THE USE OF CO-OCCURRENCE PATTERNS IN SINGLE IMAGE BASED FOOD PORTION ESTIMATION

Shaobo Fang[⋆]*, Fengqing Zhu*[⋆]*, Carol J Boushey*[†] *and Edward J Delp*[⋆]

[⋆]School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana, USA
[†]Cancer Epidemiology Program, University of Hawaii Cancer Center, Honolulu, Hawaii, USA

ABSTRACT

Measuring accurate dietary intake is considered to be an open research problem in the nutrition and health fields. Food portions estimation is a challenging problem as food preparation and consumption process pose large variations on food shapes and appearances. We use geometric model based technique to estimate food portions and further improve estimation accuracy using co-occurrence patterns. We estimate the food portion co-occurrence patterns from food images we collected from dietary studies using the mobile Food Record (mFR) system we developed. Co-occurrence patterns is used as prior knowledge to refine portion estimation results. We show that the portion estimation accuracy has been improved when incorporating the co-occurrence patterns as contextual information.

Index Terms— Dietary Assessment, Food Portion Size Estimation, Geometric Model, Food Portion Co-Occurrence Pattern.

1. INTRODUCTION

Six of the ten leading causes of death in the United States, including cancer, diabetes, and heart heart disease can be directly linked to diet. Due to the growing concern of chronic diseases and other health problems related to diet, there is a need to develop accurate methods to estimate individual's food and energy intake. Dietary assessment, the process of determining what someone eats during the course of the day, provides valuable insights for mounting intervention programs for prevention of many of the above chronic diseases. Measuring accurate dietary intake is considered to be an open research in the nutrition and health fields. Developing methods for dietary assessment and evaluation has continued to be a challenging task. Traditional dietary assessment technique, such as dietary record, requires individuals to keep detailed written reports for 3-7 days of all food or drink consumed [1, 2], hence it is a time consuming and tedious process. With smartphone quickly gaining popularity in the recent years, the use of smartphones can provide a unique mechanism for collecting dietary information of users. Mobile dietary assessment systems have been developed such as the Technology Assisted Dietary Assessment (TADA) system [3, 4, 5], FoodLog [6], FoodCam [7], DietCam [8], Tuingle [9], Im2Calories [10] to automatically determine the food types and energy consumed by a user using image analysis techniques. In the mobile dietary assessment system that we developed, we focus on the use of image analysis techniques to automatically analyze and determine the food types and energy consumed by a user [11, 5, 12] from a single food image.

Food volume estimation (also known as portion size estimation or portion estimation) is a challenging problem as food preparation and consumption process can pose large variations in food shape and appearance. To date several images analysis based techniques have been developed such as those based on single view image [13, 14, 15, 12, 10, 16, 17], multiple images [18, 8, 19], video [20], 3D range finding [21] and RGB-D image [22]. We feel that either modifying the mobile device or acquiring multiple images/videos of the eating scene is not desirable for users. We focus on food volume estimation using a single-view food image as shown in [23] which reduces user burden.

Estimating the volume of an object from a single-view image is an ill posed problem. Most of the 3D information has been lost during the projection process from 3D world coordinates onto the 2D camera sensor plane. The use of the priori information is required to estimate the food portions. In [15] food portion estimation was converted into pre-determined serving size classification hence the technique could not be generalized. In [14, 24] pre-defined 3D template matching was used however it required manual tuning hence scaling with many foods became a problem. Food portion estimation using geometric models [12] and the approach based on predicted depth map using a Convolutional Neural Network (CNN) [10, 25] overcame the scaling issue with many foods. We compared the food portion size estimation accuracy using both geometric models and depth images [26]. We showed that geometric model based approach achieved

This work was partially sponsored by the US National Institutes of Health under grant NIH/NCI 1U01CA130784-01 and NIH/NIDDK 1R01-DK073711-01A1, 2R56DK073711-04 and a Healthway Health Promotion Research Grant and from the Department of Health, Western Australia. 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 US National Institutes of Health. Address all correspondence to Edward J. Delp, ace@ecn.purdue.edu or see www.tadaproject.org.

Fig. 1. Sample food images collected by users using mFR with fiducial markers placed in the scenes.

higher accuracy compared to that using high quality depth images obtained by structured light technique [27]. In addition, the quality of depth map obtained using commercial level portable devices lead to even worse performance.

