Food Volume Computation for Self Dietary Assessment Applications

J. Dehais, S. Shevchik, P. Diem, S.G. Mougiakakou, *Member IEEE*

***Abstract*— There is great demand for easily-accessible, user-friendly dietary self-management applications. Yet accurate, fully-automatic estimation of nutritional intake using computer vision methods remains an open research problem. One key element of this problem is the volume estimation, which can be computed from 3D models obtained using multi-view geometry. The paper presents a computational system for volume estimation based on the processing of two meal images. A 3D model of the served meal is reconstructed using the acquired images and the volume is computed from the shape. The algorithm was tested on food models (dummy foods) with known volume and on real served food. Volume accuracy was in the order of 90 %, while the total execution time was below 15 seconds per image pair. The proposed system combines simple and computational affordable methods for 3D reconstruction, remained stable throughout the experiments, operates in near real time, and places minimum constraints on users.**

I. Introduction

The increasing prevalence of diet and lifestyle related chronic diseases - including obesity, diabetes mellitus, cardiovascular disease (CVD), hypertension and stroke, and some types of cancer – makes imperative the provision of tools and services that will allow the continuous and personalized support of the patients towards enhanced self-management. Diet and nutrition management are key factors of health and wellbeing.

Information and communication technologies provide a new route for intervention in lifestyle and everyday habits. Specifically, recent advances in computer vision, machine learning, sensors, multimedia technologies, wireless communications and processors permitted the design and development of computerized systems for estimating the nutritional content from meal images.

Such systems are based on the volume estimation of the pictured food items. The problem of estimating volume can be separated in two phases i) 3D shape reconstruction from images and ii) estimation of volume from the shape. The theory and algorithms of 3D shape reconstruction from images are well developed, though a general case algorithm requires additional work: the methods are sensitive to the relative camera positions, to noise in image measurement, and to the common lack of textural information. Besides these, food shapes often present strong irregularities, requiring higher spatial resolution in the 3D model. To obtain acceptable accuracy under the aforementioned circumstances, more advanced algorithms need to be applied, requiring more time and computational power.

An algorithmic approach offering a compromise between accuracy and efficiency is presented here; it is made of two processing stages in which 3D reconstruction is performed with minimal user input (two images), and volume is estimated from the shape. The algorithm was tested on models of different food types with known volume (dummy foods), and in real conditions against a depth sensor. Methods were devised to reduce the influence of lack of shape and texture in images. The proposed method was stable under a variety of camera configurations and illuminations. Scope of the presented approach is to be used as a building block in autonomous, automated dietary assessment systems.

Historically, the first approaches for volume estimation were based on food-specific shape models such as in Chae et al. [1] and Chen et al. [2]. These methods aim to align 3D shape primitives to the edges of the food captured by a photograph. Efficient and straightforward, they however do not cope with the problem of 3D reconstruction in cases of irregular shaped food items and remain semi-automatic to date.

In parallel, promising results in the computation of food volumes were given by Puri et al. [3] using multi-view geometry. The authors used three photographs of the food and reported some of the smallest errors for the volume estimation of irregularly-shaped food. The authors consider efficiency and colour shifts between images, but the proposed method has strong requirements on size and colour of the reference object.

Another system, Dietcam, was proposed for smartphones by Kong and Tan [4], wherein sparse 3D reconstruction was performed from two images before applying food specific templates, this system benefits from lightweight computations, but is limited to the use of visual features.

In this paper, an integrated framework for 3D reconstruction and volume estimation of irregular-shaped food items based on two view geometry is proposed. The system will be used by Type 1 Diabetic (T1D) patients in order to estimate the carbohydrate (CHO) content of the upcoming meal. Recent studies have shown that inaccuracies in the order of 20 g in CHO counting affect the post-prandial glucose profile

Manuscript received 31 July 2013. Research was supported from the Bern University Hospital "Inselspital" and European Union Seventh Framework Programme (FP7-PEOPLE-2011-IAPP) under grant agreement n° 286408 [www.gocarb.eu].

