2015 IEEE International Symposium on Multimedia

Single-View Food Portion Estimation Based on Geometric Models

Shaobo Fang, Chang Liu, Fengqing Zhu, Edward J. Delp
Video and Image Processing Laboratory
School of Electrical and Computer Engineering
Purdue University
West Lafayette, U.S.A.

Carol J. Boushey
Cancer Epidemiology Program
University of Hawaii Cancer Center
Honolulu, U.S.A.

***Abstract*—In this paper we present a food portion estimation technique based on a single-view food image used for the estimation of the amount of energy (in kilocalories) consumed at a meal. Unlike previous methods we have developed, the new technique is capable of estimating food portion without manual tuning of parameters. Although single-view 3D scene reconstruction is in general an ill-posed problem, the use of geometric models such as the shape of a container can help to partially recover 3D parameters of food items in the scene. Based on the estimated 3D parameters of each food item and a reference object in the scene, the volume of each food item in the image can be determined. The weight of each food can then be estimated using the density of the food item. We were able to achieve an error of less than 6% for energy estimation of an image of a meal assuming accurate segmentation and food classification.**

***Keywords*-3D Reconstruction, Dietary Assessment, Food Portion Estimation, Geometric Model**

I. Introduction

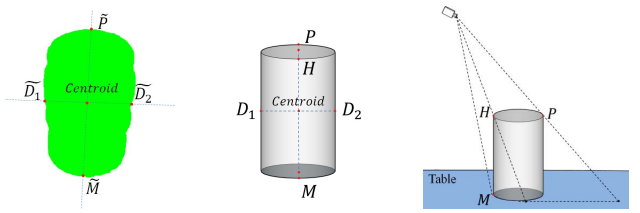
The need to develop accurate methods to measure an individual's food and energy intake has become imperative due to growing concern of chronic diseases and other health problems related to diet. Our previous studies [1] have shown that the use of images of the food eaten by an user can improve the accuracy and reliability of estimating food types and energy consumed. We have developed a mobile dietary assessment tool, the Technology Assisted Dietary Assessment (TADA) system, to determine the food type and energy consumed by a user [1], [2]. Using a mobile telephone, this system allows users to acquire images of their food and then uses image processing and analysis methods to determine the food type and energy consumed [2], [3]. If the food type and volume can be estimated from an image, then the energy (kilocalories, kcal) of the food consumed can be estimated [4], [5]. In this paper, we assume that food items in an image have been correctly identified and an accurate segmentation mask associated with each food item has been obtained [3].

Food volume estimation (also known as portion estimation) is a challenging problem since the food preparation process and the way food is consumed can cause large variation in food shape and appearance. There have been various approaches for portion estimation based on a single image [5], [6], multiple images [4], [7], video [8] and 3D range finding [9]. Our work has focused on the use of a single image for portion estimation since our studies have indicated that this reduces a user's burden [1].

3D reconstruction from a single image is an ill-posed problem and 3D objects in general can not be fully reconstructed from a single-view. However, since our goal is to estimate the volumes of foods in an image, it is not necessary to fully reconstruct the complete scene. The use of geometric models will allow for volume estimation where we can use the food label to index into a class of geometric models for single view volume estimation. The 3D model for a food type (e.g. a banana) can be reconstructed based on multiple-views using shape from silhouettes [10]. We denote the 3D graphical model that is reconstructed from multiple-views as a pre-built 3D model [11]. In addition to pre-built 3D models we have added pre-defined 3D models for conventional shapes [12]. Using the camera parameters we can project both the pre-built and pre-defined 3D models of each food item back onto the image plane then the food volume can be estimated based on a similarity measure of the back-projected region overlaid on the food image segmentation mask. We have also examined the use of prism models (an area-based volume model) that either have non-rigid shapes or do not have significant 3D structures (e.g. scrambled eggs) [12], [13]. Our previous portion estimation technique requires manual initialization of the parameters for different food types prior to use [5], [12]. Although this approach has yielded reasonable results, the manual initialization can pose issues in scaling with many foods.

