Public Health Nutrition: 22(7), 1180–1192 doi:10.1017/S136898001800054X

Image-based food portion size estimation using a smartphone without a fiducial marker

Yifan Yang[1,2], Wenyan Jia[1], Tamara Bucher[3], Hong Zhang[2] and Mingui Sun[1,*]
[1]Departments of Neurosurgery, Electrical & Computer Engineering, and Bioengineering, University of Pittsburgh, Pittsburgh, PA 15260, USA: [2]Image Processing Center, School of Astronautics, Beihang University, Beijing, People's Republic of China: [3]Priority Research Center for Physical Activity and Nutrition, Faculty of Health and Medicine, The University of Newcastle, Callaghan, New South Wales, Australia

Submitted 2 August 2017: Final revision received 26 January 2018: Accepted 9 February 2018: First published online 6 April 2018

Abstract

Objective: Current approaches to food volume estimation require the person to carry a fiducial marker (e.g. a checkerboard card), to be placed next to the food before taking a picture. This procedure is inconvenient and post-processing of the food picture is time-consuming and sometimes inaccurate. These problems keep people from using the smartphone for self-administered dietary assessment. The current bioengineering study presents a novel smartphone-based imaging approach to table-side estimation of food volume which overcomes current limitations.
Design: We present a new method for food volume estimation without a fiducial marker. Our mathematical model indicates that, using a special picture-taking strategy, the smartphone-based imaging system can be calibrated adequately if the physical length of the smartphone and the output of the motion sensor within the device are known. We also present and test a new virtual reality method for food volume estimation using the International Food Unit™ and a training process for error control.
Results: Our pilot study, with sixty-nine participants and fifteen foods, indicates that the fiducial-marker-free approach is valid and that the training improves estimation accuracy significantly ($P < 0{\cdot}05$) for all but one food (egg, $P > 0{\cdot}05$).
Conclusions: Elimination of a fiducial marker and application of virtual reality, the International Food Unit™ and an automated training allowed quick food volume estimation and control of the estimation error. The estimated volume could be used to search a nutrient database and determine energy and nutrients in the diet.

Keywords
Dietary assessment
Smartphone
Image processing
International Food Unit™
Augmented reality

Over the years, unhealthy foods with increasingly large portion sizes have been among the most popular products of fast-food chains and restaurants, and people often eat more than they need without being aware of the extra energy intake. Constructing an effective tool that quickly informs people of the amounts of energy and nutrients in a plate of food at the dining table, i.e. allowing them to eat with quantitative dietary knowledge, is highly significant in public health.

As the mobile and electronic technologies advance, numerous smart wristbands (e.g. Fitbit wristband) and watches (e.g. Apple Watch) have become available for quantitative physical activity assessment[1–4]. As a result, these devices have been used widely by the public to gauge their level of exercise, such as the distance walked and the energy expended. In contrast, there is currently no convenient way for an individual to know the energy content in a plate of food before it is consumed. Currently, to determine the energy intake of an individual, a self-reporting procedure to a dietitian is the most utilized method. This subjective method is not only unreliable and inaccurate, but also very burdensome, costly and time-consuming for both the individual being assessed and the assessor. To innovate dietary assessment, several electronic approaches have been developed. Piezoelectric sensors and microphones have been developed to measure chewing and swallowing during eating events[5–9]. However, these sensors may not be suitable for long-term use in people's daily life. Electronic sensors have been embedded within utensils, such as plates and mugs, to weigh the food or beverage[10]; these novel methods monitor eating automatically. However, their utility is limited due to the required use of specially made utensils. Wearable cameras have been used to reduce dietary assessment error in 24 h dietary recalls[11]. However, existing commercial body-worn cameras are not well suited for dietary assessment because they are mostly designed for public security or entertainment purposes.

*Corresponding author: Email drsun@pitt.edu

To conduct objective dietary assessment, the eButton, a small wearable device worn in the upper chest area, has been developed[12,13]. The eButton has a specially designed camera with a large field of view (up to 170°) and an appropriately tilted lens to observe food on a table. Although the eButton can conduct dietary assessment objectively, it is more suitable for dietary research than everyday use, mainly because people near the device may be accidently recorded and the wearer may forget to turn it off during private events. These problems can be controlled by pre-filtering data using advanced computational algorithms (e.g. deep learning-based eating recognition[14–19]) and specially trained people[20]. However, these methods cannot be adopted easily in non-research applications.