In this paper we use contextual information to further improve food portion estimation using geometric models based approach [12]. We define contextual dietary information as the data that is not directly produced by the visual appearance of an object in the image, but yields information about a user's diet or can be used for diet planning [28]. Food portion co-occurrence pattern is one type of contextual information. We estimate the patterns from food images we collected for dietary studies. The patterns we estimated provide valuable insights about a user's eating behavior. Such contextual information can not be determined by examining a single food image alone. In this work we develop a method to model the food portion co-occurrence patterns. We use the co-occurrence models to further refine the portion estimation results. We are able to obtain more accurate estimates of food portion sizes.

2. ESTIMATING FOOD PORTION CO-OCCURRENCE PATTERNS FOR PORTION ESTIMATION REFINEMENT

Volume estimation based on a single-view image is an ill-posed problem and the 3D structure of the scene can not be fully reconstructed. The correct food classification label and segmentation mask in the image alone is insufficient for 3D reconstruction of a food item. The use of geometric models will allow for volume estimation where food label is used to index into a proper class of a food type [12]. In this work we focus on the food classes that have varying shapes and appearances. We use prism model [12] to the food classes as the prism model is designed for food classes with varying shapes and appearances. We have designed a checkerboard pattern fiducial marker to be placed in the eating scene shown in Figure 1. The fiducial marker serves as a reference for both image rectification and food area sizes in world coordinates (in cm^2). We designed the fiducial marker to a credit card size for users to conveniently carry. The small size of the fiducial marker causes errors in the rectified image using computer vision techniques [29]. For example, if a food item is placed far away from the fiducial marker in the eating scene, the estimated volume for such food item may be less accurate. To improve the accuracy of portion size estimation, we rely on a user's eating behavior modeled from food images of dietary studies. By proper modeling and incorporating the food portion co-occurrence patterns into portion estimation, we are able to improve the accuracy of portion estimation. Food portion co-occurrence patterns consist of the distributions of portion sizes and the associated weighting factor. We use Gaussian distributions as they best represent the characteristics of portion sizes distributions. We then refine the food portion estimates based on the models of food portion co-occurrence patterns.

2.1. Food Portion Estimation Using Prism Model

The prism model is an area-based volume estimation technique based on the assumption that the height is the same for the entire horizontal cross-section of the food item. The 5×4 blocks color cherkerboard pattern fiducial marker is used as a reference for corner correspondences and the absolute size in world coordinates. The corners on the checkerboard pattern marker can be estimated using [30]. We obtain the 3×3 homography matrix $\mathbf{H}$ using Direct Linear Transform (DLT) [29]. Assume $\mathbf{I}$ is the original food image (as in Figure 1), the rectified image $\hat{\mathbf{I}}$ can then be obtained by: $\hat{\mathbf{I}} = \mathbf{H}^{-1}\mathbf{I}$. The segmentation mask S_j associated with food j in the original image can be projected from the pixel coordinates to rectified image coordinates. The area of segmentation mask $\hat{S}_j$ from the rectified image can then be estimated. We assume the height h_j for the entire horizontal cross-section. We use median height as the height of the same food class in our food image dataset. The volume of a food item V_j associated with segmentation mask S_j is then estimated: $V_j = \hat{S}_j \times h_j$.

2.2. Food Combination Patterns

Food combination pattern describes the frequencies of various food pairs present in the eating scenes. We use conditional probability of food items appearing in the same eating scene as the food combination patterns [28]. We estimate the food combination patterns from our food images collected from dietary studies. We define $c_{j,k}$ as the conditional probability of food category j appeared given that food category k is present as:

$$c_{j,k} = \frac{p(j,k)}{p(k)} = p(j|k) \tag{1}$$

The food combination patterns only indicate whether two food items are likely to present in the same food image, hence it is insufficient to refine portion size estimation. We need to develop a technique to model the food portion co-occurrence patterns to refine portion estimation.

2.3. The Use of Food Portion Co-occurrence Patterns for Portion Estimation Refinement

The food portion co-occurrence patterns can help refining the portion estimates as they represent the insights reflected by the entire food image dataset rather than a single image. For example, if we know that food items j and k (e.g. fries and

ketchup) usually appear in the same eating scene and the distribution of food portions j and k, we are able to refine the portion estimates based on such prior knowledge.

