HHS Public Access

Author manuscript

Meas Sci Technol. Author manuscript; available in PMC 2016 February 01.

Published in final edited form as:

Meas Sci Technol. 2015 February ; 26(2): . doi:10.1088/0957-0233/26/2/025702.

Saliency-aware food image segmentation for personal dietary assessment using a wearable computer

Hsin-Chen Chen, Ph. D.[1,2], **Wenyan Jia, Ph. D.**[2], **Xin Sun**[5], **Zhaoxin Li**[5], **Yuecheng Li, Ph. D.**[2], **John D. Fernstrom, Ph. D.**[6], **Lora E. Burke, Ph. D.**[7], **Thomas Baranowski, Ph. D.**[8], and **Mingui Sun, Ph. D.**[2,3,4]

[1] Department of Radiation Oncology, Washington University in Saint Louis, Saint Louis, MO, USA

[2] Department of Neurological Surgery, University of Pittsburgh, Pittsburgh, PA, USA

[3] Department of Bioengineering, University of Pittsburgh, Pittsburgh, PA, USA

[4] Department of Electrical Engineering, University of Pittsburgh, Pittsburgh, PA, USA

[5] School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

[6] Department of Psychiatry and Pharmacology, University of Pittsburgh, Pittsburgh, PA, USA

[7] Health and Community Systems, University of Pittsburgh, Pittsburgh, PA, USA

[8] Department of Pediatrics, Baylor College of Medicine, Houston, TX, USA

Abstract

Image-based dietary assessment has recently received much attention in the community of obesity research. In this assessment, foods in digital pictures are specified, and their portion sizes (volumes) are estimated. Although manual processing is currently the most utilized method, image processing holds much promise since it may eventually lead to automatic dietary assessment. In this paper we study the problem of segmenting food objects from images. This segmentation is difficult because of various food types, shapes and colors, different decorating patterns on food containers, and occlusions of food and non-food objects. We propose a novel method based on a saliency-aware active contour model (ACM) for automatic food segmentation from images acquired by a wearable camera. An integrated saliency estimation approach based on food location priors and visual attention features is designed to produce a salient map of possible food regions in the input image. Next, a geometric contour primitive is generated and fitted to the salient map by means of multi-resolution optimization with respect to a set of affine and elastic transformation parameters. The food regions are then extracted after contour fitting. Our experiments using 60 food images showed that the proposed method achieved significantly higher accuracy in food segmentation when compared to conventional segmentation methods.

Correspondences: Mingui Sun (drsun@pitt.edu; phone: 1-412-802-6481; address: 3520 Forbes Ave., Suite 202, Pittsburgh, PA 15213, USA), Hsin-Chen Chen (hsinlovehsin@gmail.com; phone: 1-412-245-8143).

Keywords

active contour model; food segmentation; multi-resolution; saliency map; quantitative dietary assessment

1. Introduction

Diet evaluation is important in the study of a variety of chronic conditions, such as heart disease, diabetes, and obesity [1]. Currently, self-report based on dietary recall is the most commonly used method. However, this method is inaccurate and biased, especially among overweight individuals who often under-report their caloric intake [2][3]. These problems not only cause errors in studies that include dietary assessments, but also in the development of effective dietary strategies for overweight individuals. Because of the importance of accurate dietary assessments in such situations, we have developed a wearable computer "eButton" for the objective evaluation of food intake in real-life settings [4][5][6]. The eButton is a small, unobtrusive chest fob that can be pinned to clothing on the chest. It contains a low-power, high-performance central processing unit, several communication interfaces, electronic sensors, and a Linux or Android operating system. As a wearable computer, it has many applications that are determined by the selected subset of available sensors. In this study of dietary assessment, we use its camera to take pictures automatically at a rate of one picture per second during eating events as shown in the leftmost panel in figure 1. The image data are stored on a SD memory card for later analysis. With the eButton, an individual's diet can be passively recorded without disturbing the subject during eating episodes.

Image-based dietary assessment involves several steps (figure 1). First, food objects in images are recognized. Although this task is extremely difficult for a computer to accomplish, it can be performed easily by the person who previously consumed the food within a short period of time (e.g., within several days). Moreover, food objects are segmented from a selected image. This segmentation task is currently performed manually. In contrast to food recognition, manual segmentation is very time-consuming due to the need to follow all food contours using a mouse or a screen-interactive pen. Once food objects are segmented, the volume (or portion size) of each food object is estimated. In this process, we select a computer generated wire mesh in a shape similar to the food (e.g., a sphere for an apple) and fit the mesh to the segmented food object in the image using a 3D/2D model-to-image registration algorithm [7]. Finally, the food name and volume are both provided to a standard nutrient database to obtain information about calories and nutrients. Detailed descriptions of these procedures are beyond the scope of this paper but can be found in the literature [4]-[11]. Here we focus on the critical problem of food segmentation which, when automated computationally, significantly improves the data processing efficiency in image-based dietary assessment.

Although there have been several image segmentation algorithms developed and applied to quantitative image analysis [12][13], the problem of automatic food segmentation has not been resolved due to several complex issues in images acquired from real-life settings. First,

food has a variety of colors, textures and geometries, and these features may even change substantially among different samples of the same food. It is often difficult to describe food servings mathematically and even to specify their boundaries (figures 2(a) and (b)). Second, a food container, e.g. a plate, may be printed with decorative patterns that can cause visual confusion regarding what is and what is not actual food (figure 2(b)). Third, food may be partially occluded or separated in two or more regions by a utensil (e.g., the spoon in figure 2(c)). As a result, strong edges of the utensil may interfere with the food boundary, causing instability in segmentation results. In general, automatic food segmentation from images is a very challenging problem. Our goal in this study is to overcome these difficulties and design a practical and reliable segmentation tool to support automated image analysis of visual food intake records in large-scale, population-based dietary and obesity research.

2. Previous work on food segmentation

Existing image segmentation methods for dietary assessment fall into three categories: threshold-based, region-based, and deformable model-based methods. In the first category, Mery *et al.* [15] proposed an integrated segmentation approach using Otsu's thresholding. An optimization strategy based on regional contrast was developed to extract food boundaries. Kang *et al.* [16] segmented a number of food items based on color analysis and thresholding. A polynomial equation was derived to characterize the food color distribution. Unfortunately, since the food often contains components with different and variable colors, it is difficult for threshold-based methods to select color patches automatically and group them correctly.

In the second category, Sun *et al.* [17] proposed a region-based method with Sobel boundary constraints to segment food images. Morikawa *et al.* [18] implemented a manually seeded region-growing method implemented on the smartphone for personal dietary monitoring and evaluation. In general, region-based methods depend on the intensity/color similarity during the segmentation process. Similar to the methods in the first category, region-based methods have difficulties in identifying food accurately. Frequently, these methods produce several separate regions when food is composed of multiple components or placed in a container with certain decorative patterns, typically flowers, leafs, or fruits resembling real food ingredients in the container.

