Single Image-Based Food Volume Estimation Using Monocular Depth-Prediction Networks

Alexandros Graikos[1], Vasileios Charisis[1(✉)], Dimitrios Iakovakis[1], Stelios Hadjidimitriou[1], and Leontios Hadjileontiadis[1,2]

[1] Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Thessaloniki, Greece
vcharisis@ee.auth.gr

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

Abstract. In this work, we present a system that can estimate food volume from a single input image, by utilizing the latest advancements in monocular depth estimation. We employ a state-of-the-art, monocular depth prediction network architecture, trained exclusively on videos, which we obtain from the publicly available EPIC-KITCHENS and our own collected *food videos* datasets. Alongside it, an instance segmentation network is trained on the UNIMIB2016 food-image dataset, to detect and produce segmentation masks for each of the different foods depicted in the given image. Combining the predicted depth map, segmentation masks and known camera intrinsic parameters, we generate three-dimensional (3D) point cloud representations of the target food objects and approximate their volumes with our point cloud-to-volume algorithm. We evaluate our system on a test set, consisting of images portraying various foods and their respective measured volumes, as well as combinations of foods placed in a single image.

Keywords: Food volume estimation · Monocular depth estimation · Food image processing · Deep learning

1 Introduction

People's nutritional concerns are on a rising trend over the last few years, as they become more and more involved with the health benefits, safety and environmental sustainability of the foods they consume [18]. Consequently the need for automated dietary assistants, which help users monitor their daily nutritional intake, has emerged, as they have the potential to facilitate the assessment and maintenance of a pursued diet. In order to boost adherence, a desired feature of such assistants, apart from accuracy, is requiring minimal user engagement, as users are prone to making errors when asked to estimate their own food intake [23] and giving up on using applications that require heavy interaction [7].

M. Antona and C. Stephanidis (Eds.): HCII 2020, LNCS 12189, pp. 532–543, 2020.
https://doi.org/10.1007/978-3-030-49108-6_38

To satisfy the minimal user-interaction requirement, a suggested approach is to ask users to take pictures of their meals [6], from which the nutritional information can be extracted in an automated fashion. In order to automate this process, both the different food types and their volumes must be first determined and used in conjunction with food density and nutrition databases [1,24], to approximate the total nutritional value of the meal. Whilst the food type detection has been adequately solved using widely-applied classification networks, trained on food-image datasets [3,15,21], estimating the volume remains a complex issue with varied proposed solutions.

The majority of existing works, employ well-established computer vision methodologies to estimate the target food volume. However, these approaches, for the most part, cannot be generalized and applied in complex environments, which is necessary since the assistant is expected to operate, most of the time, in non-ideal conditions. They also usually involve extra effort on the users' side, in the form of requiring them to take images from multiple views or to calibrate the camera, which contradicts the ease-of-use target originally set. On the contrary, learning-based methods may be able to surpass both the generalization and wearisome user input requirement problems, but rely heavily on the quantity and quality of the training data, which in most cases are scarce.

We aim to contribute towards a user-friendly, image-based dietary assistant, by introducing a system for estimating the volumes of the foods present in an image, using deep convolutional neural networks (CNNs). First, we train a depth prediction network, using easily obtained monocular video sequences related to food and an instance segmentation network on a labeled food-image dataset. During inference, the user provides a single image of his meal and combined with the predicted depth map, a 3D point cloud projection is created. The segmentation network identifies the different foods in the given image and by applying our point cloud-to-volume algorithm on each, we approximate the individual food volumes.

Our method incorporates the advantages of learning-based approaches, while avoiding the pitfalls related to the availability of high-quality training data. We demonstrate this property by showing how the fine-tuning of the depth prediction network, on smartphone-captured food videos, can significantly improve the volume estimations with relatively minimal effort. We finally present the performance of our system on a collected test set, where we compare the volumes estimated from images taken to ground truth volume measurements.

