Multilevel Segmentation for Food Classification in Dietary Assessment

Fengqing Zhu*, Marc Bosch*, Nitin Khanna*, Carol J. Boushey† and Edward J. Delp*
*School of Electrical and Computer Engineering
†Department of Foods and Nutrition
Purdue University
West Lafayette, Indiana, USA

***Abstract*—Given a dataset of images, we seek to automatically identify and locate perceptually similar objects. We combine two ideas to achieve this: a set of segmented objects can be partitioned into perceptually similar object classes based on global and local features; and perceptually similar object classes can be used to assess the accuracy of image segmentation. These ideas are implemented by generating multiple segmentations of each image and then learning the object class by combining different segmentations to generate optimal segmentation. We demonstrate that the proposed method can be used as part of a new dietary assessment tool to automatically identify and locate the foods in a variety of food images captured during different user studies.**

I. Introduction

Assigning predefined class labels to every pixel in an image is a highly unconstrained problem. Human vision system has the remarkable abilities to group pixels of an image into object segments without knowing *apriori* which objects are present in that image. Designing well-behaved models capable of making more informed decisions using increased spatial support is an open problem for segmentation and classification systems. It is necessary to work at different spatial scale on segments that can model either the entire objects, or at least sufficiently distinct parts of them. Recent developments in this area have shown promising results. In [1], the authors present a framework for generating and ranking plausible objects hypotheses by solving a sequence of constrained parametric min-cut problems and ranking the object hypotheses based on mid-level properties. A multiple hypothesis framework is proposed in [2] for robust estimation of scene structure from a single image and obtaining confidences for each geometric label. Sivic et. al. [3] use a probability latent semantic analysis model to discover the object categories depicted in a set of unlabeled images. The model is applied to images using vector quantization on SIFT-like region descriptors.

We are interested in developing methods to locate and identify perceptually similar food objects for dietary assessment applications. Nutritional epidemiology is concerned with quantifying dietary exposures and estimating the association of these exposures with risks for disease. Diet represents one of the most universal biological exposures; however accurate assessment of food and beverage intakes is problematic [4]. Our research focuses on developing a novel food record method using a mobile device that will provide an accurate account of daily food and nutrient intake [6]. Our goal is to identify food items using a single image acquired from the mobile device. An example of this is shown in Figure 1, where each food item is segmented and identified. The system must be easy to use and not place a burden on the user by having to take multiple images, carry another device or attach other sensors to their mobile device. A prototype system has been deployed on the Apple iPhone and its functionality has been verified with various combinations of foods. To aid with interactive design, the application has been tested by adolescents ages 11-18y and adults from various age groups in controlled meal sessions [7].

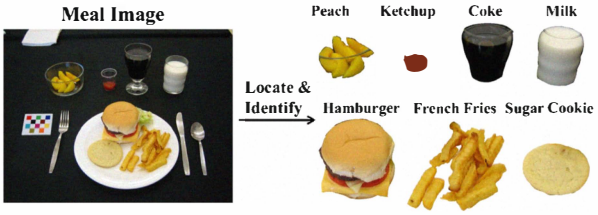

Fig. 1. An Ideal Food Image Analysis System.

The paper is organized as follows. The use of multiple segmentation hypothesis is proposed in Section 2. Methods to perform feature extraction and classification of segments are explained in Section 3. Criterion for assigning segment label and parameter selection are described in Section 4. Experimental results are presented in Section 5. Finally, we conclude with discussion of our proposed methods in Section 6.

II. Multiple Segmentation Hypothesis

Given a large, unlabeled collection of images, for each pixel in the test images, our goal is to predict the class of the object containing that pixel or declare it as "background" if the pixel does not belong to any of the specified classes. The output is a labeled image with each pixel label indicating the inferred class. We exploit the fact that segmentation methods are not stable as one perturbs their parameters, thus obtaining a variety of different segmentations. Many segmentation methods such as Normalized Cuts [8], take number of segments as one of the input parameters of the segmentation method. Since, the exact number of segments in a test image is not known a

priori, a particular choice of number of segments results in either under segmented or over segmented image. Further, for a particular choice of number of segments, some objects may be under segmented, while others may be over segmented. That is, some of the segments may contain pixels from more than one class while more than one segments may correspond to a single class. The probabilistic classifier used in our system, for classification of these segments, gives K most probable candidate classes along with their probability estimates. Consequently, probability estimates achieve smaller values for over-segmented or under-segmented classes. To overcome these problems associated with unknown number of segments in test images, we aim to generate a pool of segments for each input image to achieve high probability of obtaining "good" segments that may contain potential objects. Since we are not relying on any particular segmentation to be correct, the choice of segmentation method is not critical [5].

