Human-mimetic Estimation of Food Volume from a Single-View RGB Image using an AI System

zhengeng yang
Hunan University
Hongshan Yu
Hunan University
Shunxin Cao
University of Pittsburgh
Wenyan Jia
University of Pittsburgh
Qi Xu
Huazhong University of Science and Technology
Ding Yuan
Beihang University
Hong Zhang
Beihang University
Zhi-Hong Mao
University of Pittsburgh
Mingui Sun (drsun@pitt.edu)
University of Pittsburgh https://orcid.org/0000-0001-7948-9205

Research

Keywords: Food Volume Estimation; Nutrition; Artificial Intelligence; Dietary Assessment; Deep Learning

Posted Date: April 28th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-23998/v1

Version of Record: A version of this preprint was published at Electronics on June 28th, 2021. See the published version at https://doi.org/10.3390/electronics10131556.

RESEARCH

Human-mimetic Estimation of Food Volume from a Single-View RGB Image using an AI System

Zhengeng Yang[1,2], Hongshan Yu[1], Shunxin Cao[3], Wenyan Jia[3], Qi Xu[4], Ding Yuan[5], Hong Zhang[5], Zhi-Hong Mao[3,6] and Mingui Sun[2,3,6*]

*Correspondence: drsun@pitt.edu
[2]Department of Neurosurgery, University of Pittsburgh, Pittsburgh, USA
Full list of author information is available at the end of the article

Abstract

Background: It is well-known that many chronic diseases are associated with unhealthy diet. Although improving diet is critical, adopting a healthy diet is difficult despite its benefits being well understood. Technology is needed that allows assessment of dietary intake accurately and easily in real-world settings so that effective intervention to manage overweight, obesity and related chronic diseases can be developed. In recent years, new wearable imaging and computational technologies have emerged. These technologies are capable of objective and passive dietary assessment with much simplified procedure than traditional questionnaires. However, a critical task is required to estimate the portion size (in this case, the food volume) from a digital image. Currently, this task is very challenging because the volumetric information in the two-dimensional images is incomplete, and the estimation involves a great deal of imagination, beyond the capacity of the traditional image processing algorithms.

Method: A novel Artificial Intelligent (AI) system is proposed to mimic the thinking of dietitians who use a set of common objects as gauges (e.g., a teaspoon, a golf ball, a cup, and so on) to estimate the portion size. Specifically, our human-mimetic system "mentally" gauges the volume of food using a set of internal reference volumes that have been learned previously. At the output, our system produces a vector of probabilities of the food with respect to the internal reference volumes. The estimation is then completed by an "intelligent guess", implemented by an inner product between the probability vector and the reference volume vector.

Dataset: The datasets utilized for model validation include: 1) two virtual food datasets produced by computer simulation, and 2) two real-world food datasets collected by us.

Results: The average relative volumetric errors of our AI method were less than 9% on both virtual datasets, and 11.7% and 20.1% , respectively, on the two real-world food datasets.

Discussion: We discuss: 1) the use of AI to estimate the "relative volume" of food in a plate, 2) the case of multiple foods in a plate, and 3) the potential of AI in advancing nutrition science.

Conclusion: Our AI system is able to use the same food volume estimation strategy as the human uses.

Keywords: Food Volume Estimation; Nutrition; Artificial Intelligence; Dietary Assessment; Deep Learning

Background

As of 2016, 39.6% U.S. adults were obese (BMI≥30) [1]. In order to control obesity and related chronic diseases, there is a pressing need to assess accurately the energy and nutrient intake of individuals in their daily lives. Traditionally, dietary assessment is conducted using self-report in which individuals report their consumed foods and portion sizes. Although this method is standard and has been utilized for decades, numerous studies have indicated that it is inaccurate and biased [2, 3]. In addition, self-report does not work well in Children [4].

With the developments of smartphones and wearable devices, dietary assessment can be performed without fully depending on individuals' memory and willingness to report their own intake. For example, Arab et al. [5] developed an automated image capture method to aid dietary recall using a mobile phone; Sun et al. [6] designed a wearable camera system called eButton for objective and passive dietary assessment; Farooq et al. [7] developed an eyeglass attachment containing an accelerometer and a camera to record dietary events automatically; Liu et al. [8] performed food intake monitoring using a sensor worn on top of an ear.

With automatic image capture by a wearable device, dietary assessment can be conducted objectively and passively in four steps: food detection, food recognition, volume estimation and nutrition content analysis. The first two steps have been studied using computer vision and pattern recognition methods, notably the recently developed deep learning methods [9, 10, 11]. The last step is usually implemented using an existing food database, such as the USDA Food Composition Database [12]. Although the technological tools in all four steps need further improvements, the third step (volume estimation) is currently the least developed due to a number of challenges involved, detailed below.

A food image is usually in the unit of "pixel", rather than a real-world unit (e.g., centimeter). As a result, a scale reference is required to determine the actual food size. For example, the size of an apple must be determined by comparing it with another object with a known size in the same image. Thus, many types of fiducial markers have been used as the scale reference, such as a checkerboard [13] and a credit card [3]. However, these fiducial markers must be carried by the individual and placed near the food before the individual starts eating, which is an unwelcome procedure usually difficult to implement. Secondly, for a computer to estimate food volumes effectively, it is necessary to provide enough three-dimensional (3D) information of the food. Unfortunately, much of the 3D information is lost in the imaging process where the food as a 3D object is projected to a 2D plane. For this reason, instead of using a single-view image, many researchers turned to the use of multiple images in different views [14, 15, 13, 3, 16]. However, this approach requires the individual to move either the camera or the food in the imaging process, complicating the research effort and possibly modifying the normal behavior of diet-intake. In addition, 3D reconstruction from multi-view images involves multiple challenges, such as accurate camera calibration, feature extraction, image registration, pose estimation, etc [17].

In another approach, a depth camera or a pair of cameras has been used to obtain images with depth information [18]. Although a depth image contains more 3D information, to acquire this type of images, the wearable device must be made

larger, heavier and more expensive to accommodate additional hardware and meet the requirement of increased power consumption, which affects the wearability and practical utility of the device.

With the success of deep learning based depth-estimation [19, 20], it has been demonstrated that certain 3D information can be inferred from a 2D single-view image. Thus, there is a recent trend of applying deep learning to food volume estimation using estimated depth from the RGB image (i.e., the regular image with red (R), green (G) and blue (B) as three primary colors). Typical algorithms include im2calories [21] and deepvol [22]. Although these algorithms have achieved a certain success in improving food volume estimation, estimating depth from a single RGB image is still a very challenging problem and the consensus for this method to achieve a high accuracy is pessimistic. In addition, the deep learning system requires an excessively large number of RGB images with known depth for network training, which are difficult to obtain in practice.

In this work, we present a human-mimetic approach to estimate food volume directly from a single-view 2D RGB image without using any supplemental 3D information. Our work is highlighted as follows.

1 Over the years, dietitians have used a popular portion size estimation method by comparing the observed food with a set of common objects of known volumes. We propose to use the AI technology to mimic this mental process. In the human case, the food size is first matched with the most similar size of a known object. In AI system, we use a similar strategy formulated as an image classification step; In the second step, while the human mentally fine-tunes the estimation by portioning with respect to the known object sizes, we mimics this process using an inner product between a probability vector and a reference volume vector.
2 To validate the effectiveness of our method quickly, we used two large-scale Virtual Food Datasets (VFD), constructed by computer simulation, of different volume ranges. Our method achieves a high accuracy with an average volumetric error less than 9% on both datasets.
3 To evaluate the real-world performance of our method, two real food datasets (RFD) are collected with different degrees of difficulties in estimation tasks. Our method achieves 11.7% and 20.1% relative volumetric errors on the easy and hard food datasets, respectively.