2 Related Work

2.1 Image-Based Food Volume Estimation

Color image-based food volume estimation approaches can be separated into two groups, depending on the underlying methodology used. The first group utilizes computer vision techniques, which require calibrating the camera and taking one or more images of the target food object, while in the second and more recent group, works are based on training deep convolutional networks to process the images and estimate the volume either directly or indirectly. It should be noted, that all

methods to be discussed in this section also entail locating the food in the input images. However, we will not go into detail regarding the segmentation methods used, since authors either employ well-established algorithms or omit it altogether.

Among the first group, [2,19] use common everyday items with known sizes, such as a coin or the user's thumb, to define a pixel-to-metric distance scaling factor and combine the top and side view of the food object to approximate the volume. In [26], a food shape dictionary was composed and during inference, the user is asked to take an image, having placed a calibration checkerboard in the scene, in order to fit one of the pre-defined shapes, of known volumes, to the food object.

Other methods in this group attempt to reconstruct a 3D model of the food object from multiple images and use it to approximate the volume. Hassannejad et al. [16] ask the user to take a short video of his meal, from which certain key frames are automatically extracted and combined to create a 3D point cloud representation. Similarly, in [26], authors also propose reconstructing the food object from multiple views, taken individually by the user, while in [9] they present a method to produce a point cloud, but this time only requiring two separate images. However, for all the reconstruction methodologies mentioned, a calibration checkerboard must be present in the scene at all times, to allow for proper scaling and the detection of the camera pose transformations between views.

This restriction also highlights the main issue with approaches based on computer vision techniques; the demand for multiple actions by the user. Having to constantly carry and place a certain item for calibration or taking multiple images from various views, can place a significant burden on users and potentially lead to drop-outs from using diet-monitoring applications.

Aiming to overcome these limitations, Myers et al. [22] proposed training a deep convolutional network to predict the depth of a user-provided image and, given the camera intrinsic parameters, project each pixel to a point in 3D space. From this projection, similarly to methods discussed above, a rough reconstruction of the food object is created and used to estimate the food volume. In [11], the authors move a step further and propose training a multi-task CNN to simultaneously predict food calories, category, ingredients and cooking instructions, ignoring the need for estimating the volume completely. However, the downside of both methodologies is that the result ultimately depends on having enough quality ground truth data, which in each case are hard to collect. Food depth images need to be captured with a high-fidelity depth sensor, whereas gathering a dataset that annotates food images with calories, type, ingredients and cooking instructions is a time-consuming process.

2.2 Self-Supervised Depth Estimation

Recent research on monocular depth estimation has bypassed the obstacle of ground truth depth data scarcity, by training the network in a self-supervised manner [25]. This translates to utilizing the depth predictions in reconstructing other available views of the input scene and computing the training loss on these images instead of the depth. For example, [13,25] process an image from a stereoscopic pair to predict a disparity map and reconstruct the other view,

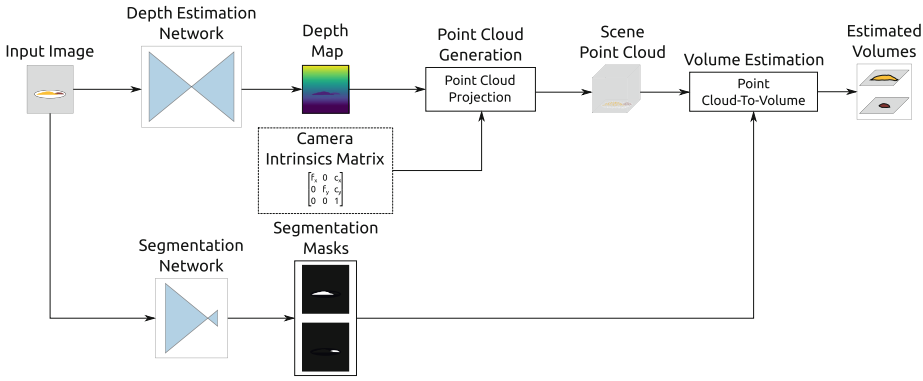


Fig. 1. Proposed system architecture.

on which the loss is computed. More recently, [14,27] reconstruct for each frame in a video, the previous and next frames, using the predicted depth map and camera pose transformations, eliminating the need for calibrated images at input. Formulating the loss for self-supervised training enables the gathering of training data with regular camera sensors and therefore, reduces the overall cost and complexity of training the depth prediction network.