A. Salient Region Detection

The proposed segmentation method includes an initial step to identify regions of interest. Unique to our application, we are interested in regions of an image containing food objects. The region of interest detection is useful to our task of assigning correct label to each pixel by rejecting non-food objects such as tablecloth, utensils, napkins, etc. and thus reducing the number of pixels to be processed.

Knowledge-based methods are used to determine these regions of interest. First step is to remove the background pixels from our search space. The images from our user studies contain uniformly colored tablecloths, therefore, we can generate a foreground-background image by labeling the most frequently occurring color in the CIE L*a*b* color space as the background pixel color. Another foreground-background image is formed by identifying strong edges present in each RGB channel of the image. In particular, we use a canny operator to extract the edges. We combine these foreground-background images and remove undesired noise, such as holes, gaps, and bulges with mathematical morphology. We then label connected components in the binary image. Since food items are generally located in a plate, bowl, or glass that have distinctive shapes, our goal is to detect these objects. We first remove known non-food objects such as the fiducial marker, currently a checkerboard pattern, used as both a geometric reference and color reference. To determine which components contain potential food items, we use the Sobel edge filter on each component and plot the normalized edge histogram. The criteria for identifying components that contain food objects is the uniformity of the edge histograms. We compute the euclidean distance between the normalized edge histogram of each salient region and a uniform distribution. Based on this criteria, a threshold is selected to determine salient region.

B. Multiscale Segmentation

By using multiscale segmentations, recent works have achieved promising segmentation results under non-trivial conditions. In [9], the authors use an algebraic multigrid to find an appropriate solution to the normalized cut measures and apply a process of recursive coarsening to produce an irregular pyramid encoding region based grouping cues. Another method, proposed in [10] constructs multiscale edge defining pairwise pixel affinity at multiple grids. Simultaneous segmentation through all graph levels is evaluated based on the average cuts criterion.

We adopted the approach proposed in [11], where multiple scales of the image are processed in parallel without iteration to capture both coarse and fine level details. The approach uses the Normalized Cut [8] graph partitioning framework. In the Normalized Cuts framework, segmentation quality depends on the pairwise pixel affinity graph, a larger graph radius generally makes the segmentation better. However, the advantage of graphs with long connections comes with a great computational cost. If implemented naively, segmentation on a fully connect graph G of size N would require at least $O(N^2)$ operations. Therefore, the ideal graph connection radius is a trade off between the computation cost and segmentation result.

In the Normalized Cut method, an image is modeled as a weighted, undirected graph. Each pixel is a node in the graph with an edge formed between every pair of pixels. The weight of an edge is a measure of the similarity between the two pixels, denoted as $W_I(i,j)$. The image is partitioned into disjoint sets by removing the edges connecting the segments. The optimal partitioning of the graph is the one that minimizes the weights of the edges that were removed (the cut). The method in [8] seeks to minimize the Normalized Cut, which is the ratio of the cut to all of the edges in the set. Two simple yet effective local groups cues are used to encode the pairwise pixel affinity graph. Since close-by pixels with similar intensity value are likely to belong to the same object, we can represent such affinity by:

$$W_I(i,j) = \exp\left[-\left(\frac{\|I_i - I_j\|_2^2}{\sigma_I^2} + \frac{\|X_i - X_j\|_2^2}{\sigma_X^2}\right)\right]. \quad (1)$$

where I_i and X_i denote pixel intensity and location. Image edges are also strong indicator of potential object boundary. The affinity between two pixels can be measured by the magnitude of image edges between them,

$$W_C(i,j) = \exp\frac{-\max_{x \in line(i,j)}\|Edge(x)\|^2}{\sigma_C^2} \quad (2)$$

where $line(i,j)$ is the line joining pixel i and j, and $Edge(x)$ is the edge strength at location x. We can combine these two grouping cues with tuning parameter α by

$$W_{comb}(i,j) = \sqrt{W_I(i,j) \times W_C(i,j)} + \alpha W_C(i,j). \quad (3)$$

The graph affinity $W(i,j)$ exhibits very different characteristics at different ranges of spatial separation. Therefore, we can separate the graph links into different scales according to their underlying spatial separation,