Method

Motivation

Instead of acquiring multiple images and reconstructing depth explicitly, we use only a single-view image to estimate food volume. Our approach to this difficult problem is to mimic the human thinking in volume estimation using the AI technology. Over the years, dieticians have used an intuitive method comparing the food (either in the physical world or in an image) of an unknown size with a number of sizes of common objects, such as a thumb tip, a golf ball, a deck of cards, and a baseball. The sizes of the objects close to the size of the food are mentally extrapolated to produce an estimate. This method, although not very accurate, is proven to be highly effective. Studies in psychology provide an explanation of the effectiveness

in terms of the Stroop effect where the size difference of two familiar objects in an image can be rapidly perceived when their sizes are congruent with those of the real-world [23]. In addition, over the process of evolution, the human becomes highly capable of not only selecting a particular object (e.g., a larger one) among the same type of objects of different sizes, but also estimating the size of one type of object in reference to to another object of a different type with a known size. This is illustrated in Fig. 1(a) where we can easily tell that the Food No. 1 appears larger than the Food No.2, assuming that the plates in the two images have the same sizes. We can also roughly estimate the volume of any one food in the pair if the actual volume of the other food is known, provided that the plates are of the same size (bottom row). The estimation is facilitated even further if more than one reference foods of known volumes are available (Fig. 1(b)), assuming that all plate sizes are identical.

These observations inspired us to adopt a human-mimetic strategy for food volume estimation from a single-view image using an AI system. This new strategy, shown in Fig. 2, consists of two steps. In the first step, our AI system roughly classifies which reference volume (the value is known) that the observed food matches the best, just as a dietitian does in portion size estimation. In the second step, our system provides a fine-tuned volumetric estimation by comparing with multiple volumetric references and extrapolating the result, in the same process as that illustrated in Fig. 1.

If we treat food images with similar volumes as an abstract class, the first step mentioned above can be interpreted as finding the closest volume class for the input image, which can be formulated as an image classification problem. For the sake of clarity, we call the volume used for class division as reference volume and the abstract class associated with it as the reference class (see Fig. 1(b)). Since deep Convolutional Nerual Networks (CNNs) such as the ResNet [24] and DenseNet [25] have shown great success in image classification, we choose from these deep learning architectures for reference class classification. In particular, we set multiple reference volumes (e.g., 200 ml, 300 ml, 400 ml) for reference class division. Each reference class is associated with numerous training food images whose volumes lie within a small range (e.g., $\pm$50 ml) with respect to its reference volume. Thus, if a food dataset has the maximum volume of 1,000 ml, for instance, we can formulate the food volume estimation as a 10-class classification problem if 100 ml is adopted as the unit of reference.

Neural Network for Food Classification

In order to deploy our AI technology to an embedded system within the wearable device for real-world food volume estimation in the future, we adopt the MobileNetV2 [26], which is a well-known real-time classification network, to construct food volume classification network. This network outputs the probabilities of the reference classes to which the food in the input image belongs, which mimics the mental process of the dietitian to determine the reference object that best-matches the food volumetrically.

For the classification with respect to reference classes, we use hard labels, i.e., one-hot encoding, to supervise network training. Specifically, each training image

is associated with a binary label vector that contains only one element equals to 1. The index of "1" indicates the reference class closest to the training image. However, our goal is volume estimation rather than volume size classification. A simple way to achieve this goal is to use the closest reference volume as the estimated volume. However, this could result in a significant information loss. For example, given a test image, its distances to the closest reference class and the second closest class could be very similar. Thus, we use soft predictions instead, i.e., the probabilities of reference classes that a food in the input image belongs to, to perform volume estimation. The soft predictions not only tell the closest reference class, but also give information about the relative closeness to other reference classes. This process again mimics the mental process of the dietitian who uses degrees of likeliness among a set of known objects to arrive at an estimate. Specifically, the food volume is computed by the following inner product

$$\hat{V} = \sum_{i=1}^{N} p(i)V(i) \tag{1}$$

where $V(i)$ is the reference volume of class i, and $p(i)$ is the probability of the i^{th} reference class that the food belongs to.

References Volume Normalization

For the AI-based volume estimation system described above, it is assumed that all the plates in images have the same size because the plate acts as a scale reference in this case. As a result, we need to train multiple models if plate sizes are different, which degrades the generality of our method. To solve this problem, we propose to crop the food along with the plate from the images and then re-size the cropped sub-images to a fixed size (see Fig. 3). Such operation can be viewed as normalizing different sizes of plates to 1 since it enforces different plates having the same radius in pixel unit. Since the maximum food volume a plate can placed is usually can be pre-defined (e.g., 1000ml) according to the plate's size, thus, the maximum reference volume within the normalized plate is also normalized to 1 (which becomes unit free due to the normalization process). After this normalization, food images with large differences in volume can have similar normalized volume and thus can be placed in the same class. As a result, we can collect the dataset using the same plate to train the classification network and need only change the reference volumes for different plates during volume estimation. Specifically, denoting the original volume of a food as V, and the normalized volume as $\overline{V}$, we have

$$\overline{V} = \frac{V - V_{min}}{V_{max} - V_{min}} \tag{2}$$

where V_{max} and V_{min} are the maximum and minimum reference volumes, respectively. Thus, the normalized volume lies in closed interval of $[0, 1]$. If we use N references classes for volume estimation, each reference class will cover a volume range of $1/N$. Then, the normalized reference volume of the i^{th} class equals to

$i/N - 1/2N$. According to (1), the normalized volume estimation can be computed by

$$\overline{V} = \sum_{i=1}^{N} p(i)(\frac{i}{N} - \frac{1}{2N}) \tag{3}$$

Finally, the estimated volume can be obtained by

$$\hat{V} = V_{min} + (V_{max} - V_{min}) \sum_{i=1}^{N} p(i)(\frac{i}{N} - \frac{1}{2N}) \tag{4}$$

according to (2) and (3). Thus, once the model has been trained, only the maximum and minimum reference volumes are required for de-normalization, i.e., for estimating the actual volumes in unseen plates. Next, we describe how to obtain these two values.

Without loss of generality, supposing that the nth reference volume of the training set is obtained from a 3D model with irregular surface shown in Fig. 4(a), Then, the reference volume of class n can be computed by a triple integral defined over the 3D model

$$V_{ref\_n}^{r_t} = \iiint_{\omega} dxdydz \tag{5}$$

where r_t is the plate radius in the training images and ω is the region enclosed by the 3D model. Then, for an unseen (with respect to the training set) plate $r_{new} = sr_t$, the 3D model for nth reference volume computation can be obtained by scaling the model used in training set with a factor s in all three dimensions (Fig. 4(b)), thus, the nth reference volume of plate r_{new} can be computed by

$$V_{ref\_n}^{r_{new}} = \iiint d(sx)d(sy)d(sz) = s^3 V_{ref\_n}^{t} \tag{6}$$

Thus, given food images placed in an unseen plate, to obtain the maximum and minimum reference volume for volume estimation, we only need the plate radius to compute s.

Datasets

It is well-known that the deep network requires large amount of annotated data for training. However, most current food datasets (e.g., PFID [27], UECFOOD-100 [28], Food-101 [29]) are designed for food recognition, rather than volume estimation since these datasets has no information about food volumes. For this reason, most existing deep learning-based studies on food volume used self-collected datasets for training [22, 21]. Unfortunately, these self-collected datasets cannot be utilized to train our AI system because they are not publicly available, highly specialized

for certain foods (e.g., fruits [22]), or focused on depth [21]. Since it is extremely time-consuming to measure a large number of foods as volumetric truths for neural network training, we first generated two virtual food datasets using computer simulation to validate our human-mimetic method quickly. After the effectiveness our method was proven, we then applied our method to the real-world food images. This two-step approach allowed us to circumvent the volume truth measurement problem initially and accelerated the design of our AI system.

Virtual Food Dataset

In order to evaluate our AI method in handling foods with different volumetric ranges, two virtual datasets were generated with different minimum and maximum volumes. In the first dataset, 15 classes were utilized, and the minimum and maximum volumes were 400 ml and 3,400 ml, respectively. In the second dataset, 15 classes were utilized again but the volumetric range was smaller, between 200 ml and 1,700 ml. Note that the minimum volumes of the two VFDS were not zero because random processes that we utilize tend to produce fewer samples when the volume was close to zero. For convenience, we called these two datasets VFDL-15 and VFDS-15, respectively ("L" and "S" represents large and small, respectively). Finally, we divided each dataset into a training set and a test set with a roughly 2:1 ratio (10,003:4,889 for VFDL-15 and 9,205:4,489 for VFDS-15).

Real Food Dataset

We first established a Real Food Dataset (RFD) consisting of 1,500 images captured by a stationary camera. This RFD contained 50 Chinese foods of a university cafeteria in China. In addition, for each food, multiple images were taken by turning around the table where the food placed, providing an ideal dataset for training and testing our AI-based volume estimation system.

