A Multimedia Database for Automatic Meal Assessment Systems

Dario Allegra[1], Marios Anthimopoulos[2,3], Joachim Dehais[2], Ya Lu[2], Filippo Stanco[1], Giovanni Maria Farinella[1], and Stavroula Mougiakakou[2,4(✉)]

[1] Department of Mathematics and Computer Science, University of Catania, Catania, Italy
{allegra, fstanco, gfarinella}@dmi.unict.it
[2] ARTORG Center for Biomedical Engineering Research, University of Bern, Bern, Switzerland
{marios.anthimopoulos, joachim.dehais, ya.lu, stavroula.mougiakakou}@artorg.unibe.ch
[3] Department of Emergency Medicine, Bern University Hospital, Bern, Switzerland
[4] Department of Endocrinology, Diabetes and Clinical Nutrition, Bern University Hospital, Bern, Switzerland

Abstract. A healthy diet is crucial for maintaining overall health and for controlling food-related chronic diseases, like diabetes and obesity. Proper diet management however, relies on the rather challenging task of food intake assessment and monitoring. To facilitate this procedure, several systems have been recently proposed for automatic meal assessment on mobile devices using computer vision methods. The development and validation of these systems requires large amounts of data and although some public datasets already exist, they don't cover the entire spectrum of inputs and/or uses. In this paper, we introduce a database, which contains RGB images of meals together with the corresponding depth maps, 3D models, segmentation and recognition maps, weights and volumes. We also present a number of experiments on the new database to provide baselines performances in the context of food segmentation, depth and volume estimation.

1 Introduction

Automatic diet assessment refers to the use of information technology for the ad-hoc translation of food intake into nutrient information in an accurate and intuitive way. Over the last years there have been a number of systems that use visual meal information to output nutrient content, mainly calories and carbohydrates [1–5], with only few of them being validated by end-users [6, 7]. Typically, once the visual information is available, a number of computer vision steps is executed: food detection, segmentation, recognition, and volume estimation. By knowing the food type and its volume and by using food composition databases the contained nutrients are estimated. Key element in the development and technical validation of the computer vision steps is the

S. Battiato et al. (Eds.): ICIAP 2017 International Workshops, LNCS 10590, pp. 471–478, 2017.
https://doi.org/10.1007/978-3-319-70742-6_46

data availability. However, the currently available food image datasets addresses needs related to the food recognition step.

Scope of the paper is to introduce a database that contains annotated and labelled RGB and RGB-D images from 80 different central-European meals served on a round dish accompanied by accelerometer data. Each meal consists of two to four different food items (e.g. vegetables, meat) of know weight, volume and nutrient composition. The newly introduced database offers resources to improve the current methods, compare among different approaches and hopefully progress the field of automatic diet assessment.

2 Food Image Datasets

One of the first datasets to address food recognition is the PFID [8]. It includes 4545 still images, 606 stereo image pairs, and 303 videos that cover a 360° angle around the food. It contains meals from 11 fast food restaurants that belong to 101 different food categories. In [9], the UECFOOD-100 was proposed: a dataset, which includes 9060 Japanese food images across 100 classes. This dataset was extended in [10] with 156 new classes (UECFOOD-256). In [11], a dataset of 50 classes with 100 Asian food images per class was introduced. The authors also presented preliminary results on food quantity estimation by using depth maps acquired through a Microsoft Kinect. In [12], the UNICT-FD889 was presented. The dataset intended to be used for near duplicate image retrieval and includes 3583 food images of real-life meals belonging to 889 classes, whereas in [13] the dataset was extended to 1200 classes and 4754 images. In [14], two new datasets to address food vs non-food classification were introduced, while in [15], a large-scale dataset with 101 classes and 1000 images per class was proposed (Food-101). The dataset was created by downloading images from the website foodspotting.com. This dataset was then partially labeled and annotated to perform food segmentation and recognition. The same 101 classes have been considered in the UPCM Food-101 dataset [16], which includes images combined with a textual description. A smaller dataset named FooDD has been proposed in [17] and consists of 3000 images of various meals acquired in restaurants under different illumination conditions. Finally, in [18], a dataset containing food images from, and geolocation of, six restaurants in Asia was presented.