3 Method

The proposed food volume estimation system can be separated into three distinct parts: (a) the depth estimation network (b) the segmentation network and (c) the point cloud-to-volume algorithm. An overall illustration of the system is shown in Fig. 1.

3.1 Depth Estimation Network

The depth network architecture used in the proposed system is the one presented in the work of Godard et al. [14], in which they demonstrate training a depth estimation network using only monocular video sequences. In each training step, three consecutive frames I_{t-1}, I_t, I_{t+1} of the video are used. The depth prediction network infers a depth map D_t of the input frame I_t, while a pose estimation network generates the camera pose transformations $T_{t\to t-1}$, $T_{t\to t+1}$ between the current and its adjacent frames. The predicted depth map, pose transformations and known camera intrinsic matrix K are used to synthesize the center frame by sampling from the previous and next as

$$\begin{aligned} I_{t-1\to t} &= I_{t-1}\langle proj(D_t, T_{t\to t-1}, K)\rangle \\ I_{t+1\to t} &= I_{t+1}\langle proj(D_t, T_{t\to t+1}, K)\rangle, \end{aligned} \tag{1}$$

where *proj* is the coordinate projection described in [27] and $\langle\ \rangle$ the sampling operator. The final training loss is the sum of a photometric loss L_p between the synthesized and original images and a depth smoothness error L_s, given as

$$L = L_p + \lambda L_s. \tag{2}$$

This approach fails to generate a meaningful training signal when the true pose transformations $T^{gt}_{t\to t-1}$, $T^{gt}_{t\to t+1}$ are zero, since in this case, the predicted depth values do not affect the image synthesis whatsoever. This limits the videos we can train the network with to those having sufficient motion between frames, which is not as common among food videos, since they are mostly captured from stationary cameras.

We train the network on the EPIC-KITCHENS dataset [8], which includes more than fifty hours of egocentric food-handling videos. The selected cooking and meal preparation sequences were taken from six different kitchens, with the consideration of incorporating variations in the environment, actions and lighting conditions. This dataset, despite being sub-optimal for the task, with it focusing around the miscellaneous actions of preparing food in a kitchen, was chosen in lieu of other food-specific video datasets, due to being the only one to satisfy the camera motion limiation. For instance, in the Pittsburgh fast-food image dataset [4], the authors employ a stationary camera and a rotating platter to capture their videos, which would not be suitable for training with this architecture.

To further improve the depth predictions, we fine-tune the network weights on a dataset of food-specific videos we collected, which we will refer to as *food videos*. In total, 38 videos were captured on daily occasions, using commercial smartphone cameras and portray a number of different food types and environments. Training with this dataset further backs the argument on the ease of training depth estimation networks in a self-supervised manner, considering that gathering an equivalent dataset of food-image depth data would require substantially more resources and time.

3.2 Food Segmentation Network

In order to be able to locate the food objects present in a given image, we train an instance segmentation network on a dataset of labeled food images. The model utilized for this purpose is the Mask R-CNN [17], which extends the object detection Fast Region-based CNN [12] architecture, by additionally predicting individual segmentation masks for each object found. With this approach, we are able to discern between the different foods present in the input, including multiple instances of the same food type and produce a segmentation mask for each, to estimate their volume separately.

We initialize the network using weights pre-trained on the COCO dataset [20] and fine-tune for food object segmentation on the UNIMIB2016 dataset [5]. The later, is composed of 1,027 meal tray images, labeled with bounding boxes, food types and segmentation masks for the food objects depicted. Since we are not interested in predicting the specific food types, we ignore the food-type labels.