The representative method in the third category was proposed by Zhu *et al.* [19] in which an active contour model was applied to food segmentation by regional variance minimization of RGB color components. He *et al.* [20] improved segmentation results using a background removal procedure. In general, deformable model-based methods incorporate contour shape properties to reduce the algorithm sensitivity to non-food regions with similar intensity properties. However, this approach is effective only if the background (e.g., a tablecloth or a container) in the food image has a single color. This requirement is certainly too restrictive in practice.

These food segmentation methods were designed directly in the color domain. While color features are useful, we suspect that foods and containers can present similar colors and textures, thereby creating significant ambiguities in defining food boundaries. In this paper

we propose a novel approach based on “awareness of food saliency” to solve the food segmentation problem. An integrated saliency model is developed mimicking a biological procedure that the human cognitive system utilizes to automatically select visually attended locations. Spatial, color and statistical features of food regions are utilized to constitute a saliency domain that enhances food locations and suppresses non-food regions. Then, a saliency-aware active contour model (ACM) is developed to automatically segment the food boundary. Besides the elastic transformation commonly used in the existing methods, we utilize an affine transformation to adjust the model pose in a registration-assisted strategy to improve segmentation results.

This paper is organized as follows: Sections 3 and 4 describe details of the proposed methods. Section 5 states experimental results and discusses findings of this work before a conclusion in the last section.

3. Integrated saliency estimation for food presence

In our design, we utilize a computational model consisting of two stages, a top-down stage and a bottom-up stage. In the first stage, we do not detect food directly from images and instead, we detect food containers. In this way, the search for edible items from numerous input images, which is an extremely difficult task, is reduced to the search for containers of simple shapes, which can be managed by the computer. Although the container assumption is not always satisfied (foods are occasionally served without a container, in most cases, this assumption is valid. Once the food container is detected, we implement the bottom-up stage in which various image features are utilized to discover food-like regions, producing a food saliency map for further segmentation. The details of the two-stage approach are described as follows.

3.1 Top-down stage

3.1.1 Container detection by shape convexity—Except in some rare cases, food containers are convex in shape. We thus employ this property for detecting food containers [21]. In general, for a convex object, it is always on the same side of the tangent line at each boundary point of the object. To carry out this concept, we firstly employ the Canny edge detector to obtain the edge information from a given image [22]. Then, a number of square windows centered at edge pixels are randomly selected (see figure 3(b)). A trained support vector machine with the histogram of gradient as the classifier input is used to discard the non-container edge windows, increasing the detection efficiency [21].

Let $A = \{a_i\}$, $i = 1,...n$, denote the windows consisting of possible container edges, and $P = \{p_j | j = 1,...m\}$ the set of containers. Without loss of generality, we assume that the containers are circular plates. Assumed a_i belongs to p_j, we can divide the entire image into two parts b_i and $\overline{b_i}$ based on the above-mentioned convexity property using the tangent line of a_i. The line is located at the center and perpendicular to the major gradient direction of a_i, as shown in figure 3(c). Note that b_i is the area on the side of the tangent line in the direction of the gradient of the center point in region a_i. It thus follows that, for a window a_k, $k \neq i$, as long as a_k belongs to p_j as well, the condition $(a_i \in b_k \wedge a_k \in b_i)$ is satisfied, where $\wedge$ represents the intersection between two conditions. An upper triangular matrix $\mathbf{C}_{n\times n}$, which is used to

characterize the relationships between regions, is formed where the (*i*, *k*)-th element is given by:

$$c_{ik}=\begin{cases} 1, & if \quad a_i \in b_k \wedge a_k \in b_i \\ 0, & otherwise \end{cases} . \quad (1)$$

Since matrix **C** can be unfolded as an undirected graph (a_i and a_k are connected if c_{ik} equals one), cliques[1] in the graph can be identified to represent the set of candidate plates, denoted by $H = \{h_q | \ q = 1,...l\}$. This identification is performed using a likelihood function $L(h_q)$:

$$L(h_q)=\frac{1}{1+|Dis(h_q)-s_{expected}|}+\frac{|h_q|}{Max\{|h_1|,\ldots,|h_l|\}}, \quad (2)$$

where $s_{expected}$ is the expected plate diameter which is given by prior knowledge of user-serving plate; and $Dis(h_q)$ is the maximum distance between two regions a_i and a_k in h_q. Each value of likelihood function $L(h_q)$ in equation (2), chosen from $q = 1,..., l$, indicates a recognized convex shape. To work on multiple plates, a threshold can be assigned empirically to identify plates. In the current study, we focus on the serving plate closest to the wearable camera (worn on the eater's chest), from which the foods are the most likely to be taken by the eater. Therefore, the clique with the maximal likelihood is selected as the target plate for personal dietary assessment. Note that, $s_{expected}$ is usually assigned as the diameter of the serving plate that is the closest one to the eater. We then apply the least square fitting method to approximate an ellipse to h_q [23], as shown in figure 3(d)).

3.1.2 Food location prior—When a food is placed in a plate, it is more likely (in a higher probability) located at the central region of the plate. In addition, when people view a plate of food, they have a tendency to pay more attention to the central region of the image [24]. Due to these phenomena, we employ a broad Gaussian weighting function F_P to suppress regions near the border of the detected plate:

$$F_P(x,y)=\frac{1}{\xi_1}exp\left(-\left(\frac{(x-c_x)^2}{2\sigma_x}+\frac{(y-c_y)^2}{2\sigma_y}\right)\right), \quad (3)$$

where c_x and c_y represent the *x*-*y* coordinate of the container center; σ_x and σ_y determine the range of suppression region, and ξ_1 is a normalization value. In our study, σ_x and σ_y were empirically set to $3l_x/8$ and $3l_y/8$ respectively, where l_x and l_y are the lengths of major and minor axes of the detected ellipse.

3.2 Bottom-up stage

With food container detected and its central region emphasized, we implement the bottom-up stage to construct a salient map.

3.2.1 Color contrast—We use color contrast to characterize the attractiveness of a region since it indicates the difference of the objects in the region against their surroundings [25].

[1]A clique is made if for every two vertices, there exists an edge connecting them.

First, four broadly-tuned color channels including red $R = r - (g + b) / 2$, green $G = g - (r + b) / 2$, blue $B = b - (r + g) / 2$, and yellow $Y = (r + g) / 2 - r - g / 2 - b$ are created (negative values are set to zero), where r, g, b are the color values in the original image. Each of the broadly-tuned channels yields maximal response for pure, fully saturated hue to which it is tuned, and yields zero response to both black and white inputs. For example, if a pixel contains both red and green color values it will have a smaller broadly-tuned red response than if it only has red color. The four channels are utilized to create Gaussian pyramids $R(\sigma)$, $G(\sigma)$, $B(\sigma)$, and $Y(\sigma)$, where $\sigma \in [1...5]$, for center-surround operations. Then, we use the four color channels to construct the double-opponent color images RG and BY:

$$RG(c,s) = |(R(c) - G(c)) \{-\} (G(s) - R(s))|, \quad (4)$$

$$BY(c,s) = |(B(c) - Y(c)) \{-\} (Y(s) - B(s))|, \quad (5)$$

where $\{-\}$ denotes across-scale difference between two images which is obtained by interpolation to the finer scale and point-by-point subtraction, and $c \in \{1,2\}$ and $s \in \{3,4,5\}$ indicate the scales of the center and surround respectively. Finally, the color contrast map C_C is computed through across-scale addition $\{+\}$, where each map is reduced to scale two and added point-by-point:

$$C_C = \bigoplus_{c=1}^{2} \bigoplus_{s=3}^{s=5} [N(RG(c,s)) + N(BY(c,s))], \quad (6)$$

where N is a normalization operator which globally promotes maps with a small number of strong peaks, while globally suppressing maps with numerous comparable peaks ("C_C" panel in figure 4). The details of implementing the normalization operator can be referred to Itti's study [25].