In order to evaluate the performance of our system for real life cases where a single-view image is taken at an unrestricted view angle, we established another RFD using personal mobile phones (brands unrestricted) to capture food images. This dataset consisted of 416 images. Unlike the previously case, these images were taken with user-determined view angles although views take directly above the food were discouraged. For convenience, we call this dataset a general RFD (GRFD). Likewise, the previous dataset is called an ideal RFD (IRFD). The volume of IRFD ranges from 110 to 410 ml, and the volume of GRFD ranges from 66 to 630 ml. As shown in Fig. 5, images in GRFD have considerable differences in view angles than the images in IRFD.

Examples of our two VFDs and two RFDs are shown in Fig. 5.

Results

Experimental Setups

Training Policy

We trained our deep neural network using the standard stochastic gradient descent (SGD) algorithm. The batch size was set to $128 \times 224 \times 224$. We set the learning rate to 0.01 and divided it by 10 after every 5,000 steps. The total training steps were set to 15K and 5K for VFD and RFD, respectively.

Data Augmentation

Because our food volume estimation system needs to see the entire 2D food, we employed only "random mirror" for data augmentation.

Evaluation Protocol

We computed top1 and top3 classification error to evaluate the classification accuracy, where "top" refers to the classes that are closest to the truth class. For example, for volume class 3, the three classes closest to it are the class 2, class 4, and 3 itself. Thus, for class 3, the top3 accuracy computes the ratio of samples that are classified into class 2,3,4 to the total samples. We also computed the mean relative volumetric error (mRVE) as a measure of accuracy for volume estimation. For each estimation, the RVE was computed by

$$RVE = \frac{|V_p - V_t|}{V_t} \tag{7}$$

where the V_p and V_t are the computed and measured volumes, respectively. The mRVE is then obtained by averaging the RVE value of all test samples.

Computing Systems

All the experiments were conducted on a computer equipped with the Intel Xeon E5-1630 (8 cores, 3.7 GHz) CPU, Titan X (12G) GPU and 32G RAM. We trained our AI system using the Tensorflow software platform [30].

Table 1 Experimental results on VFDL and VFDS with 15 classes division.

class	VFDL-15				VFDS-15			
	top1	top3	mRVE		top1	top3	mRVE	
1	63.8	100	15.1	15.9	54.2	95.8	19.6	19.8
2	67.6	98.3	11.1	12.3	60.3	96.9	12.8	13.7
3	56.1	96.8	10.6	11.9	58.8	96.7	10.7	11.4
4	46.7	96.1	9.9	11.1	50.3	95.4	10.1	11.1
5	41.3	93.4	9.7	10.7	43.7	92.2	9.4	10.7
6	44.5	89.6	9.1	9.8	37.0	85.5	10.2	10.8
7	35.7	84.3	9.3	10.2	40.4	82.0	9.2	10.3
8	39.6	83.2	8.0	8.9	40.0	85.0	8.4	8.7
9	30.1	83.3	7.6	8.6	34.7	79.8	8.1	8.6
10	36.9	78.0	7.3	8.0	32.4	75.9	8.2	9.0
11	28.6	76.7	7.3	9.0	25.4	74.2	8.4	9.2
12	28.7	75.5	6.8	7.6	25.9	69.3	7.8	8.7
13	35.4	81.2	5.9	6.2	19.7	69.1	7.2	8.0
14	36.8	83.2	5.0	5.7	31.4	85.4	5.4	6.2
15	16.9	55.9	8.2	8.4	50.8	80.1	5.3	5.1
overall	42.1	86.7	8.7	9.6	39.5	83.7	8.7	9.4

Experiments on VFD

15 Reference Classes

We first tested our human-mimetic method using the VFDL-15 and VFDS-15 datasets separately. The experimental results are shown in Table 1. The first and second columns of mRVE displays the performance of using soft predictions and hard predicted label for volume estimation, respectively. Several important observations and conclusions are described as follows.

1 Our human-mimetic AI system achieved 86.7% and 83.7% top3 accuracy on VFDL-15 and VFDS-15, respectively, which demonstrated that this system is able to find three closest volume reference classes of the input, in a similar way that a human uses to compare a food size with the sizes of a set of reference objects. In addition, according to the histograms shown in Fig. 6, it is unlikely for our method to classify the input images to the reference classes far away from their true reference class. Thus, although the top1 accuracy was relative low, our method still achieved 8.7% and 8.7% mRVE on VFDL-15 and VFDS-15, respectively.

2 Most top1 classification accuracies achieved on large volume classes were fewer than 40%, suggesting that the food volume classification model cannot distinguish large volume classes very well. This was mainly because the relative volumetric changes between large volume classes were much smaller than the changes across small classes, which made large volume classes less differentiable. Nevertheless, the volume estimation errors in large volume classes were typically smaller than the ones of small volume classes. This is reasonable since the RVE is more sensitive to the absolute error at the small classes according to (7). For example, given an absolute error equal to 100, for the VFDL-15 dataset, the mRVE of class 1 (400-600 ml) is in the range of 16.7% to 25%, while the mRVE of class 15 (3200-3400 ml) lies between 2.9% to 3.1%.

3 Using soft predictions for volume estimation achieved lower mRVE than using hard predictions on both VFDL-15 and VFDS-15, proving the effectiveness of our soft predictions based volume estimation mentioned in Method.

4 Not surprisingly, our method achieved better performance on VFDL-15 dataset for the reference classification task according to the top1 and top3 accuracy measures. However, for the volume estimation task, our method showed similar mRVE on the VFDS-15 and VFDL-15. Together with the observation 2, it can be concluded that a better reference classes classification accuracy, which usually requires larger interval between neighbor classes, does not imply a better volume estimation result.

5 We achieved 8.7% mRVE on the VFDS-15 dataset that has a similar volume range to that in the real-world, which demonstrated strongly the effectiveness of our human-mimetic approach.

Increased 30 References Classes

According to Observation 3 mentioned above, a better accuracy in reference class classification does not imply a better volume estimation. While fewer reference classes usually result in better classification accuracy, our goal is volume estimation rather than classification. Therefore, we further increased the number of reference classes and studied the resulting volume estimation performances. In particular, we used 30 reference classes for the original VFDL-15 and VFDS-15, forming VFDL-30 and VFDS-30 in the new experiment. The volume intervals between neighboring reference classes were chosen as 100ml and 50ml for VFDL-30 and VFDS-30, respectively. Our experimental results are shown in Table 2.

As expected, both top1 and top3 classification accuracies decreased significantly when the number of reference classes was increased from 15 to 30 for both VFDL

and VFDS. Since each class of VFD-15 was split into two classes in VFD-30, the top1 and top3 errors for classes in VFD-15 should correspond to top2 and top6 errors for classes in VFD-30, respectively. In this context, the top1/top3 accuracy of VFDL-30 and VFDS-30 is 40.8%/80.9%, 38.0%/77.3%, respectively. In addition, the number of training samples of each class in VFD-30 was decreased significantly compared with those in the VFD-15. In other words, a better classification accuracy is expected for VFD-30 if more samples are available.

Table 2 Experimental results on VFDL and VFDS with 30 classes division.

class	VFDL-30			VFDS-30		
	top1	top3	mRVE	top1	top3	mRVE
1	41.6	80.6	17.7	31.5	68.4	23.6
2	35.4	89.0	13.2	31.0	82.3	15.9
3	42.9	85.8	11.6	31.5	81.6	13.9
4	31.7	85.7	10.5	32.5	83.7	10.9
5	29.5	76.7	11.1	37.7	80.6	10.1
6	26.7	71.2	10.5	32.3	79.2	9.9
7	24.7	68.4	10.0	29.1	70.3	10.4
8	25.1	60.1	10.1	18.9	71.9	9.2
9	20.7	57.3	10.5	28.5	59.0	10.2
10	29.1	67.0	8.4	24.3	63.5	9.2
11	27.5	68.2	7.9	19.5	53.4	10.8
12	16.2	56.6	9.1	17.9	56.5	9.7
13	15.3	52.0	9.4	19.9	57.7	8.2
14	21.2	54.1	8.4	21.3	53.9	8.7
15	18.9	55.1	8.1	16.3	52.5	9.1
16	24.2	50.2	8.0	16.6	50.6	8.1
17	12.0	51.3	7.5	15.9	43.9	8.9
18	18.5	53.4	7.0	20.0	51.9	7.6
19	15.1	49.2	7.0	14.4	48.1	8.7
20	18.2	51.2	7.0	15.1	41.7	9.6
21	11.2	32.0	8.7	18.2	40.9	8.1
22	13.1	43.9	6.8	13.5	37.6	8.9
23	18.5	45.9	6.9	14.4	36.8	7.9
24	16.5	42.6	6.0	13.6	39.0	7.5
25	16.2	48.6	5.9	13.4	37.8	7.5
26	16.7	49.3	5.9	14.0	34.0	6.9
27	14.2	49.6	5.8	18.3	43.8	5.5
28	25.8	49.5	5.5	19.7	50.6	4.8
29	7.9	36.8	7.8	23.3	64.4	5.0
30	0.0	23.8	8.8	27.9	50.6	5.8
overall	**22.2**	**58.7**	**8.7**	**21.2**	**55.1**	**8.6**

Bias Analysis

In order to check whether our human-mimetic AI system produced biased volumetric estimates, we investigated the distribution of volumetric errors shown in Fig. 7. It appears that the error distribution, after a normalization, can be approximated by a Gaussian distribution centered at zero, suggesting that our AI-based volume estimator is unbiased.