Nowadays the smartphone has become the most ubiquitous personal electronic device. A recent survey indicated that 77 % of Americans now own a smartphone, up from 35 % in 2011[21]. Dietary assessment using the smartphone has been previously investigated[22–26]. Food pictures provide an excellent means to refresh people's memory about their dietary intake, and motivated people have formed social media groups to exchange food pictures and discuss their dietary experience (e.g. 'Food' in Facebook[24]). However, when smartphone pictures are used to evaluate energy/nutrients quantitatively, a number of limitations exist: picture taking must be volitionally initiated by an individual for each plate of food; two pictures are needed when there is a leftover for volume subtraction; it is difficult to use this method if multiple people fetch food from shared containers; and the process of picture taking may disrupt normal eating habits. Despite these problems, considering the widespread availability of the smartphone, we believe that this device is currently the most practical tool for self-motivated food energy/nutrient analysis in real life.

Obtaining the energy/nutrients from a food image requires two important pieces of information: food name and food volume. There are several ways to let the smartphone know the food name, such as speech recognition of the user's announcement[27,28] and deep learning-based computer vision[14–17]. In addition to food name, the volume, or portion size, of the food must be entered into a food database to determine energy and nutrients. The volume problem is more difficult because: (i) the ordinary food image does not have a dimensional reference, preventing the size of the food from being gauged; and (ii) the two-dimensional image lacks information about the three-dimensional surface of the food which defines its volume.

Currently, the dimensional reference is obtained by using a fiducial marker that must be present in a food picture. Clearly, this is a very strong requirement. Many objects have been suggested as fiducial markers, including: (i) a colour checkerboard[29]; (ii) a business card[30,31]; (iii) a specially designed physical cube[32,33]; (iv) a circular plate of a known diameter[34]; (v) a specially designed tablecloth[25]; and (vi) a laser-generated pattern or mesh[35,36]. Methods (i) to (iii) require people to carry a physical object and place it beside the food before taking the picture; methods (iv) and (v) require special dining setups; and method (v) requires a complex laser device attached to the smartphone.

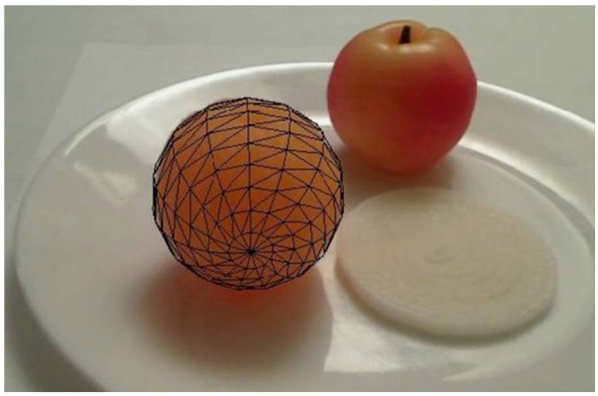

Fig. 1 (colour online) Wireframe method for food volume estimation

A common approach to estimating food volume is by human visualization. This method is often used when the fiducial marker is missing, such as in the case of a finger food. To achieve better accuracy, a wireframe volume estimation method was developed[37–40]. Since a two-dimensional image does not carry the complete three-dimensional information, the observed food is assumed to have a specific shape, such as a cuboid, wedge, cylinder or partial ellipsoid. Once a wireframe shape is selected, the food is fit by scaling in different dimensions until the best fit is visually achieved (Fig. 1). Then, the volume of the scaled wireframe provides a volumetric estimate. This method has been well studied by both our and other groups[20,38]. The results indicate that the method has a high accuracy when the shape of food and the wireframe are well matched. However, if this condition is not met, the estimation error can be large. In addition, the wireframe method is ineffective when multiple foods occlude each other or are mixed.

In the current paper, we present a new smartphone-based dietary evaluation method that does not require the use of any fiducial marker placed with food. Instead, the length (or the width) of the smartphone itself is used to calibrate the imaging system. We also present a new interactive portion size estimation method based on virtual reality technology, where a virtually generated cube is used as the unit of estimation. Instead of finding a numerical value of the food volume by comparing the sizes of food and cube, which is mentally demanding, we ask the estimator to scale the cube up or down until the volumes of the cube and the food are visually equivalent. This scaling procedure can be easily and quickly performed on the smartphone screen by sliding two fingers

https://doi.org/10.1017/S136898001800054X Published online by Cambridge University Press

inward or outward. To gauge the estimation error for the estimator, we also present an automated training procedure to evaluate the individual's skill level. These new technologies will allow people to perform table-side estimation of energy and nutrients before making a dietary decision.

Methods

In this section, we first present a fiducial-marker-free method to calibrate the camera system. Then, a virtual reality method for food volume estimation is described.

Calibration of imaging system

The key concepts that enable us to eliminate the external fiducial marker are: (i) making use of the motion sensor within the smartphone to determine camera orientation; (ii) utilizing the length or width of the smartphone to uniquely determine the location of any visible point on the tabletop; and (iii) using a special way to photograph a food (setting the bottom of the smartphone on the tabletop during picture taking, see Fig. 3 below). We show, mathematically, that these concepts provide an adequate calibration of the imaging system, allowing volume estimation of the food on the tabletop without any fiducial marker in the image.