3.2.2 Color abundance—Color is important in food preparation and presentation since a colorful food receives more visual attention. In order to reflect this attention, we estimate color abundance using the entropy of local chromatic distribution:

$$C_A(x,y) = -\sum_{i=-W/2}^{W/2} \sum_{j=-W/2}^{W/2} f_{\mathbf{c}}(x+i, y+j) \log f_{\mathbf{c}}(x+i, y+j), \quad (7)$$

where W is the window size used for evaluating the color abundance, and $f_{\mathbf{c}}$ is the probability of a color $\mathbf{c}$ appearing in the local window. Since, in the eButton case, the variation of the distance between the camera and food is limited, a fixed W can be empirically determined (W was set to 20 in our experiments). The calculation of $f_{\mathbf{c}}$ is performed on the 128-bin RGB color space. Equation (7) is used to characterize the color abundance of a local region so that boundary structures that separate colored patches can be identified ("C_A" panel in figure 4).

3.2.3 Spatial arrangement—Although not universally true, a food element, in terms of spatial distribution, is usually centralized and compact, as compared with non-food regions,

e.g. decorating patterns on a plate. Therefore, a regional compactness map R_C is defined based on color-weighted spatial variation [26]:

$$R_C(x,y) = \sum_{j=1}^{Q} w(\mathbf{c}(x,y), \mathbf{c}(x_j,y_j)) \cdot \sqrt{x - x'^2 + y - y'^2}, \quad (8)$$

where Q is the number of image pixels; $\mathbf{c}(x, y) = [r(x, y)\ g(x, y)\ b(x, y)]^T$ represents the color vector at pixel (x, y) in the RGB color space; x' and y' indicate the averaged position of a region with similar color to the pixel at (x, y) (Equation (9)); and w is a 3-D Gaussian function measuring the color similarity between two pixels (Equation (10)).

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \sum_{j=1}^{Q} w(\mathbf{c}(x,y), \mathbf{c}(x_j,y_j)) \begin{bmatrix} x_j \\ y_j \end{bmatrix}, \quad (9)$$

$$w(\mathbf{c}(x,y), \mathbf{c}(a,b)) = \frac{1}{\xi_2} exp\left(-\frac{\|\mathbf{c}(x,y) - \mathbf{c}(a,b)\|^2}{2\sigma_c^2}\right), \quad (10)$$

where σ_c controls the bandwidth of the weighting function (empirically assigned as 12 in our experiments), and ξ_2 is a normalization factor to ensure the summation of pixel weights to be one. In our implementation, images were down-sampled by 2^3 to accelerate computation. Our experiments indicate that the compactness map is helpful in enhancing food elements and suppressing printed patterns on a plate ("R_C" panel in figure 4).

3.3 Integrated saliency map

After implementing the top-down and bottom-up saliency estimation stages, we generate the final saliency map F_S by

$$F_S = \frac{1}{3} F_P \quad C_C + C_A + R_C. \quad (11)$$

The entire saliency estimation process is exemplified in figure 4 in which the effects of each step of computation can be observed. The original food image presents complicated colored patterns on the container and a certain interference from a fork. It can be seen that the top-down stage can directly put eyes' focus on the food container, and each of the three visual attention features capture respective salient information from the image. A region is hence considered food-like if it is centrally located, attractive, structural and compact in the resulting saliency map.

4. Saliency-aware food segmentation

In this section we present a registration-assisted deformation approach to segment the food boundary based on the saliency map.

4.1 Primitive representation

The conventional ACM manipulates a parametric curve by adjusting its shape parameters (i.e. the coordinates of contour points) based on gradient descent optimization [27]. However, the local search process for solutions is highly dependent on the given initial primitive. We therefore use an ellipse as the primitive but add a set of pose parameters to gain flexibility in curve manipulation. The proposed method parameterizes a deforming primitive contour as

$$\mathbf{\Omega}^{i+1}(s) = T(\mathbf{\Theta}) \cdot \mathbf{\Omega}^{i}(s) + \mu_s \cdot \left(\mathbf{n}\left(T(\mathbf{\Theta}) \cdot \mathbf{\Omega}^{i}(s)\right)\right), \quad s \in [0,1], \quad (12)$$

where $\mathbf{\Omega}^i$ and $\mathbf{\Omega}^{i+1}$ represent the evolving contours in iterations i and i+1; $\mathbf{\Theta} = (t_x, t_y, \theta, s_x, s_y)$ is the set of pose parameters including two displacements, one rotation angle and two axial scales of the elliptic primitive; T is an affine transformation matrix; $\mathbf{n}(\mathbf{\Omega}(s))$ represents the outer-pointing norm (in radial direction) of the primitive contour at $\mathbf{\Omega}(s)$, and μ_s is the displacement along $\mathbf{n}(\mathbf{\Omega}(s))$. Note that, in equation (12), the global pose is determined by $\mathbf{\Theta}$, and the local shape is controlled by a set of shape parameters $\boldsymbol{M} = \{\mu_0, ..., \mu_s, ..., \mu_1\}$. In the first iteration, $T(\mathbf{\Theta}) = \mathbf{I}_3$, $\boldsymbol{M} = \{0\}$, and the primitive is an ellipse with center position (c_x, c_y) and axes $(0.5*l_x, 0.5*l_y)$. The regions inside and outside the contour $\mathbf{\Omega}$ are denoted as $\mathbf{\Omega}_{in}$ and $\mathbf{\Omega}_{out}$ respectively. The proposed method iteratively and dynamically optimizes the contour by modifying $\mathbf{\Theta}$ and $\boldsymbol{M}$ individually in two separate steps as follows.

4.2 Primitive registration using regional saliency

A regional energy for deriving the pose parameters is designed based on the minimization of saliency inhomogeneity inside and outside the contour with respect to $\mathbf{\Theta}$:

$$E_1(\mathbf{\Theta}) = \left\{ \int_{T(\mathbf{\Theta})\cdot\mathbf{\Omega}_{in}} (F_s(x,y) - \mu_{in})^2 dxdy + \int_{T(\mathbf{\Theta})\cdot\mathbf{\Omega}_{out}} (F_s(x,y) - \mu_{out})^2 dxdy - \kappa \cdot \mu_{in} \right\}, \quad (13)$$

where (x, y) represents the 2D coordinate of a pixel, and μ_{in} and μ_{out} are calculated by

$$\mu_{in} = \frac{\int_{T(\mathbf{\Theta})\cdot\mathbf{\Omega}_{in}} F_s(x,y)\, dxdy}{\int_{T(\mathbf{\Theta})\cdot\mathbf{\Omega}_{in}} dxdy}, \quad (14)$$

$$\mu_{out} = \frac{\int_{T(\mathbf{\Theta})\cdot\mathbf{\Omega}_{out}} F_s(x,y)\, dxdy}{\int_{T(\mathbf{\Theta})\cdot\mathbf{\Omega}_{out}} dxdy}. \quad (15)$$

Equation (13) is viewed as a modified Chan-Vese contouring energy with additional emphasis on the inner region (food) saliency [28]. It measures the variance of regional contrast rather than the gradient strength at local boundaries. Moreover, the optimization is performed on the parametric domain by using the Powell's multidimensional direction set method [29]. We are hence able to obtain good registration results and to provide a good initial condition for the subsequent contour deformation.