Training with Normalized Reference Class

In order to demonstrate the effectiveness of our normalization approach described previously, we mixed the training data of the VFDL-15 and VFDS-15 and obtained a combined training set consisting of 19,208 images. Then, we adjusted the reference volume for each dataset during the test stage according to (4). The experimental results are shown in Table 3. By comparing Table 1 with Table 3, it can be observed that training with mixed datasets achieved better performance than training with

Table 3 Experimental results with mixed training but separate tests for VFDL and VFDS.

class	VFDL-15			VFDS-15		
	top1	top3	mRVE	top1	top3	mRVE
1	65.3	100	14.7	62.5	95.8	19.1
2	68.5	98.9	10.9	58.8	96.4	13.1
3	60.0	98.0	9.7	60.6	95.7	10.7
4	49.5	94.9	10.5	48.1	94.6	10.4
5	50.0	92.8	8.8	47.6	90.9	9.6
6	44.2	90.1	9.0	32.9	89.3	10.2
7	32.8	85.8	9.2	39.5	87.7	8.2
8	39.9	85.2	8.0	40.0	83.8	8.2
9	32.1	84.5	7.6	33.7	81.7	7.8
10	35.5	78.3	7.5	35.5	76.9	7.7
11	30.6	80.0	6.8	26.1	75.2	7.8
12	33.9	81.5	6.1	28.8	73.1	7.3
13	34.0	83.3	5.7	21.9	72.9	7.1
14	38.6	82.3	5.5	29.9	83.1	5.3
15	15.3	62.7	7.7	53.3	80.1	5.3
overall	**43.8**	**88.1**	**8.5**	**40.1**	**84.6**	**8.5**

each dataset individually. In particular, top1 accuracy, top3 accuracy and mRVE on the VFDL-15 dataset were improved by 1.7, 1.4 and 0.2 percentage points, respectively, these three quantities were improved by 0.6, 0.9 and 0.2 percentage points, respectively, for the VFDS-15. These improvements may have resulted from the increase in the number of training samples for each class.

Our experiments demonstrated that food images can be placed in the same class as long as they share similar normalized volumes, regardless of their actual volumes, as we stated previously.

Experiments on RFD

We first tested our human-mimetic method using the IRFD dataset. Similar to the VFD results, we also experimented two reference units to divide the reference classes. When 100 ml was adopted as the reference unit, the available food images could only be divided into 3 classes. Thus, we did not list the top3 accuracy since they were 100% in this case. It can be observed from Table 4 that our AI system produced similar mRVEs when using different reference units. In addition, our AI system produced an mRVE less than 15% for most individual classes and 11.6% overall, which are satisfactory results in food volume estimation from single-view images.

Table 4 Experimental results on IRFD dataset

		classes					
		1	2	3	4	5	overall
100ml	top1	89.0	82.4	60.7	-	-	79.6
	mRVE	13.0	10.8	11.8	-	-	11.7
50ml	top1	78.4	27.5	79.4	35.0	75.4	67.4
	top3	97.3	100	87.6	100	88.4	93.5
	mRVE	13.2	17.7	8.9	11.6	8.9	11.6

Next, we applied our method to the GRFD dataset which contained 416 images with measured volumes. We divided these images into the training (242 images) and testing (174 images) set. Since the GRFD had significantly fewer training samples (for IRFD v.s. 242 for GRFD), and the view angles of the images in GRFD had much higher variabilities than those in IRFD, the estimation error was larger.

Table 5 Experimental results on GRFD dataset

		classes					
		1	2	3	4	5	overall
100ml	top1	100.0	35.0	40.0	0.0	59.6	42.5
	top3	100	100	85.0	97.6	100	96.0
	mRVE	25.8	27.3	20.9	19.1	15.3	20.1

Nevertheless, a reasonable performance was achieved with a 20.1% mRVE (Table 5).

Table 6 Experimental results on IRFD. The foods in the test set were unseen in the training set.

		classes			
		1	2	3	overall
100ml	training samples	350	400	210	960
	test samples	210	200	130	540
	top1	88.1	76.0	78.5	
	mRVE	13.6	14.0	8.3	12.5

Since the images were randomly divided into the training and test sets in previous experiments, the same type of food could be found in both the training and test sets. To investigate the capability of our human-mimetic system in handling unseen foods, we performed an experiment where the images of IRFD were divided into the training or test sets with different food types. In other words, all food types in the test set cannot be found in the training set. Note that we did not conduct this experiment on the GRFD dataset because a large number of training samples would be required to enable the deep network to handle unseen foods. With the same consideration, we used three classes rather than five classes. We finally obtained 960 training images and 540 test images. It can be observed from Table 6 that our AI system produced 12.5% mRVE even all the test foods are new to the network. This experiment demonstrated the our human-mimetic volume estimation system was able to focus on the its "mental activity" on the food volume, rather than other food features.

Discussion

In this section, we discuss several important issues related to the automatic approach to food volume estimation.

Relative Food Volume

As mentioned previously, the most stringent requirement in image-based volume estimation is to provide a scale reference (or a fiducial marker) in the image, such as a checkerboard card. Although a person can place this reference within the view of the camera, it is inconvenient in practice. Since, in many parts of the world, foods are usually placed in a plate (mostly a circular shape) for serving, using the plate as a scale reference is a more suitable choice. For a circular plate, in particular, only its diameter needs to be known. However, this parameter still requires human effort for measurement. As a result, billions of food images existing on the websites cannot be utilized for volume estimation because the plate diameter is unknown, which is a great waste of resources. In this work, we presented a normalization method where a food image is cropped and normalized. As a result, the plate, regardless its

actual size, becomes standardized and unitless. For such a normalized image, our AI system is able to estimate the relative volume for food images without information about the plate size. The relative volume can be later converted easily to the true volume when the plate diameter becomes available.

Multiple Foods in One Plate

Most experiments in this paper were designed for evaluating the effectiveness of the AI approach to food volume estimation. For simplicity, we limited to the case where each plate contains only one type of food. In practice, however, more than one foods are occasionally placed in a single plate. Although automatic separation of multiple foods is beyond the focus of this paper, we briefly discuss the computational procedure called image segmentation. Decades of research in this field have produced a rich set of algorithms to label and separate objects. Since food objects have complex shapes and textures, the traditional algorithms have limited success. Recently, deep learning based semantic segmentation algorithms [31, 32, 33] have emerged. Using these algorithms, different foods can be first recognized and separated by a deep neural network [18, 21]. Then, the human-mimetic method presented in this paper can be applied to each food for volume estimation.

AI Perspective

Although using AI for dietary assessment is still in its initial infancy, we believe that this approach has a great potential to advance nutrition science and dietetics significantly. As the research on this approach progresses, it becomes increasingly clear that at least some, and perhaps the entire, previously time-consuming self-reporting tasks can be passed to a robotic system which is unbiased, objective and highly accurate. If successful, this new approach will exert a strong impact on public health in producing quantitative, unbiased dietary data for preventing and controlling diet-rated chronic diseases.