In this paper we propose to develop a volume estimation technique that uses prior knowledge of the "container shape" as geometric contextual information. For example, the most commonly used containers that have significant 3D structure either can be modeled as cylinders or can be approximated to be cylinders. Knowing that a specific food is likely to be served in a cylindrical shaped container (e.g. milk served in a glass or lettuce in a bowl), using the estimated radius and height of the cylinder, the volume of the food can be obtained. Glasses, cups or even bowls can all be approximated as cylinders. More specifically we focus in this paper on estimating the locations of points of interest in 3D

978-1-5090-0379-2/15 $31.00 © 2015 IEEE
DOI 10.1109/ISM.2015.67



(a) Points of interest estimated from the segmentation mask in pixel coordinates.

(b) Front view of selected points of interest in world coordinates.

(c) Side view of selected points of interest in world coordinates.

Figure 1. Points of interest viewing from world coordinates and partial correspondences estimated in pixel coordinates.

world coordinates of the container based on the projection of 3D containers onto a 2D image plane. The points of interest are selected so that they have sufficient information with respect to the radius and height of the food item as shown in Figure 1(a). We use the prism model which is an area-based volume estimation method for food items that do not have significant 3D structure, such as scrambled egg on a plate with the plate size serving as a reference [5]. After obtaining the volume of each food, we can estimate the weight using the food density (measured in grams/cubic centimeter [14]). The food energy (in kilocalories) can then be obtained from the United States Department of Agriculture (USDA) Food and Nutrient Database for Dietary Studies (FNDDS) [15].

II. The Use Of Geometric Models

Since foods can have large variation in shapes, there does not exist a single geometric model that would be suitable for all types of foods. The correct food classification label and segmentation mask in the image is alone insufficient for 3D reconstruction of a food item, hence the use of geometric models will allow for volume estimation where we can use the food label to index into a class of geometric models for single view volume estimation.

A. The Cylinder Model

If we assume the food item is "cylinder-like" such as liquid in a glass or a bowl of lettuce then we know that the cylinder can be defined by its radius and height. We cannot estimate the radius and height of this cylinder solely based on the segmentation mask which is essentially a projection of a cylinder in world coordinates onto the camera sensor. Three coordinates systems are involved in the estimation of parameters for a cylinder model: the 3D world coordinates, the 2D pixel coordinates which is the original 2D image, and the 2D rectified image coordinates. The 2D rectified image coordinates have the projective distortion removed from the original image.

Camera Parameters and Coordinates Systems: Since the camera parameters are essential for both image rectification and 3D to 2D projection, the intrinsic parameters of the camera and the extrinsic parameters for a specific image must be known. This requires that we have some known structure in the scene. To provide essential reference information, we have designed a checkerboard pattern or fiducial marker (FM) in the TADA system. The fiducial marker is printed and is included in the scene by the user to serve as a reference for the estimation of scale and pose of the objects in the scene [16]. The FM is also used to estimate the camera parameters. Based on the detected corners on the checkerboard and their correspondences in world coordinates, the intrinsic and extrinsic parameters can be obtained [12], [17]. Using the intrinsic parameter $\mathbf{K}$, extrinsic parameters of rotation matrix $\mathbf{R}$ and displacement vector $\vec{t}$, the 3D to 2D projection process for a given point in 3D world coordinates $X : (x_w, y_w, z_w, 1)^T$ to the corresponding point $\widetilde{X} : (\tilde{x}, \tilde{y}, 1)^T$ in the pixel coordinates in an image can be described as:

$$s \begin{bmatrix} \tilde{x} \\ \tilde{y} \\ 1 \end{bmatrix} = \mathbf{K}[\mathbf{R} \quad \vec{t}] \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} \tag{1}$$

more specifically:

$$s \begin{bmatrix} \tilde{x} \\ \tilde{y} \\ 1 \end{bmatrix} = \begin{bmatrix} \alpha & \gamma & x_0 \\ 0 & \beta & y_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{bmatrix} \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} \tag{2}$$

where (x_0, y_0) is the coordinates of the principal points, α, β are the scale factors of x and y axes and γ describes the skew between two image axes. $(\tilde{x}, \tilde{y}, 1)^T$ is homogeneous, and s is a scale factor. Based on the projection described above, although there is a unique projection in pixel coordinate $\widetilde{X} : (\tilde{x}, \tilde{y}, 1)^T$ for any point in 3D world coordinates $X : (x_w, y_w, z_w, 1)^T$, the converse is false.

