MODEL-BASED FOOD VOLUME ESTIMATION USING 3D POSE

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

[1] School of Electrical and Computer Engineering, Purdue University
[2] Department of Electronics and Communication Engineering, Graphic Era University, Dehradun, India
[3] Cancer Epidemiology Program, University of Hawaii Cancer Center

ABSTRACT

We are developing a dietary assessment system to automatically identify and quantify foods and beverages consumed by analyzing meal images captured with a mobile device. After food items are segmented and identified, accurately estimating the volume of the food in the image is important for determining the nutrient content of the food. In this paper, we proposed a novel food portion size estimation method for rigid food items using a single image. First, we create a 3D graphical model during the training step using 3D reconstruction from multiple views. Then, for each food image, we determine the translation and elevation parameters of each of the food items, which are relative to the camera coordinate through camera calibration. Using these geometric parameters we project the pre-built 3D model of each food item back to the image plane. Subsequently, the remaining degrees-of-freedom (DOF) for the final pose is estimated by image similarity measure. The experimental results of our volume estimation method for four food categories validate the accuracy and reliability of our model-based approach.

***Index Terms*—** dietary assessment, 3D reconstruction, 3D model rendering, image segmentation, pose estimation

1. INTRODUCTION

Accurately measuring dietary intake is considered to be an open research problem in the nutrition and health fields. Traditional dietary assessment is composed of written and orally reported methods that are time consuming and tedious, which makes them not widely acceptable or feasible for everyday monitoring. Recently, a number of dietary assessment systems utilizing images /videos of eating occasions have been proposed [1, 2, 3, 4]. These systems provide unique mechanisms for improving the accuracy and reliability of dietary assessment. Most of these approaches involve manual or automatic food identification. Portion size of the food items is then estimated through volume estimation. Once food portion size is estimated, the energy and nutrient information of food eaten is obtained. We are developing a system, known as the mobile device food record (mdFR), to automatically identify and quantify foods and beverages consumed by analyzing a single meal image captured with a handheld mobile device [3]. This system is designed to be easy to use and not place a burden on users by having to acquire multiple images or a video of the eating occasions.

Portion size estimation is extremely difficult since many foods have large variations in shape and appearance due to eating or food preparation conditions. Most image-based dietary assessment systems use a single image [5, 6], multiple images [7], video [8], or 3D rangefinding [9]. For example, "DietCam" [2] is a mobile application where automatic food intake assessment is based on images acquired from multiple views. It requires users to acquire three images separated by about 120° which increases user burden. A mobile structured light system (SLS) to measure daily food intake is being developed by Sheng et al. [9]. A laser device which attaches to a mobile telephone is used to capture depth images of the food objects. This system seems burdensome and not suitable for daily use. Jia et al. [10] developed a wearable camera device to collect eating occasion information. It makes use of a known-size plate as the geometric reference. They define several simple geometric shapes to model food shapes and manual adjustment is required. Chen et al. [6] proposed a 3D/2D model-based image registration method for quantitative food intake assessment. The method utilizes a global contour to solve the position, orientation and scale of the user-selected 3D shape model. It obtains reliable food volume estimation for most simple-model food items. However, it does not have a solution for foods that do not fit a simple model (e.g. banana, pear) or complex structured food items (e.g. fries, salad). In addition, it only uses the outline of the object and discards the internal structure (lines, curves, and ridges) of the segments, which could lead to low accuracy in pose registration.

This work was sponsored by grants from the National Institutes of Health under grants NIDDK 1R01DK073711-01A1 and NCI 1U01CA130784-01. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Institutes of Health. Address all correspondence to Edward J. Delp, ace@ecn.purdue.edu or see www.tadaproject.org.

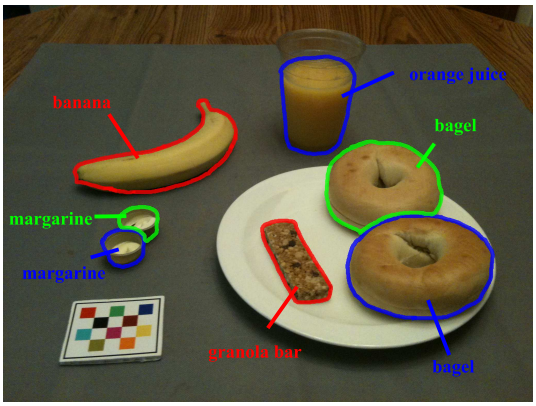

Fig. 1. An example of a food image with checkerboard in the scene. Each food item is segmented and identified using our dietary assessment system.