3 Data Collection and Processing

Each of the 80 meals was placed on a table with a fully visible reference card next to it for color and geometric calibration. The acquisition procedure was conducted in the environment of a laboratory following two setups: (i) constrained and (ii) unconstrained. For each setup, the following systems were used: Intel® RealSense™ Camera SR300 and GoCARB App [7] installed on a Samsung Galaxy S4. Finally, the LG Nexus 5X was used to get a 3D multiview reconstruction used as ground truth for computing the food items' volume.

3.1 Constrained Setup

The dish was placed in a small table with a rotating bracket mount with limited degrees of freedom, in order to control distance and angle. The acquisition device was attached at the top of the bracket. Data were acquired at two different distances (40 cm and 60 cm) and four angles (0°, 30°, 60°, 90°). Thus, for each dish a total of eight captures / device was acquired.

- *Intel® RealSense™ Camera SR300:* From the depth sensor, we got four different types of images per capture:
  1. A 24-bit RGB image at 1920 × 1080;
  2. A 16-bit depth 640 × 480 image, where the pixel values is the distance from the sensor in tenth of millimeters;
  3. A depth image aligned with the RGB one;
  4. An RGB image masked with the related aligned depth map.
- *GoCARB App installed in a Samsung Galaxy S4:* From the GoCARB system for each capture we get a 4128 × 3096 RGB image and the information about calibration, as well as the gravity vector.

3.2 Unconstrained Setup

In the unconstrained setup, the device was placed freely in front of the dish and data were acquired at a randomly chosen distance and angle in the range of 40 cm to 60 cm and 45° to 90° respectively.

- *Intel® RealSense™ Camera SR300:* From the depth sensor, 200 consecutive RGBD frames at 10 fps were captured.
- *GoCARB App installed in a Samsung Galaxy S4:* Three image pairs were captured, each of them with the characteristics mentioned in the constrained setup.

Finally, approximately 50 images with resolution 4032 × 3024 were captured from all possible angles above the table (360° view) using the LG Nexus 5X. The images were used to build the ground truth 3D model of each meal. The acquisition information is summarized in Table 1.

3.3 Data Processing

Image labeling and annotation: For a subset of the acquired RGB and RGB-D images, segmentation and recognition maps are provided after manual manipulation. Details are presented in Table 1, while a sample of the proposed database in given in Fig. 1.

Ground truth estimation: The set of photos obtained with the LG Nexus 5X has been used to create a 3D reconstruction of the dish, through the online Autodesk Recap 360 service. The resulting 3D models were manually cleaned, rotated to a horizontal alignment, and rescaled by using the real size as obtained from the calibration card. In this clean model, we manually separated the food items. Hence, we have computed the individual volumes, in order to use it as ground truth for volume estimation algorithms.

Table 1. A summary of the acquired RGB and RGB-D data, along with the provided maps. For each meal served on a round dish the weight and volume of each food items is available, as well as information from smartphone's accelerometer.

Sensors	Setup	Images	Distance (cm)	Angle	Maps	
					Recognition	Segmentation
Intel® RealSense™ Camera SR300	Constrained	8 RGB-D	40	0°	-	
				30°	-	
				60°	X	1
				90°	X	1
			60	0°	-	
				30°	-	
				60°	X	1
				90°	X	1
	Unconstrained	200 RGB-D	[40–60]	[45° –90°]	X	2
Samsung Galaxy S4 (using GoCARB)	Constrained	8 RGB	40	0°	-	
				30°	-	
				60°	X	1
				90°	X	1
			60	0°	-	
				30°	-	
				60°	X	1
				90°	X	1
	Unconstrained	6 RGB	[40–60]	[45° –90°]	X	1
LG Nexus	Unconstrained	∼50 RGB	[40–60]	360° view	X	1
Total/ dish		**~272 (208 RGB-D; ~64 RGB)**			X	**12**
Total for the 80 dishes		**21807 (16640 RGB-D; 5167 RGB)**			X	**960**

4 Baseline Methods

The images of the proposed database were used to benchmark some state of the art methods for food segmentation, depth and volume estimation.