Let us assume that the tabletop is a level surface. The camera within the smartphone, the tabletop and the food on the table form an imaging system, as shown in Fig. 2. For this system, we can establish four coordinate systems, including the world coordinates, camera coordinates, optical image coordinates and pixel coordinates, illustrated in Fig. 2.

Let $[X, Y, Z]^t$ be the coordinates in the world coordinate system. We define the camera coordinate system $[U, V, W]^t$ in such a way that its W-axis is parallel to the optical axis of the lens and its origin is located at the centre of the lens. With these definitions and assuming that the food shape does not change during the estimation process, the relationship between coordinate systems $[U,V,W]^t$ and $[X,Y,Z]^t$ can be represented as a rigid body transformation[41,42]:

$$\begin{bmatrix} U \\ V \\ W \end{bmatrix} = \boldsymbol{R} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} + \boldsymbol{T}, \tag{1}$$

where $\boldsymbol{T}$ is a 3×1 translation vector determined by the choice of origin for the world coordinates and $\boldsymbol{R}$ is a 3×3 rotation matrix with its entries being sinusoidal functions of three angles, i.e. pitch θ, yaw φ and roll ψ, given by:

$$\boldsymbol{R} = \begin{bmatrix} \cos\varphi\cos\psi & \sin\theta\sin\varphi\cos\psi - \cos\theta\sin\psi & \cos\theta\sin\varphi\cos\psi + \sin\theta\sin\psi \\ \cos\varphi\sin\psi & \sin\theta\sin\varphi\sin\psi + \cos\theta\cos\psi & \cos\theta\sin\varphi\sin\psi - \sin\theta\cos\psi \\ -\sin\varphi & \sin\theta\cos\varphi & \cos\theta\cos\varphi \end{bmatrix}. \tag{2}$$

Note that the values of θ, φ and ψ can be measured using the inertial measurement unit within the smartphone. These parameters determine the orientation, or pose, of the smartphone when the food picture is taken.

Now, let us define the image coordinate system in the imaging sensor plane and the camera coordinate system at the plane of the optical lens. The origin of the image

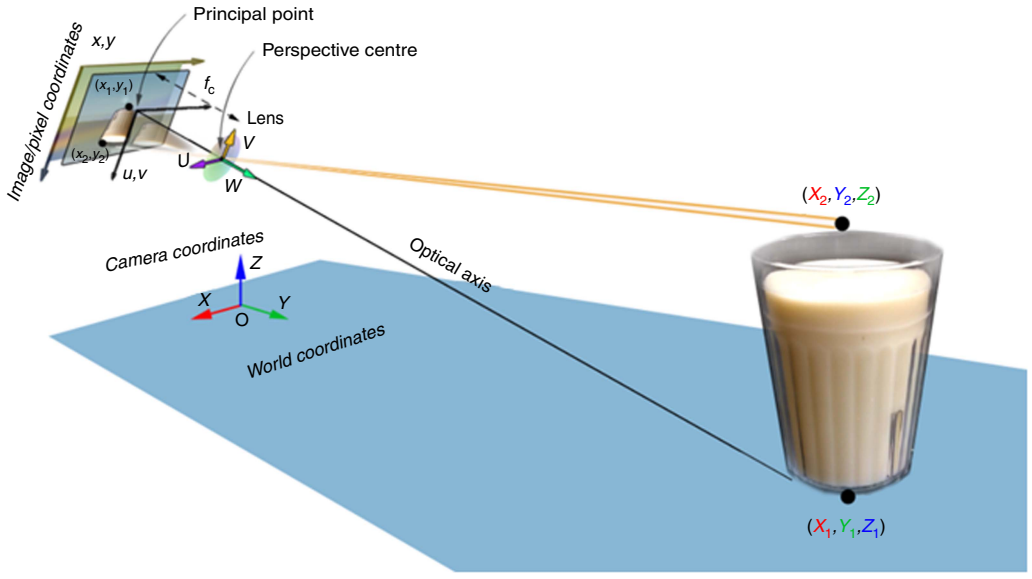

Fig. 2 (colour online) Model of the smartphone imaging system

https://doi.org/10.1017/S136898001800054X Published online by Cambridge University Press