$$W_{full} = W_1 + W_2 \approx W_1 + {C_{1,2}}^T W_2 C_{1,2} = W_{reconstruction}, \quad (4)$$

where W_i contains affinity between pixels with certain spatial separation range and can be compressed using a recursive sub-sampling of the image pixels such as the use of interpolation matrix $C_{1,2}$ between two scales. This decomposition allows one to study behaviors of graph affinities at different spatial separations. The small number of short-range and long-range connections can have virtually the same effect as a large fully connected graph. This method is able to compress a large fully connected graph into a multiscale graph with $O(N)$ total graph weights. The combined grouping cues are applied to CIE L*a*b* color space. Selections of Normalized Cut parameters to generate multiple segmentation hypothesis are discussed in Section IV.

C. Fast Rejection

Having a large pool of segments makes our overall methods more reliable, however many segments are redundant and not good. These segments are results of selecting inappropriate clustering number in the segmentation step reflecting accidental image grouping. We deal with these problems using a fast rejection step. We first filter out small segments (up to 500 pixels in area) in our implementation as these segments do not contain significant feature points to represent the object classes. We then assign label to segments in each salient region which belong to background as detected previously. The number of segments that passes the fast rejection step is indicative of how rich or cluttered a salient region is.

III. Feature Extraction and Classification

In the proposed system, segmentation step is followed by feature extraction, classification and feedback to the segmentation step. Feature extraction and classification is independently performed on each segment. Feature extraction step estimates suitable features to visually characterize the segments, while the classification step categorizes the segments by using appropriate classifiers on these features.

A. Feature Extraction

This section briefly describes the features used in our experiments. We have proposed a framework that combines global and local features with late decision fusion [12]. Global features are the features that incorporate statistics of the overall distribution of visual information in the object. For the global features we considered three types of color features namely *color statistics, entropy statistics, and predominant color statistics* and three types of texture descriptors namely *Entropy Categorization and Fractal Dimension estimation (EFD), Gabor-Based Image Decomposition and Fractal Dimension Estimation (GFD), and Gradient Orientation Spatial-Dependence Matrix (GOSDM)*. Finally, 4 types of local feature are also extracted. Extracting local features consists of describing visual information from a neighborhood around points of interest in the segment. The local features used in this paper are: *SIFT, Haar wavelets, Steerable filters, and color statistics*.

Color features: Color statistics includes the 1st and 2nd order moment estimates of the *R, G, B, Cb, Cr, a, b, H, S, V* channels for the entire segment. For *Entropy statistics* each segment is first divided into smaller blocks ($N \times N$ pixels) and then 1st and 2nd order moment statistics of the entropy in the *R, G, B* channels are estimated for each block. The average values for all the blocks is used as the final entropy features. *Predominant color statistics* describes the distribution of four most representative colors (in RGB space) for an object [13].

Texture features: EFD is an extension of multifractal analysis framework [16], [15]. We select the entropy of the image as a measure to define a point categorization. The Fractal dimension is, then, estimated for every point set according to this categorization. This approach attempts to characterize the variation of roughness of homogenous parts of the texture in terms of complexity. The Fractal dimension is estimated using the Box-counting method [14]. *GFD* is another variant of multifractal theory, in this case the image is decomposed into primitives in its spatial frequency dimension. For each filtered response the fractal dimension is estimated. Finally, our third texture descriptor, *GOSDM* consists of a set of gradient orientation spatial-dependence matrices to describe textures by determining the probability of occurrence of quantized gradient orientations at a given spatial offsets.

As previously stated, we also extract local features to describe visual information from smaller portions of the segment/object. These include *SIFT* descriptor introduced by Lowe in 2004 [17], *Haar wavelets*, which capture the distribution of gradients within the neighborhood around the point of interest, and *Steerable filters* which refer to randomly oriented filters synthesized using a linear combination of the basis filters [18]. The feature vectors consists of 1st and 2nd order moment statistics of the response of the filtered patch. Finally, *local color statistics* features are also considered. The 1st and 2nd of the *R, G, B, Cb, Cr, a, b, H, S, V* channels around each point of interest are estimated to capture local color information. The Differential-of-Gaussians approach (DoG) [17] is used for detecting the points of interest.

