Published in final edited form as:

Proc (IEEE Int Conf Multimed Expo). 2013 July ; 2013: .

FOOD IMAGE ANALYSIS: SEGMENTATION, IDENTIFICATION AND WEIGHT ESTIMATION

Ye He[1], **Chang Xu**[1], **Nitin Khanna**[2], **Carol J. Boushey**[3], and **Edward J. Delp**[1]

[1]School of Electrical and Computer Engineering, Purdue University

[2]Department of Electronics and Communication Engineering, Graphic Era University

[3]Cancer Epidemiology Program, University of Hawaii Cancer Center

Abstract

We are developing a dietary assessment system that records daily food intake through the use of food images taken at a meal. The food images are then analyzed to extract the nutrient content in the food. In this paper, we describe the image analysis tools to determine the regions where a particular food is located (image segmentation), identify the food type (feature classification) and estimate the weight of the food item (weight estimation). An image segmentation and classification system is proposed to improve the food segmentation and identification accuracy. We then estimate the weight of food to extract the nutrient content from a single image using a shape template for foods with regular shapes and area-based weight estimation for foods with irregular shapes.

Index Terms

Dietary Assessment; Image Segmentation; Object Identification; Weight Estimation

1. INTRODUCTION

There is a health crisis in the US related to diet that is further exacerbated by our aging population and sedentary lifestyles. Dietary assessment, the process of determining what someone eats during the course of a day, is essential for understanding the link between diet and health. Preliminary studies have shown that mobile telephones with built-in digital cameras and network connectivity provide unique mechanisms for improving the accuracy and reliability of dietary assessment [1]. Previous research advances have been made in this area including plate detection [2], food image enhancement [3, 4] and volume estimation with a laser attachment [5]. We are developing an image analysis system to estimate the foods consumed at an eating occasion from food images acquired by mobile telephones. The system is designed to be easy to use and not place a burden on users by having to acquire multiple images or a video of the food. Thus our goal is to automatically identify foods and estimate the weight of foods from a single image (see Figure 1).

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

In our approach, once a food image is acquired, image segmentation methods are used to locate object boundaries for food items in the image. We propose an integrated image segmentation and identification method which is similar to [6] in the way that we iteratively use the identification feedback to refine the segmentation and identification results, but our method is not based on the assumption of particular food location or a uniform color tablecloth. After food items are segmented and identified, it is important to accurately estimate the weight of each food item in order to determine the nutrient content. For foods with regular shapes, we use food specific shape templates to estimate the volume similar to the approach described in [7]. Once the volume is estimated the nutrient content is obtained using the density information for that particular food type [8]. For foods which do not conform to a regular shape (e.g. scrambled eggs) or foods which have large variations in shape due to eating or food preparation conditions (e.g. cut carrots in a salad), we investigate a direct area-based weight estimation method.

2. FOOD SEGMENTATION AND IDENTIFICATION

There are three main components of our food segmentation and identification approach: image segmentation, feature extraction and classification, and segmentation refinement (see Figure 2). Traditional segmentation methods usually require user-specified input parameters which can result in under-segmentation or over-segmentation. To obtain a stable segmentation, we propose to combine image segmentation and classification using a segmentation refinement step in which the classification feedback can be used to refine image segmentation until maximal classification confidence has been achieved.

2.1. Image Segmentation

Several popular image segmentation methods have been evaluated in [9] based on 300 images from the Berkeley Segmentation Database by estimating precision and recall with regard to human ground-truth boundaries. The ranges for the input parameters of each segmentation method were determined experimentally so as to produce segmentations that go from under-segmented to over-segmented. According to the experimental results in [9] we adopt the framework of Local Variation [10] to generate a list of initial segmentations in our system because this segmentation method is more stable to the change of input parameters. For example, from under-segmentation to over-segmentation, the input parameter range of Normalized Cuts [11] is within [2, 128] while the input parameter ranges of Local Variation [10] is within [10, 1800].

Local Variation [10] is a graph based image segmentation method with weight on each edge measuring the dissimilarity between pixels. This method segments the image based on the degree of variability in neighboring regions of the image. The internal difference of a segmented region is defined to be the largest weight in the minimum spanning tree, MST(A, E), of the region:

$$\mathrm{Int}(A) = \max_{e \in MST(A,E)} w(e) \quad (1)$$

where $w(e)$ is the weight of an edge e. The difference between two segmented regions is defined to be the minimum weight edge connecting the two regions:

$$\mathrm{Dif}(A,B) = \min_{p \in A, q \in B, (p,q) \in E} w(p,q) \qquad (2)$$

Two regions are segmented if the difference between the two regions $Dif(A, B)$ 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 k:

$$\mathrm{minInt}(A,B) = \min(\mathrm{Int}(A) + \frac{k}{|A|}, \mathrm{Int}(B) + \frac{k}{|B|}) \qquad (3)$$

where $|A|$ denotes the size of A.