We use x_i^j, x_i^k to denote the food portions for food classes j, k estimated from food image with index i, where $i \in \{1, 2, 3, \cdots, N\}$. N is the size of our food image dataset, and $j, k \in \{1, 2, 3, \cdots, M\}$ where M is the number of the food classes we use. $\mathcal{C}_{j,k}$ is a combination pair that represents food items j and k are present in the same image. For the combination $\mathcal{C}_{j,j}$, the associated conditional probability is always $c_{j,j} = 1$. We denote $\mathcal{S}_i$ as the set containing all the combination pairs $\mathcal{C}_{j,k}$ exist in food image i. We use 2D Gaussian to model the distributions of portion sizes $\{x_i^j, x_i^k\}$ from our user food image data:

$$g_{j,k}(x_i^j, x_i^k) \sim N(\mu_{j,k}, \sigma_{j,k}) \tag{2}$$

Similarly, we use 1D Gaussian to model the distribution of portions x_i^j estimated for food class j:

$$g_{j,j}(x_i^j, x_i^j) = g_j(x_i^j) \sim N(\mu_j, \sigma_j) \tag{3}$$

where $\mu_{j,k}$, μ_j are the means and $\sigma_{j,k}$, σ_j are the standard deviations of the food estimates we obtained from our user food image data.

As the frequency of combination $\mathcal{C}_{j,k}$ appearing in our user food image data is different, we assign different weighting factors. For example, for the combination $\mathcal{C}_{j,k}$ that appears often across in food image dataset, we assign a heavier weight as it contributes more to the refinement of the portion estimates. Otherwise, we assign a lighter weight for combination $\mathcal{C}_{j,k}$. Furthermore, as the $\mathcal{S}_i$ is different for each food image i, the same $\mathcal{C}_{j,k}$ can carry different weight in different food image. We define the weighting factor $w_i^{j,k}$ of the combination $\mathcal{C}_{j,k}$ in image i as:

$$w_i^{j,k} = \frac{c_{j,k}}{\sum_{\forall \mathcal{C}_{j,k} \in \mathcal{S}_i} c_{j,k}} \tag{4}$$

Note that for each image i the weighting factor $w_i^{j,k}$ is different depending on the food combinations present. The food portion co-occurrence patterns consist of both the weighting factor $w_i^{j,k}$ and the distribution of the portion size estimates as shown in Equation 2 and 3. To refine the portion estimation results obtained using geometric models, we aim to minimize the cost function defined as:

$$f(x_i^j) = 1 - \sum_{\forall \mathcal{C}_{j,k} \in \mathcal{S}_i} w_i^{j,k} \cdot g_{j,k}(x_i^j, x_i^k) \tag{5}$$

In the cost function we weight the probability of portion estimates: (x_i^j, x_i^k) in the image i based on co-occurrence patterns. Our goal is to minimize the cost such that the refined portion size best reflect the co-occurrence patterns we estimate from our food images collected in dietary studies. The refined food portion $\hat{x}_i^j$ in the eating scene can then be obtained by:

$$\hat{x}_i^j = \arg\min_{x_i^j} \{f(x_i^j)\} \tag{6}$$

3. EXPERIMENTAL RESULTS

We divide our food image data into testing and training subsets. We tested on a total of 40 food classes. To reduce the errors propagate from the automatic classification and segmentation, we use ground truth food labels and segmentations masks. We use a subset of our food images for testing and leave the rest food images for training. Geometric model-based technique [12] is used to estimate the portion sizes from our food images. The median height of each food class is estimated from the training dataset. We model the food portion co-occurrence patterns based on the training subset for portion estimation refinement.

We refine the portion estimation results using food portion co-occurrence patterns. We compare the refined portion estimation error of each food class obtained using geometric models to the portion size errors without refinement. The errors in portion size estimation in Figure 2 are defined as:

$$Error = \frac{|\text{Estimated Portion Size} - \text{Ground Truth Portion Size}|}{\text{Ground Truth Portion Size}} \tag{7}$$

The ground truth portion sizes for each food item are provided by nutrient professionals.

The average errors per food class are obtained based on average error of 20 trials. In each trial we randomly sample 5% of our food images as testing subset. We use sampling with replacement technique so that the sizes of the training and testing subsets are the same for each trial. The original error in Figure 2 is the error of food portion estimates obtained using geometric model based [12] approach where the refined error is obtained by incorporating food portion co-occurrence patterns. For most food classes, we are able to improve portion estimation accuracy significantly using refinement technique, such as turkey meal. For some food classes the refinement technique was not sufficient to improve the estimation accuracy significantly (such as grapes). For a few food classes (garlic bread and rice krispy bar) the portion estimates become less accurate with refinement. This is due to the co-occurrence patterns we estimated from training dataset do not generalize well for these specific food classes. Such issue can be addressed by increasing the size of the food image dataset collected from future dietary studies as refinement is fundamentally adding biasness to our system based on past observations. If the past observations include variety of scenarios for most user behavior patterns, we can further improve the estimation accuracy. Our geometric model based portion estimation technique becomes less sensitive to noise by incorporating co-occurrence patterns. It has been shown in Figure 2 that the co-occurrence patterns we estimated generalize well for most of the food classes in our dietary studies. We define the improvement rate as:

$$Improvement = \frac{\text{Original Error} - \text{Refined Error}}{\text{Ground Truth Portion Size}} \tag{8}$$