A correspondence point that provides the reference location of the same object in the different coordinates must be defined in the segmentation mask. We denote such a reference point as locator M, as illustrated in Figure 1(b)(c). In world coordinates we define the locator M to be the closest point to the camera on the bottom surface of the cylinder, which has direct contact with the table. Furthermore, we define $z_w = 0$ for all the points in 3D world coordinates that are contacting the table directly or on the same elevation level. The locator M would be on the $z_w = 0$ surface accordingly. In order to detect the corresponding locator point $\widetilde{M}$ in pixel coordinates, we approximated it to be the lowest point in the column of pixels that is along the centroid of the segmentation mask as illustrated in Figure 1(a). With the assumption that $z_w = 0$, the corresponding point M in world coordinates can be determined based on M using back

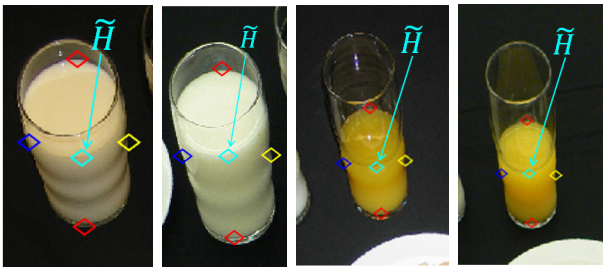


Figure 2. Examples of the estimated $\widetilde{H}$ (cyan $\diamond$) in rectified image coordinates.

projection from 2D to 3D as:

$$s\begin{bmatrix}\tilde{x}\\ \tilde{y}\\ 1\end{bmatrix} = \mathbf{K}[\mathbf{R} \quad \vec{t}]\begin{bmatrix}x_w\\ y_w\\ z_w=0\\ 1\end{bmatrix} = \underbrace{\mathbf{K}[\vec{r_1}\ \vec{r_2}\ \vec{t}]}_{\text{3 by 3 matrix}}\begin{bmatrix}x_w\\ y_w\\ 1\end{bmatrix} \quad (3)$$

where $\vec{r}_1$ and $\vec{r}_2$ are the first and second column vectors of the rotation matrix $\mathbf{R}$. Back projection from 2D to 3D is only valid under the constraint where $z_w = 0$.

Height and Radius Estimation for Cylinder Model: Knowing the locations of the locator alone is insufficient to estimate the radius and the height of the cylinder. Hence more points of interest must be selected and estimated on the segmentation mask in the pixel coordinates. Based the assumption that the food item is "cylinder-like" model, the points of interest are selected such that the line connecting D_1 and D_2 would represent the diameter and the line connecting H and locator M would represent the height in world coordinates as shown as in Figure 1(b)(c). D_1 and D_2 are defined to be on the same elevation level of cylinder's centroid.

Similar to the way we obtained the locator M, we estimate the diameter of the cylinder using the number of row pixels along the centroid of the segmentation mask in the pixel coordinates. The diameter can be determined based on the estimated $\widetilde{D_1}$ and $\widetilde{D_2}$ as shown in Figure 1(a). However the point of interest $\widetilde{H}$ is lost in the 2D pixel coordinates. Instead of estimating $\widetilde{H}$ directly in the pixel coordinates, $\widetilde{P}$ can be estimated by assigning the highest point (away from M) in the column of pixels along the centroid in the segmentation mask (Figure 1(a)). We can infer the location of $\widetilde{H}$ by subtracting the diameter in the $\widetilde{P} \rightarrow \widetilde{M}$ direction from $\widetilde{P}$. The estimation of interest point $\widetilde{P}$ would be performed in the rectified image coordinates as illustrated in Figure 2, where the top of the cylinder is a circle with projective distortion removed. The rectified image coordinates can be obtained by projecting the original image back to 3D world coordinates, under the assumption of $z_w = 0$, using the inverse projection operation of (3). All the points of interest estimated directly from the segmentation mask in 2D pixel coordinates can be projected onto rectified image coordinates. With the locations of the points of interest