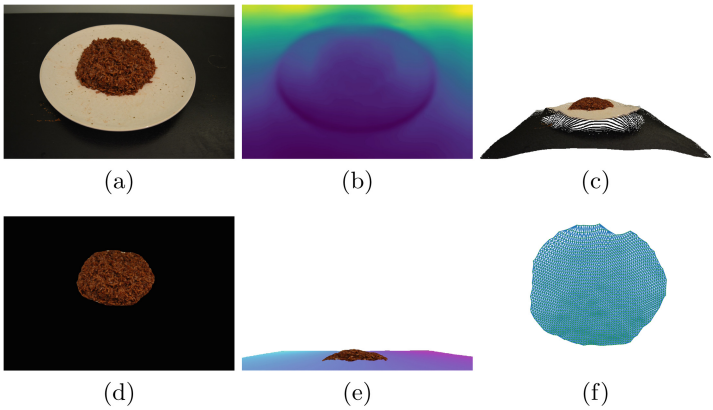


Fig. 2. System stages of estimating the volume of a food object. (a) Input image. (b) Depth prediction. (c) Point cloud representation of the input. (d) Generated segmentation mask. (e) Base plane estimation. (f) Triangulation.

3.3 Volume Estimation

Having inferred the depth map D of the input image and given the camera intrinsic matrix K, we project each pixel (x, y) to a corresponding point in 3D space, using its homogeneous coordinates and the inverse projection model

$$P_{xy} = K^{-1} \begin{bmatrix} x & y & 1 \end{bmatrix}^T D_{xy}, \tag{3}$$

to form a point cloud representation P. In turn, the segmentation masks generated by the instance segmentation network, are applied onto the image to distinguish between the different foods depicted and split P into subsets of food object points.

For each of these sets of points, we first remove any outliers using a statistical outlier removal (SOR) filter and then determine the base plane on which the food is placed upon by applying principal component analysis. The eigenvector of the least important component is set as the plane normal vector, while the plane is also adjusted to be at the bottom of the object. The final step for approximating the volume contained between the base and food is projecting the food points onto the plane and partitioning the covered area into triangles by computing an α-complex from the Delaunay triangulation [10]. The total volume is made up of the triangular prisms, defined by each triangle and the average distance of its vertices from their corresponding food points. The stages of estimating the volume of a food object in an image, as described above, are portrayed in Fig. 2.

This approach, only works in cases where the food is placed on top of a planar plate. We cannot estimate the volume of foods that are inside containers, such as a bowl, in which case the volume is not defined by the food object but instead by the container itself.

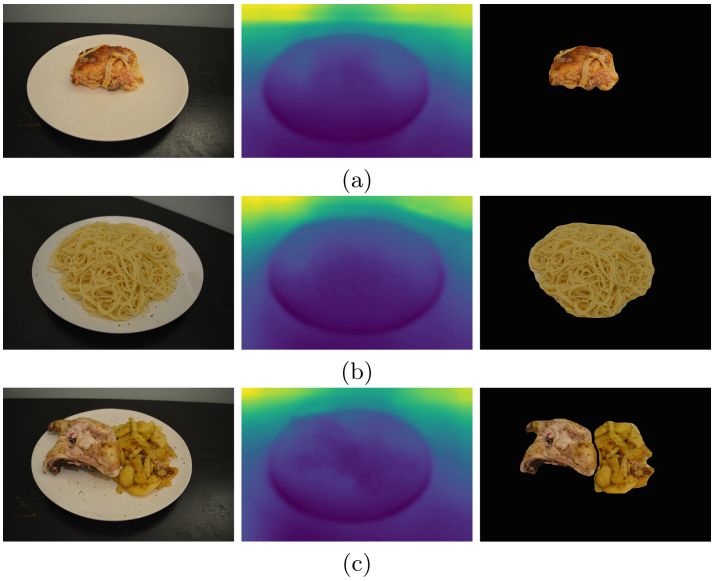


Fig. 3. Samples from the test set used for evaluation, along with the depth and segmentation mask predictions. (a) Sample from the *Souffle* food type. (b) Sample from the *Spaghetti* food type. (c) Sample from the *Potatoes2*/*Chicken2* food type.