Joachim Dehais is with the Graduate school of Cellular and Biomedical Sciences and the ARTORG Center for Biomedical Engineering Research, University of Bern, CH-3010 Bern (Email: joachim.dehais@artorg.unibe.ch).

Sergey Shevchik is with the ARTORG Center for Biomedical Engineering Research, University of Bern, 3010 Bern, Switzerland (Email: sergey.schevchik@artorg.unibe.ch).

P. Diem is with the Bern University Hospital, Inselspital, Dep. of Endocrinology, Diabetes and Clinical Nutrition, 3010 Bern, Switzerland (email: peter.diem@insel.ch).

S. G. Mougiakakou is with the ARTORG Center for Biomedical Engineering Research, University of Bern, 3010 Bern, Switzerland (Corresponding author; phone: +41 31 632 7592; fax: +41 31 632 7576; e-mail: stavroula.mougiakakou@artorg.unibe.ch).

978-1-4799-3163-7/13/$31.00 ©2013 IEEE

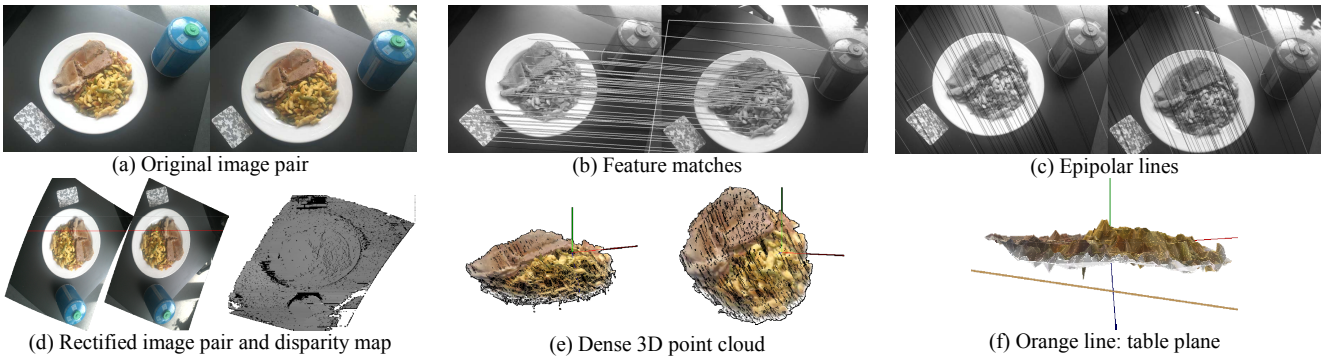

Figure 1. Pipeline and intermediate results of the algorithm

substantially [5], resulting in short- and long-term diabetes complications.

The presented research assumes that: i) the food is wholly contained in the plate, limiting variations in servings, ii) the food shape is irregular and cannot be approximated by high level shape priors, and iii) the surfaces have no visible underside.

- General case dense 3D reconstruction with minimal input from the user;
- High accuracy volume estimation to fulfill the special needs of diabetic patients.
- Low complexity on general purpose hardware.

II. Algorithmic Framework

The two main stages of the algorithm are: i) 3D shape reconstruction from one image pair and ii) volume estimation from the shape. The entire pipeline is presented in Fig. 1 and described in the following paragraphs.

A. Shape reconstruction

A 3D model of the dish was created by applying two-view geometry on the captured food images (Fig. 1a). The first step of two-view geometry consists of finding points in both images. Scale Invariant Feature Transform (SIFT) and Speeded Up Robust Features (SURF) detectors were used for key point extraction, while point matching was performed using a kD-Tree (Fig. 1b). Point correspondences are quickly filtered for outliers using the LO-RANSAC paradigm [6] combined with a *Tdd* test [7].

RANSAC based methods are iterative algorithms which use sample subsets to compute representative models for data sets containing outliers. RANSAC samples are generated using the 5-point algorithm. To improve candidate quality, samples are chosen to maximize constraint independence using Gramm-Schmidt orthonormalisation.