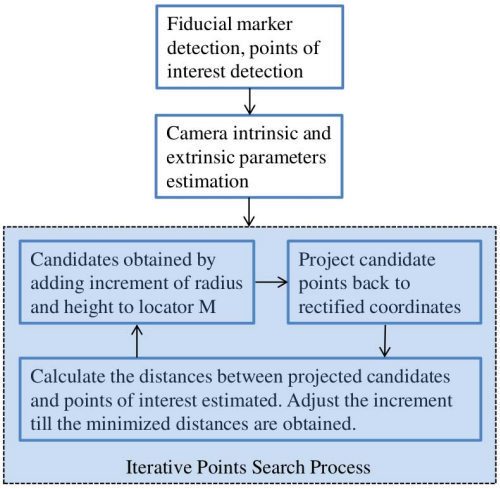


Figure 3. The point of interest search process for radius and height estimation.

in both pixel coordinates and rectified image coordinates estimated, a points search process can be used in 3D world coordinates based on locator M to estimate the radius and height. The process of searching for points in 3D world coordinates whose projections are in 2D coordinates would correspondingly find the best match of $\widetilde{D}_1$, $\widetilde{D}_2$ and $\widetilde{H}$ in the segmentation mask (Figure 3).

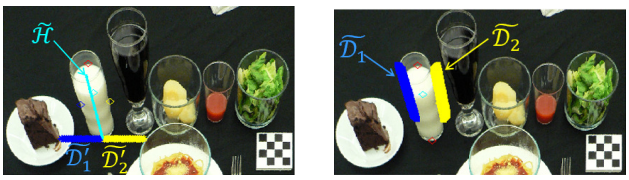
Candidates sets are generated for the purpose of radius and height estimation in 3D world coordinates based on (1). A set of candidate points $\mathcal{H}$ can be obtained in 3D world coordinates by an incremental search along the vertical direction starting from M where each candidate point $H_h \in \mathcal{H}$ is associated with a specific height increment h. The candidates set $\mathcal{H}$ becomes $\widetilde{H}_h \in \widetilde{\mathcal{H}}$ when projected from 3D to 2D as shown in Figure 4(a). The estimated height is obtained by:

$$\hat{h} = \underset{H_h \in \mathcal{H}}{\arg\min} ||\widetilde{H}_h - \widetilde{H}|| \quad (4)$$

Similarly, two sets of candidate points $\mathcal{D}'_1$ and $\mathcal{D}'_2$ that represent the vertical projections of points D_1 and D_2 onto $Z_w = 0$ plane can be obtained by incremental search along horizontal direction of M (Figure 1), where $D'_{1r} \in \mathcal{D}'_1$ and $D'_{2r} \in \mathcal{D}'_2$ are points associated with a specific candidate radius r. The projected candidates sets in 2D pixel coordinates are denoted as $\widetilde{\mathcal{D}}'_1$, $\widetilde{\mathcal{D}}'_2$ and are shown in Figure 4(a). Therefore, the radius can be estimated based on the following:

$$\hat{r} = \underset{D'_{1r}\in\mathcal{D}'_1, D'_{2r}\in\mathcal{D}'_2}{\arg\min} \{\frac{1}{2}||\widetilde{D}'_{1_r} - \widetilde{D}_1|| + \frac{1}{2}||\widetilde{D}'_{2_r} - \widetilde{D}_2||\} \quad (5)$$

The errors in the estimated radius will be reflected in the estimating volume exponentially, we propose a refinement method estimated radius. As shown in Figure 4(a), the searching regions are the vertical projection of D_1 and D_2



(a) Initial search region for radius and height in rectified image coordinates.