4 Evaluation

4.1 Implementation Details

In our experiments, we trained the depth estimation network on an input resolution of 128×224 and a batch size of eight-frame triplets per step. We set the depth outputs to be in the range of 0.01 to 10 units and the smoothness term λ to 10^{-2}. We performed the data augmentations suggested by [14] on brightness, contrast, saturation, hue and left-right image flipping.

Training on the EPIC-Kitchens dataset, we used the sequences P01_10, P05_08, P10_04, P15_12, P20_01, P30_10, selecting frames from the original recordings with a stride of 10, to produce a total 42,066 frames to train on. We trained the network for 20 epochs with a learning rate of 10^{-4}, halved at epoch 15.

For our own *food videos* dataset, we extracted four frames per second for all 38 sequences, generating 1,712 frames to fine-tune the depth estimation network on. As before, we trained for 20 epochs, with the same initial learning rate and scheduling.

The depth predictions of the network are not on a metric scale, since there is no ground truth depth value fed to it during training with monocular videos. To overcome this, we apply the median ground truth rescaling proposed by [27], where the predicted depth map D is multiplied by a scalar

$$s = \frac{median(D^{gt})}{median(D)}. \tag{4}$$

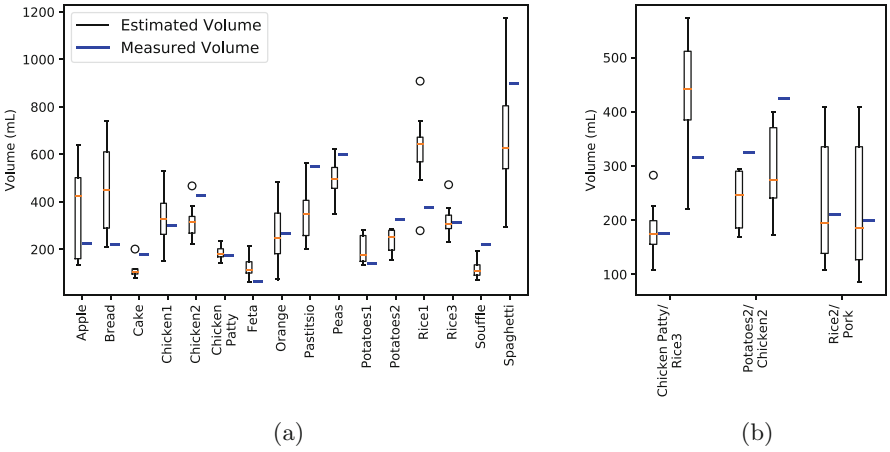


Fig. 4. Measured volume and distribution of system estimations on (a) the 16 test foods (b) the combined meals.

The median ground truth depth value $median(D^{gt})$ used for our experiments, is set as the estimated distance between the camera sensor and the food object, which we defined for all test cases at 0.35 m.

To train the Mask R-CNN network, we set the input batch size to 2, learning rate to 10^{-3} and performed data augmentations in brightness, contrast, saturation and hue, uniformly, in the ranges of ± 40, ± 0.2, ± 40, ± 20, respectively, as well as left-right and up-down flipping, with a probability 0.5. We initialized the network with pre-trained weights on the COCO dataset and trained on a split of 925 training and 102 validation inputs, with all non-packaged food classes aggregated into a single *food* class. The network was trained for eight epochs in total, freezing all but the top layers for the first three, allowing for steady adaptation of weights.

During volume inference, the camera intrinsic matrix is generated from the field-of-view angle of the camera sensor used, which we set at an estimated average value of 70°. The Z-Score for the SOR filter was computed from the maximum distance of each point to its neighbors and the α value for constructing the α-complex was set at 0.01.

4.2 Results and Discussion

To evaluate our system, we measured the volumes of 16 foods, using the water displacement method and captured eight images of each, from varying angles. The distributions of our system volume estimations, grouped by food type, are demonstrated, alongside the measured volumes, in Fig. 4 (a). In addition to that, we also present the mean absolute percentage errors (MAPE) of our estimations, per food, in Table 1 (a). Due to the low resolution of our measurement setup, we add a ± 20 mL measurement error to the ground truth volumes we present.

Table 1. Volume measurements and mean absolute percentage error (MAPE) of estimations for (a) the 16 test foods (b) the combined meals.