4.1 Segmentation

Food segmentation is a challenging task due to the great variability in food types, shapes and colors. Here, we investigate whether the use of depth-map information could improve the segmentation result. To this end, we applied a method similar to [19] and compared the results with and without considering the depth as input. The method consists of two main steps: border maps extraction by a convolutional neural network

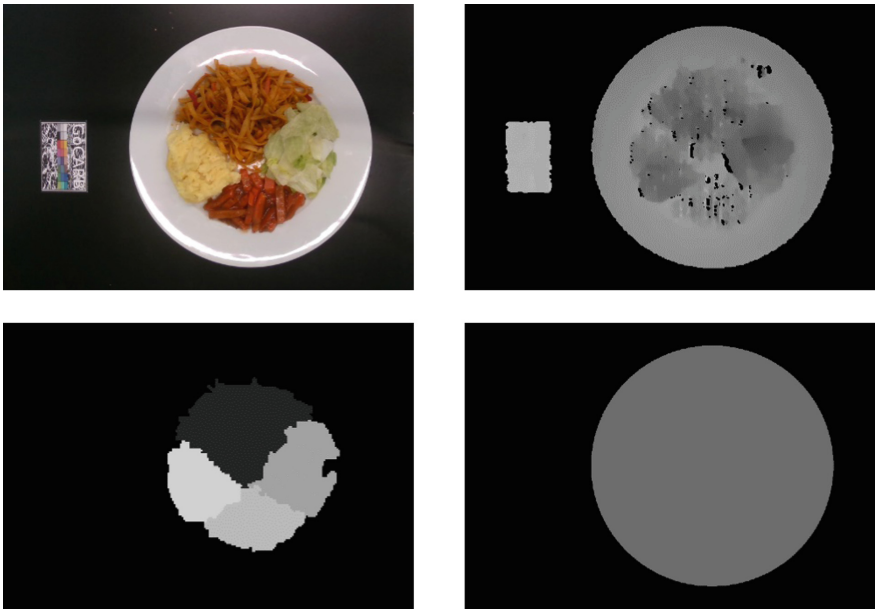

Fig. 1. On the top-left, the RGB acquisition performed with Real Sense at 40cm and 90° (top view). On the top-right, the corresponding depth map. On the bottom-left, the segmentation map for the image in the top-left. On the bottom-right, the plate map for the same images, got under the same constrains. Intensities were rescaled for visualization.

(CNN) and region growing segmentation. In our experiments, we altered the first step by utilizing different CNN architectures (SegNet [20] and U-Net [21]) and using the depth map as additional input. Specifically, the top view (40 cm and 90°) acquisition performed with Intel® RealSense™ Camera SR300 was used, after being inverted to represent the distance from the table. Finally, all the data have been normalized and automatically cropped to 256 $\times$ 256 by employing the plate map. We used 60 images for training, 10 for validation and 10 for testing. To increase the image variability, we augmented the dataset by considering two flips and four rotations. The results, confirm that depth-map information can reduce the error for borders extraction step and consequently for segmentation. Online augmentation (column AUG RGB-D in Table 2), has been performed by randomly modifying for each image at each iteration of the training data. Specifically, we add a random number from a normal distribution with mean 0 and standard deviation 0.01, to the color channels, while we multiply the depth map with a random number with mean 1 and standard deviation 0.1. The metrics used to assess the performance are the same used in [19]. We tested different architectures (SegNet, Unet), loss functions (mean square error - MSE and mean absolute error - MAE) and batch normalization strategy (per feature-map, mode 0; per batch, mode 2). Best results have been achieved with Unet, with no batch normalization e by training the CNN with online augmentation. As expected the best result has been obtained with depth information, moreover data augmentation with the smallest standard deviation has increased the training generalization.

Table 2. Segmentation results (MSE: Mean square error; MAE: Mean absolute error).