In the early development of our system, we used a shape template method for 3D reconstruction of some specific shaped food objects [11]. We utilized the feature/corner points from the segmented image to compute the geometric information of the shape template. However, this method is highly dependent on the accuracy of the segmentation mask and the feature points detection is not robust. Moreover, it fails when the food item has a complex or amorphous shape.

In this paper, we propose a novel and reliable volume estimation method based on geometric constraints and a contextual 3D model. We first obtain a 3D graphical model of the food object from multiple training images. We then compute a segmentation mask and a food label using our image segmentation and food identification techniques described in [12, 13, 14]. The segmentation mask provides the location of a food item and the food label indicates the food identification as shown in Figure 1. A credit card sized colored checkerboard is used as the fiducial marker, which is included in every image as a geometric reference for the scale of the world coordinates and to provide color calibration information [15]. We then estimate the camera pose from the checkerboard and establish the world coordinates. The degrees-of-freedom of the pose for different foods are obtained based on the food identification. We utilize several geometric constraints and the food placement regularities to solve the pose registration problem. After the pose of a food item is determined, we are able to estimate the volume of the food. Once the volume is estimated, the nutrient content of the food is obtained using the density for that particular category of food [16].

2. MODEL-BASED FOOD VOLUME ESTIMATION USING 3D MODELLING AND POSE ESTIMATION

3D shape recovery involves a one-to-many mapping from a 2D image to a 3D space. Therefore, single view 3D reconstruction in general is an ill-posed problem. However, if the location and identification of the food is known, and a 3D model of this food item is provided, we can simplify the 3D reconstruction problem to a 3D to 2D pose registration problem. Therefore, the 3D models of food items need to be trained and stored in the database prior to volume estimation.

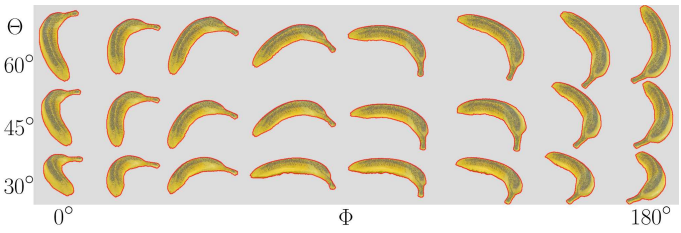

Fig. 2. An example of projecting a 3D banana model to a 2D image plane with two pose angles. The elevation angle ϕ is varied from 0° to 180°, and the azimuth angle θ is varied from 30° to 60°.

2.1. 3D Model Generation

In order to obtain the 3D model, we first need to reconstruct the model of the food from multiple images or a video sequence. After taking multiple images or a video of the food items, we used a variation of a multi-view shape recovery method - *Shape from Silhouettes* [17], also known as *Back-projection Reconstruction* [18]. This method reconstructs a 3D model of an object from the set of contours that outline the projection of the object onto a sequence of 2D image planes. The ideal image acquisition step for shape from carving is to acquire images of the object from different view angles around it such as a turn-table. The typical number of images required for most food items are 15 to 20 images. These images can be obtained from video frames or by capturing multiple still images. The selection of frames can be automatically done by checkerboard detection and camera pose detection. Then, the intrinsic and extrinsic camera parameters need to be determined for each image. In our case, in order to calibrate the images, each image needs to include the checkerboard in the scene. After computing the camera calibration matrix, each camera image is converted to a binary mask which indicates the object silhouette using "1" for object pixels and "0" for background and other contents. Shadow, blur and specular reflection effects could decrease the segmentation accuracy. To make our method more robust to segmentation noise, we use morphological operators to clean up the boundary and avoid small holes (less than 1% of the segmented area of a food item) in the object mask.

Next, the bounding box of the object masks from each image is back-projected onto 3D world coordinates using the camera projection matrix. Based on this 3D bounding box, we fill it with a 3D grid of volume voxels, V, for "carving." The next crucial step is to repeatedly project every $v \in Surf(V)$

onto all the camera images $c_1, c_2, ...c_n$, where $Surf(V)$ is the surface of the volume formed by V. Any voxel that lies outside the object mask in c_i needs to be removed or carved away. As the number of projection and carving steps increases, the object 3D boundary becomes tighter. The iteration is terminated if no non-photo-consistent voxel is found. After carving away every voxel that does not belong to the 3D object model, we obtain the 3D voxel-based model for this food item. We also estimate the volume of the food object by counting how many voxels are left and the size of the voxels from the world coordinate. As shown in Figure 2, we project the reconstructed 3D banana model onto the 2D image plane with various rotation angles. This is a training step and it is applied prior to the volume estimation experiments.