The input parameter k roughly controls the size of the regions in the resulting segmentation. Smaller values of k yield smaller regions and favor over-segmentation. In our implementation we use $k = 150$. Examples of initial segmentation results using Local Variation are presented in Figure 5 (b).

2.2. Food Classification

Food classification is particularly difficult because foods can dramatically vary in appearance such as shape, texture, color and other visual properties. An essential step in solving the food classification problem is to adequately represent the visual information of foods. This is commonly known as feature extraction. In our implementation, we extract color and texture features for food classification. Four color descriptors namely Scalable Color Descriptor (SCD), Color Structure Descriptor (CSD), Dominant Color Descriptor (DCD) and Color Layout Descriptor (CLD) in MPEG-7 standard [12] are used in our system.

SCD is a color histogram descriptor in HSV Color Space with a uniform quantization of the HSV space to 256 bins, which includes 16 levels in H, 4 levels in S, and 4 levels in V. CSD expresses local color structure in HMMD color space using an 8 × 8 grid scanning the image. Suppose the image contains K structuring elements (s_0, s_1, …, s_{K-1}) and N quantized colors (c_0, c_1, …, c_{N-1}). Then if $c_n \in s_k$, $x_{n,k} = 1$; otherwise $x_{n,k} = 0$. A color structure histogram can be denoted by $h(n)$, $n = 0, 1, \ldots, N - 1$ where the value in each bin $h(n) = \sum_{k=0}^{K-1} x_{n,k}$. DCD uses color clustering to extract a small number of representing colors and their percentages from a segmented region in the perceptually uniform CIE LUV color space. CLD is used to capture the spatial distribution of color in a segmented region. The segmented region is divided into small blocks. The average color of each block in YCrCb color space is calculated to form the descriptor.

Texture, similar to color, is a very descriptive low-level feature for image search and matching applications. In our system, we used the following three texture descriptors [13] for food classification:Gradient Orientation Spatical-Dependence Matrix (GOSDM), Entropy-Based Categorization and Fractal Dimension Estimation (EFD) and Gabor-Based Image Decomposition and Fractal Dimension Estimation (GFD).

GOSDM consists of a set of gradient orientation spatial dependence matrices to describe the texture by the occurrence rate of the spatial relationship of gradient orientations for different neighborhood size. EFD can be seen as an attempt to characterize the variation of roughness of homogeneous parts of the texture in terms of complexity. In general regions of the image corresponding to high complexity (high level of detail) tend to have higher entropy, thus entropy can be seen as a measure of local signal complexity. Once the entropy is estimated for pixels in the texture image, the regions with similar entropy values are clustered to form a point categorization. The fractal dimension descriptor is, then, estimated for every point set according to this categorization. GFD is also based on fractal dimension. Instead of using entropy categorization, the image is decomposed into sub-images in its spatial frequency dimension using Gabor filter-bank which consists of a set of Gabor filters. The fractal dimension is estimated for each filtered response.

In our system, each segmented region is regarded as a stand-alone image by masking and zero padding the original image. After extracting color and texture features from a segmented region, we assign a category label to the segment based on a majority vote rule of the nearest neighbors. Let I_i be the training image index and c_i be the category index of the i −th training image. Let ϕ be the feature space and $s_0, s_1, \ldots, s_{N-1}$ be the set of segmented objects in the test image. For each s_n in the feature channel ϕ_f, we find K nearest neighbors of s_n among all the training images using the normalized L1 norm distance (sum of absolute differences). Denote the K nearest distances as:

$$d_{0,\ldots,K-1}(s_n, i, f) = \min_{i}{}_{0,\ldots,K-1} \|\phi_f(s_n) - \phi_f(I_i)\| \qquad (4)$$

$d_0(s_n, i)$ is the minimum distance of the test segment s_n to all the training images and the i-th image is the best match of s_n in the feature space ϕ_f. In order to combine different feature channels and find the best matching category, we introduce the following category labeling confidence score of the segment s_n in the m−th category:

$$p(s_n, m) = \sum_{f=0}^{4} \sum_{k=0}^{K-1} \sum_{c_i=m} \frac{d_0(s_n, i, f)}{d_k(s_n, i, f)} \qquad (5)$$

In our system, each segment is assigned to 4 food categories with the top 4 largest category labeling confidence score. The original image is sent back to the participant to confirm the top food category, select from the next three, or designate a food category from a larger list of choices.