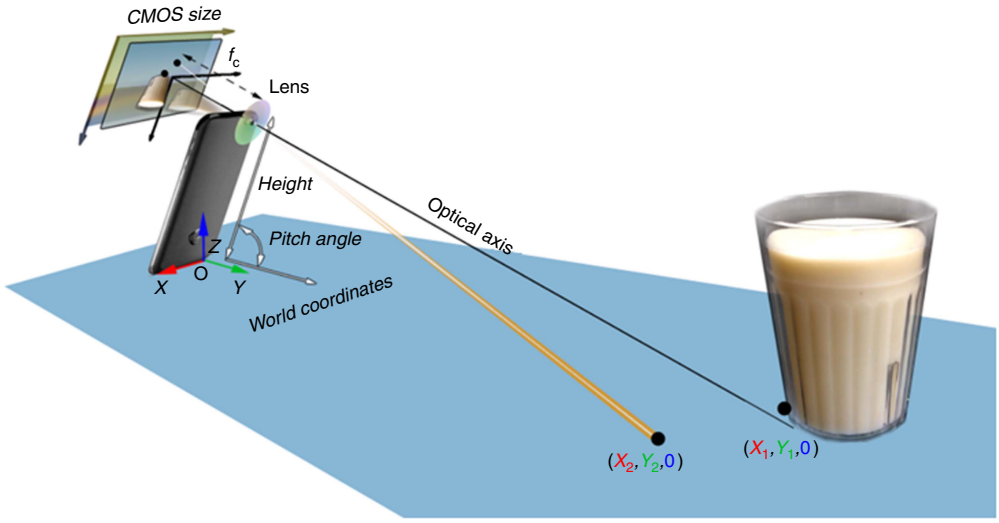

Fig. 3 (colour online) Mathematical model for reconstructing the world coordinates of the tabletop (CMOS, complementary metal oxide semiconductor)

coordinate system is located at the centre of the imaging sensor, which is also the intersection between the imaging plane and the optical axis. Particularly, let the u- and v-axes in the image coordinate system be parallel with the U- and V-axes in the camera coordinate system, and the origin of the camera coordinate system be the centre of the optical lens. Then, the projection from the camera coordinate system to the image coordinate system is given by:

$$\begin{bmatrix} u \\ v \end{bmatrix} = \frac{f_c}{W} \begin{bmatrix} U \\ V \end{bmatrix}, \tag{3}$$

where f_c is the focal length of the camera.

Within the camera, the optical image of the food object at the imaging plane is sampled digitally by a CMOS (complementary metal oxide semiconductor) sensor, resulting in a rectangular array of pixels as a digital image. We thus perform another coordinate transformation to associate the optical image with the digital image, given by:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} s_c^x & 0 \\ 0 & s_c^y \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} + \begin{bmatrix} c_x \\ c_y \end{bmatrix}, \tag{4}$$

where c_x and c_y are offsets which are the pixel coordinates of the origin in the optical image coordinates, and s_c^x and s_c^y are scale factors determined by the resolution of the particular imaging sensor.

Combining Eqs (1) to (4), we have a compact form of the coordinate transformation:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \mathrm{f}(X, Y, Z). \tag{5}$$

Reconstruction of tabletop coordinates

Eq. (5) provides a mathematical model describing the projection of any visible point in the real world to a specific pixel in the food image. Conversely, given a food image, we are interested in the inverse function of Eq. (5) for the purposes of volume estimation, i.e.

$$(X, Y, Z) = \mathrm{f}^{-1}(x, y). \tag{6}$$

However, there is a fundamental problem in defining Eq. (6). Because, in general, for surface points on the food, depth W cannot be determined from the two-dimensional image, Eq. (5) is not invertible and thus Eq. (6) does not exist. However, if certain constraints on the imaging system are imposed, Eq. (6) can be well defined, allowing the world coordinates to be reconstructed.

Let us assume that the food image is taken while the smartphone is set on the tabletop as shown in Fig. 3. With the known smartphone orientation provided by the inertial measurement unit, a right angle between the surface of the smartphone and the optical axis, and the projection relationship in Eq. (5), it can be shown (detailed in the Appendix) that the inverse of function f in Eq. (6) exists for the tabletop, i.e.

$$(X, Y, 0) = \mathrm{f}^{-1}(x, y). \tag{7}$$

Note that $Z = 0$ in Eq. (7) represents the plane equation of the tabletop. Since, according to Fig. 3, the angles φ and ψ, which represent yaw and roll, respectively, are both zero, Eq. (2) becomes:

$$\boldsymbol{R} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix}. \tag{8}$$

From Eq. (4), the world coordinates of the tabletop are related to the pixel coordinates by:

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} 1/s_c^x & 0 \\ 0 & 1/s_c^y \end{bmatrix} \begin{bmatrix} x - c_x \\ y - c_y \end{bmatrix}, \tag{9}$$

https://doi.org/10.1017/S136898001800054X Published online by Cambridge University Press

$$W = \frac{h \cdot \sin\theta}{\cos\theta - (v/f_c) \cdot \sin\theta}, \tag{10}$$

$$\begin{bmatrix} U \\ V \end{bmatrix} = \frac{W}{f_c} \begin{bmatrix} u \\ v \end{bmatrix} \tag{11}$$

and

$$\begin{bmatrix} X \\ Y \\ 0 \end{bmatrix} = \boldsymbol{R}^{-1} \left(\begin{bmatrix} U \\ V \\ W \end{bmatrix} - \boldsymbol{T} \right), \tag{12}$$

where the derivation of Eq. (10) is included in the Appendix.