4.3 Shape deformation using boundary saliency

Since there are still certain shape deviations between the registered primitive and the food boundary, we hence perform model deformation by solving the shape parameter vector $\boldsymbol{M}$ that minimizes the energy function E_2:

$$E_2(\boldsymbol{M}) = \int_0^1 (1-\alpha-\beta) E_{edge}\left(\boldsymbol{\Omega}^i(s)\right) + \alpha E_{smooth}\left(\boldsymbol{\Omega}^i(s)\right) + \beta E_{shape}\left(\boldsymbol{\Omega}^i(s)\right) ds, \quad (16)$$

where E_{edge} , E_{smooth}, and E_{shape} are, respectively, the edge energy, smooth energy, and shape energy; α and β are weights determining the relative importance of the energies. We formulate the edge energy based on weighted saliency gradients around the food boundary:

$$E_{edge}\left(\boldsymbol{\Omega}^i(s)\right) = \int_1^L F_s\left(\boldsymbol{\Omega}^i(s) + a\mathbf{n}\left(\boldsymbol{\Omega}^i(s)\right)\right) da - \left(1 - exp\left(-\frac{\varpi}{2\sigma_d^2}\right)\right) \cdot \int_{-L}^0 F_s\left(\boldsymbol{\Omega}^i(s) + a\mathbf{n}\left(\boldsymbol{\Omega}^i(s)\right)\right) da, \quad (17)$$

$$\varpi = \frac{\int_{-L}^0 F_s\left(\boldsymbol{\Omega}^i(s) + a\mathbf{n}\left(\boldsymbol{\Omega}^i(s)\right)\right) da}{\int_1^L da}, \quad (18)$$

where L is used to define the neighborhood around the contour, and σ_d is a factor determining the weight for the saliency of inner neighborhood. The edge energy is used to attract the contour to the position having both a strong saliency gradient and a high saliency value inside the model contour.

The smooth energy indicating the local shape curvature of the deformable model is defined as

$$E_{smooth}\left(\boldsymbol{\Omega}^i(s)\right) = \left|\boldsymbol{\Omega}^{i-1}(s+1) - 2\cdot\boldsymbol{\Omega}^i(s) + \boldsymbol{\Omega}^{i-1}(s-1)\right|_2, \quad (19)$$

where Ω(s−1) and Ω(s+1) represent the neighboring points of Ω(s) on the contour model. Finally, the shape energy is given by

$$E_{shape}\left(\boldsymbol{\Omega}^i(s), \boldsymbol{\Omega}^{i-1}(s)\right) = \sum_{j\in Nei(i)} \left| \left(\frac{\boldsymbol{\Omega}^j(s) - \boldsymbol{\Omega}^i(s)}{\|\boldsymbol{\Omega}^j(s) - \boldsymbol{\Omega}^i(s)\|}\right) \cdot \left(\frac{\boldsymbol{\Omega}^j(s) - \boldsymbol{\Omega}^{i-1}(s)}{\|\boldsymbol{\Omega}^j(s) - \boldsymbol{\Omega}^{i-1}(s)\|}\right) - 1 \right|. \quad (20)$$

This energy aims to maintain the model curvature similarity in adjacent iterations. Otherwise, a large value is produced so that E_2 in equation (16) will be penalized. The edge energy aims to accurately capture the desired food boundary, while the smooth and shape energies constrain the degree of contour deformation to avoid excessive distortions.

In implementation, we speed up the contour fitting process by a multi-resolution Gaussian pyramid. A 4-level pyramid of the saliency map is constructed by scaling down the original map by half between adjacent levels. We let the primitive evolve in the coarsest image using the previously described method. Once the primitive stabilizes at the current resolution, we proceed to the adjacent finer scale and let the primitive evolve until it stabilizes again at this resolution, as shown in figure 5. This process repeats until the finest resolution is reached.

Along with the change of coarse-to fine image pyramid, the value of α (weight of the smooth energy) is decreased progressively from 0.4 to 0.2, the value of β (weight of the shape energy) is decreased progressively from 0.4 to 0.2, and the weight of the edge energy is set to $(1-\alpha-\beta)$. In each level of the pyramid, the shape evolution is stopped when the sum of the changed points in the contour between the previous and current iterations is less than 3, or when the number of iterations reaches 40. Our experience indicates that the primitive in each scale stabilizes quickly, normally in only a few iterations.

5 Results and discussions

5.1 Data sets

With institutional review board (IRB) approval, five adult human subjects participated in our experiments. Each of them wore an eButton on top of the chest to record eating activity as shown in the leftmost panel in figure 1. To ensure the food(s) can be imaged at such a short distance, a 105° wide angle camera was equipped within the eButton. The acquired image sequence was saved on an SD memory card within the eButton. After recording, the data were loaded to a computer where images without motion blur artifacts and missing parts in the plate of food were selected to test the proposed algorithm. In order to gain a balance between power consumption and image quality, the eButton was programmed to take one picture per second in resolution of 640 × 480 pixels.

Besides the eButton-recorded images representing the 30 eating events, additional 30 food images were randomly selected from the Jawbone database [14]. We used this database to test the feasibility of our algorithm for images acquired by other devices such as cell phones since the database contained a large number of food images taken by cell phones all over the world. For each of the 60 test images, we applied our algorithm to segment food automatically without human involvement. Then, we evaluated food segmentation accuracy, compared to the results of conventional methods, and analyzed the performances in different components of the algorithm, to be detailed below.

5.2 Accuracy evaluation of food segmentation

A qualitative evaluation was performed by visually inspecting the quality of fit of the automatic results to the food boundaries. Figure 6 includes both the eButton (top part) and JawBone (bottom part) images, where the original and the segmented images are shown on the left and right sides, respectively. Rectangles and arrows are utilized to indicate the difficulties in segmentation described in Section 1. A rectangle, solid arrow, and dashed arrow indicate, respectively, complex components and appearance, plate with printed patterns, and occlusion by non-food objects.