2.3. Integrating Food Segmentation and Classification

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 (see Figure 3). We seek to overcome the segmentation problem by using the confidence scores iteratively to refine the segmentation results. To detect under-segmentation, we first scan all the segments ($s_0, s_1, \ldots, s_{N-1}$) in the image to filter out small segments. In our implementation, we define "small segments" as segments that contain less than 1/50 pixels of the original image. Each remaining segment is iteratively segmented and classified until the maximum confidence score has achieved. If the food category labeling 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 renew the label of segments as $s_0, s_1, \ldots, s_q$ and the corresponding food category label as $c_{0,0}, c_{0,1}, \ldots, c_{0,K-1}, \ldots, c_{q,0}, c_{q,1}, \ldots, c_{q,K-1}$. For each adjacent pair of segments (s_i, s_j), if a food category label $c_{i,m}$ in the set $c_{i,0}, c_{i,1}, \ldots, c_{i,K-1}$ equals to a food category label $c_{j,n}$ in the set $c_{j,0}, c_{j,1}, \ldots, c_{j,K-1}$, the corresponding food category labeling confidence scores $p_{i,m}$ and $p_{j,n}$ are added to form a new confidence score p. If $p > p_{i,0}$ and $p > p_{j,0}$, we combine these two segments s_i and s_j 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 maximum overall confidence score is achieved. An example of segmentation refinement is shown in Figure 4.

After under-segmentation and over-segmentation examination, we still have redundant segments, such as segments in checkerboard area and segments in background area. We deal with these redundant segments using a fast rejection step. The segments in the checkerboard and image boundary area are filtered out. The checkerboard is designed for color correction and image calibration. Besides food items, our training dataset also has a "non-food" category which contains the training segments of non-food items such as plates, glasses, and forks. Any segment that falls into this category will be filtered out. For the final step, we filter out the segments with low food category labeling confidence scores. Examples of the final image segmentation results are shown in Figure 5.

Our proposed integrated image segmentation and classification method was tested on 1453 food images with 96 unique food items. The food images were acquired by 45 participants in natural eating conditions (also referred to "free-living" or "community-dwelling" such as home and on the go). The food identification accuracy is defined as:

$$\text{accuracy}=\frac{TP}{TP+\frac{FP}{K}+TN} \qquad (6)$$

where TP indicates True Positives (correctly detected food segments); FP indicates False Positives (incorrectly detected food segments or misidentified foods); TN indicates True Negatives (food not detected). Finally, K refers to the identification accuracy order. If we are interested in knowing the identification accuracy using the top 4 outputs (food category labels) of the classifier, K is set to 4. The identification accuracy of all the images using the top 1 and top 4 food category outputs from the classifiers is 34%and 63%respectively. Some foods (e.g. yogurt, water) in our free-living images are difficult to identify because they are served in non-transparent containers; some others (e.g. coke and coffee) are inherently difficult to identify due to their visual similarity in the feature space. We hope to further improve the food identification accuracy by exploring contextual information in addition to visual characteristics.

3. FOOD WEIGHT ESTIMATION

Estimating the weight of food is a critical component in both clinical and research dietary studies. In our system, after foods are segmented and identified, we estimate the weight in order to determine the nutrient content. In general, there are two ways to estimate the weight of foods from a food image. One way is to estimate the volume of the food, then use the density information [8] for that particular food type to estimate weight. The other way is to directly estimate the weight of food using area and training data. In our system, we use the shape template 3D reconstruction method for foods with regular shapes described in [7], and an area-based weight estimation method for foods which do not conform to a regular shape or foods which have large variations in shape.