We point out that all the parameters in Eqs (8) to (12) are known from smartphone specifications or are available from the output of the inertial measurement unit (containing an accelerometer, a gyroscope and a magnetometer) within the smartphone. For example, for the iPhone 6 Plus, the specifications and calibration parameters used in Eqs (8) to (12) are listed in Tables 1 and 2, respectively.

Table 1 Smartphone (iPhone 6 plus) parameters used in Eqs (8) to (12)

Focal length	4·15 mm
CMOS size	Height: 4·8 mm Width: 3·6 mm
Image resolution	Height: 1280 pixels Width: 960 pixels
Angle of pitch	Real-time readout from accelerometer sensor unit: rad
Phone height (from camera to the bottom)	152 mm

CMOS, complementary metal oxide semiconductor.

Table 2 Calibration parameters used in Eqs (8) to (12) for the smartphone (iPhone 6 plus)

f_c	4·15 mm
s_c^y	1280 pixels/4·8 mm
s_c^x	960 pixels/3·6 mm
c_y	1280 pixels/2
c_x	960 pixels/2
$\boldsymbol{T}$	$[0, -152\,\text{mm}, 0]^t$
$\boldsymbol{R}$	$\begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix}$, where θ is obtained from the accelerometer in the IMU

IMU, inertial measurement unit.

Background of food image-based volume estimation

Although Eqs (8) to (12) allow the world coordinates $[X, Y, 0]^t$ of any visible points in the tabletop to be reconstructed, the information provided about the food volume is still very sparse because the surface function of the food is unknown except along the observable intersection (if it exists) between the food and the tabletop. Previously, the wireframe method[38–40] solved this problem by assuming that the food shape is close to one of a set of predefined shapes (e.g. cuboid, wedge, cylinder, partial ellipsoid) rendered in deformable wireframes (Fig. 1 and related text). This method produces accurate results for well-matched shapes, but shows only limited performance for complex shaped, obscured, mixed or hand-held foods. In these cases, visual estimation becomes the only way to provide an estimate. This is a difficult task because the estimator must come up with a volumetric value for each food. If the volume is in the unit of cm^3 (or equivalently, ml), the value is usually very large and the estimate is not intuitive. Although the 'cup' is often used as the unit for measuring real-world foods, it is not very suitable for image-based measures because its height and shape are not standardized, and its mental images differ for different people. In many parts of the world, the Western 'cup' is an unfamiliar object.

International food unit

To improve the robustness of food volume estimation from images, we propose to use the International Food Unit™ (IFU™)[32], which is a 4 cm × 4 cm × 4 cm cube (64 cm^3; Fig. 4). The result of food volume measurement is in 'F'. For example, an apple of 128 cm^3 is 2 F. The IFU™ cube can be dyadically divided into 2 cm and 1 cm subcubes, in units of 'dF' (where 'd' follows 'divided')[32]. Thus, 1 F = 8 dF = 64 ml. The features of the IFU™, including its cubic shape, dyadic division in length and octave division in volume, appear to be more convenient for human visualization from images than other forms of divisions. In addition, they are also preferred by computer processing because they are parallel to the binary and octave number systems utilized by the computer. The IFU™ has many other attractive properties. First, since it is a cube of known size, its world coordinates (Eq. (6)) are well defined from pixels in images so long as it sits on the tabletop, either physically or virtually. Second, although other convex shapes (e.g. a sphere) may also invert Eq. (6), the corners of the cube, when projected into a two-dimensional image, produce a well observable sense of orientation. Third, previously, a significant practical difficulty was to estimate the height of the food from an image when a planar fiducial marker is used. The three-dimensional IFU™ facilitates height estimation significantly by providing a clearly observable height reference. Finally, the IFU™ is connected to both the metric millilitre (1 F = 64 cm^3 = 64 ml) and the commonly used cup measures (1 F ≈ ¼ cup, whereas the cup has multiple confusing definitions: 1 metric cup = 250 ml, 1 US legal cup = 240 ml, 1 US customary cup = 236·6 ml and 1 imperial cup = 284 ml). Previous studies on IFU™ have shown a significantly higher accuracy in food volume estimation using the IFU™ than the cup for untrained individuals[32,33].

