SINGLE IMAGE-BASED FOOD VOLUME ESTIMATION USING MONOCULAR DEPTH-PREDICTION NETWORKS

Alexandros Graikos[1]**, Vasileios Charisis**[1]**, Dimitrios Iakovakis**[1]**, Stelios Hadjidimitriou**[1]
Leontios Hadjileontiadis[1,2]
graikosal@gmail.com, vcharisis@ee.auth.gr, dimiiako12@gmail.com,
stellios22@gmail.com, leontios.hadjileontiadis@ku.ac.ae

[1]Department of Electrical and Computer Engineering
Aristotle University of Thessaloniki
Thessaloniki, Greece

[2]Department of Electrical and Computer Engineering
Khalifa University of Science and Technology
Abu Dhabi, UAE

Extended Abstract

People's nutritional concerns have increased over the last few years [1] and the need for an automated dietary assisntant, who can accurately provide nutritional information on daily meals, is rapidly growing. An important feature of such an assistant, is having minimal user input as users are prone to making errors when estimating their food intake [2] and give up on using applications that require heavy interaction [3].

To incorporate this property, a plethora of approaches attempt to extract the nutrient information from images depicting the target meal. Both the specific food type and volume must first be determined and used in conjunction with food density and nutritional information databases [4, 5], to provide the nutritional data. The food type detection has been effectively solved using deep convolutional networks, also applied on other classification tasks. On the contrary, estimating the volume remains an intricate issue, with varied proposed solutions.

Color image-based food volume estimation approaches have mostly relied on well-established computer vision methodologies. In [6, 7], the volume estimation algorithms require adding a known-size object in the scene, such as a coin or human thumb, to define a pixel to distance scale and apply pre-defined primitive volume formulas. Similarly, in [8, 9], a calibration checkerboard is placed alongside the meal and is used in combining multiple views of the target food to construct an approximate 3D model. These approaches however, contraditct the minimal user input requirement, since the calibration object has to be carried and placed in the scene at all times. In addition, [8, 9] require the user to snap multiple shots or a video of the target meal, extending the interaction between assistant and user.

Aiming to overcome these limitations, Myers et al. [10] proposed training a deep convolutional network, to predict the depth of the input image and knowing the camera intrinsics, assign each pixel to a point in 3D space. The resulting point cloud is used to estimate the food volume, similar to 3D reconstruction methods, but now relying only on a single input image. However, training this network requires quality ground truth depth data of food images, which are hard to collect due to the increased cost of high-fidelity depth sensors.

The proposed method utilizes the latest advancements in monocular depth estimation to develop a more functional approach to food volume estimation (Figure 1). A depth-predicting network is trained using video sequences with camera motion, as suggested by Godard et al. [11] (Figure 2) and produces a depth map of the given image. The input is also passed through a semantic segmentation module that infers a segmentation mask, designating the food-associated pixels. Combining the predicted depth map and segmentation mask with the known camera intrinsics, a point cloud of the depicted meal is generated. The total volume is calculated as the volume contained between the food surface and an estimated plate surface, on which the meal is assumed to be placed upon. This approach, apart from removing most user burdens, also eliminates the need for scarce food depth image data, since training is done with easier to collect, food related videos.

In the current implementation, the depth network is trained using the EPIC-Kitchens 2018 dataset [12] which includes more than fifty hours of egocentric, food handling videos. This dataset is sub-optimal for the given task, since it is composed of general kitchen scenes and not food-specific videos, but was chosen in lieu of other food video datasets due to being the only to have sufficient camera motion. Other food video datasets, such as PFID [13], use a stationary camera and employ a rotating platter to capture multiple views of the meal, which would not generate a training signal in this architecture. The segmentation module used is composed of a VGG-16 network, modified to generate class activation maps [14], pre-trained on ImageNet and fine-tuned on Food-101 [15]. The GrabCut algorithm is applied to the output of the network, producing the output segmentation mask. The plate surface is estimated using the random sample consensus (RANSAC) algorithm. To calculate the total volume, the 3D food points are projected onto the plate surface and triangulated, where each triangle's volume is computed using the average vertex distance from the plate.

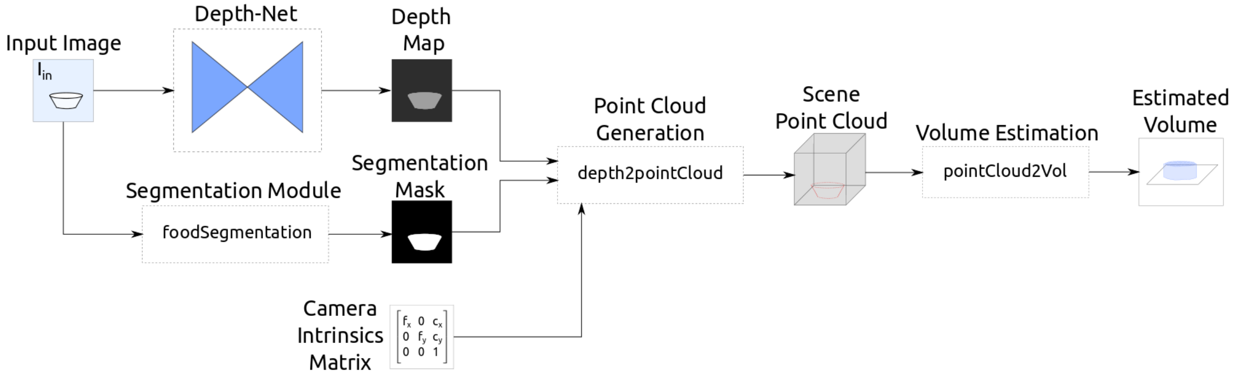


Figure 1: Proposed food volume estimation method diagram.