Food	Volume (mL)	MAPE
Apple	225	89.094
Bread	220	108.30
Cake	180	39.030
Chicken1	300	29.285
Chicken2	425	27.677
Chicken Patty	175	13.737
Feta	65	99.798
Orange	265	40.362
Pastitsio	550	37.283
Peas	600	18.512
Potatoes1	140	42.661
Potatoes2	325	26.925
Rice1	375	71.288
Rice3	315	15.852
Souffle	220	47.019
Spaghetti	900	35.395

(a)

Food	Volume (mL)	MAPE
Chicken Patty	175	21.235
Rice3	315	45.380
Potatoes2	325	26.790
Chicken2	425	30.766
Rice2	210	45.119
Pork	200	52.160

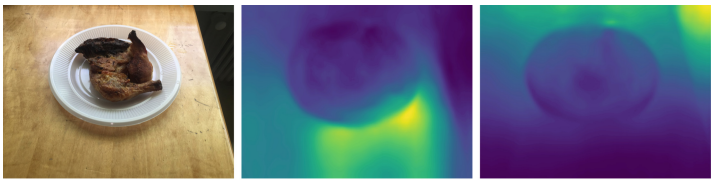
(b)

In cases where the segmentation network partitioned a single food object into portions that the object is made up of, we compensated by simply adding up the individual predicted volumes.

We also tested the ability of our system to estimate multiple volumes simultaneously, by arranging two different foods in a single plate and predicting their individual volumes from a single image of the combined meal. For this scenario, we evaluate our system on three meals of eight images each, with results presented in the same format in Fig. 4 (b) and Table 1 (b). Concerning the *Rice2* and *Pork* meal, we mixed the two foods together beforehand, to intentionally stress the segmentation network, and recorded any results where the foods were not separated, as having estimated the same volume for both. Therefore, their estimation distributions are similar. We present samples from our test set, along with the depth and segmentation mask predictions in Fig. 3.

In Fig. 5, we showcase the benefits gained when fine-tuning the depth prediction network on our own collected *food videos*. If we were to use the original prediction, we would struggle to create an adequate point cloud representation of the scene and subsequently estimate the volume, since it has reduced precision in the areas of peak interest, i.e., on and around the food object, and portrays several depth inaccuracies.

Although the depth network manages to infer accurate depth unit maps, the volume estimation algorithm also depends on applying the correct median distance rescaling, since it operates on metric depth values. We acknowledge this



(a) (b) (c)

Fig. 5. Comparison of predictions, before and after fine-tuning the depth network on the *food videos* dataset. (a) Input image. (b) Predicted depth before fine-tuning. (c) Predicted depth after fine-tuning.

weakness of our system and demonstrate it in the volume predictions for *Rice1*, where all but one images used were taken from a much closer distance than the one set for scaling. This resulted in noticeably larger volume estimations and exhibits the need to communicate effectively the camera to food object distance to the user, or bypass it completely by finding a way to extract the median depth automatically.

Our point cloud-to-volume algorithm is mostly hindered by its limitation to generalize, as we previously mentioned for most computer vision-based methodologies. The algorithm was designed for and operates best, when the target food object is stacked on top of a plate and most of it is visible from the angle the image is taken from. If this is not the case, the predicted base plane can be incorrectly placed and results in computing a volume that does not properly represent the depicted object. As an example, when estimating the volumes of *Apple* and *Orange*, the images that portray only a side of the objects will lead the algorithm to place the base plane perpendicular to the plate and therefore, only compute a slice of the total volume.

5 Conclusion

In this work, we presented our approach for image-based food volume estimation, aiming to contribute towards the development of a burden-free and accurate dietary assistant application. We capitalize on the latest advancements in monocular depth estimation, to show that using both the publicly available and our easily collected data, we are able to train a depth prediction network and employ it in estimating food volumes, requiring only a single image at inference. Nevertheless, there are still a plethora of issues to resolve before claiming success, but we firmly believe that following in the direction of learning-based methods paves the way for fruitful results.