Volume estimation using the IFU™

Based on the fiducial-marker-free food image estimation technique and the IFU™ concepts, we present a fast,

https://doi.org/10.1017/S136898001800054X Published online by Cambridge University Press

interactive method for portion size estimation using the smartphone. Since, by Eqs (8) to (12), the world coordinates of the tabletop are known, we can manipulate the image properly and utilize virtual reality technology on the tabletop to assist the estimation task shown in Fig. 5(a). First, we extend the tabletop virtually, thus providing more spaces in the image. Then, we create a virtual 'tablecloth' in the form of a grid. The side length of each square in this grid is exactly 4 cm in the world coordinates, the same as the side length of the IFU^{TM}. This grid both marks the location and orientation of the tabletop in the image and provides a scale reference, in both horizontal and vertical directions, for all foods on the table. Next, we place a virtual IFU^{TM} cube on the tabletop. During the estimation process, the estimator moves the cube to a desired place (sliding with one finger on the smartphone screen), scales it (sliding with two fingers in an inward or outward motion), while comparing it side-by-side with any food item of his/her interest until the volume of the scaled cube is visually equivalent to that of the food (what we call 'volumetric equivalence' (VE)), as seen in Fig. 5(b) and 5(c).

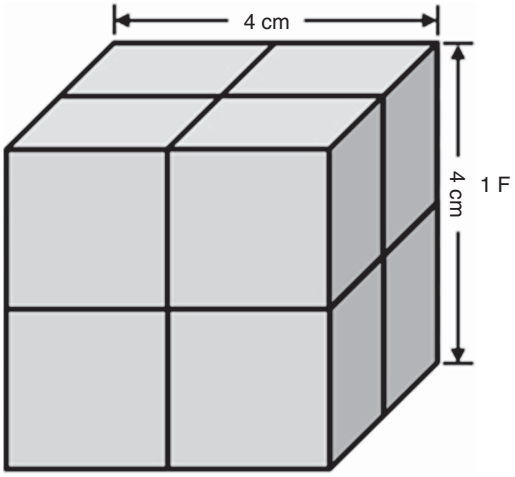

Fig. 4 International Food Unit^{TM}

It is important to point out that we ask the estimator to achieve a VE rather than coming up with a numerical value because, we believe, the VE task is easier than the numerical task for most people, although this assertion still needs a proof. A distinct advantage of the IFU^{TM} approach over the wireframe approach is that it is more robust in handling complex real-world cases, as shown in Fig. 6 where the use of wireframes would be very difficult.

Nevertheless, to correctly judge VE requires human experience. In practice, a person who has no experience in VE needs to acquire skills, and it is important to gauge a person's skill level by measuring his/her average portion estimation error. Thus, we propose an automated training process using an app to conduct a set of exercises. The person first estimates the volume of a food in an image, then a feedback with the true volume is provided so he/she can learn from the error. If this error is larger than a pre-set bound (e.g. 20 %), the person will need to retry the exercise until the error bound is reached.

Experiments

We conducted two experimental studies to: (i) compare the effect of training; and (ii) evaluate whether the pre-set error bound for food volume estimation is achieved successfully.

Study 1

In this experiment, we implemented the described algorithms in MATLAB® with a graphical interface (MathWorks, Natick, MA, USA). Fifteen commercial food models of known volumes (Nasco Life/form® Food Replica; shown in Fig. 7(a)) were used as the test objects.

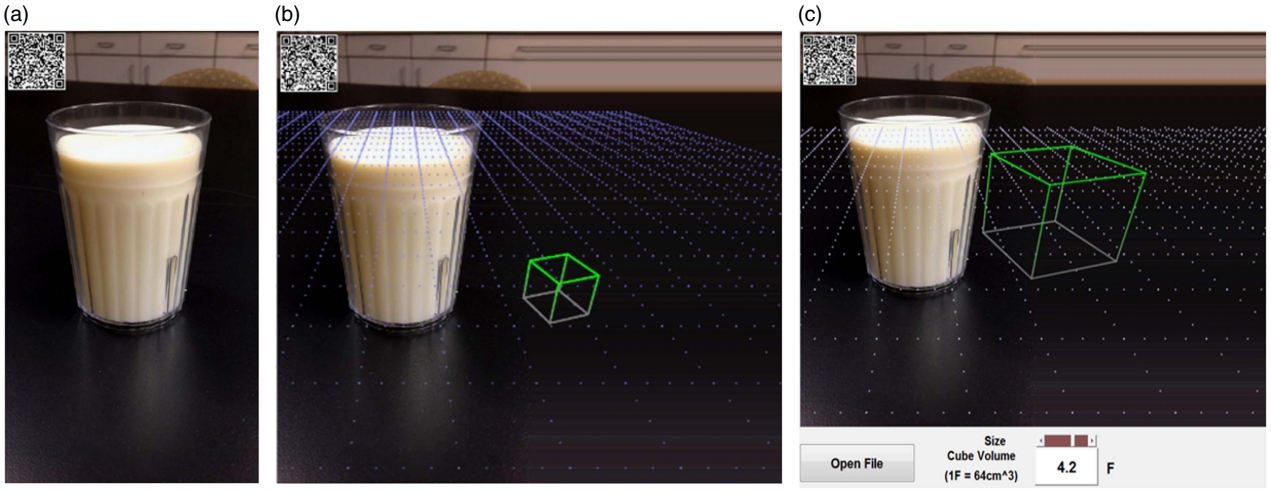

Fig. 5 (colour online) Virtual reality-based volume estimation using the International Food Unit^{TM} (IFU^{TM}): (a) fiducial-marker-free image; (b) extended image with a virtual grid of 4 cm spacing and an IFU^{TM}; (c) final image in which the volume of the scaled IFU^{TM} cube is visually equivalent to the volume of the milk, yielding an estimate of 4·2 F

https://doi.org/10.1017/S136898001800054X Published online by Cambridge University Press