To discriminate inliers from outliers, an adaptive symmetric epipolar distance was computed, wherein the distance is scaled up when closer to the epipole. To improve the quality of the epipolar geometry the sum of distances between a match pair and its representative epipolar line is minimized by using the Levenberg-Marquard algorithm ("gold standard" minimization algorithm [9]). Epipolar lines after the optimization can be seen in Fig. 1c.

To search for dense pixel correspondences between the two images, the image pair is rectified using the radial decomposition of [10], guaranteeing successful rectification for all relative camera configurations. Following this, the rectified image pair is used to compute a stereo disparity map using hierarchical dynamic programming as described in [11] (Fig. 1d).

Matching quality was improved using dynamic penalization costs inversely proportional to the local signal-to-noise ratio of the matching windows.

B. Volume estimation

Multi-view geometry defines shapes up to scale. A flat, fixed size (5 cm x 8 cm), easily recognised object with a known pattern was placed in the field of view in advance to provide a scale reference. The object is recognized by the feature matches between the image and the known object pattern, while the scale of the scene is given by the size of the object. Furthermore, the object's position and orientation adds constraints on the scene depths. In parallel, the reference object provides additional texture information which stabilizes the extraction of the epipolar geometry whenever the number of extracted features is low.

Figure 2. (a, c) Examples of image pairs for dishes models, and (b, d) corresponding subsampled 3D models.

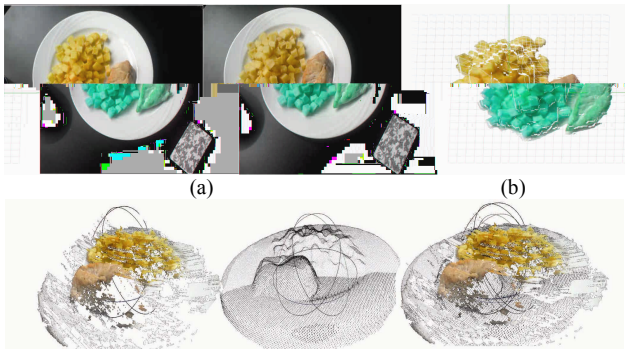

Figure 3. Evaluation process for the reconstruction, (a) image pair, (b), (c) point cloud, (d) ground truth, and (e) alignment.

Assuming that the different food items of the meal image are already segmented, each food segment is projected on the 3D model to calculate the corresponding volume. The volume can be defined as the integral of the distance between the surface of each segment and either plate, or the table (Fig. 1f). The former is identified by the reference object, the latter by its rim and its reconstructed shape.

III. Evaluation Procedure

A two-step evaluation procedure was followed: i) evaluation of the 3D reconstruction and ii) evaluation of volume estimation. In both cases, segmentation was done manually on one image before reconstruction to prevent eventual errors of an automatic step, and only the dense 3D models were used. For both steps the images were taken from a variety of viewpoints using a standard smart phone, generating many different camera configurations. Pairs of views were selected at random in each image set to build the 3D models and estimate the volume. Image resolution was readily translates into error bounds for the volume estimation artificially reduced to one Megapixel to reduce computational time. Examples of image pairs with the corresponding 3D models are presented in Fig. 2. The complete procedure was run multiple times on a desktop computer equipped with a 64-bit Intel i7-3770 (3.4 GHz, 16GB RAM).

A. Evaluation of 3D reconstruction

The evaluation of 3D reconstruction was carried out using real served food. Ground truth was obtained using the Kinect Fusion algorithm, which is sufficiently accurate for verification purposes [12], [13]. The pairs of 3D models were aligned using the Point Cloud Library [14]; one such alignment can be seen in Fig. 3.