Conclusion

In this paper, we have presented an image-based automatic method for food volume estimation, aiming at solving a long-standing problem in nutrition science where dietary assessment is subjective and time-consuming. We took advantage of the recently developed AI technology and developed a human-mimetic system that imitates a dietitian's mental process by comparing the food size with the sizes of commonly known objects. We proposed a novel idea for food volume estimation. In particular, we showed that food volume estimation can be formulated as an image classification problem if we treat images with similar volumes as a reference class. Moreover, we showed food images with different volumes can be also placed into the same class for network training as long as they have similar normalized volume. Then, we only need to adjust real volumes of the normalized reference classes for different plates during the testing process. Based on this approach, we pre-defined a number of references with ascending volumes acting: as virtual volumetric gauges stored in computer's memory. Our AI system then classify the observed food into a set of probabilities of the reference volumes. Finally, our AI system produces the best-guess volume based on the stored volumes and the computed probability

vector. Our experimental results have shown that this human-mimetic approach is both accurate and robust, capable of producing a reasonable estimate from a 2D image which contains only partial 3D information. We have also developed a new normalization procedure allowing a collection of different food volumes into the same reference class, which greatly facilitates the training process for the deep neural network. In addition, we introduced the relative volume concept based on the normalization procedure for the practical cases where the plate diameter is not available. Our human-mimetic method has a potential to pass the time-consuming food portion size estimation task to an unbiased and well-trained robotic system, liberating humans from the time-consuming portion size estimation task and allowing them to improve their diet based on automatically and objectively performed dietary assessment results.

List of abbreviations

Artificial Intelligent (AI)
Virtual Food Datasets (VFD)
Real Food Datasets (RFD)
Convolutional Neural Network (CNN)
mean Relative Volumetric Error (mRVE)

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Availability of data and materials
The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests
The authors declare that they have no competing interests.

Author's contributions
MS and ZY proposed the idea of "Human-mimetic food volume estimation". ZY and SC generated the virtual food dataset under supervision of MS and WJ. QX provided the general real-world food dataset. DY and HZ collected the ideal real food dataset. ZY and HY designed the deep learning network and conducted experiments under supervision of MS and ZHM. ZY, WJ and MS wrote the paper and all authors provided feedback and approved the final manuscript.

Funding
This research was funded by: National Natural Science Foundation of China OF grant number 61973106; China Scholarship Council of grant number 201806130030; Gates Foundation (OPP1171395); and National Institutes of Health (R56DK113819).

Acknowledgements
APC charges for this article were fully paid by the University Library System, University of Pittsburgh.

Author details
[1]College of Electrical and Information Engineering, Hunan University, Changsha, China. [2]Department of Neurosurgery, University of Pittsburgh, Pittsburgh, USA. [3]Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, USA. [4]School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China. [5]Image Processing Center, Beihang University, Beijing, China. [6]Department of Bioengineering, University of Pittsburgh, Pittsburgh, USA.

References

1. Hales, C.M., Carroll, M.D., Fryar, C.D., Ogden, C.L.: Prevalence of obesity among adults and youth: United states, 2015–2016. NCHS Data Brief (288) (2017)
2. Chen, H.-C., Jia, W., Yue, Y., Li, Z., Sun, Y.-N., Fernstrom, J.D., Sun, M.: Model-based measurement of food portion size for image-based dietary assessment using 3d/2d registration. Measurement Science and Technology **24**(10), 105701 (2013)

3. Dehais, J., Anthimopoulos, M., Shevchik, S., Mougiakakou, S.: Two-view 3d reconstruction for food volume estimation. IEEE transactions on multimedia **19**(5), 1090–1099 (2016)
4. Livingstone, M.B.E., Robson, P., Wallace, J.: Issues in dietary intake assessment of children and adolescents. British Journal of Nutrition **92**(S2), 213–222 (2004)
5. Arab, L., Estrin, D., Kim, D.H., Burke, J., Goldman, J.: Feasibility testing of an automated image-capture method to aid dietary recall. European journal of clinical nutrition **65**(10), 1156 (2011)
6. Sun, M., Burke, L.E., Baranowski, T., Fernstrom, J.D., Zhang, H., Chen, H.-C., Bai, Y., Li, Y., Li, C., Yue, Y., *et al.*: An exploratory study on a chest-worn computer for evaluation of diet, physical activity and lifestyle. Journal of healthcare engineering **6**(1), 1–22 (2015)
7. Farooq, M., Sazonov, E.: A novel wearable device for food intake and physical activity recognition. Sensors **16**(7), 1067 (2016)
8. Liu, J., Johns, E., Atallah, L., Pettitt, C., Lo, B., Frost, G., Yang, G.-Z.: An intelligent food-intake monitoring system using wearable sensors. In: 2012 Ninth International Conference on Wearable and Implantable Body Sensor Networks, pp. 154–160 (2012). IEEE
9. Kagaya, H., Aizawa, K., Ogawa, M.: Food detection and recognition using convolutional neural network. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 1085–1088 (2014). ACM
10. Mezgec, S., Koroušić Seljak, B.: Nutrinet: a deep learning food and drink image recognition system for dietary assessment. Nutrients **9**(7), 657 (2017)
11. Aguilar, E., Remeseiro, B., Bolaños, M., Radeva, P.: Grab, pay, and eat: Semantic food detection for smart restaurants. IEEE Transactions on Multimedia **20**(12), 3266–3275 (2018)
12. U.S. Department of Agriculture, A.R.S.: FoodData Central (2019). https://fdc.nal.usda.gov/ Accessed March 9, 2019
13. Hassannejad, H., Matrella, G., Ciampolini, P., Munari, I., Mordonini, M., Cagnoni, S.: A new approach to image-based estimation of food volume. Algorithms **10**(2), 66 (2017)
14. Puri, M., Zhu, Z., Yu, Q., Divakaran, A., Sawhney, H.: Recognition and volume estimation of food intake using a mobile device. In: 2009 Workshop on Applications of Computer Vision (WACV), pp. 1–8 (2009). IEEE
15. Rahman, M.H., Li, Q., Pickering, M., Frater, M., Kerr, D., Bouchey, C., Delp, E.: Food volume estimation in a mobile phone based dietary assessment system. In: 2012 Eighth International Conference on Signal Image Technology and Internet Based Systems, pp. 988–995 (2012). IEEE
16. Woo, I., Otsmo, K., Kim, S., Ebert, D.S., Delp, E.J., Boushey, C.J.: Automatic portion estimation and visual refinement in mobile dietary assessment. In: Computational Imaging VIII, vol. 7533, p. 75330 (2010). International Society for Optics and Photonics
17. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge university press, ??? (2003)
18. Lo, F., Sun, Y., Qiu, J., Lo, B.: Food volume estimation based on deep learning view synthesis from a single depth map. Nutrients **10**(12), 2005 (2018)
19. Liu, F., Shen, C., Lin, G.: Deep convolutional neural fields for depth estimation from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5162–5170 (2015)
20. Chen, W., Fu, Z., Yang, D., Deng, J.: Single-image depth perception in the wild. In: Advances in Neural Information Processing Systems, pp. 730–738 (2016)
21. Meyers, A., Johnston, N., Rathod, V., Korattikara, A., Gorban, A., Silberman, N., Guadarrama, S., Papandreou, G., Huang, J., Murphy, K.P.: Im2calories: towards an automated mobile vision food diary. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1233–1241 (2015)
22. Li, H., Han, T.: Deepvol: Deep fruit volume estimation. In: International Conference on Artificial Neural Networks, pp. 331–341 (2018). Springer
23. Konkle, T., Oliva, A.: A familiar-size stroop effect: real-world size is an automatic property of object representation. Journal of Experimental Psychology: Human Perception and Performance **38**(3), 561 (2012)
24. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
25. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
26. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018)
27. 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 (2009). IEEE
28. Matsuda, Y., Hoashi, H., Yanai, K.: Recognition of multiple-food images by detecting candidate regions. In: 2012 IEEE International Conference on Multimedia and Expo, pp. 25–30 (2012). IEEE
29. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101–mining discriminative components with random forests. In: European Conference on Computer Vision, pp. 446–461 (2014). Springer
30. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., *et al.*: Tensorflow: A system for large-scale machine learning. In: 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16), pp. 265–283 (2016)
31. Yu, H., Yang, Z., Tan, L., Wang, Y., Sun, W., Sun, M., Tang, Y.: Methods and datasets on semantic segmentation: A review. Neurocomputing **304**, 82–103 (2018)
32. Yang, Z., Yu, H., Sun, W., Mao, Z., Sun, M.: Locally shared features: An efficient alternative to conditional random field for semantic segmentation. IEEE Access **7**, 2263–2272 (2018)
33. Yang, Z., Yu, H., Feng, M., Sun, W., Lin, X., Sun, M., Mao, Z., Mian, A.: Small object augmentation of urban scenes for real-time semantic segmentation. IEEE Transactions on Image Processing **29**, 5175–5190 (2020)

Figures

Figure 1 (a) Motivation of our method Human can easily tell that Food NO.1 is larger than Food NO.2 if the plates in the two images have the same size. Moreover, we can roughly estimate the volume of Food No. 1 or 2 if some reference volumes (Foods No. 3 and 4) are given. **(b) Definitions of terms**The reference class is an abstract food class that have similar volumes, the reference volume is the center volume of a reference class, and the reference unit is the interval between two neighboring reference volumes.