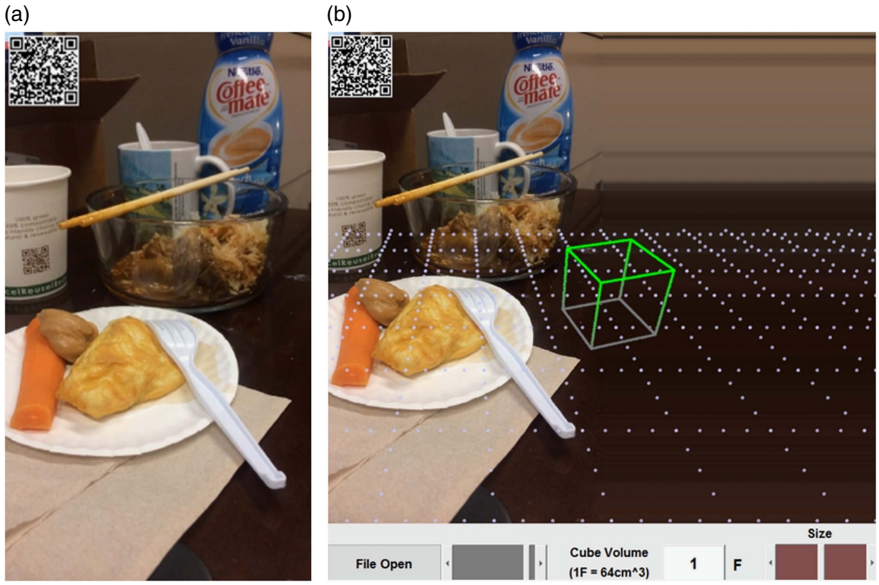

Fig. 6 (colour online) (a) A complex real-world fiducial-marker-free image; (b) extended virtual reality image from which any food item can be estimated by moving and scaling the International Food Unit™

Forty volunteers, who were first- and second-year undergraduate students, participated in the study. These students had never estimated food portion sizes from images. During the experiment, each participant manually adjusted the size of the virtual cube to reach the VE as described previously. At this point, the participant had to click a control bar to move to the next food image.

Study 2

In this experiment, twenty-nine volunteers, who were again first- and second-year undergraduate students without experience in estimating portion sizes from images, participated in the study. The same fifteen food models (Fig. 7(a)) were used for tests. However, before the tests, we implemented a training process. We selected fourteen foods, beverages and non-food objects (different from the fifteen foods for testing), as shown in Fig. 7(b). It can be observed that many of these objects or containers have irregular shapes, unsuitable for using the wireframe method. The volumes of the training items were measured using cup measures (for liquids and grains), by water displacement (for submergible ones) or with a ruler (for cuboids). They were photographed in different containers, such as plates, glasses and bowls, resulting in forty-five training images. The training was provided by our self-developed MATLAB software. For each image, the participant scaled the virtual cube until he/she believed that a VE was reached between the food and the cube. Once he/she clicked the 'submit' button, the software provided a feedback of the true volume in a message box. If the absolute estimation error was larger than 20 %, the participant had to retry the volume estimation for this specific item until the error was less than 20 %. The participant was asked to estimate as many training images as possible until he/she felt confident about his/her skills for food volume estimation. After training, the test was conducted the same as that in Study 1.

Results

Table 3 lists the name of each food (column 1) and its ground truth volume (ml or, equivalently, cm^3; column 2) as measured by cup, water displacement or ruler. For each test with or without training, we list the median, interquartile range (defined as the difference between the medians of the lower and higher halves of the data values after sorting in an ascending order[43]), average absolute error and root-mean-square error. Each of these values was calculated over the number of estimates provided by the research participants. In the last column, we also list the probability of the null hypothesis (i.e. 'no difference' with and without training) from the Mann–Whitney *U* test for each food.

Several important observations can be made from Table 3. First, all the measures indicate that the training process improves volume estimation performance. Second, in fourteen out of fifteen cases (except for egg), there is a statistically significant reduction in the relative error (defined as (estimate – ground truth)/ground truth × 100 %) after training (last column). Third, without training, there is a large estimation bias as measured by the median (average 87·97 %). After training, it is nearly eliminated (average −1·14 %).