In line with the assumptions made by this work on surface types, two error metrics were used: the Absolute Height Difference (AHD) and the Height Difference (HD) between corresponding points of the aligned point clouds (Fig. 4). The AHD has for lower bound the geometric distance,

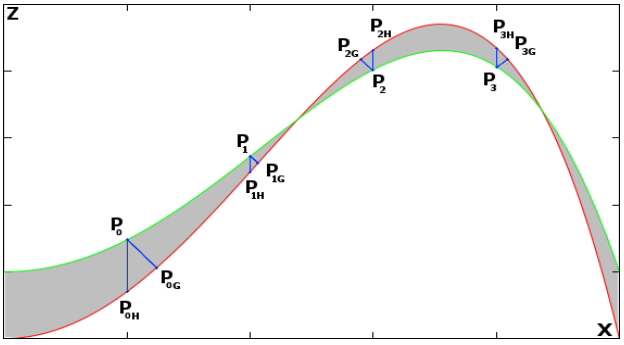

Figure 4 – Red, green: aligned height curves, $P_{i(G,H)}$: corresponding points using the geometric distance and AHD respectively, grey: integral of HD

and it readily translates into error bounds for the volume estimation. The sum of AHD is the sum of absolute differences - widely used in image patch comparison. The integral of HD over the surface is the volumetric difference between the surfaces.

B. Evaluation of volume estimation

In order to evaluate the algorithm's ability to estimate the volume, a series of food models (dummy foods) was used. Each model corresponds to common food items and has the same color, shape, and texture. Due to their artificial nature, the volume of these items is known. The set of models included single items and collections of those placed alone or in combinations on a plate. Some of the samples with reconstructed 3D shapes are presented in the Fig. 2.

IV. Experimental Results and Discussion

A. Evaluation of 3D reconstruction

Table I summarizes the results for three types of food objects, each with special characteristics of texture and shape. Qualitatively, weak texture is often characterized by low gradient magnitudes in the image, difficult stereo matching, and strong continuity constraints. Similarly, shape can be characterized by the height variance, where greater surface variance requires weaker continuity constraints.

From Table I it can be seen that the average AHD was 5mm or less overall. While this result indicates that there is a lot that can still be improved on, the average distance, corresponding to a more volumetric type of error, remains very low, indicating greater stability in volume estimation that in shape.

TABLE I
Results of Shape Evaluation. Average ± Standard Deviations values are presented
(AHD: Absolute Height Difference, HD: Height Difference)

Texture	Shape	AHD (mm)	HD (mm)
Strong	Medium	3.28 ± 0.62	-0.03±0.58
Weak	High	5.3 ± 1.32	1.23±0.88
Mixed	Low	3.68 ± 0.91	-0.72±0.8

TABLE II
RESULTS OF VOLUME EVALUATION AND CHO ESTIMATION. AVERAGE ± STANDARD DEVIATIONS VALUES ARE PRESENTED

Dummy food item	Real Volume (ml)	Estimated Volume (ml)	Estimated CHO (g)	Real CHO (g)	CHO Error (g)
Chicken	75	72 ± 12.1	0	0	0
Spaghetti	375	373.6 ± 21.4	114.9± 6.59	115.4	4.9±4.09
Fusilli	125	123.1 ± 17.3	29.01±4.06	28.6	3.45±1.84

B. Evaluation of volume estimation

Table II contains the statistical data of the volume estimates evaluation. It can be observed that the complexity of the surface place an important role in the error values, where higher complexity generates greater errors. This can be partly justified by the difficult of matching complex surfaces in stereo vision. Lack of texture remains a major source of errors. However, visual evaluation showed that our modification of the stereo matching method have greatly reduced the problem.

To obtain a final picture of the system's usability we have calculated the CHO content for each dish based on the estimated volumes by using the USDA Food and Nutrient Database [15]. From Table II can been seen that the system was able to estimate the CHO content with inaccuracy in the order of 5 g, which is below the threshold affecting the postprandial glucose profile in individuals with T1D.