2.2. Pose Initialization

Once the model of the food has been reconstructed, we use it to estimate the geometrical state of a food object in the world coordinate. In general, an indoor and small scale object has 9 degrees of freedom (DOF), $W = (X, Y, Z, \Theta_X, \Theta_Y, \Theta_Z, s_x, s_y, s_z)^T$, consisting of the object translation along three coordinate axes, three rotation angles to the axes, and three relative scale parameters (see Figure 3).

When estimating the pose of a food item on the table, we put geometric constraints on the pose. The first constraint is that we assume the food object is adjacent to the table plane. We define the world coordinate with the checkerboard: one corner of the checkerboard pattern is assigned as the world origin O_w as shown in Figure 3. The x-axis and y-axis are aligned with the lines on the checkerboard thus x-y plane approximates the table plane. As a result, the 3D point P on the object bottom surface has $Z_w = 0$. Also, most food objects have only one placement position on the table. For example, a banana usually lays on the table on its side instead of standing up on the table. Accordingly, two remaining rotation angles of the food object on the table are represented by the azimuth angle ϕ and the elevation angle θ. The azimuth ϕ, is the horizontal rotation about the z-axis from the negative y-axis. The elevation θ is the vertical elevation of the viewpoint moving above the x-y plane (Figure 3).

We also observe that most non-rigid foods and beverages have isotropic shape (e.g. an orange) or symmetric and balanced shape (e.g. orange juice and donut). For these food items there is not much variation in the ratio of the width to the length of the foods. Therefore, we can safely fix the ratio of s_x to s_y with respect to the ratio we obtained from the 3D model for these foods. For food where the ratio of s_x to s_y has considerable variation from sample to sample (e.g. a banana), we currently use the average ratio of two dimensions s_x, s_y to approximate the 3D model of a banana. For other foods, we will obtain the possible range of the s_x to s_y ratio for the specific food and then project the 3D reconstructedmodel by varying the ratio along with the azimuth angle ϕ as well.

After determining the three DOF, the coordinate representation of the object can be written as $E = (x, y, \phi, \theta, s_x, s_z)^T$. First, we use the bottom center 2D point p within the object contour on the image to define the object location. We find two displacement parameters x, y by back-projecting the 2D point $p(p_x, p_y)$ to its corresponding 3D surface point $P(x, y, 0)$ using $(p_x, p_y, 1)^T = K[R|T](x, y, 0, 1)^T$ [19], where K, R, T are respectively the intrinsic matrix, the rotation matrix, and the translation vector obtained during the camera calibration step using the checkerboard. Elevation angle θ will be determined by the right triangle consisting of the camera center C, the object point P, and the projection point C' that C projects onto the x-y plane as shown in Equation (1).

$$\theta = arcsin(\frac{C_z}{\|(C_x, C_y, C_z) - (P_x, P_y, P_z)\|}) \quad (1)$$

Two scale parameters s_x, s_z are given by the length and height of the bounding box on the segmented food.

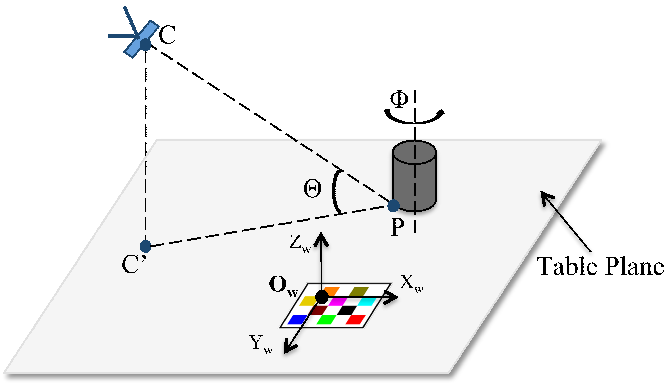

Fig. 3. The geometric relationships between the camera center and the food object.

