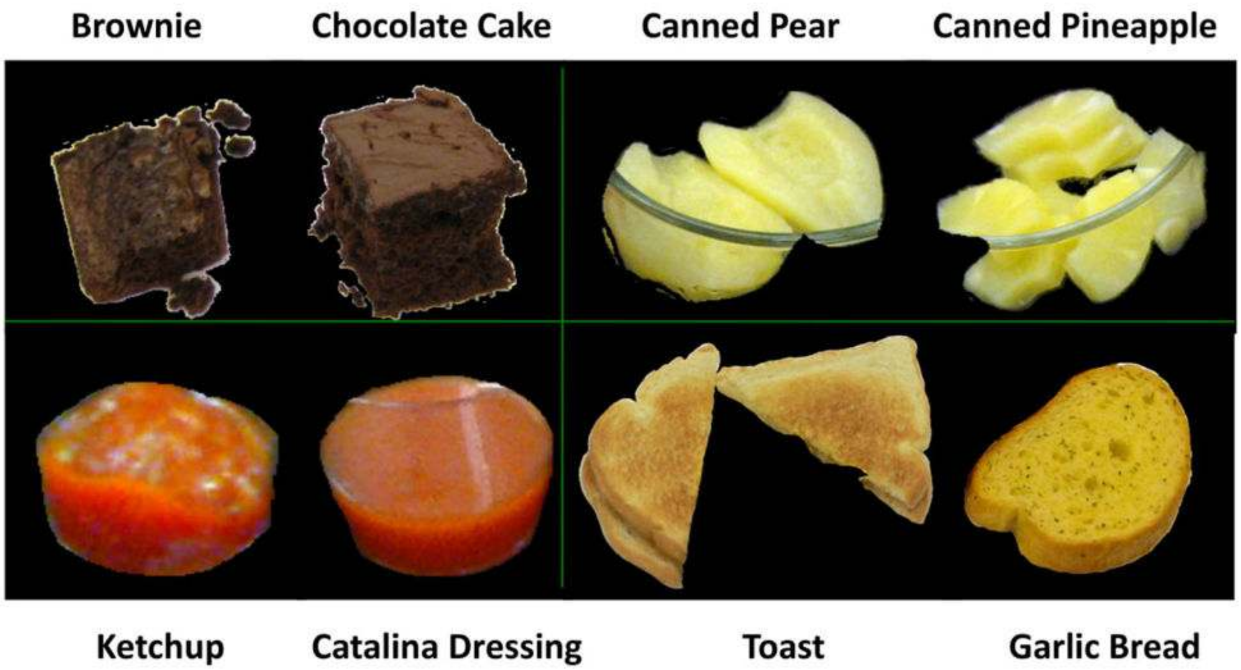

Fig. 2.
Examples of Misclassified Food Items.

Table 1

Correct classification accuracy averaged for all categories for each type of feature independently and combined. (Using half of the data for training and half for testing. Experiments were repeated 10 times with random selection of training and testing data. L indicates local feature and the dimension is the length of the feature vector per point of interest. G means global feature and the dimension is the length of the feature vector for the entire object.)

Feature type	Dimension	Dataset 1	Dataset 2	Datasets 1 and 2
Color (L)	20/point	0.861	0.845	0.792
Entropy (L)	6/point	0.313	0.279	0.262
Tamura (L)	3/point	0.472	0.452	0.411
Gabor (L)	48/point	0.326	0.311	0.291
SIFT (L)	128/point	0.714	0.702	0.652
Haar Wavelet (L)	64/point	0.687	0.675	0.641
Steerable (L)	50/point	0.548	0.532	0.517
DAISY (L)	200/point	0.694	0.678	0.603
Color (G)	20/object	0.826	0.817	0.786
Predominant (G)	20/object	0.674	0.639	0.606
Entropy (G)	6/object	0.867	0.858	0.782
Gabor (G)	48/object	0.484	0.465	0.402
Final		**0.981**	**0.972**	**0.861**