In addition to qualitative evaluation, the computational results were compared against the manual results. In each manual case, two research participants segmented the same image independently using a mouse on a 24-inch computer screen. The results were averaged as the ground truth. The comparison was based on three metrics: the mean error (ME), relative mean error (RME) and the dice similarity coefficient (DSC) [30], given by

$$ME=\frac{1}{N}\sum_{i=1}^{N} min\left\{\|\mathbf{a}_i-\mathbf{b}_j\|, j=1,2,\ldots,P-1,P\right\}, \quad (21)$$

$$RME=ME/|\mathbf{b}|, \quad (22)$$

$$DSC=\frac{2|\mathbf{A}\cap\mathbf{B}|}{|\mathbf{A}|+|\mathbf{B}|}, \quad (23)$$

where $\mathbf{a}_i$ represents the *i*-th point of the total *N* points on the automatic contour; $\mathbf{b}_j$ is the *j*-th point of the total *P* points on the manual contour closet to $\mathbf{a}_i$; **A** and **B** are the sets of pixels classified as the foods in the automatic result and the ground truth, respectively. Briefly, the ME, RME and DSC evaluate, respectively, the contour distance deviation, the ratio of distance deviation to the food boundary, and the spatial dependency between the two segmented food regions. The evaluation results are summarized in Table 1. The average MEs were less than 7 pixels and the RMEs were smaller than 1%, which indicated a small segmentation error. Moreover, the obtained average DSCs for both data sets (93.46% and 95.32%) were much greater than 70%, which has been reported to indicate a strong overlap between automatic results and ground truth [30].

5.3 Comparison study of segmentation methods

Although there have been numerous color based image segmentation methods, those specifically designed for food segmentation are rare. Three methods were found from the literature, including Otsu's thresholding (OT) [31], Chan-Vese level set evolution (CVLSE) [28], and grab cut (GC) [32]. We thus implemented them to compare with our method. Since the existing methods do not provide solutions for dining container extraction, they will very likely detect background objects falsely as food, thus disturbing the comparison. To avoid this problem, we limit the domain of segmentation to the dining plate detected using the algorithm described in Section 3.1.1. As in the previous case, we compared the new and existing methods both visually and quantitatively. The row in figure 7(a) shows four input images from eButton (first two images) and the JawBone database (last two images). The segmentation results using OT, GC, and our methods are shown, respectively, in rows (b), (d) and (e) with corresponding DSC values. It can be observed that, in both visual and quantitative measures, the proposed method is clearly superior.

In this comparative study, the OT method was based on the optimal binarization of the image intensity histogram, while the CVLSE method evolved the level set curve relying on the color homogeneity. Both methods were sensitive to the color distribution. For example, in the 1st, 3rd and 4th cases in figure 7, they produced fragmental regions. On the other hand, the GC method targeted an optimal cut of the undirected graph to minimize an image-based energy function. It further considered the classification consistency between pixels in the foreground and background, so that the segmentation results were improved over the OT and CVLSE methods. Nevertheless, the GC method required a careful manual determination of a region of interest that must cover the target object and exclude the background. This is

time-demanding for users to handle a large number of food images. In addition, the three major problems described in Section 1 were not solved since the food boundary was still defined purely in the color domain in terms of GC energy. The segmentation results were hence prone to deviations in the presence of complex container patterns or textures.

Compared to the three existing methods, the proposed method combines top-down and bottom-up features, including location prior, color contrast, information abundance, and spatial arrangement, to properly characterize food saliency in the image. For example, the color contrast in Section 3.2.1 can effectively reduce the saliency of strong lighting reflection or silver utensils which have quite limited color attention (see the 1st, 2nd, and 3rd examples in figure 7). Moreover, the spatial arrangement feature worked promisingly on eliminating the saliency of container decorating patters (see the 3rd and 4th examples in figure 7). Hence, the proposed method yielded more desirable segmentation results, as indicated by a high values of the DSC between the ground truth and computed boundaries.

5.4 System performance evaluation

In this experiment we study how segmentation accuracy is influenced by the individual processing components utilized in the proposed method. Two key components, primitive registration and ACM deformation, were examined using four input images shown in the left column of figure 8, and the ME, RME and DSC were evaluated using the standard ACM and the registration-assisted ACM. Both methods were given the same initial conditions. The segmentation results (after average for the four input images) of the standard ACM were 8.443 ± 5.203 mm, 0.8 ± 0.5 %, and 92.385 ± 4.974, respectively, for ME, RME, and DSC, while the corresponding values for the registration-assisted ACM are 3.215 ± 0.323 mm, 0.3 ± 0.1 %, and 96.703 ± 1.239 %. Clearly, these results indicate that that improvement by the registration step is significant.

The conventional ACM framework searches for solutions mostly based on gradient descent or greedy optimization within a local region [33]. If the initial contour is far away from the target boundary, while the descending step is not well assigned or the search line does not reach the target, the conventional ACM tends to fail to converge. In the proposed method, an elliptical primitive is initialized automatically at the center of container. Then, the method registers the primitive with the salient region by adjusting its pose parameters in the parameter space. Because the registration step is able to drive the primitive close to the food boundary, the subsequent standard ACM deformation thus converges to the desired solution more easily. The results of the performance evaluation have also confirmed this argument.

5.5 Limitations and future works

We pointed out that our segmentation algorithm is not universally effective. It may have certain limitations. The proposed plate detector is designed based on the combination of both convex shape and edge information. Theoretically, as long as the edge segments of the plate can be extracted from the background, the detector should be capable of identifying the plate, even in the case of white plate on white table where the plate contour may be observed partially from the presence of a certain shadow or a small contrast difference. In the extremely tough case where the contrast between plate and background is so poor that the

plate edge is not observable, the proposed edge-based detector may fail. Preprocessing steps such as color contrast enhancement could be incorporated, prior to plate edge detection, in order to handle weak image contrast. Moreover, the used convex shape property is applicable to not only the circular shape, but also rectangular and many other shapes. The current circular plate detection protocol could be extended to detect rectangular plates by replacing the least square ellipse fitting method [23] with other geometric shape detection methods (e.g. using tiled or generalized Hough transform to approximate a parallelogram to candidate convex shapes). A future investigation can be addressed with the hope to expand the function of our system to detecting containers with ellipse, rectangular, and other convex shapes.

In addition, the method introduced in this paper can automatically generate accurate food contours from an eating image by starting from dining plate detection. However, many foods are consumed without using a circular plate. In these cases (e.g., figure 8(a)), our segmentation method can still be used but requiring a simple manual sketching step (e.g., figure 8(b)). Instead of using equation (3), the food location prior was calculated using the distance transform as the weighting function [34]. Following the same segmentation algorithm, food boundaries can be identified satisfactorily (figure 8(c)). These examples indicate that our method is still applicable when a plate is absent from the image. However, due to the use of hand-sketching, the segmentation process is not strictly automatic in these cases.

When the food is held by fingers during the eating process, a large segmentation error may occur. In this case, the hand itself represents a compact region in the image, as the case of some food. Therefore, the algorithm will likely assign a high saliency value to the hand, confusing the food segmentation process. In order to make the system more practical and convenient, a user-friendly interface, not discussed in this paper, has been developed in our software to allow a fast visual scan and make adjustments on the automatic segmentation results. In the future, the method presented in this work will be combined with our ongoing investigation on automatic food recognition, aiding in more efficient and objective dietary studies for a large population using advanced wearable computers, and ultimately informing people about their caloric intake automatically.