2.3. Pose Finalization

For rotation symmetric food items (e.g. bagel and orange juice), the object pose is invariant to the azimuth angle ϕ. In other cases, the self-rotation angle can be estimated by image patch matching. Previously, we have trained the food 3D model in our database and determined the elevation angle θ and two size factors s_x, s_z. Based on the fixed θ, we sample the 2D projection images by varying the self-rotation angle ϕ as shown in Figure 2. The images are sampled at 10° intervals and normalized by the size factors s_x, s_z.The last DOF ϕ is chosen by measuring the similarity between each sampled image G_ϕ with the segmented food image F as shown in Equation (2).

$$\phi = \arg\max_{\phi}(\frac{\sum_{i,j}\left[F(x_i, y_j) - \bar{F}\right]\left[G_\phi(x_i, y_j) - \bar{G_\phi}\right]}{\sigma_F \sigma_G}) \quad (2)$$

Where $F(x_i, y_i)$ is the gray scale value of a point (x_i, y_i) in the segmented food image patch. $\bar{F}$ is the average intensity value of the image F. $G_\phi(x_i, y_i)$ is the gray scale value of the point (x_i, y_i) on the sampled image after normalization in reference to ϕ. $\bar{G_\phi}$ is the average intensity value of the image G_ϕ. i, j represents the pixel index. σ_F, σ_G are the standard deviation of $F(x, y)$ and $G_\phi(x, y)$, respectively. The similarity score between F and G_ϕ is based on the normalized cross correlation (NCC) of these two image patches.

After the object pose is finalized, we render the predefined 3D model of the food into the world coordinate in the eating occasion. Food volume will be estimated based on the volume of the 3D rendering model.

3. EXPERIMENTAL RESULTS

The initial performance evaluation of our proposed single view volume estimation method was done by conducing an experiment with five food items (orange juice, bagel, orange, rice krispy treat and banana). We are interested in comparing template based methods [11] and our new proposed 3D model based method. Orange juice, rice krispy and orange treat can be reconstructed using a single view in an efficient manner since they have very regular shapes (cylinder, square box, and sphere). A bagel could also be considered as a regular shaped object, but due to the ambiguity of its color homogeneity, height, and depth information, it cannot be clearly distinguished. Moreover, its textureless uniform color composition does not allow us to use shape information to distinguish height from depth. A banana has a complex shape and there is no regular 3D geometrical template that can be used from the 2D segmentation mask. Orange juice and rice krispy treats are examples of foods where the template based approach [11] can be used, whereas bagels and bananas require a more complex model for their volume reconstruction.

The images were captured using the integrated camera available on the iPhone 3GS. We obtained 15 to 20 images for training as discussed in Section 2.1 and acquired 35 images per food from various view angles and estimated their corresponding volume using our method. The images are captured using different food items of the same type in training and evaluation steps. The results of the estimated volume (mean and standard deviation) for four food items are shown in Table 1 in terms of milliliters and compared with ground truth volume obtained from a water displacement measurement. The estimation error is determined by $|V_e - V_g|/V_g$, where V_e is the estimated volume and V_g is the ground truth volume.

The results of banana and bagel for the template based method [11] are not available, because this method cannot be used for the foods without regular shapes. The results of orange juice, orange, and the rice krispy treat using template based method are satisfactory, but our method further improves the volume estimation accuracy. In addition, we observed the template based approach is sensitive to the segmentation noise, since it is based on feature extraction. Overall, our new 3D model based method achieves an average volume estimation error of 10%. Given that portion size estimation errors of more than 50% from human observation have been reported in the use of traditional dietary assessment methods [20, 21], our results are reasonable and exceed traditional approaches.

Table 1. Comparison of a template based method and our model based method for food 5 items. The mean and standard deviation values are reported $\mu(\sigma)$.

Food Item	**Template Based Method(ml)**	**Model Based Method(ml)**	**Ground Truth (ml)**	**Error of Our Method**
Banana	N/A	182.6(15.9)	170	7.4%
Bagel	N/A	151.2(14.3)	145	4.3%
Orange	179.5(43.2)	215(24.2)	244	12.3%
Orange Juice	179.9(26.6)	183.1(19.1)	200	8.5%
Rice Krispy Treat	78.8(13.6)	72.5(5.3)	70	3.6%

4. CONCLUSION

In this paper we proposed a single view volume estimation method to automatically estimate food portion size. Based on the experimental results, we observed that our model-based volume estimation method not only improves the volume estimation accuracy for foods with simple shapes, but also provides a quantitative approach to estimate volume for foods with irregular shapes. Compared with other methods, it appears to be robust to segmentation noise.

We plan to further investigate our method by testing it with other food items and examine the accuracy for foods with non-rigid shape (e.g. scrambled eggs) or foods with large variations in shape due to eating or food preparation conditions (e.g. cut carrots in a salad).