V. CONCLUSION

In this paper, an algorithmic framework for 3D reconstruction and volume estimation of various food items contained on a plate was presented. Failure rate was low in the experiments, while the current results have shown that two-view geometry functions well on most food shapes, and thus can be applied in practical food management applications, putting minimum constraints on the user. With computational time below 15 seconds per image pair, the potential speed improvements, and hardware specific optimizations, the method is a good candidate for a mobile port. Future work on the problem will include: i) additional constraints on epipolar geometry, and distortion robust dense matching for improved accuracy ii) Shape quality can be further improved through the use of low level shape priors, iii) computational time can be reduced by automatically and/or dynamically choosing the desired shape reconstruction accuracy and density, and iv) automatic segmentation of food contents using shape and colour.

REFERENCES

[1] J. Chae, I. Woo, S. Ye Kim, R. Maciejewski, F. Zhu, J E. Delp., Carol J. Boushey,b and D. S. Ebert, "Volume estimation using food Specific shape templates in mobile image-based dietary assessment," *in Proc SPIE*, 7873: 78730K, 2011.

[2] Chen, H. C., Jia, W., Yue, Y., Li, Z., Sun, Y. N., Fernstrom, J. D., & Sun, M. (2013). Model-based measurement of food portion size for image-based dietary assessment using 3D/2D registration. *Measurement Science and Technology,* 24(10), 105701.

[3] M. Puri, Zhu Z., Q. Yu, A. Divakaran, and H. Sawhney, "Recognition and volume estimation of food intake using a mobile device*," in 2009 Workshop on Applications of Computer Vision (WACV)*, pp. 1 –8, December, 2009.

[4] F. Kong, J. Tan. "DietCam: Automatic dietary assessment with mobile camera phones." *Pervasive and Mobile Computing 8.1* (2012): 147-163.

[5] C. E. Smart, B. R. King, P. McElduff, C. E. Collins "In children using intensive insulin therapy, a 20-g variation in carbohydrate amount significantly impacts on postprandial glycaemia", *Diabet Med.,* vol. 29, no. 7, pp. 21-24, 2012.

[6] O. Chum, J. Matas, J. Kittler, "Locally optimized RANSAC," in *Pattern Recognition Lecture Notes in Computer Science (B. Michaelis, G. Krell, Eds.), vol*. 2781, pp. 236-243, 2003.

[7] J. Matas, O. Chum, "Randomized RANSAC with Tdd test," *Image and Vision Computing*, vol. 22, no. 10, pp. 837-842, 2004.

[8] J. T. Botterill, S. Mills, R. Green, "Fast RANSAC hypothesis generation for essential matrix estimation," in *International Conference on Digital Image Processing: Techniques and Applications,* pp. 565 – 569, 2011.

[9] R.I. Hartley, A. Zisserman, *Multiple view geometry in computer Vision.* Cambridge University Press (3rd Ed.), 2003.

[10] M. Pollefeys, R. Koch, L. Van Gool, "A simple and efficient rectification method for general motion," in *Proc. International Conference on Computer Vision*, pp. 496-501, 1999.

[11] G. Van Meerbergen, M. Vergauwen, M. Pollefeys, L. Van Gool. "A hierarchical symmetric stereo algorithm using dynamic programming," *International Journal on Computer Vision*, vol. 47, no. 1-3, pp. 275-285, 2002.

[12] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon, "KinectFusion: Real-time dense surface mapping and tracking," *in 2011 10th IEEE International Symposium on Mixed and Augmented Reality* vol. 7,pp. 127–136, 2011

[13] S. Meister, P. Kohli, S Izadi, M. Hämmerle, C. Rother, D. Kondermann, "When Can We Use Kinect-Fusion for Ground Truth Acquisition?" *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, Workshop on Color-Depth Camera Fusion in Robotics*, 2012.

[14] RB Rusu, S Cousins, "3D is here: Point Cloud Library (PCL)," *IEEE International Conference Robotics and Automation (ICRA)*, 2011.

[15] M.D. Beltsville USDA Food and Nutrient Database for Dietary Studies, 3.0., *U.S. Dep of Agriculture, Agricultural Research Service, Food Surveys Research Group,* (2008)