The overall improvement rate for our dataset is 36.9%.

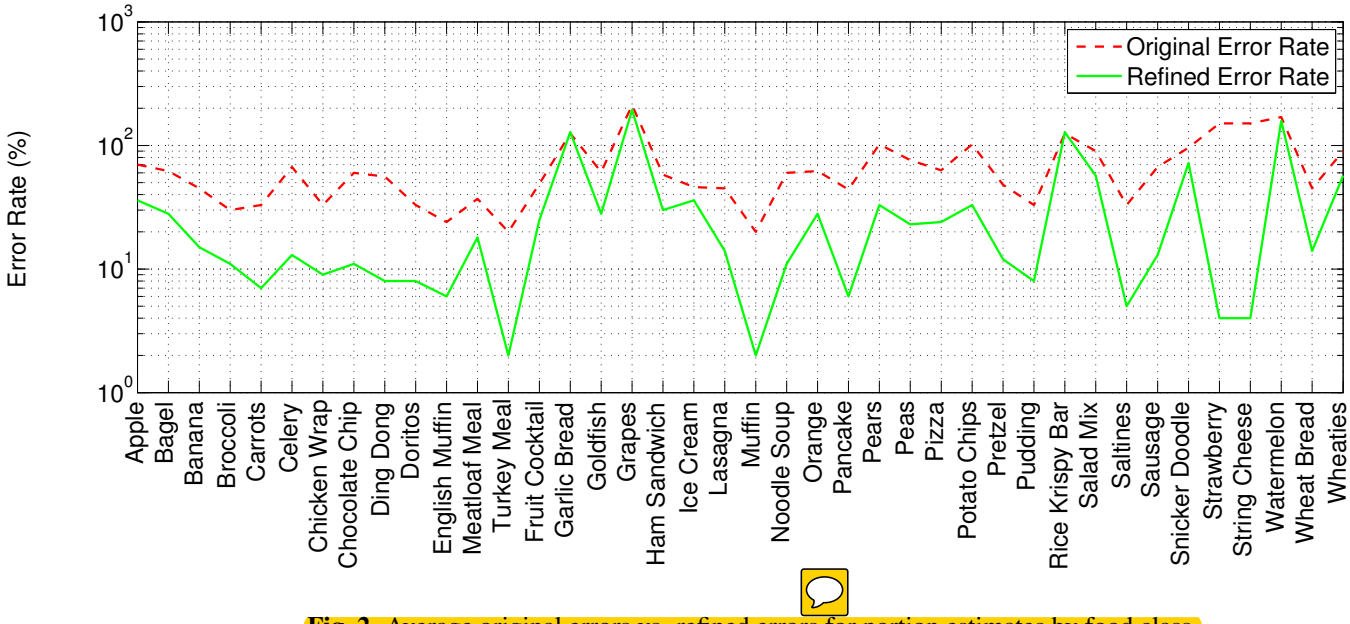

Fig. 2. Average original errors vs. refined errors for portion estimates by food class.

4. CONCLUSION AND FUTURE WORK

In this work we model the food portion co-occurrence patterns based on the food images we collected in dietary studies. The food portion estimation is refined by incorporating the portion co-occurrence patterns. We have shown that with the refinement we significantly improve the estimation accuracy for most of the food classes. In the future we are interested in developing predictive models that extend to food classes that we do not currently have sufficient knowledge on their co-occurrence patterns.

5. REFERENCES

[1] B. Six, T. Schap, F. Zhu, A. Mariappan, M. Bosch, E. Delp, D. Ebert, D. Kerr, and C. Boushey, "Evidence-based development of a mobile telephone food record," *Journal of the American Dietetic Association*, vol. 110, no. 1, pp. 74–79, January 2010.