Figure 2 Overview of the proposed food volume estimation system which contains two stages. In the first stage, an image classification network outputs a vector of the probability values with respect to a pre-selected set of reference classes. In the second stage, the food volume is estimated by an inner product between the probability vector and a volume vector consisting of the volumes of reference classes.

Figure 3 Concept of normalized references Different food volumes can be normalized to the same or a similar reference volume by first cropping the foods from the input image and then resizing the foods to the same size.

Figure 4 Supposed 3D models for reference volume computation.

Figure 5 Examples of a) VFDL b) VFDS, c) IRFD and d) GRFD

Figure 6 Histograms (40 ml for bin width) of classification results for VFDL-15. White distributions indicate the test set (for clarity, all classes are shown and separated by a blank bin). Orange distributions (one class for each panel) represent classification results for classes 1 through 15.

Figure 7 Histograms of errors on VFDS-30.

Figures

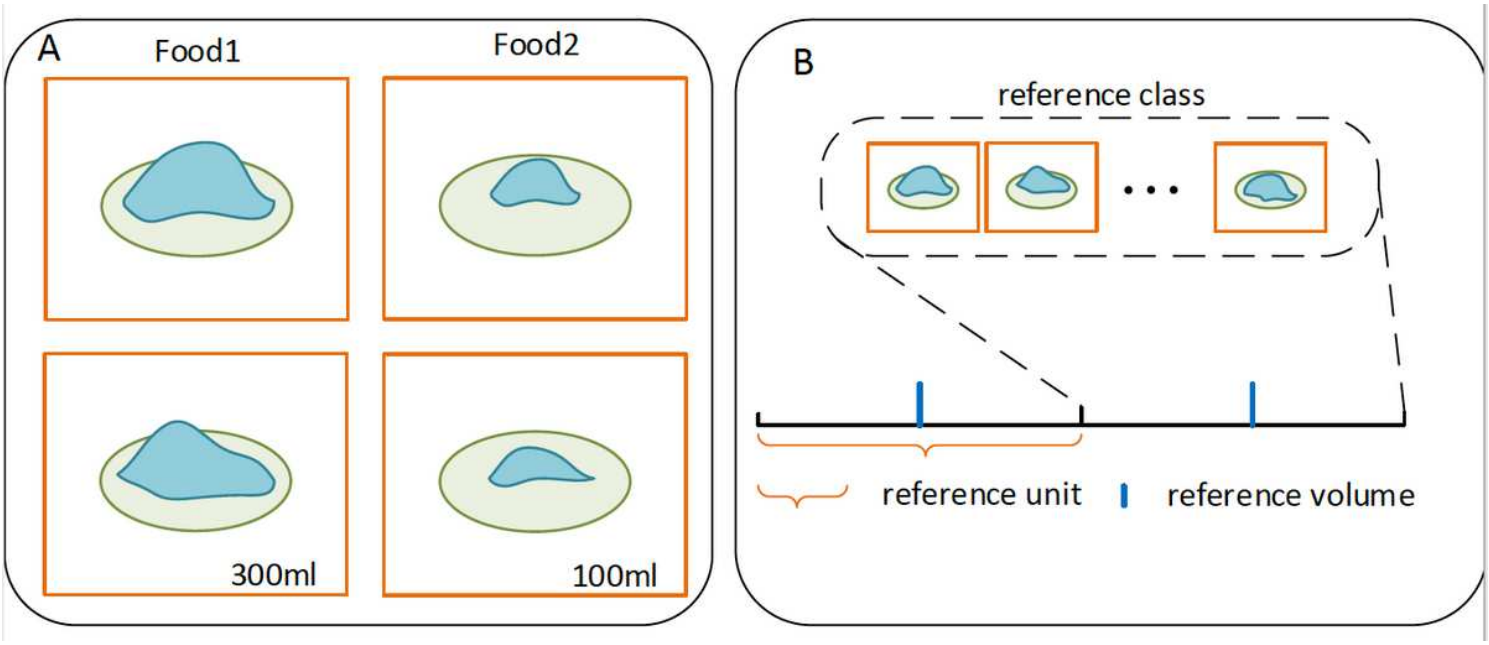

Figure 1

(a) Motivation of our method Human can easily tell that Food NO.1 is larger than Food NO.2 if the plates in the two images have the same size. Moreover, we can roughly estimate the volume of Food No. 1 or 2 if some reference volumes (Foods No. 3 and 4) are given. (b) Definitions of termsThe reference class is an abstract food class that have similar volumes, the reference volume is the center volume of a reference class, and the reference unit is the interval between two neighboring reference volumes.

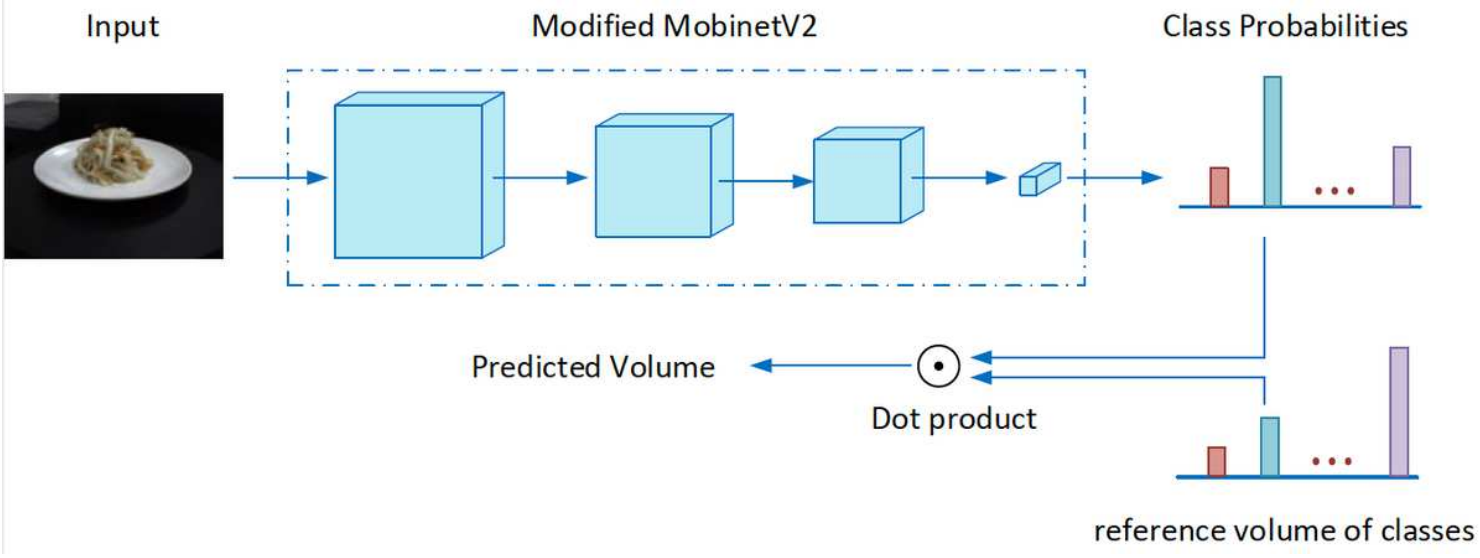

Figure 2

Overview of the proposed food volume estimation system which contains two stages. In the first stage, an image classification network outputs a vector of the probability values with respect to a pre-selected

set of reference classes. In the second stage, the food volume is estimated by an inner product between the probability vector and a volume vector consisting of the volumes of reference classes.