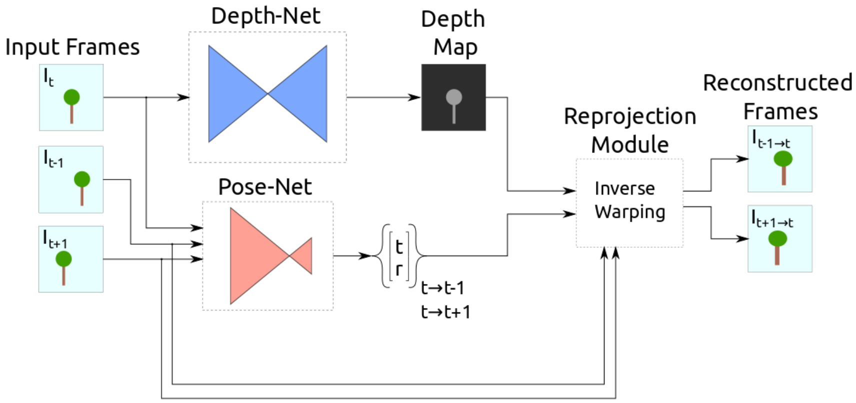
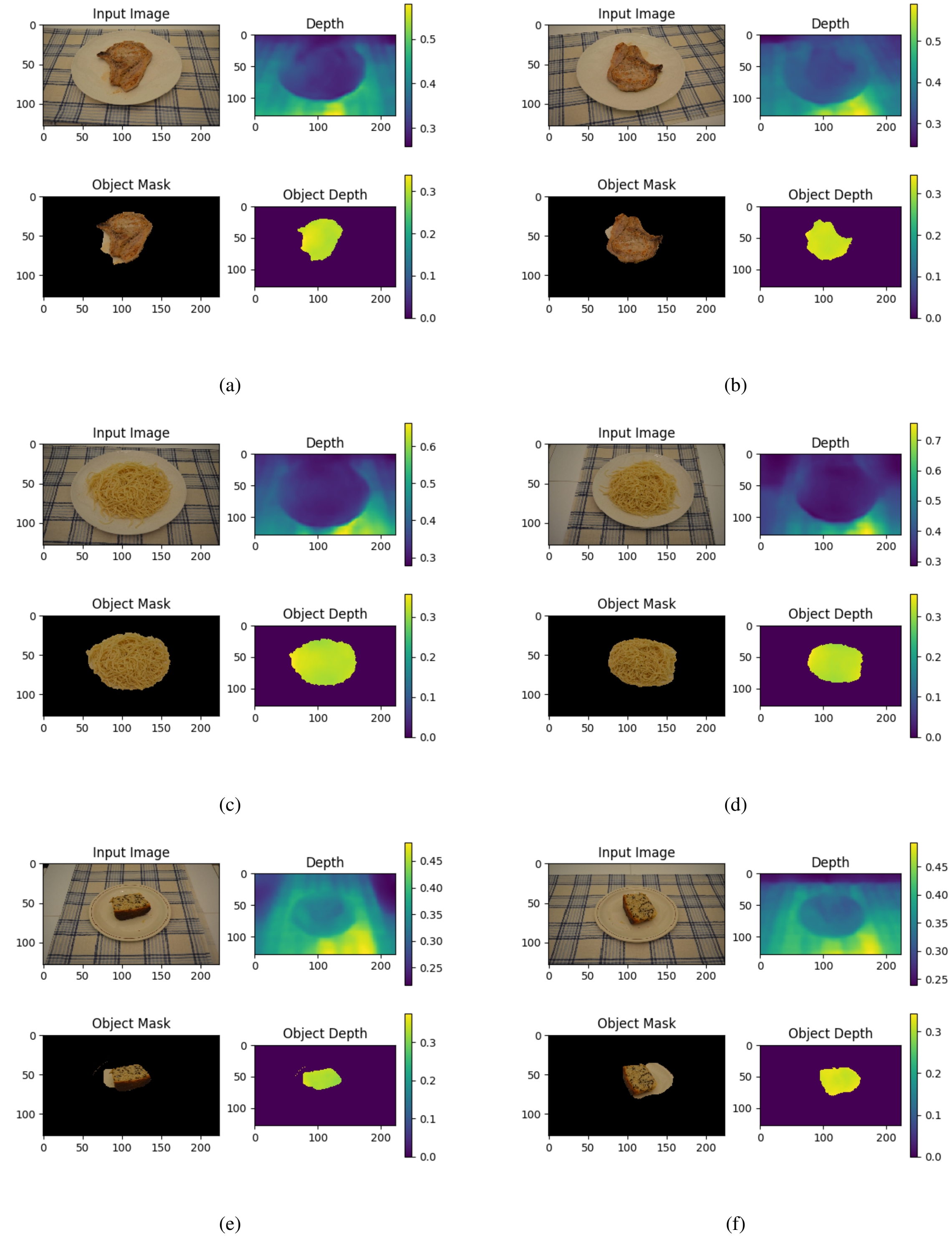


Figure 2: Depth network training architecture.