B. Classification

The above features are independently classified for each of the 12 feature channels ($l = 1, \cdots, 12$). Global features are classified using Support Vector Machines(SVM) [19]. Radial Basis Function are used as kernels of our SVM implementation. Local features are represented by the frequency histogram of visual words obtained by assigning each descriptor of the segment to the closest visual word. Visual words are formed from the training set by using hierarchical k-means clustering. For each of the four local features, the signature of the segment is defined as follows:

$$\phi_j = \{(t_1, m_1), ..., (t_i, m_i), .., (t_N, m_N)\} \quad (5)$$

where ϕ_j represents the signature of the object for the j[th] local feature channel, t_i represents the frequency term and m_i is the *medoid* of the i[th] cluster. N is the number of visual words. These signatures are applied to a nearest neighbor search algorithm to select the final class for each channel.

Both classification methods used, namely SVM for global features and nearest neighbors for local features, output K ($K = 4$ for these experiments) candidate decision categories for each segment (S_q), and each feature channel l, $(c_k^{(l)}(S_q))$. The final decision, $C_k(S_q)$, is made by applying a majority vote rule on $c_k(S_q) = [c_k^{(1)}(S_q), \cdots, c_k^{(L)}(S_q)]$. This means that the top K categories are selected as final candidates for segment S_q.

In our framework, not only the classifier's decision is used as "side" information for the segmentation module, but also the confidence score from the classifier. Regardless what classification method is used for each individual feature channel, the confidence score is estimated in the same fashion for all channels. The confidence score describes the classifier's confidence in its inferred label being correct. The confidence score $\psi_l(S_q, c)$, for assigning segment S_q to class c in the feature channel l is defined as:

$$\psi_l(S_q, c) = \frac{1}{T}\sum_{i=1}^{T} exp(-d(S_q, S_c^i)), \text{for each } c \in C_K(S_q), \tag{6}$$

where $d(S_q, S_c^i)$ represents the distance between normalized feature vector of the query segment S_q and the normalized feature vector of the i^{th} nearest neighbor training segment belonging to class c. T is set to 5 in our experiments. The final confidence score for all feature channels of each candidate class c is defined as:

$$\Psi(S_q, c) = \frac{1}{L}\sum_{i=1}^{L} \psi_l(S_q, c), \tag{7}$$

where $L = 12$ is the total number of feature channels, and $\psi_l(.,.)$ represents the confidence score per feature channel, and $\Psi(.,.)$ the final confidence score of the classifier to label segment S_q with label c.

As a result of this, each pixel in the segment is mapped to four confidence scores corresponding to the four candidate categories predicted by the classifier. The next section describes the approach followed to achieve segmentation stability and robustness by using the confidence scores.

IV. Optimal Segmentation

Different segmentation hypotheses vary in the number of segments and class labels, thus making errors in different regions of the image. Our challenge is to determine which parts of the hypotheses are likely to be correct and combine different hypotheses to accurately determine the class labels. Since the "correct" number of segments Q yielding to "optimal" segmentation is unknown *a priori*, we would like to explore all possible parameter settings. Nonetheless, we are still left with defining the optimal segmentation. In [20], the authors built upon stability-based approaches to develop methods for automatic model order selection. The approach includes the ability to detect multiple stable clusterings instead of only one, and a simple means of calculating stability that does not require training a classifier.

We propose an iterative stability framework for joint segmentation and classification. To produce multiple segmentations, we vary the number of segments Q in two ways, depending on the size of the salient region. For our image database, $Q = 3$ is used as the initial number of segments for regions less than 250-pixels in length or breadth of the bounding box, and $Q = 7$ for larger regions. Figure 2 shows multiple segmentations for each salient region of the input image shown in Figure 1. Let $S_{(i,j)}^m$ denote the segment corresponding to the pixel $I(i,j)$, for the m^{th} iteration of segmentation and classification steps. $C_K(S_{(i,j)}^m)$ denotes the set of K best class labels for segment $S_{(i,j)}^m$. The set of K best class labels for pixel $I(i,j)$, after M iterations is denoted by $C_K^M(I(i,j))$. Each of the candidate classes $c_k^M(I(i,j))$ is estimated based on the cumulative scores $\Psi^M(I(i,j), c)$ defined in Equation 8.

$$c_k^M(I(i,j)) = \underbrace{argmax}_{c \notin C_{(k-1)}^M} \Psi^M(I(i,j), c), \text{ where,}$$

$$\Psi^M(I(i,j), c) = \frac{\sum_{m=1}^{M} \sum_{c_i \in C_K(S_{(i,j)}^m)} 1_{(c_i = c)} \Psi(S_{(i,j)}^m, c)}{\sum_{c_i \in C_K(S_{(i,j)}^m)} 1_{(c_i = c)}} \tag{8}$$

Two stopping criterion are used for deciding the total number of iterations (M), either the percentage of pixels labels being updated is less than 5% or there is no improvement in the confidence scores $\Psi^M(I(i,j), c)$. In general, we achieve the "optimal" results after 4 iterations. The ouput is a labeled map with each pixel assigned to the best class label.