(b) Refined search region for radius in rectified image coordinates.

Figure 4. The projections of candidates sets from 3D world coordinates to rectified 2D coordinates.

onto $z_w = 0$ plane: $\mathcal{D}_1'$ and $\mathcal{D}_2'$. With the initial estimate of the radius $\hat{r}$ and height $\hat{h}$, we can refine our search region so the candidates sets match D_1 and D_2 in Figure 4(b).

B. Prism Model

The prism model is an area-based volume estimation method that can be used for food types that do not have significant 3D structures such as scrambled eggs on a plate or toast. For the prism model we assume that the height is the same for the entire horizontal cross-section [5]. In order to accurately estimate the food areas, the original 2D image should be rectified so that the projective distortion can be removed. The fiducial marker can serve as a reference to obtain the 3×3 homography matrix $\mathbf{H}$ used for projective distortion removal. The projective transformation matrix can be estimated using the Direct Linear Transform (DLT) method based on the estimated corners and correspondence pattern [18]. The segmentation mask can be projected from the pixel coordinates of the original 2D image to the coordinates of rectified image. The area of segmentation mask $\hat{S}$ can be estimated in the rectified image. In order to have a better estimation of area of the food, we utilize the area of the plate. If the plate size is consistent across images, we choose the median of the estimated plate size $\hat{P}$ and use it as a scale reference in addition to fiducial marker. In our experimental data used here, the plate size is consistent and is estimated to be $441cm^2$ in world coordinates. The refined area estimation results demonstrated improvement with the estimated plate size serving as a reference: Refined $\hat{S} = \hat{S} \div \hat{P} \times 441cm^2$. The median height for each food item can be estimated based on the ground truth volume and median area estimated for the same type of food: Median Height $= \frac{\text{Ground Truth Volume}}{\text{Median Area}}$.

III. Experimental Results

We used food images from various user studies we conducted as part of the TADA project to test our portion size estimation methods [1]. For these images we had ground truth information for the food types and portion sizes. We assume we have accurate segmentation and food classification. We used 19 food types in our experiments. We used the cylinder model for 9 types of food and the prism model for the rest of the 10 types of food. For the cylinder model with estimated radius $\hat{r}$ and height $\hat{h}$, the volumes $\hat{V}$ can be obtained by $\hat{V} = \pi \times \hat{r}^2 \times \hat{h}$. Although a glass containing a soft drink is more of a semi-cone in a single view than a cylinder, we use the radius and height to estimate the volume of the semi-cone. As another example, chocolate cake is not a cylinder, however since it has significant 3D structures we can approximately use the width and height of the cake to estimate the volume. For the prism model, the volume is the estimated area $\hat{S}$ of segmentation mask in the rectified image multiplied by the estimated median height $\hat{h}$ for the same type of food.

With the food density ρ (in grams/cubic centimeter), the food weight can be computed based on the volume as: $\hat{W} = \rho \times \hat{V}$ [14]. For our test data the same type of food has approximately the same ground truth weight [5]. We compare the estimated average weight for each type of food to the ground truth weight as shown in Table I. The ratio of estimate food weight to ground truth food weight is used as an indicator to determine the accuracy of the estimates. The ratios are obtained by dividing the mean of the estimated weight $\hat{W}$ to the mean of the ground truth weight W. We have compared our results here to those we previously reported [5], [12]. We discussed in [5] that a 15% error or less (i.e. the ratio shown in Table I being from 0.85 to 1.15) would be considered to be an acceptable range for most foods. Out of the 19 food types, only 3 types of food: lettuce, French dressing and ketchup have estimated errors larger than 15%. Although lettuce has a ratio of 1.26 (26% error), it is an improvement compared to the results of 4.61 we presented in [5] and 1.70 we presented in [12]. Given the low energy density of lettuce, the error represents approximately 2 additional kilocalories. For the ketchup and French dressing, the errors generated are due to the height estimates using the cylinder model. Since one would not consume a large amount of ketchup or French dressing in a typical meal, the large errors would not result in a significant impact on the estimate of energy consumed for the entire meal.