3.1. Volume Estimation

After camera calibration and obtaining the camera pose information, we implemented a food specific shape template method to reconstruct a 3D model of the food item [7]. Two inputs are used for this approach, the segmentation mask that indicates the location of each food item and the food label obtained from food identification. We obtain the specific shape template for each food item based on its food label. For example, we use a sphere shape to approximate the 3D shape of an apple or an orange and a cylinder shape for a liquid. We then detect the feature/corner points from the segmented region of each food item. We define the feature points as the dominant points or corner points which can be used to compute the geometric information of the shape template. For example, the feature points of a cylinder shape template are the top-left, top-right, bottom-left, bottom-right corners from the segmentation mask. These points can be used to estimate the top radius, bottom radius and the height of the cylinder shape. The feature points are then back projected onto 3D world coordinate using the camera matrix P obtained from the camera calibration as shown in Equation 7. In Equation 7, we represent the points x_i on the image plane as $(u, \upsilon, 1)^T$ and the 3×4 homogeneous camera projection matrix P is equal to $C[R|T]$. We also denote the 3D world coordinate points as $(X_w, Y_w, Z_w, 1)^T$. The origin of the world coordinate is located at one of the checkerboard corners. Then, we use the assumption that all the bottom points which are adjacent to the table have $Z_w = 0$. We also use other assumptions such as the top point of a cylinder shaped object has the same X_w, Y_w values with the bottom point. Using these assumptions, we obtain the 3D coordinate of all the feature points. Finally, we use the coordinates of the feature points in 3D world coordinate to obtain the food volume of the regularly shaped object. By using our pre-defined shape templates, we are able to estimate the food portion size. Based on the experiments for beverage images using cylinder shape template, the average relative error was about 11% [7]. This result shows that estimation of the food volume from single image can be accomplished using the computational solution.

However, further study is needed to expand our method as to improve the accuracy of volume estimation for more food items.

$$\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = C[R|T] \begin{pmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{pmatrix} \quad (7)$$

3.2. Area-Based Weight Estimation

When a food image is acquired, the distance, direction and pose of the camera are not restricted, thus the food image is usually geometrically distorted. To accurately estimate the food area, we convert a geometric distorted image to a front view image. We model the camera as a perspective camera which is commonly utilized in photography. Projective geometry models the geometric distortion which arises when a plane is imaged by a perspective camera. Twelve corner points are detected on the checkerboard and labeled as $p_{src}^1, p_{src}^2 \ldots p_{src}^{12}$ in the following order: top to bottom, left to right. Each corner point is represented using their homogeneous coordinates $(x^i, y^i, 1)^T$, where i ranges from 1 to 12. We define the 3 × 3 projective transformation matrix H as a mapping from the corner points in the distorted image to the undistorted image.

$$p_{dst} = Hp_{src} \quad (8)$$

where p_{dst} is the corresponding corner points in the distortion free image. The transformation matrix H has nine elements with only their ratio significant, so the transformation is specified by 8 parameters. We use the Direct Linear Transformation (DLT) method [14] to estimate H using the corner points (Equation 9).

$$\begin{pmatrix} 0^T & -(p_{src}^i)^T & y_i^{'}(p_{src}^i)^T \\ (p_{src}^i)^T & 0^T & -x_i^{'}(p_{src}^i)^T \end{pmatrix} \begin{pmatrix} h^1 \\ h^2 \\ h^3 \end{pmatrix} = 0 \quad (9)$$

where p_{dst}^i is denoted as $(x_i^{'}, y_i^{'}, 1)^T$; h^j is the i-th row of H. Figure 6 shows an example of the original image and the corresponding corrected front view image after applying H to remove both perspective distortion and affine distortion in the original image. The area of the segmented food item is measured in the corrected image as the ratio between the segmented food area and the area of one color block in the checkerboard (fiducial marker).

After the food in the image is identified and the area is estimated, we estimate the weight of the food using the area-weight relation in the training data of the corresponding food item. Linear interpolation of the N nearest areas in the training data is used to estimate the weight of the requested food item.

We use images of fried razor clams in the experiment to evaluate the accuracy of our area-based weight estimation method. We choose to use fried razor clams mainly because they can have large variations in shape and size due to different food preparation conditions or different number of razor clams on the plate. Figure 6 shows sample images of fried razor clams used in our experiment. We have acquired 143 images in total and divided them evenly into 13 data sets. Weight of the food items in the images ranged from 20 gram to 320 gram. Each data set contains the images of fried razor clams with the same appearance, size and weight, but the images are taken in different distance or direction. Images in different data sets vary in appearance, size and weight. We randomly selected some of the data sets as training data and others as testing data to test the accuracy of our area-based weight estimation method. The estimation error is obtained using Equation 10.

$$e=\frac{1}{N}\sum_{i=1}^{N}\frac{|w_i - w_g|}{w_g} \qquad (10)$$

where N is the number food items in testing images; w_i is the estimated weight of food item i; w_g is the ground-truth weight of the corresponding food item. The weight estimation error reduces to about 10% as the number of training data sets increases which is reasonable comparing to the 11% error of the shape templates method [7]. The area-based weight estimation method is used as a simple approach when the foods do not conform to a regular shape or have large variations in shape due to eating or food preparation conditions. We investigate to further improve our approach using geometric constraints and context information.