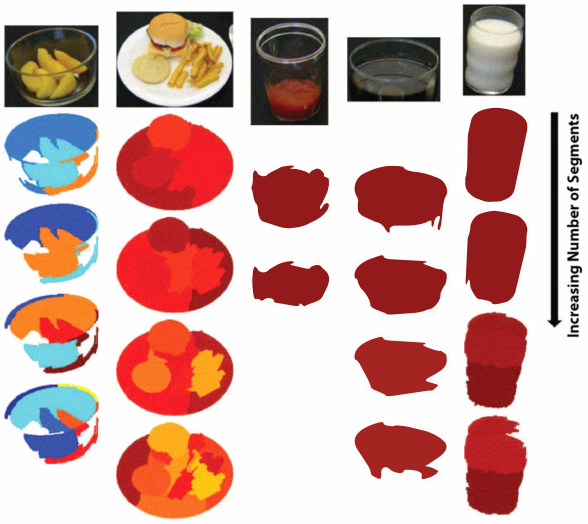

Fig. 2. Multiple Segmentation Results of Salient Regions.

V. Experiments

The proposed system is tested on a collection of food images acquired by the participants during nutritional studies

Fig. 3. Sample Images of Food Objects Used in Our Experiments.

conducted by the Department of Foods and Nutrition at Purdue University. We also developed groundtruth data for the images including the segmentation mask for each food item and the corresponding class label. We have developed an integrated database system containing three interconnected database: an image database that contains data generated by food images, an experiments database that contains data related to nutritional studies and results from the image analysis, and finally an enhanced version of a nutritional database by including both nutritional and visual descriptions of each food [21].

For these experiments we considered images from two nutritional studies. This set of images have a total of 32 food classes from 200 images, each image consists of 6-7 food classes. We divide the dataset into training and testing, for each category approximately half of the images are training data and the other half are testing data. We use a minimum of 15 training samples per category. Examples of food objects used in our experiments are shown in Figure 3.

The performance of proposed segmentation and classification method is evaluated by the confusion matrix for 26 food classes and one background class, averaged over multiple instances of training and testing datasets. Let Δ denotes the $N_f \times N_f$ confusion matrix (for these experiment, the number of classes $N_f = 27$). Then, $\Delta(i,j)$ is the number of pixels predicted as class j while they actually belonged to class i, divided by the total number of pixels actually belonging to class i. Thus, higher values on the main diagonal of confusion matrix indicate good performance. Table I shows the diagonal entries of the confusion matrix, that is the average accuracy for all the classes. Final average classification accuracy for 32 food classes is 44%. Some foods are inherently difficult to classify due their similarity in the feature space; some others are difficult to segment due to faint boundary edges that camouflage the food item; as well as the non-homogeneous nature of certain foods. For example, yellow cake has an average classification accuracy of 11% because its appearance is very similar to that of cream cheese (Figure 3). Similarly, coke has an average classification accuracy of 16% due to its visual similarity to coffee (Figure 3). Since foods are generally served in certain combinations, we hope to explore contextual information in addition to visual characteristics to increase accuracy of classification.

VI. Conclusion

We have described a segmentation and classification framework based on generating multiple segmentation hypothesis. It does so by selecting optimal segmentations using confidence scores assigned to each segment. The approach uses effective methods to generate multiple partitions of segmentation, and proposes an iterative stability measure to assign best class label to each pixel in an image. We have shown the proposed framework is able to generate segments that successfully represent food objects in dietary assessment applications.

Acknowledgment

This work was sponsored by grants from the National Institutes of Health under grants NIDDK 1R01DK073711-

TABLE I
SEGMENTATION ACCURACY FOR EACH CLASS AND AVERAGE PERFORMANCE.