We also estimated the energy for each meal as captured by the food images. There are a total of 45 images corresponding to 45 different individual eating occasions reported by participants. More specifically, for this particular dataset we only have 3 different combinations of food, with each combination having approximately the same energy for different images. Examples of each combination of food items are illustrated in Figure 5. For each image the total energy (kilocalories) can be obtained by summing the energy for each food item based on the estimated weight using the FNDDS database [15]. We compare the estimated energy to the ground truth energy (in kilocalories) and then determine the ratio of estimates to the ground truth, for each type of combination as shown by Figure 5. We were able to achieve an error of less than 6%. Therefore our method appears to be

Food Name[a]	n[b]	Estimated radius: $\hat{r}$ (mm $\pm$ SD)	Estimated height: $\hat{h}$ (mm $\pm$ SD)	Estimated weight: $\hat{W}$ (g $\pm$ SD)	Ground truth weight: W (g $\pm$ SD)	Ratio of estimates $\hat{W}$ to ground truth W[c]
Milk(C)	45	34.1 $\pm$ 1.6	66.0 $\pm$ 5.0	235.9 $\pm$ 26.8	220.0 $\pm$ 0.0	1.07
Orange Juice(C)	15	31.1 $\pm$ 1.3	40.1 $\pm$ 2.5	122.0 $\pm$ 10.6	124.0 $\pm$ 0.0	0.98
Strawberry Jam(C)	15	17.9 $\pm$ 0.8	18.2 $\pm$ 11.8	22.1 $\pm$ 15.3	21.1 $\pm$ 1.1	1.05
Margarine(C)	15	18.8 $\pm$ 2.2	29.4 $\pm$ 10.4	32.0 $\pm$ 13.1	27.8 $\pm$ 0.6	1.15
Lettuce(C)	15	51.1 $\pm$ 3.5	24.3 $\pm$ 13.0	61.1 $\pm$ 34.0	48.3 $\pm$ 4.8	1.26
Coke(C)	30	39.8 $\pm$ 2.5	64.8 $\pm$ 9.8	225.9 $\pm$ 43.5	227.2 $\pm$ 2.3	0.99
Chocolate Cake(C)	15	36.7 $\pm$ 4.3	28.3 $\pm$ 16.7	77.0 $\pm$ 41.1	81.5 $\pm$ 12.5	0.95
French Dressing(C)	15	22.6 $\pm$ 1.5	12.6 $\pm$ 4.7	22.1 $\pm$ 7.7	35.7 $\pm$ 1.0	0.62
Ketchup(C)	15	17.7 $\pm$ 1.1	9.6 $\pm$ 2.6	10.9 $\pm$ 3.7	15.5 $\pm$ 0.4	0.70
Food Name	n	Estimated area: $\hat{S}$ (cm^2 $\pm$ SD)	Median height: $\hat{h}$ (mm)	Estimated weight: $\hat{W}$ (g $\pm$ SD)	Ground truth weight: W (g $\pm$ SD)	Ratio of estimates $\hat{W}$ to ground truth W
Sausage(P)	15	32.5 $\pm$ 2.5	17.0	47.8 $\pm$ 3.6	41.5 $\pm$ 2.8	1.03
Scrambled Egg(P)	15	50.5 $\pm$ 4.5	10.8	61.3 $\pm$ 5.5	61.5 $\pm$ 0.7	1.00
White Toast(P)	15	141.2 $\pm$ 16.2	13.0	50.5 $\pm$ 5.8	47.7 $\pm$ 3.4	1.06
Garlic Bread(P)	15	79.8 $\pm$ 12.2	9.3	42.1 $\pm$ 6.4	41.1 $\pm$ 3.0	1.02
Sugar Cookie(P)	15	44.2 $\pm$ 5.4	7.1	26.8 $\pm$ 3.3	27.8 $\pm$ 1.9	0.97
Spaghetti(P)	15	137.0 $\pm$ 10.6	26.0	237.8 $\pm$ 18.4	240.3 $\pm$ 2.6	0.99
French Fries(P)	15	79.6 $\pm$ 6.6	37.8	72.5 $\pm$ 6.0	70.5 $\pm$ 4.3	1.03
Peaches(P)	15	62.2 $\pm$ 14.9	12.3	73.0 $\pm$ 17.5	69.3 $\pm$ 9.9	1.05
Pear Halves(P)	15	52.8 $\pm$ 11.5	13.5	74.5 $\pm$ 16.2	75.6 $\pm$ 4.9	0.99
Cheeseburger(P)	15	122.6 $\pm$ 16.9	26.1	191.7 $\pm$ 26.4	198.8 $\pm$ 11.5	0.96