4. CONCLUSION

In this paper we proposed an integrated image segmentation and identification system to determine the regions in an image where food items are located and identify the food category. After food segmentation and identification, we estimated the weight of food to extract the nutrient content from a single image using shape template for foods with regular shapes and area-based weight estimation for foods with irregular shapes.

Acknowledgments

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

REFERENCES

1. Boushey CJ, Kerr DA, Wright J, Lutes KD, Ebert DS, Delp EJ. Use of technology in children's dietary assessment. European Journal of Clinical Nutrition. 2009 Feb; 63(Suppl 1):S50–S57. [PubMed: 19190645]
2. Jia, W., Yue, Y., Fernstrom, J., Zhang, Z., Yang, Y., Sun, M. 3d localization of circular feature in 2d image and application to food volume estimation. Proceedings of Annual International Conference of the IEEE Engineering in Medicine and Biology Society; August 2012; San Diego, CA. p. 4545-4548.

3. Xu, C., Zhu, F., Khanna, N., Boushey, C., Delp, E. Image enhancement and quality measures for dietary assessment using mobile devices. Proceedings of the IS&T/SPIE Conference on Computational Imaging X; February 2012; San Francisco, CA. p. 82 960Q-82 910Q.
4. He, Y., Khanna, N., Boushey, C., Delp, E. Specular highlight removal for image-based dietary assessment. Proceedings of the 2012 IEEE International Conference on Multimedia and Expo Workshops; Australia, July 2012; Melbourne. p. 424-428.
5. Shang, J., Duong, M., Pepin, E., Zhang, X., Sandara-Rajan, K., Mamishev, A., Kristal, A. A mobile structured light system for food volume estimation. Proceedings of the 2011 IEEE International Conference on Computer Vision Workshops; November 2011; Barcelona, Spain. p. 100-101.
6. Zhu, F., Bosch, M., Khanna, N., Boushey, C., Delp, E. Multilevel segmentation for food classification in dietary assessment. Proceedings of the 7th International Symposium on Image and Signal Processing and Analysis; September 2011; Dubrovnik, Croatia. p. 337-342.
7. Chae, J., Woo, I., Kim, S., Maciejewski, R., Zhu, F., Delp, E., Boushey, C., Ebert, D. Volume estimation using food specific shape templates in mobile image-based dietary assessment. Proceedings of the IS&T/SPIE Conference on Computational Imaging IX; February 2011; San Francisco, CA. p. 78730K
8. Kelkar S, Stella S, Boushey C, Okos M. Developing novel 3d measurement techniques and prediction method for food density determination. Procedia Food Science. 2011; 1:483–491.
9. Estrada FJ, Jepson AD. Benchmarking image segmentation algorithms. International Journal of Computer Vision. 2009 Nov; 85(2):167–181.
10. Felzenszwalb, P., Huttenlocher, D. Image segmentation using local variation. Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition; Jun 1998; Washington, DC, USA. p. 98-104.
11. Shi J, Malik J. Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2000 Aug; 22(8):888–905.
12. Manjunath B, Ohm J-R, Vasudevan V, Yamada A. Color and texture descriptors. IEEE Transactions on Circuits and Systems for Video Technology. 2001 Jun; 11(6):703–715.
13. Bosch, M., Zhu, F., Khanna, N., Boushey, C., Delp, E. Combining global and local features for food identification and dietary assessment. Proceedings of the International Conference on Image Processing; September 2011; Brussels, Belgium. p. 1789-1792.
14. Hartley, RI., Zisserman, A., et al. Multiple View Geometry in Computer Vision. 2nd. Cambridge University Press; 2004.