5. REFERENCES

[1] W. Jia, Y. Yue, J. Fernstrom, N. Yao, R. Sclabassi, M. Fernstrom, and M. Sun, "Imaged based estimation of food volume using circular referents in dietary assessment," *Journal of Food Engineering*, vol. 109, no. 1, pp. 76–86, 2012.

[2] F. Kong and J. Tan, "Dietcam: Automatic dietary assessment with mobile camera phones," *Pervasive and Mobile Computing*, vol. 8, no. 1, pp. 147–163, 2012.

[3] F. Zhu, M. Bosch, I. Woo, S. Kim, C. Boushey, D. Ebert, and E. Delp, "The use of mobile devices in aiding di-

etary assessment and evaluation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 4, pp. 756 –766, August 2010.

[4] M. Bosch, "Visual feature modeling and refinement with application in dietary assessment," Ph.D. dissertation, Purdue University, West Lafayette, IN, USA, May 2012.

[5] I. Woo, K. Ostmo, S. Kim, D. S. Ebert, E. J. Delp, and C. J. Boushey, "Automatic portion estimation and visual refinement in mobile dietary assessment," *Proceedings of the IS&T/SPIE Conference on Computational Imaging VIII*, San Jose, CA, January 2010, p. 75330O.

[6] H. Chen, W. Jia, Z. Li, Y. Sun, and M. Sun, "3d/2d model-to-image registration for quantitative dietary assessment," *Proceedings of 38th Annual Northeast Bioengineering Conference*, Philadelphia, USA, March 2012, pp. 95–96.

[7] F. Kong and J. Tan, "Dietcam: regular shape food recognition with a camera phone," *Proceedings of the International Conference on Body Sensor Networks*, London, UK, May 2011, pp. 127 –132.

[8] M. Sun, J. D. Fernstrom, W. Jia, S. A. Hackworth, N. Yao, Y. Li, C. Li, M. H. Fernstrom, and R. J. Sclabassi, "A wearable electronic system for objective dietary assessment," *Journal American Dietetic Association*, vol. 110, pp. 45–47, 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," *Proceeding of IEEE International Conference on Computer Vision Workshops*, Barcelona, Spain, November 2011, pp. 100–101.

[10] W. Jia, Y. Yue, J. Fernstrom, Z. Zhang, Y. Yang, M. Sun, *et al.*, "3d localization of circular feature in 2d image and application to food volume estimation," *Proceeding of 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Philadelphia, USA, March 2012, pp. 4545–4548.

[11] J. Chae, I. Woo, S. Kim, R. Maciejewski, F. Zhu, E. Delp, C. Boushey, and D. Ebert, "Volume estimation using food specific shape templates in mobile image-based dietary assessment," *Proceedings of the IS&T/SPIE Conference on Computational Imaging IX*, vol. 7873, San Francisco, CA, February 2011, p. 78730K.

[12] M. Bosch, F. Zhu, N. Khanna, C. Boushey, and E. Delp, "Combining global and local features for food identification and dietary assessment," *Proceedings of the International Conference on Image Processing*, Brussels, Belgium, September 2011, pp. 1789 – 1792.

[13] F. Zhu, M. Bosch, N. Khanna, C. Boushey, and E. Delp, "Multilevel segmentation for food classification in dietary assessment," *Proceedings of the 7th International Symposium on Image and Signal Processing and Analysis*, Dubrovnik, Croatia, September 2011, pp. 337–342.

[14] Y. He, N. Khanna, C. Boushey, and E. Delp, "Snakes assisted food image segmentation," *Proceedings of IEEE International Workshop on Multimedia Signal Processing*, Banff, Canada, September 2012, pp. 181–185.

[15] 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, San Francisco, USA, February 2012, p. 82960Q.

[16] 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, 2011.

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

[18] C. Foshee, "Goal-driven three-dimensional object inspection from limited view backprojection reconstruction," Ph.D. dissertation, Purdue University, West Lafayette, IN, USA, December 1991.

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

[20] 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. 63, pp. S50–57, 2009.

[21] T. Schap, B. Six, E. Delp, D. Ebert, D. Kerr, and C. Boushey, "Adolescents in the united states can identify familiar foods at the time of consumption and when prompted with an image 14 h postprandial, but poorly estimate portions," *Public Health Nutrition*, vol. 14, no. 7, pp. 1184–1191, 2011.