[2] T. Aflague, C. Boushey, R. Guerrero, Z. Ahmad, D. Kerr, and E. Delp, "Feasibility and use of the mobile food record for capturing eating occasions among children ages 3–10 years in Guam," *Nutrients*, vol. 7, no. 6, pp. 4403–4415, 2015.

[3] C. Boushey, D. Kerr, J. Wright, K. Lutes, D. Ebert, and E. Delp, "Use of technology in children's dietary assessment," *European Journal of Clinical Nutrition*, vol. 63, pp. S50–S57, February 2009.

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

[5] 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.

[6] K. Kitamura, T. Yamasaki, and K. Aizawa, "Foodlog: Capture, analysis and retrieval of personal food images via web," *Proceedings of the ACM multimedia workshop on Multimedia for cooking and eating activities*, pp. 23–30, November 2009, Beijing, China.

[7] T. Joutou and K. Yanai, "A food image recognition system with multiple kernel learning," *Proceedings of the IEEE International Conference on Image Processing*, pp. 285–288, October 2009, Cairo, Egypt.

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

[9] "Tuingle," [Online]. Available: http://tuingle.com/.

[10] A. Meyers, N. Johnston, V. Rathod, A. Korattikara, A. Gorban, N. Silberman, S. Guadarrama, G. Papandreou, J. Huang, and K. Murphy, "Im2calories: Towards an automated mobile vision food diary," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1233–1241, December 2015, Santiago, Chile.

[11] Y. He, C. Xu, N. Khanna, C. Boushey, and E. Delp, "Analysis of food images: Features and classification," *Proceedings of the IEEE International Conference on Image Processing*, October 2014, Paris, France.

[12] S. Fang, C. Liu, F. Zhu, E. Delp, and C. Boushey, "Single-view food portion estimation based on geometric models," *Proceedings of the IEEE International Symposium on Multimedia*, pp. 385–390, December 2015, Miami, FL.

[13] 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.

[14] H. Chen, W. Jia, Z., 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.

[15] K. Aizawa, Y. Maruyama, H. Li, and C. Morikawa, "Food balance estimation by using personal dietary tendencies in a multimedia Food Log," *IEEE Transactions on Multimedia*, vol. 15, no. 8, pp. 2176 – 2185, December 2013.

[16] P. Pouladzadeh, S. Shirmohammadi, and R. Almaghrabi, "Measuring calorie and nutrition from food image," *IEEE Transactions on Instrumentation and Measurement*, vol. 63, no. 8, pp. 1947–1956, August 2014.

[17] W. Zhang, Q. Yu, B. Siddiquie, A. Divakaran, and H. Sawhney, "Snap-n-Eatfood recognition and nutrition estimation on a smartphone," *Journal of Diabetes Science and Technology*, vol. 9, no. 3, pp. 525–533, April 2015.

[18] 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.

[19] J. Dehais, M. Anthimopoulos, and S. Mougiakakou, "Dish detection and segmentation for dietary assessment on smartphones," pp. 433–440, 2015.

[20] 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.

[21] 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*, pp. 100–101, November 2011, Barcelona, Spain.

[22] M. Chen, Y. Yang, C. Ho, S. Wang, S. Liu, E. Chang, C. Yeh, and M. Ouhyoung, "Automatic chinese food identification and quantity estimation," *Proceedings of SIGGRAPH Asia 2012 Technical Briefs*, pp. 29:1–29:4, 2012, Singapore, Singapore.

[23] 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, pp. e58, April 2012.

[24] 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, 2013, Barcelona, Spain.

[25] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," *Proceedings of the IEEE International Conference on Computer Vision*, December 2015, Santiago, Chile.

[26] S. Fang, F. Zhu, C. Jiang, S. Zhang, C. Boushey, and E. Delp, "A comparison of food portion size estimation using geometric models and depth images," *Proceedings of the IEEE International Conference on Image Processing*, pp. 26 – 30, September 2016, Phoenix, AZ.

[27] S. Gorthi and P. Rastogi, "Fringe projection techniques: Whither we are?," *Optics and Lasers in Engineering*, vol. 48, no. 2, pp. 133–140, Feburuary 2010.

[28] Y. He, C. Xu, N. Khanna, C. Boushey, and E. Delp, "Context based food image analysis," *Proceedings of IEEE International Conference on Image Processing*, pp. 2748–2752, September 2013, Melbourne, Australia.

[29] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, second edition, 2004.

[30] G. Bradski and A. Kaehler, *Learning OpenCV: Computer vision with the OpenCV library*, O'Reilly Media, Inc., 2008.