A sample of the tests conducted is presented in Figure 3, with the corresponding food volume measurements and estimation results summarized in Table 1. Two tests are presented for each case, with images taken from different angles. The overall test set mean absolute percentage error (MAPE) is 14.72%. Examining the inferred depth maps, they may lack in granularity as expected with the chosen training dataset, but nevertheless manage to capture the overall shape of the food object and provide adequate volume estimations.

		Volume Estimation	
Food	Measured Volume	Test 1	Test 2
Steak	297mL	308mL	293mL
Spaghetti	518mL	568mL	540mL
Cake	131mL	143mL	180mL

Table 1: Volume measurements and estimations for the test sample presented.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 3: Tests performed with Steak, Spaghetti and Cake. (a), (c), (e) correspond to column Test 1 of Table 1, whereas (b), (d), (f) to column Test 2. In (e), (f), a failure of the segmentation module includes parts of the plate into the segmentation mask, impairing to an extent the volume estimation.

Acknowledgment

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 817732.

References

[1] International Food Information Council. 2019 food and health survey, 05 2019. URL https://foodinsight.org/wp-content/uploads/2019/05/IFIC-Foundation-2019-Food-and-Health-Report-FINAL.pdf.

[2] Dale A. Schoeller, Linda G. Bandini, and William H. Dietz. Inaccuracies in self-reported intake identified by comparison with the doubly labelled water method. *Canadian journal of physiology and pharmacology*, 68 7: 941–9, 1990.

[3] Felicia Cordeiro, Daniel A. Epstein, Edison Thomaz, Elizabeth Bales, Arvind K. Jagannathan, Gregory D. Abowd, and James Fogarty. Barriers and negative nudges: Exploring challenges in food journaling. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 1159–1162, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3145-6. doi: 10.1145/2702123.2702155. URL http://doi.acm.org/10.1145/2702123.2702155.

[4] David Haytowitz U. Ruth Charrondiere and Barbara Stadlmayr. Fao / infoods databases, density database version 2.0, 2012. URL http://www.fao.org/3/ap815e/ap815e.pdf.

[5] Agricultural Research Service U.S. Department of Agriculture. Fooddata central, 2019. URL https://fdc.nal.usda.gov/.

[6] Yanchao Liang and Jianhua Li. Deep learning-based food calorie estimation method in dietary assessment. *CoRR*, abs/1706.04062, 2017. URL http://arxiv.org/abs/1706.04062.

[7] Rana Almaghrabi, Gregorio Villalobos, Parisa Pouladzadeh, and Shervin Shirmohammadi. A novel method for measuring nutrition intake based on food image. 05 2012. doi: 10.1109/I2MTC.2012.6229581.

[8] Chang Xu, Ye He, Nitin Khannan, Albert Parra, Carol Boushey, and Edward Delp. Image-based food volume estimation. In *Proceedings of the 5th International Workshop on Multimedia for Cooking &#38; Eating Activities*, CEA '13, pages 75–80, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2392-5. doi: 10.1145/2506023.2506037. URL http://doi.acm.org/10.1145/2506023.2506037.

[9] Hamid Hassannejad, Guido Matrella, Paolo Ciampolini, Ilaria De Munari, Monica Mordonini, and Stefano Cagnoni. A new approach to image-based estimation of food volume. *Algorithms*, 10, 06 2017. doi: 10.3390/a10020066.

[10] Austin Myers, Nick Johnston, Vivek Rathod, Anoop Korattikara, Alex Gorban, Nathan Silberman, Sergio Guadarrama, George Papandreou, Jonathan Huang, and Kevin Murphy. Im2calories: towards an automated mobile vision food diary. In *ICCV*, 2015. URL http://www.cv-foundation.org/openaccess/content_iccv_2015/papers/Meyers_Im2Calories_Towards_an_ICCV_2015_paper.pdf.

[11] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. *CoRR*, abs/1806.01260, 2018. URL http://arxiv.org/abs/1806.01260.

[12] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018.

[13] Mei Chen, Kapil Dhingra, Wen Wu, and Rahul Sukthankar. Pfid: Pittsburgh fast-food image dataset. volume 289-292, pages 289–292, 11 2009. doi: 10.1109/ICIP.2009.5413511.

[14] B. Zhou, A. Khosla, Lapedriza. A., A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. *CVPR*, 2016.

[15] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.