6 Conclusion

This paper has presented a new saliency-aware segmentation method for personal dietary assessment using food images. An integrated approach with top-down and bottom-up visual attention mechanisms has been developed to generate a food saliency feature domain that automatically enhances food regions. The boundary saliency values and container shape constraints have been incorporated into a deformable model in the segmentation process. A registration-assisted deformation scheme has been presented to improve the segmentation results. This saliency-aware segmentation method can overcome the difficulties caused by complex food configurations and confusing non-food regions, such as decorating patterns on the food container and the use of utensils (e.g., spoons or forks). Our experiments have also shown that the proposed method improves food segmentation results significantly when

compared to the existing methods. The proposed saliency-aware method eliminates the tedious manual segmentation task, providing a powerful tool to support obesity research.

Acknowledgements

This work was supported in part by the National Institutes of Health grant R01CA165255 and R21CA172864.

References

1. Mozaffarian D, Hao T, Rimm EB, Willett WC, Hu FB. Changes in diet and lifestyle and long-term weight gain in women and men The New England. Journal of Medicine. 2011; 364:2392–2404.
2. Goris AH, Westerterp-Plantenga MS, Westerterp KR. Undereating and underrecording of habitual food intake in obese men: selective underreporting of fat intake. The American Journal of Clinical Nutrition. 2000; 71:130–134. [PubMed: 10617957]
3. Livingstone MBE, Robson PJ, Wallace JMW. Issues in dietary intake assessment of children and adolescents. British Journal of Nutrition. 2004; 92:S213–S222. [PubMed: 15522159]
4. Sun, M.; Yao, N.; Hackworth, SA.; Yang, J.; Fernstrom, JD.; Fernstrom, MH.; Sclabassi, RJ. A human-centric smart system assisting people in healthy diet and active living. Proceedings of International Symposium of Digital Life Technologies: Human-Centric Smart Living Technology; Tainan, Taiwan. 2009.
5. Bai, Y.; Li, C.; Jia, W.; Li, J.; Mao, ZH.; Sun, M. Designing a wearable computer for lifestyle evaluation. Proceedings of 38th Annual Northeast Bioengineering Conference; Philadelphia, PA, USA. 2012.
6. Zhang, H.; Li, Y.; Hackworth, SA.; Yue, Y.; Li, C.; Yan, G.; Sun, M. The design and realization of a wearable embedded device for dietary and physical activity monitoring. Proceedings of 3rd International Symposium on Systems and Control in Aeronautics and Astronautics; Harbin, China. 2010.
7. Chen HC, Jia W, Li Z, Yue Y, Sun YN, Sun M. Model-based measurement of food portion size for image-based dietary assessment using 3D/2D registration. Measurement Science and Technology. 2013; 24:105701, 11.
8. Jia W, Chen HC, Yue Y, Li Z, Fernstrom J, Sun M. Accuracy of food portion size estimation from digital pictures acquired by a chest-worn camera. Public Health Nutrition. 2013; 4:1–11.
9. Jia W, Yue Y, Fernstrom JD, Yao N, Sclabassi RJ, Fernstrom MH, Sun M. Imaged based estimation of food volume using circular referents in dietary assessment. Journal of Food Engineering. 2012; 109:76–86. [PubMed: 22523440]
10. Sun, M.; Burke, L.; Mao, ZH.; Chen, Y.; Chen, HC.; Bai, Y.; Li, Y.; Li, C.; Jia, W. eButton: a wearable computer for health monitoring and personal assistance. Proceeding of 51st ACM Design Automation Conference; San Francisco, CA. 2014.
11. Sun M, Fernstrom JD, Jia W, Hackworth SA, Yao N, Li Y, Li C, Fernstrom MH, Sclabassi RJ. A wearable electronic system for objective dietary assessment. Journal of the American Dietetic Association. 2010; 110:45–47. [PubMed: 20102825]
12. Chen HC, Yang TH, Thoreson A, Zhao C, Amadio PC, Sun YN, Su FC, An KN. Automatic and quantitative measurement of collagen gel contraction using model-guided segmentation. Measurement Science and Technology. 2013; 24:085702.
13. Lu H, Pan Y, Mandal B, Eng HL, Guan C, Chan DWS. Quantifying limb movements in epileptic seizures through color-based video analysis. IEEE Transactions on Biomedical Engineering. 2013; 60:461–469. [PubMed: 23192478]
14. Jawbone, inc.; 99 Rhode Island St, San Francisco CA: p. 94103www.jawbone.com
15. Mery D, Pedreschi F. Segmentation of colour food images using a robust algorithm. Journal of Food Engineering. 2005; 66:353–360.
16. Kang SP, Sabarez HT. Simple colour image segmentation of bicolour food products for quality measurement. Journal of Food Engineering. 2009; 94:21–25.

17. Sun DW, Du CJ. Segmentation of complex food images by stick growing and merging algorithm. Journal of Food Engineering. 2004:6117–26.
18. Morikawa, C.; Sugiyama, H.; Aizawa, K. Food region segmentation in meal images using touch points. Proceeding of ACM multimedia 2012 workshop on Multimedia for cooking and eating activities; Nara, Japan. 2012.
19. Zhu F, Bosch M, Woo I, Kim SY, Boushey CJ, Ebert DS, Delp EJ. The use of mobile devices in aiding dietary assessment and evaluation IEEE. Journal of Selected Topics in Signal Processing. 2010; 4:756–76.
20. He, Y.; Khanna, N.; Boushey, CJ.; Delp, EJ. Snakes assisted food image segmentation. IEEE 14th International Workshop on Multimedia Signal Processing; Banff, Canada. 2012.
21. Nie, J.; Wei, Z.; Jia, W.; Li, L.; Fernstrom, JD.; Sclabassi, RJ.; Sun, M. Automatic detection of dining plates for image-based dietary evaluation. Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); Buenos Aires. 2010.
22. Canny J. A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1986; 8:679–698. [PubMed: 21869365]
23. Fitzgibbon AW, Pilu M, Fisher RB. Direct least squares fitting of ellipse. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1999; 21:476–480.
24. Tatler BW. The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor biases and image feature distributions. Journal of Vision. 2007; 7:1–17. [PubMed: 18217799]
25. Itti L, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1998; 20:1254–1259.
26. Perazzi F, Krähenbühl P, Pritch Y, Hornung A. Saliency filters: contrast based filtering for salient region detection. IEEE CVPR. 2012
27. Levi Z, Gotsman C. D-snake: image registration by as-similar-as-possible template deformation. IEEE Transactions on Visualization and Computers Graphics. 2013; 19:331–343.
28. Chan TF, Vese LA. Active contours without edges. IEEE Transactions on Image Processing. 2001; 10:266–277. [PubMed: 18249617]
29. Press, WH.; Teukolsky, SA.; Vetterling, WT.; Flannery, BP. Numerical Recipes in C. 2nd ed.. Cambridge University Press; Cambridge: 1992.
30. Zou KH, Warfield SK, Bharatha A, Tempany CM, Kaus MR, Haker SJ, Wells WM 3rd, Jolesz FA, Kikinis R. Statistical validation of image segmentation quality based on a spatial overlap index. Academic Radiology. 2004; 11:178–189. [PubMed: 14974593]
31. Otsu N. Threshold selection method from grey-level histograms. IEEE Transactions on Systems, Man, and Cybernetics. 1979; 9:62–66.
32. Rother C, Kolmogorov V, Blake A. “GrabCut”: interactive foreground extraction using iterated graph cuts. ACM SIGGRAPH. 2004
33. Kass M, Witkin A, Terzopoulos D. Snake: active contour models. International Journal of Computer Vision. 1988; 1:321–331.
34. Toet A. Target detection and recognition through contour matching. Technical Report TNO.WP31.AT.95b CALMA project. 1994 1994.