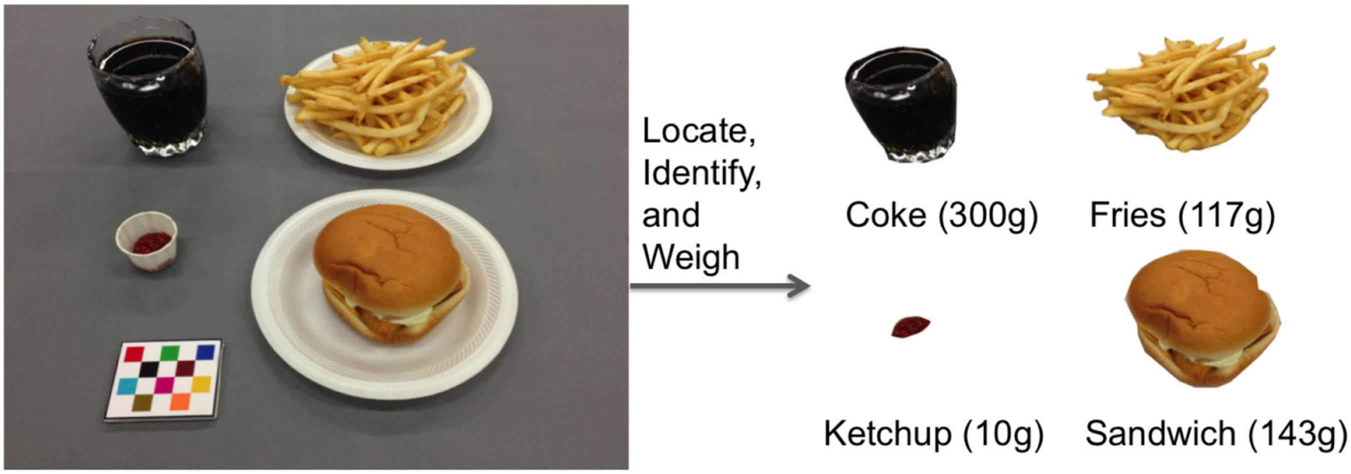

Fig. 1.
An ideal food image analysis system for dietary assessment.

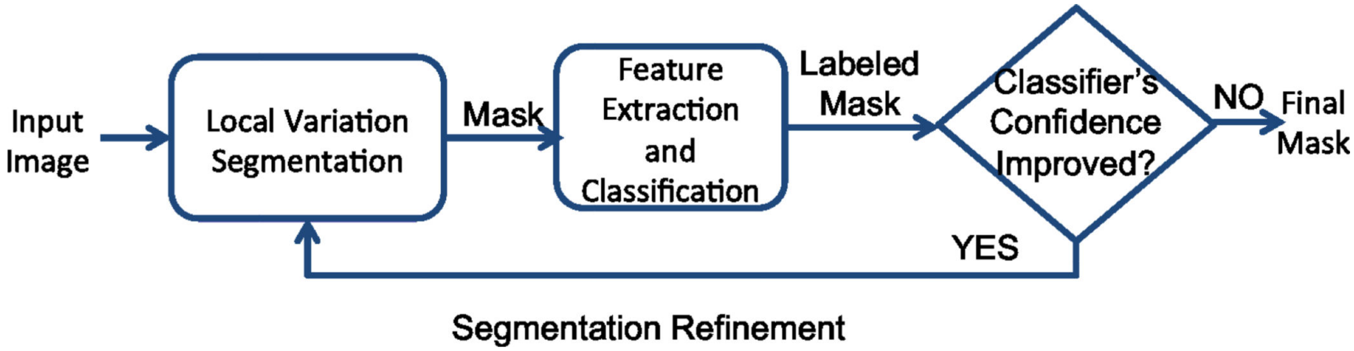
Input Image
Local Variation Segmentation
Mask
Feature Extraction and Classification
Labeled Mask
Classifier's Confidence Improved?
NO
Final Mask
YES
Segmentation Refinement

Fig. 2.
Our food location and identification system.