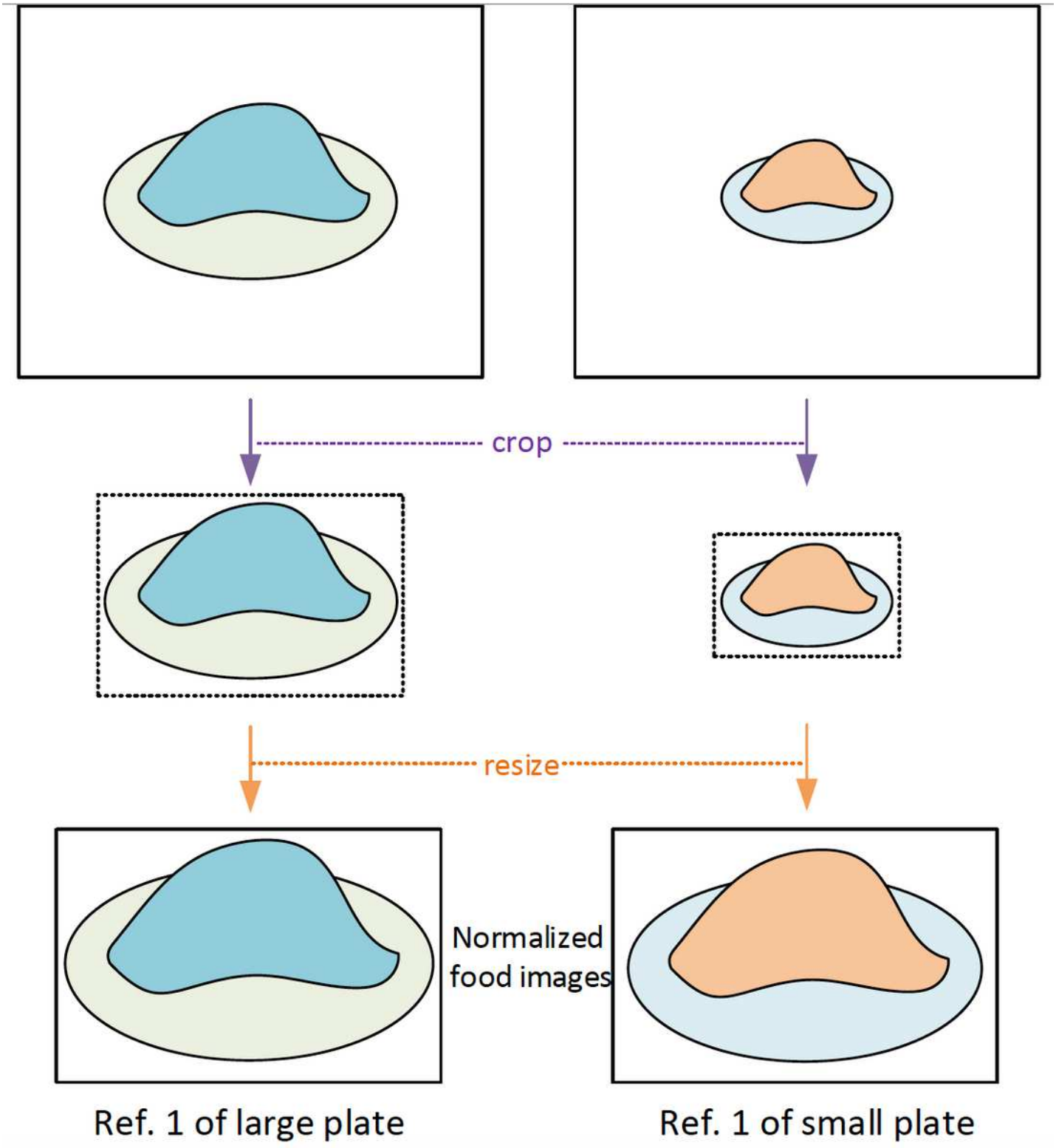

Figure 3

Concept of normalized references Different food volumes can be normalized to the same or a similar reference volume by first cropping the foods from the input image and then resizing the foods to the same size.

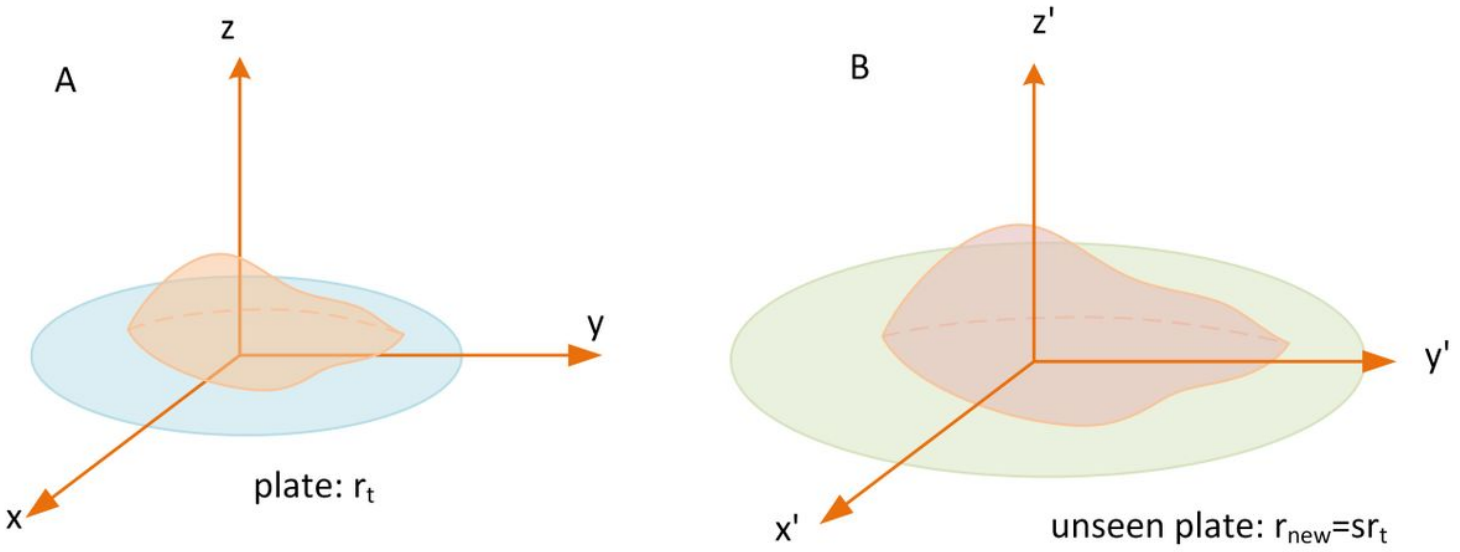

Figure 4

Supposed 3D models for reference volume computation.