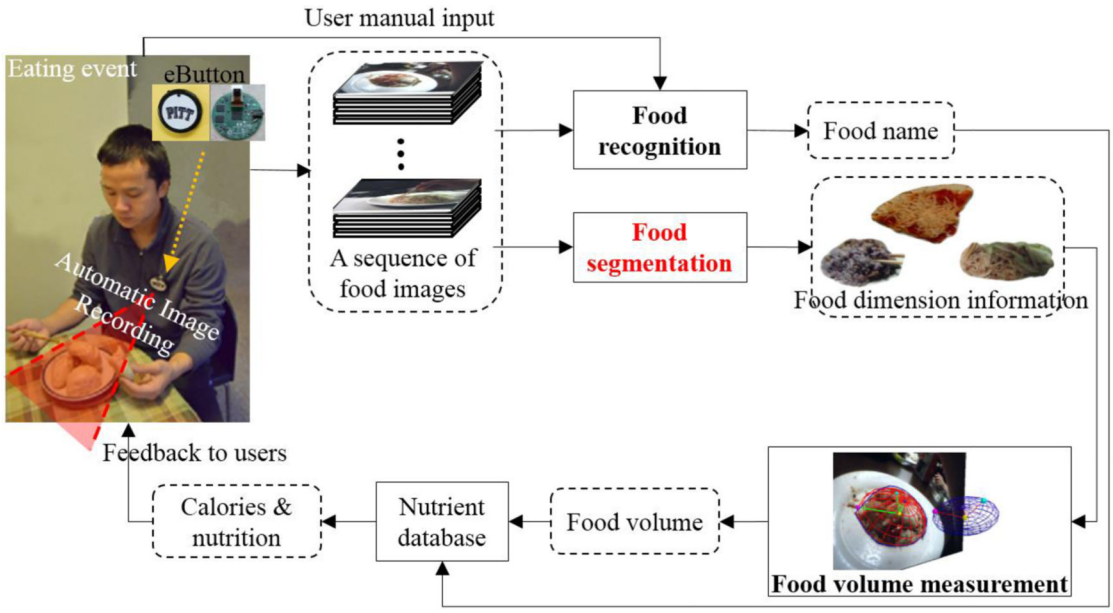

Figure 1.

Personal dietary assessment using a wearable computer "eButton." A subject wears an eButton on his shirt to record eating event. The acquired food image sequence was then processed by food recognition, segmentation, and volume measurement to generate a dietary report. Patient-specific feedback can thus be provided to the user.

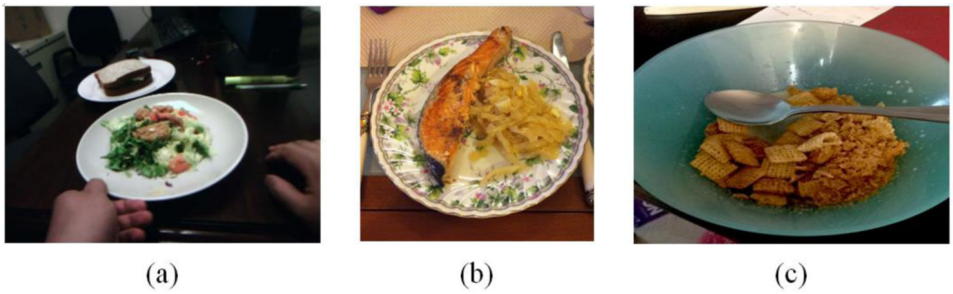

Figure 2.

Major difficulties in automatic food segmentation; (a) multiple food components in complex and varying configurations; (b) colored decorative patterns on plates; (c) occlusion by non-food objects. Photos (b) and (c) are from [14].

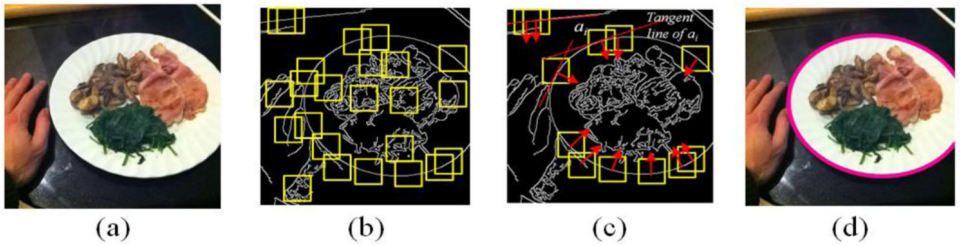

Figure 3.

Container detection using shape convexity; (a) input food image; (b) randomly selected edge windows; (c) candidate window selection based on tangent line separation; (d) detection result of the plate.

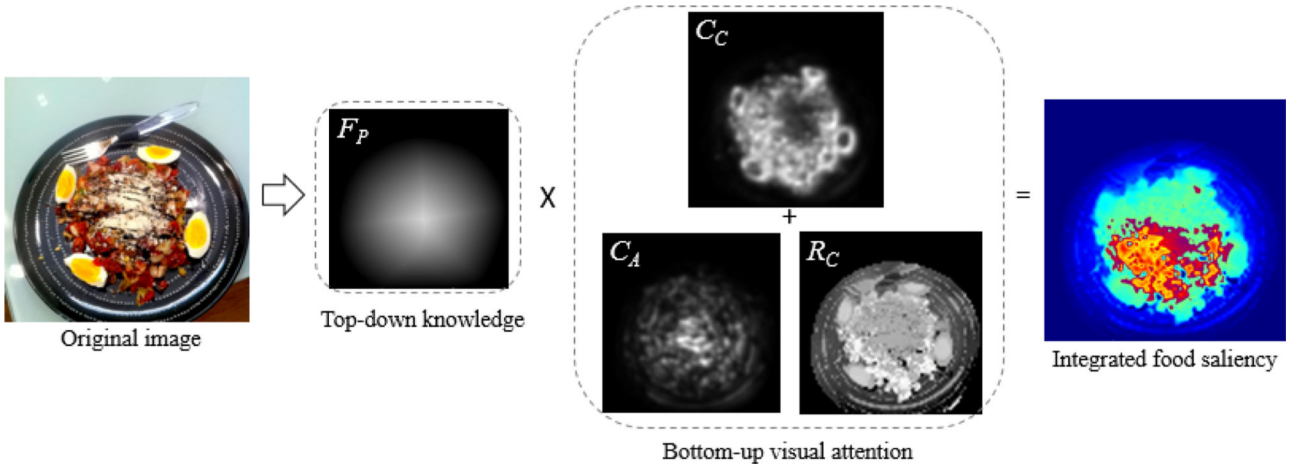

Figure 4.