CNN	Loss	AUG RGB-D	RGB		RGB-D	
			Min Fscore	Total Fscore	Min Fscore	Total Fscore
Segnet	MSE	No	0.7329	0.9326	0.7059	0.9288
Unet	MAE	No	0.6889	0.9268	0.7351	0.9342
Unet	MSE	No	0.6875	0.9247	0.7332	0.9328
Unet	MAE	Yes	0.6893	0.9281	**0.7426**	**0.9369**

4.2 Depth Estimation

Calculating the depth of a food image is a significant component in understanding the 3D geometry of a meal, which is essential for food volume estimation. However, depth prediction is usually performed by stereo images or motion as in [22]. Here, we present a method that performs depth estimation by using just one RGB image, as input to a CNN. The method is inspired by [23], although some modifications have been made to adapt it to our problem. The considered task focuses on the estimation of the depth of near distance objects (within 1 meter), whereas the scenarios in [23] range within several meters. The dataset is composed by two images per dish acquired by the Intel® RealSense™ Camera SR300 in unconstrained setup: 60 dishes used for training, 10 for validation and 10 for testing, resulting in 120, 20 and 20 images, respectively. The training data were augmented by flipping the images left-to-right. The chosen network architecture is similar to Segnet-Basic [20], which consists of four encoding and four decoding convolutional layers. Each convolutional layer has 64 kernels and is followed by batch normalization and a ReLU activation, while the last layer uses the sigmoid activation function as a loss function, we use the mean absolute difference (MAD) between the estimated depth map and the ground truth from the depth sensor. For optimization, we used Adam with learning rate of 0.0005.

Table 3 reports the quantitative comparison of the depth prediction on the proposed dataset, where only the pixels inside the plate are evaluated. Apart from MAD value, the absolute relative difference (ARD) with respect to ground truth is also provided for the sake of clarity. As expected, in food depth prediction scenario, the result obtained using standard algorithm of [23] shows relatively poor performance. However, by using the proposed method, the performance is significantly improved. For further demonstration, the prediction result performed by our method is shown in Fig. 2(b), revealing a good agreement with the ground truth depicted in Fig. 2(a).

Table 3. Comparison on the proposed dataset (MAD: Mean absolute difference; ARD: Absolute relative difference).

Method	MAD (mm)	ARD (%)
NYU [23]	37.09	7.53
Proposed	**8.64**	**1.76**

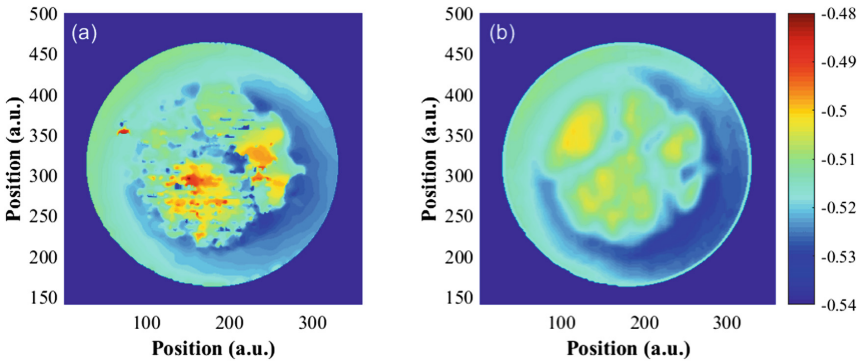

Fig. 2. (a): Depth map captured by the Intel® RealSense™ Camera SR300; (b): Depth map predicted by the proposed network. Color bar is in meter. (Color figure online)

4.3 Volume Estimation

Knowing the food volume is critical to estimate its nutritional value. In this experiment, we compare the performance of the GoCARB system in two different scenarios. As first we estimate the volume of each food item by reconstructing a 3D model as in [22]. The second experiment is aimed to assess the importance of depth map in volume estimation. We replace the depth estimation step as calculated in [22] with the depth obtained from the RGB-D images captured by depth sensor. In this case, we have to estimate the vertical direction and the table plane from the depth map to calculate the volume. To do so, we modify the table plane estimation method of [22]. First, we detect the plate through RGB channels, then sample the depth map at its border, and fit a plane to the selected points to find the ellipse plane. To find the table plane, we select all the points outside of the plate, and shift the ellipse plane to their modal height. To measure the performance the mean absolute percentage error, as defined in [22], was used.

In these conditions, the average error using stereo reconstruction was 13.8%, and 14% using RGB-D images. The two methods provide comparable results, however it has to be noted that the RGB-D sensor baseline (distance between the two elements of the stereo reconstruction module) is quite small, reducing its accuracy, while already developed algorithms were employed without any prior optimization the specific problem. However, these results indicate that a monocular RGB-D image can replace stereo pairs for volume estimation without performance drop.