[a] 'C' indicates cylinder model, 'P' indicates prism model

[b] Number of food images that contains a particular food item

[c] A value '> 1' indicates weight is overestimated, where a value '< 1' indicates the weight is underestimated.

Table I
THE ESTIMATED FOOD WEIGHT (IN GRAMS $\pm$ STANDARD DEVIATION) USING THE CYLINDER AND PRISM MODELS.

very promising for estimating the energy for a meal based on using a single image.

IV. CONCLUSION

In this paper we proposed a method to estimate food portion size from a single-view image. Instead of relying on manual initialization estimation parameters, our method can automatically do volume estimation using the geometric contextual information from the scene. We no longer have the issues in scaling with many foods due to manual initialization of parameters. In the future, we plan to use more contextual information for volume estimation. We are also interested in developing a more robust scheme for energy estimation so that the impact of segmentation and food classification errors (or food portion estimation) can be minimized.

ACKNOWLEDGMENT

This work was sponsored by a grant from the National Institutes of Health under grant NIEH/NIH 2R01ES012459-06. Address all correspondence to E. J. Delp: ace@ecn.purdue.edu or see www.tadaproject.org.

REFERENCES

[1] B. Daugherty, T. Schap, R. Ettienne-Gittens, F. Zhu, M. Bosch, E. Delp, D. Ebert, D. Kerr, and C. Boushey, "Novel technologies for assessing dietary intake: Evaluating the usability of a mobile telephone food record among adults and adolescents," *Journal of Medical Internet Research*, vol. 14, no. 2, p. e58, April 2012.

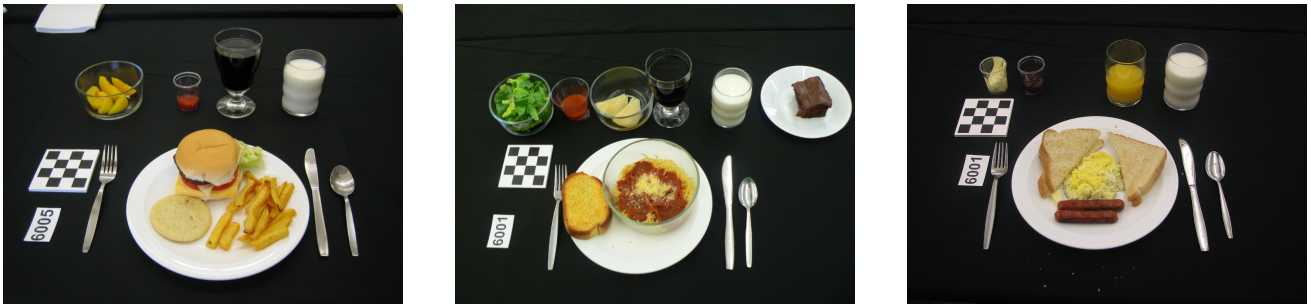
[2] F. Zhu, M. Bosch, I. Woo, S. Kim, C. Boushey, D. Ebert, and E. Delp, "The use of mobile devices in aiding dietary assessment and evaluation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 4, pp. 756–766, August 2010.

[3] F. Zhu, M. Bosch, N. Khanna, C. Boushey, and E. Delp, "Multiple hypotheses image segmentation and classification with application to dietary assessment," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 1, pp. 377–388, January 2015.