Integrated food saliency estimation; given a food image, a top-down saliency search is performed, followed by a bottom-up visual attention calculation. The results are integrated to obtain the final saliency information. The original food photo is from the JawBone database [14].

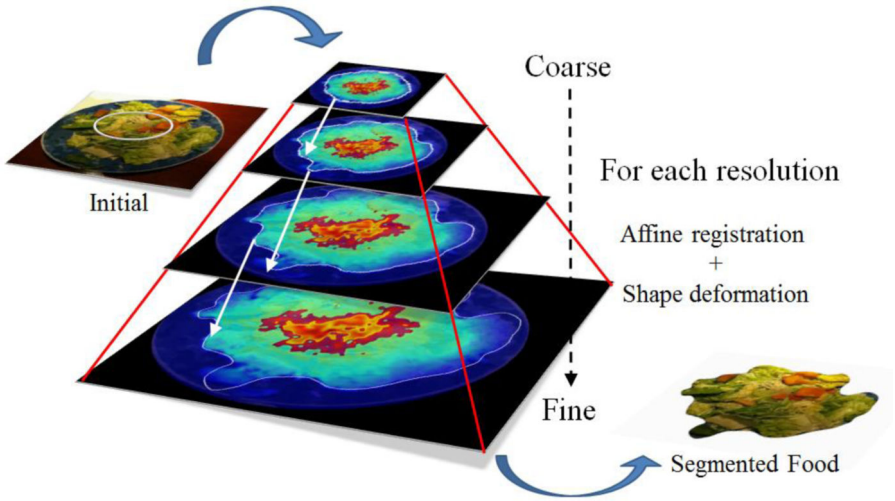

Figure 5.
Saliency-aware food segmentation on multi-resolution pyramid.

Figure 6.
Automatic food segmentation results obtained by applying the proposed saliency-aware approach to the eButton and JawBone data sets.

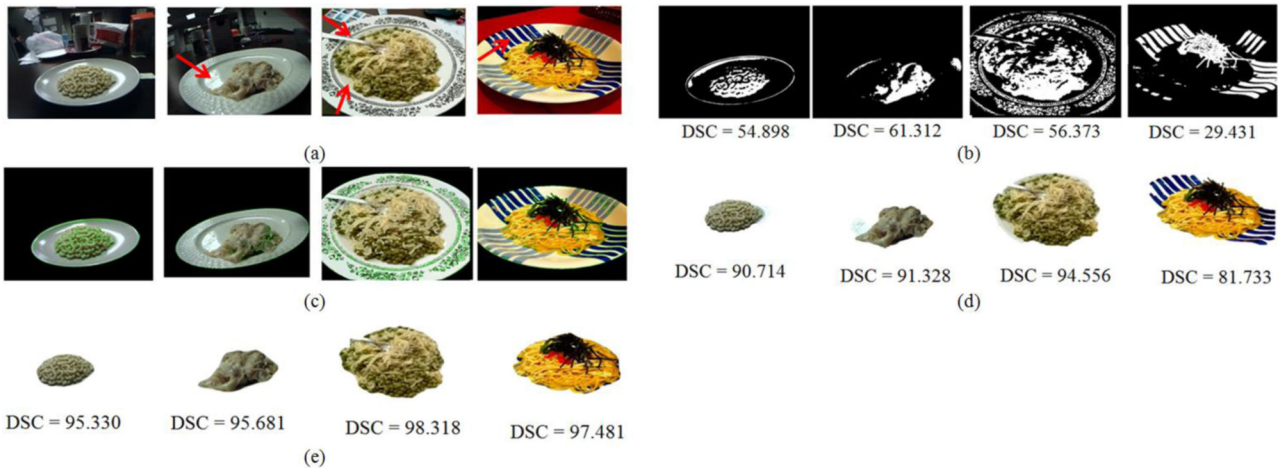

Figure 7.

Comparison of segmentation methods; (a) original images, in which the first two rows belong to eButton data and the others are from JawBone database; (b) results of Otsu's thresholding method; (c) results of Chan-Vese level set evolution; (d) results of grab cut; (e) results of the proposed method. Arrows indicate the difficulties appearing in these examples.

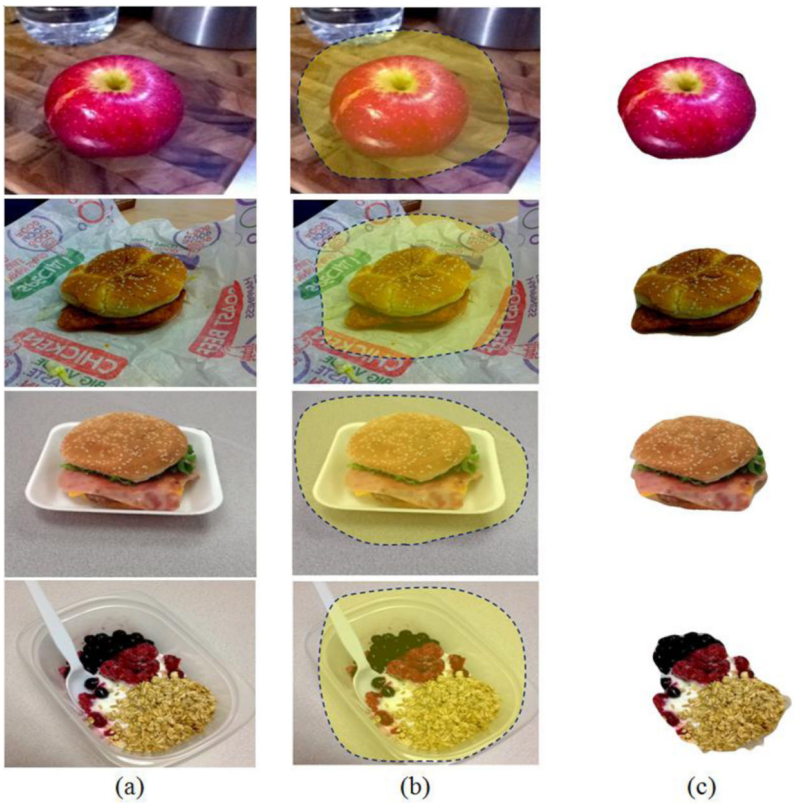

Figure 8.

Examples of applying the proposed method to food images (JawBone) without a contianer or with an unknown shaped container; (a) original images; (b) easy user drawing for regions of interest; (c) final segmentation results.

Table 1

Segmentation accuracy for the eButton and JawBone data sets evaluated using the mean error, relative mean error, and dice similarity coefficient in (mean, standard deviation).

Data Set	ME (pixel)	RME (%)	DSC (%)
eButton (30 images)	(4.397, 1.708)	(0.751, 0.259)	(93.463, 2.025)
JawBone (30 images)	(6.542, 3.133)	(0.483, 0.237)	(95.329, 1.906)