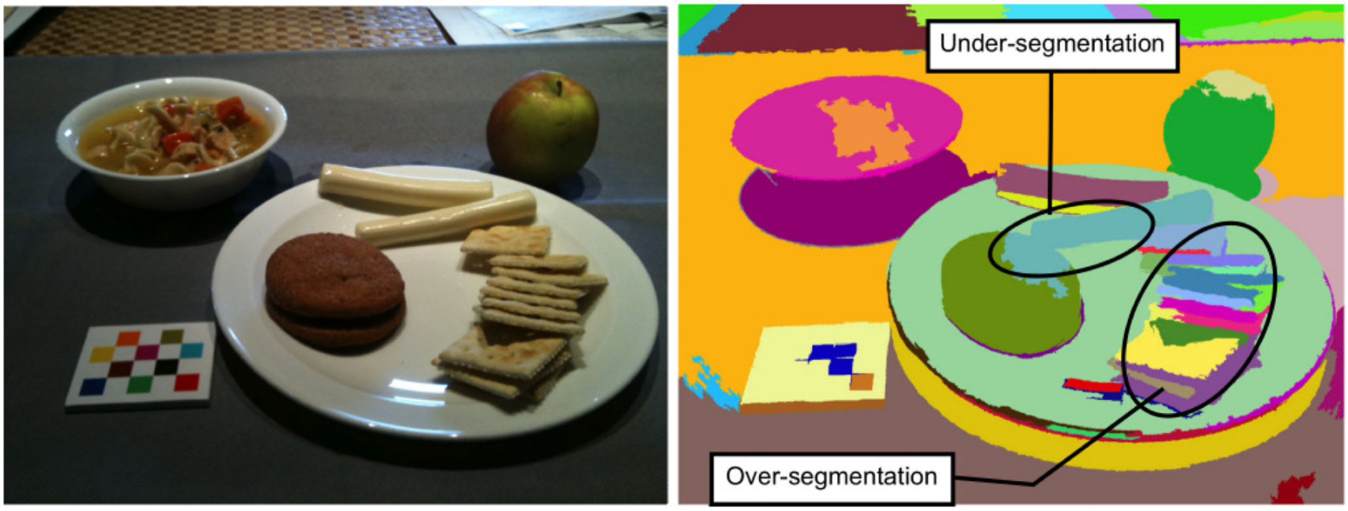

Fig. 3.

An example of under-segmentation and over-segmentation. Left column shows the original image; Right column shows the initial segmentation result.

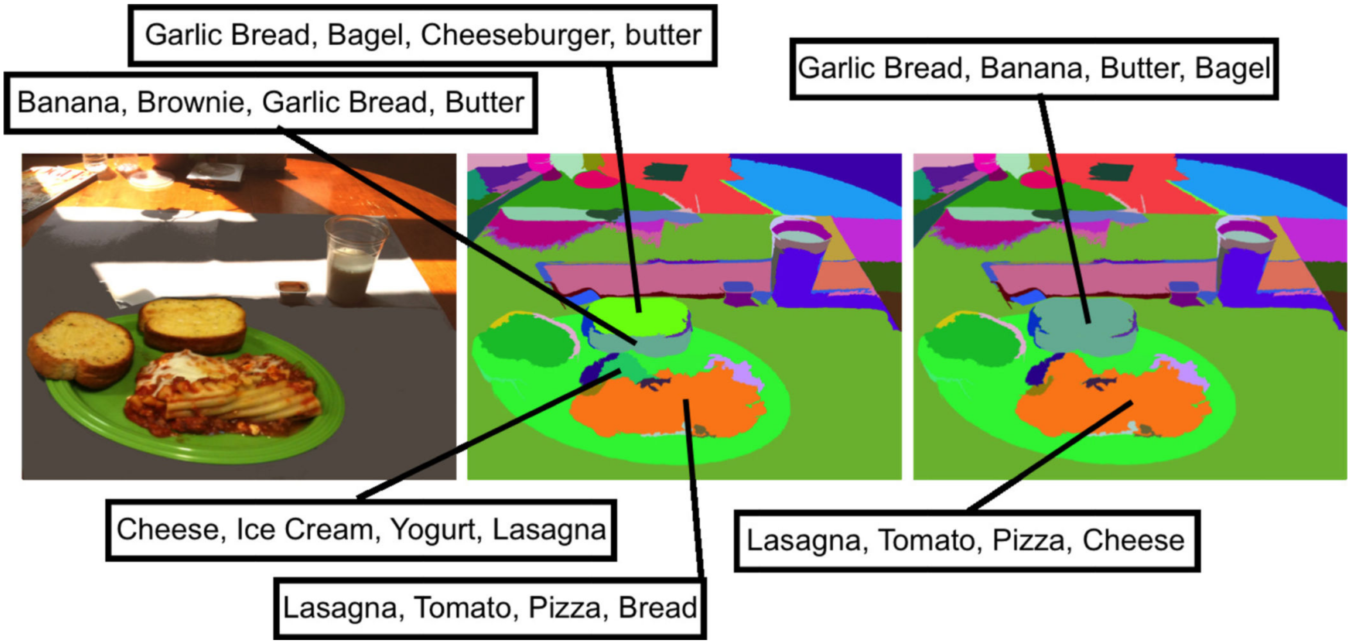
Garlic Bread, Bagel, Cheeseburger, butter
Banana, Brownie, Garlic Bread, Butter
Garlic Bread, Banana, Butter, Bagel
Cheese, Ice Cream, Yogurt, Lasagna
Lasagna, Tomato, Pizza, Cheese
Lasagna, Tomato, Pizza, Bread

Fig. 4.