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

References

1. U.S. Department of Agriculture, A.R.S.: FoodData central (2019). https://fdc.nal.usda.gov/
2. Almaghrabi, R., Villalobos, G., Pouladzadeh, P., Shirmohammadi, S.: A novel method for measuring nutrition intake based on food image. In: 2012 IEEE International Instrumentation and Measurement Technology Conference Proceedings, pp. 366–370. IEEE (2012)
3. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101 – mining discriminative components with random forests. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8694, pp. 446–461. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10599-4_29
4. Chen, M., Dhingra, K., Wu, W., Yang, L., Sukthankar, R., Yang, J.: PFID: pittsburgh fast-food image dataset. In: 2009 16th IEEE International Conference on Image Processing (ICIP), pp. 289–292. IEEE (2009)
5. Ciocca, G., Napoletano, P., Schettini, R.: Food recognition: a new dataset, experiments, and results. IEEE J. Biomed. Health Inform. **21**(3), 588–598 (2016)
6. Cordeiro, F., Bales, E., Cherry, E., Fogarty, J.: Rethinking the mobile food journal: exploring opportunities for lightweight photo-based capture. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pp. 3207–3216 (2015)
7. Cordeiro, F., et al.: Barriers and negative nudges: Exploring challenges in food journaling. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pp. 1159–1162 (2015)
8. Damen, D., et al.: Scaling egocentric vision: the epic-kitchens dataset. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11208, pp. 753–771. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01225-0_44
9. Dehais, J., Anthimopoulos, M., Shevchik, S., Mougiakakou, S.: Two-view 3D reconstruction for food volume estimation. IEEE Trans. Multimedia **19**(5), 1090–1099 (2016)
10. Edelsbrunner, H., Harer, J.: Computational Topology: An Introduction. American Mathematical Society, Providence (2010)
11. Ege, T., Yanai, K.: Image-based food calorie estimation using recipe information. IEICE Tran. Inf. Syst. **101**(5), 1333–1341 (2018)
12. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
13. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 270–279 (2017)
14. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3828–3838 (2019)
15. Hassannejad, H., Matrella, G., Ciampolini, P., De Munari, I., Mordonini, M., Cagnoni, S.: Food image recognition using very deep convolutional networks. In: Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management, pp. 41–49 (2016)
16. Hassannejad, H., Matrella, G., Ciampolini, P., Munari, I.D., Mordonini, M., Cagnoni, S.: A new approach to image-based estimation of food volume. Algorithms **10**(2), 66 (2017)

17. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)
18. International Food Information Council (IFIC) Foundation: 2019 Food and Health Survey (2019). https://foodinsight.org/wp-content/uploads/2019/05/IFIC-Foundation-2019-Food-and-Health-Report-FINAL.pdf
19. Liang, Y., Li, J.: Deep learning-based food calorie estimation method in dietary assessment. arXiv preprint arXiv:1706.04062 (2017)
20. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
21. Martinel, N., Foresti, G.L., Micheloni, C.: Wide-slice residual networks for food recognition. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 567–576. IEEE (2018)
22. Myers, A., et al.: Im2calories: towards an automated mobile vision food diary. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1233–1241 (2015)
23. Schoeller, D.A., Bandini, L.G., Dietz, W.H.: Inaccuracies in self-reported intake identified by comparison with the doubly labelled water method. Can. J. Physiol. Pharmacol. **68**(7), 941–949 (1990)
24. U. Ruth Charrondiere, D.H., Stadlmayr, B.: FAO/INFOODS databases, density database version 2.0 (2012) http://www.fao.org/3/ap815e/ap815e.pdf
25. Xie, J., Girshick, R., Farhadi, A.: Deep3D: fully automatic 2D-to-3D video conversion with deep convolutional neural networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 842–857. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_51
26. Xu, C., He, Y., Khannan, N., Parra, A., Boushey, C., Delp, E.: Image-based food volume estimation. In: Proceedings of the 5th International Workshop on Multimedia for Cooking & Eating Activities, pp. 75–80 (2013)
27. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1851–1858 (2017)