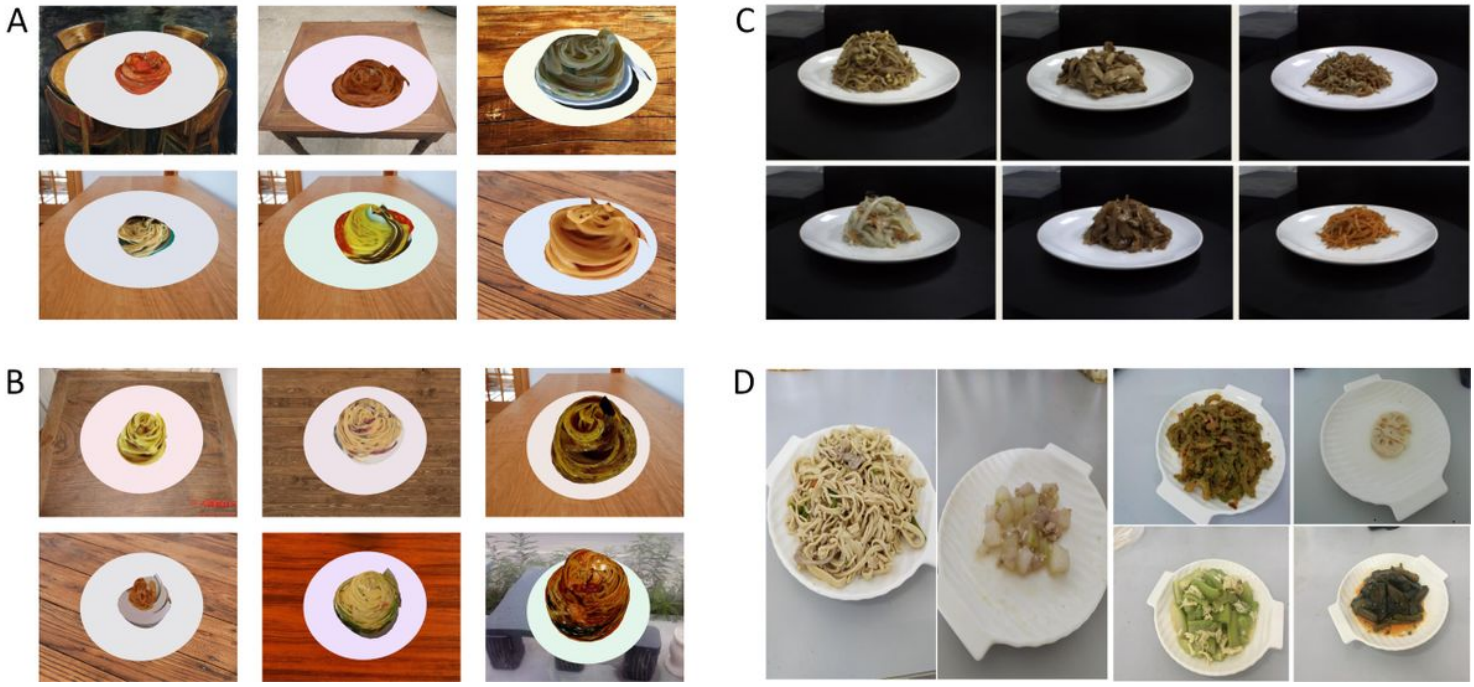

Figure 5

Examples of a) VFDL b) VFDS, c) IRFD and d) GRFD

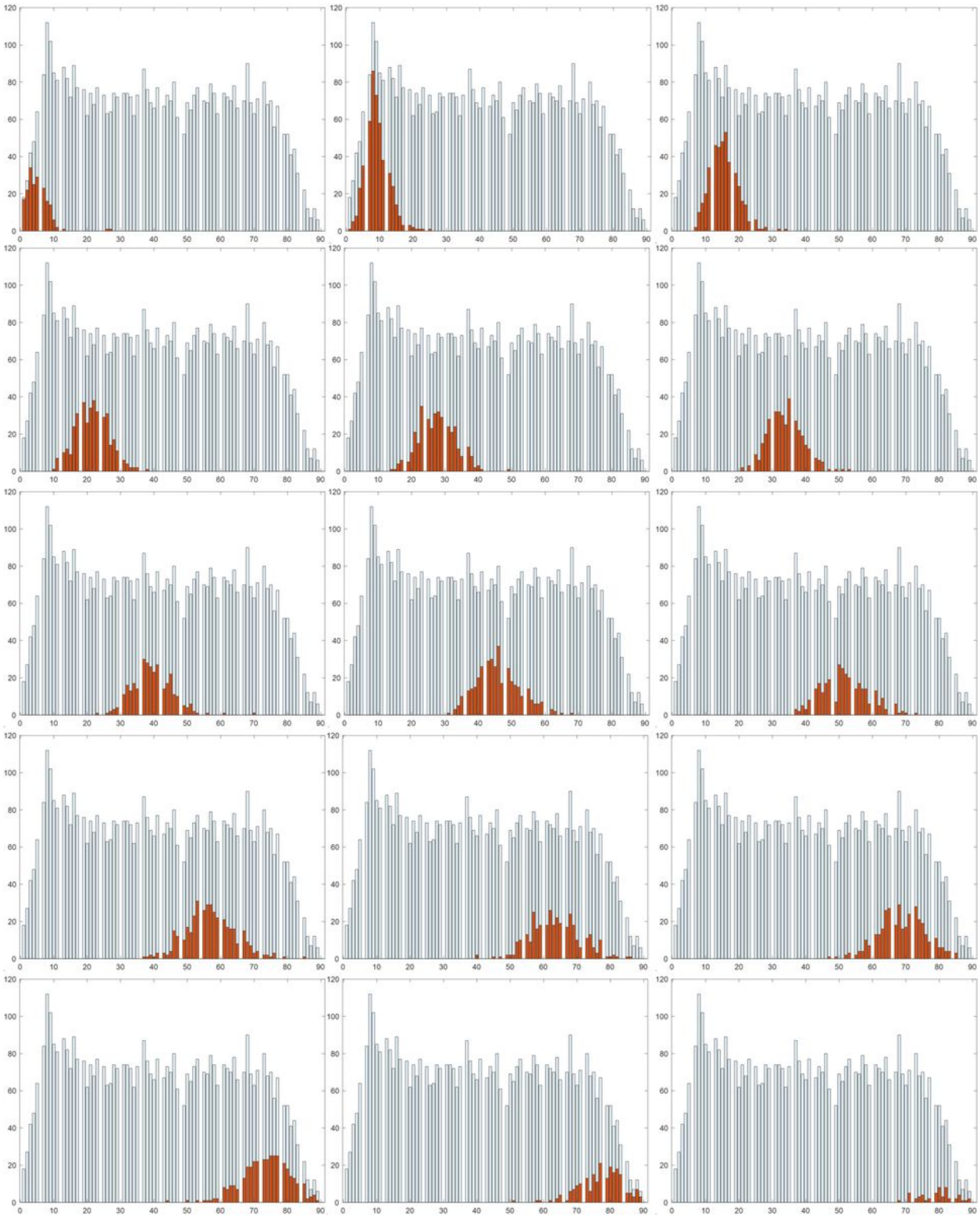

Figure 6

Histograms (40 ml for bin width) of classification results for VFDL-15. White distributions indicate the test set (for clarity, all classes are shown and separated by a blank bin). Orange distributions (one class for each panel) represent classification results for classes 1 through 15.

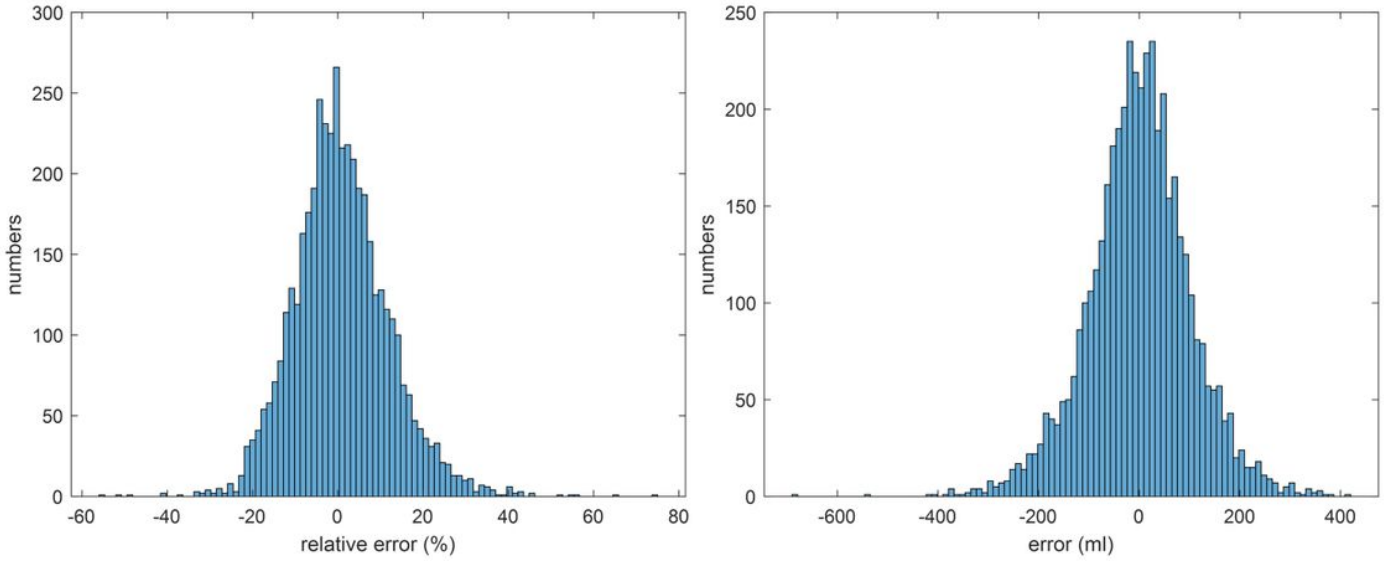

Figure 7

Histograms of errors on VFDS-30.