Background	Apple Juice	Bagel	Barbecue Chicken	Broccoli	Brownie	Cheeseburger
97%	29%	98%	28%	66%	13%	17%
Chocolate cake	Coffee	Coke	Cream Cheese	Scrambled Egg	French Dressing	French Fries
14%	93%	16%	95%	35%	35%	73%
Fruit Cocktail	Garlic Bread	Green Beans	Lettuce	Mac&Cheese	Margarine	Mashed Potato
78%	29%	17%	34%	57%	31%	49%
Milk	Orange Juice	Peach	Pear	Pineapple	Pork Chop	Sausage Links
11%	48%	24%	24%	38%	11%	30%
Spaghetti	Sugar Cookie	Vegetable Beef Soup	White Toast	Yellow Cake		**Average**
43%	69%	63%	64%	13%		**44%**

01A1 and NCI 1U01CA130784-01. Address all correspondence to Edward J. Delp, ace@ecn.purdue.edu or see www.tadaproject.org.

REFERENCES

[1] J. Carreira and C. Sminchisescu, "Constrained parametric min-cuts for automatic object segmentation, " *Proceedings of the IIEEE Conference on Computer Vision and Pattern Recognition*, June 2010, pp. 3241-3248.

[2] D. Hoiem, A. Efros, and M. Hebert, "Geometric context from a single image, " *Proceedings of the Tenth IEEE International Conference on Computer Vision*, vol. 1, Oct. 2005, pp. 654-661.

[3] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman, "Discovering objects and their location in images, " *Proceedings of the Tenth IEEE International Conference on Computer Vision*, vol. 1, Oct. 2005, pp. 370-377.

[4] C.J. Boushey, D.A. Kerr, J. Wright, K.D. Lutes, D.S. Ebert, and E.J. Delp, "Use of Technology in Children's Dietary Assessment", *European Journal of Clinical Nutrition*, vol. 1, pp. S50-S57, 2009.

[5] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie, "Objects in Context, " *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2007.

[6] F. Zhu, M. Bosch, I. Woo, S. Kim, C.J. Boushey, D.S. Ebert, and E.J. Delp, "The Use of Mobile Devices in Aiding Dietary Assessment and Evaluation", *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 4, pp. 756-766, August 2010.

[7] B. Six, T. Schap, F. Zhu, A. Mariappan, M. Bosch, E. Delp, D. Ebert, D. Kerr, and C. Boushey, "Evidence-based development of a mobile telephone food record", *Journal of American Dietetic Association*, pp. 74-79, January 2010.

[8] J. Shi and J. Malik, "Normalized cuts and image segmentation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888-905, 2000.

[9] E. Sharon, A. Brandt, and R. Basri, "Fast multiscale image segmentation, " *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, June. 2000, pp. 70-77.

[10] S. Yu, "Segmentation using multiscale cues, " *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, June 2004, pp. 247-254.

[11] T. Cour, F. Benezit, and J. Shi, "Spectral segmentation with multiscale graph decomposition, " *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, June 2005, pp. 1124-1131.

[12] M. Bosch, F. Zhu, N. Khanna, C.J. Boushey, and E.J. Delp, "Combining global and local features for food identification and dietary assessment, " *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, Brussels, Belgium, 2011.

[13] B. Manjunath and W. Ma, "Texture Features for Browsing and Retrieval of Image Data", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 837-842, August 1996.

[14] N. Sarkar, and B. Chaudhuri, "An Efficient Differential Box-Counting Approach to Compute Fractal Dimension of Images", *IEEE Transactions on Systems, Man and Cybernetics*, vol. 24, pp. 115-120, 1994.

[15] J. Vehel, P. Mignot, and J. Merriot, "Multifractals, Texture, and Image Analysis, " *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 661-664, 1992.

[16] K. Falconer, *Fractal Geometry: Mathematical Foundations and Applications*, Wiley, England, 1990.

[17] D. Lowe, "Distinctive Image Features from Scale-Invariant Key-points", *International Journal on Computer Vision*, vol. 2, no. 60, pp. 91-110, 2004.

[18] W. Freeman and Y. Adelson, "The design and use of steerable filters", *IEEE Transactions on Systems, Man and Cybernetics*, pp. 460-473, 1978.

[19] V. Vapnik, "The nature of statistical learning theory", *Springer-Verlag* Nwe York, NY, 1995.

[20] A. Rabinovich, S. Belongie, T. Lange, and J. Buhmann,, "Model order selection and cue combination for image segmentation, " *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, June 2006, pp. 1130-1137.

[21] M. Bosch, T. Schap, N. Khanna, F. Zhu, C.J. Boushey, and E.J. Delp, "Integrated databases system for mobile dietary assessment and analysis, " *Proceedings of the 1st IEEE International Workshop on Multimedia Services and Technologies for E-health in conjunction with the International Conference on Multimedia and Expo*, Barcelona, Spain, July 2011.