[4] M. Puri, Z. Zhu, Q. Yu, A. Divakaran, and H. Sawhney, "Recognition and volume estimation of food intake using a mobile device," *Proceedings of the IEEE Workshop on Applications of Computer Vision*, pp. 1–8, December 2009, Snowbird, UT.

[5] C. Lee, J. Chae, T. Schap, D. Kerr, E. Delp, D. Ebert, and C. Boushey, "Comparison of known food weights with image-based portion-size automated estimation and adolescents' self-reported portion size," *Journal of Diabetes Science and Technology*, vol. 6, no. 2, pp. 428–434, March 2012.

[6] H. Chen, W. Jia, Z. Li, Y. Sun, and M. Sun, "3d/2d model-to-image registration for quantitative dietary assessment," *Proceedings of the IEEE Annual Northeast Bioengineering Conference*, pp. 95–96, March 2012, Philadelphia, PA.



(a) Combination type A: ground truth energy: 834.9 kcal, average estimated energy: 843.2 kcal, ratio of estimate to ground truth: 1.01.

(b) Combination type B: ground truth energy: 1142.8 kcal, average estimated energy: 1107.6 kcal, ratio of estimate to ground truth: 0.97.

(c) Combination type C: ground truth energy: 745.9 kcal, average estimated energy: 788.3 kcal, ratio of estimates to ground truth: 1.06.

Figure 5. Examples of three combinations of food items. Ground truth energy is based on a single serve.

[7] F. Kong and J. Tan, "Dietcam: Automatic dietary assessment with mobile camera phones," *Pervasive and Mobile Computing*, vol. 8, pp. 147–163, February 2012.

[8] M. Sun, J. Fernstrom, W. Jia, S. Hackworth, N. Yao, Y. Li, C. Li, M. Fernstrom, and R. Sclabassi, "A wearable electronic system for objective dietary assessment," *Journal of the American Dietetic Association*, p. 110(1): 45, January 2010.

[9] J. Shang, M. Duong, E. Pepin, X. Zhang, K. Sandara-Rajan, A. Mamishev, and A. Kristal, "A mobile structured light system for food volume estimation," *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pp. 100–101, November 2011, Barcelona, Spain.

[10] K. Kutulakos and S. Seitz, "A theory of shape by space carving," *International Journal of Computer Vision*, vol. 38, no. 3, pp. 199–218, July 2000.

[11] C. Xu, Y. He, N. Khanna, C. Boushey, and E. Delp, "Model-based food volume estimation using 3D pose," *Proceedings of the IEEE International Conference on Image Processing*, pp. 2534–2538, September 2013, Melbourne, Australia.

[12] C. Xu, Y. He, N. Khannan, A. Parra, C. Boushey, and E. Delp, "Image-based food volume estimation," *Proceedings of the International Workshop on Multimedia for Cooking & Eating Activities*, pp. 75–80, October 2013, Barcelona, Spain.

[13] Y. He, C. Xu, N. Khanna, C. Boushey, and E. Delp, "Food image analysis: Segmentation, identification and weight estimation," *Proceedings of the IEEE International Conference on Multimedia and Expo*, pp. 1–10, July 2013, San Jose, CA.

[14] S. Kelkar, S. Stella, C. Boushey, and M. Okos, "Developing novel 3D measurement techniques and prediction method for food density determination," *Procedia Food Science*, vol. 1, pp. 483 – 491, May 2011.

[15] "USDA food and nutrient database for dietary studies, 3.0." Beltsville, MD: Agricultural Research Service, Food Surveys Research Group, 2008.

[16] C. Xu, F. Zhu, N. Khanna, C. Boushey, and E. Delp, "Image enhancement and quality measures for dietary assessment using mobile devices," *Proceedings of the IS&T/SPIE Conference on Computational Imaging X*, vol. 8296, pp. 82 960Q–82 960Q–10, January 2012, San Francisco, CA.

[17] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330–1334, November 2000.

[18] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, 2004.