5 Conclusions

In this paper, we have introduced a new multimedia food database that contains images, depth maps, weight/volume measurements of served meals, nutrient content together with the corresponding annotations, labels and accelerometer data. Furthermore, the results of some baseline methods on food segmentation and depth/volume estimation have been presented.

References

1. Merler, M., et al.: Snap, Eat, RepEat: a food recognition engine for dietary logging. In: Proceedings of the MADiMa 2016, pp. 31–40 (2016)
2. Myers, A., et al.: Im2Calories: towards an automated mobile vision food diary. In: Proceedings of the ICCV 2015, pp. 1233–124 (2015)
3. Anthimopoulos, M., et al.: Computer vision-based carbohydrate estimation for type 1 patients with diabetes using smartphones. J. Diabetes Sci. Technol. **3**, 507–515 (2015)
4. Zhu, F., et al.: The use of mobile devices in aiding dietary assessment and evaluation. IEEE J-STSP **4**(4), 756–766 (2010)
5. Miyazaki, T., et al.: Image-based calorie content estimation for dietary assessment. In: Proceedings of the IEEE ISM (2011)
6. Bally, L., et al.: Carbohydrate estimation supported by the GoCARB system in individuals with type 1 diabetes – a randomized prospective pilot study. Diabetes Care **40**(2), e6–e7 (2016). dc162173
7. Rhyner, D., et al.: Carbohydrate estimation by a mobile phone-based system versus self-estimations of individuals with type 1 diabetes mellitus: a comparative study. J. Med. Internet Res. **18**(5), e101 (2016)
8. Chen, M., et al.: PFID: Pittsburgh fast-food image dataset. In: Proceedings of the ICIP 2009, pp. 289–292 (2009)
9. Matsuda, Y., et al.: Recognition of multiple-food images by detecting candidate regions. In: Proceedings of the ICME 2012, pp. 25–30 (2012)
10. Kawano, Y., et al.: Automatic expansion of a food image dataset leveraging existing categories with domain adaptation. ECCV **2014**(8927), 3–17 (2014)
11. Chen, M.-Y., et al.: Automatic chinese food identification and quantity estimation. In: Proceedings of the SIGGRAPH Asia Technical Briefs, pp. 1–4 (2012)
12. Farinella, G.M., et al.: A benchmark dataset to study the representation of food images. ACVR **2014**(8927), 584–599 (2014)
13. Farinella, G.M., et al.: Retrieval and classification of food images. Comput. Biol. Med. **77**, 23–39 (2016)
14. Farinella, G.M., et al.: On the exploitation of one class classification to distinguish food vs non-food images. MADiMa **2015**(9281), 375–383 (2015)
15. Bossard, L., et al.: Food-101 - mining discriminative components with random forests. ECCV **8694**, 446–461 (2014)
16. Wang, X., et al.: Recipe recognition with large multimodal food dataset. In: Proceedings of the IEEE ICMEW 2015, pp, 1–6 (2015)
17. Pouladzadeh, P., et al.: FooDD: food detection dataset for calorie measurement using food images. MADiMa **2015**(9281), 441–448 (2015)
18. Herranz, L., et al.: A probabilistic model for food image recognition in restaurants. In: Proceedings of the ICME 2015, pp. 1–6 (2015)
19. Dehais, J., et al.: Food image segmentation for dietary assessment. In: Proceedings of the MADiMa 2016, pp. 23–28 (2016)
20. Badrinarayanan, V., et al.: SegNet: a deep convolutional encoder-decoder architecture for scene segmentation. In: Proceedings of the IEEE TPAMI (2017)
21. Ronneberger, O., et al.: U-Net: convolutional networks for biomedical image segmentation. MICCAI **9351**, 234–241 (2015)
22. Dehais, J., et al.: Two-View 3D reconstruction for food volume estimation. IEEE TMM **19** (5), 1090–1099 (2017)
23. Eigen, D., et. al.: Depth map prediction from a single image using a multi-scale deep network. In: Proceedings of the NIPS 2014, pp. 2366–2374 (2014)