An example of segmentation refinement. Left column shows the original image; Middle column shows the initial segmentation result; Right column shows the segmentation refinement result.

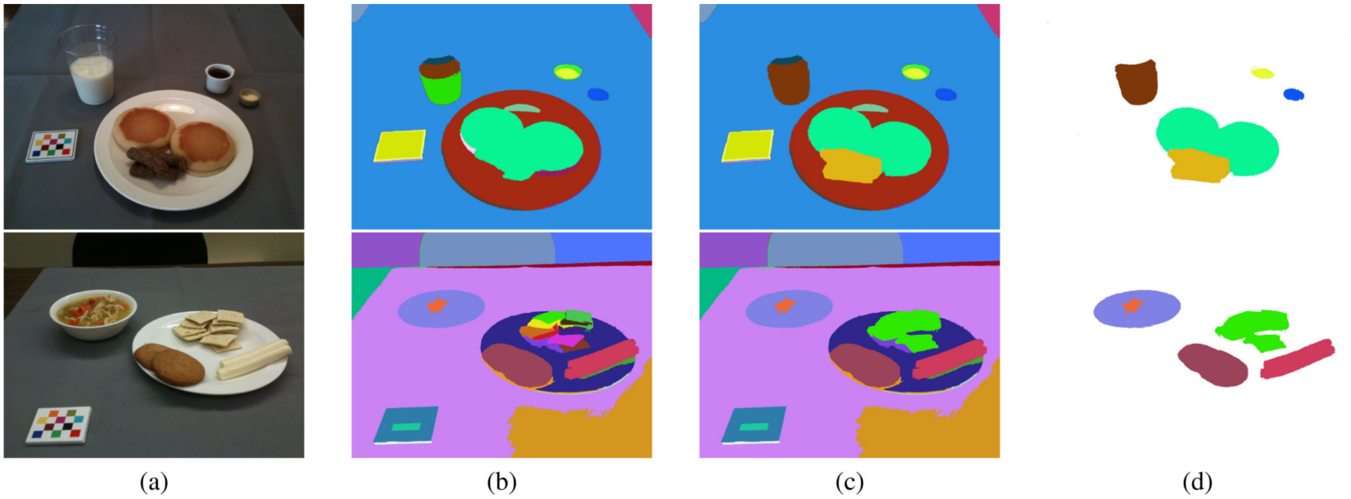

Fig. 5.

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

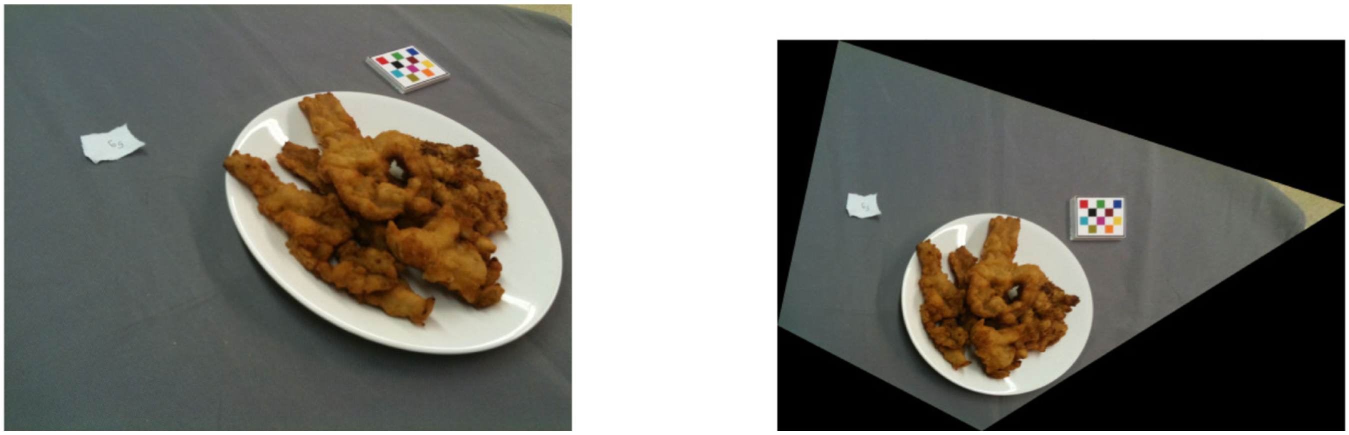

Fig. 6.

An example of image transformation: the original image (left) and the corresponding corrected front view image (right).