Figure 8(a) and 8(b) further compare the relative error (with the standard derivation indicated) and the root-mean-square error, respectively. It is noticeable that, from both measures, the errors tend to be much larger in the last

https://doi.org/10.1017/S136898001800054X Published online by Cambridge University Press

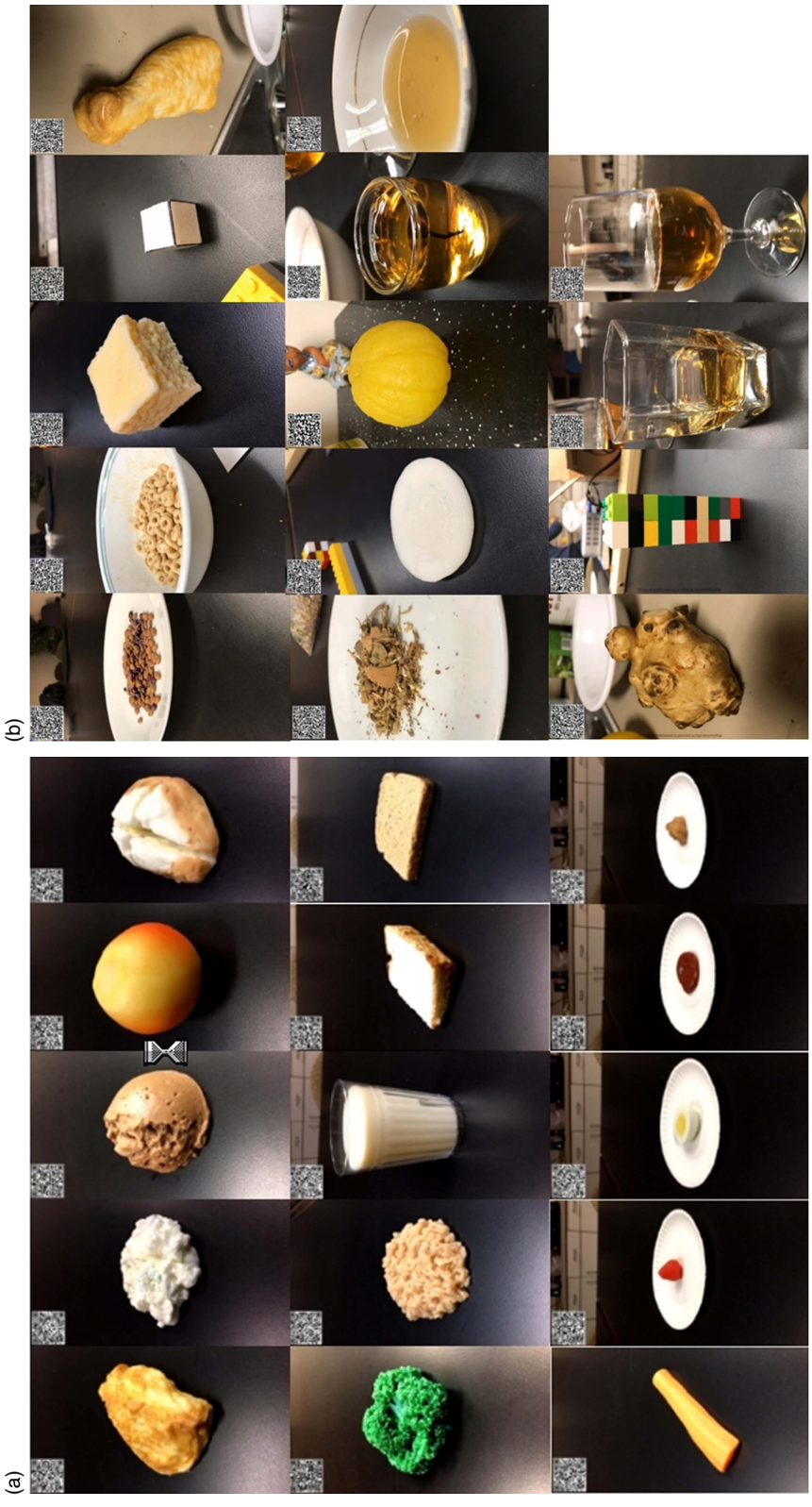

Fig. 7 (colour online) (a) Fifteen food models tested in both Studies 1 and 2; (b) fourteen foods, beverages and non-food objects utilized for generating forty-five training images in different containers for Study 2. Note that, in each image, the two-dimensional barcode records the motion sensor data at the time of image acquisition

https://doi.org/10.1017/S136898001800054X Published online by Cambridge University Press

Table 3 Experimental results of the fifteen foods, tested with and without training, using the new approach of fiducial-marker-free image-based food portion size estimation using a smartphone

		Without training (*N* 40)				With training (*N* 29)				
Food name	Ground truth (ml)	Median (%)	IQR (%)	Absolute error (%)	RMSE (%)	Median (%)	IQR (%)	Absolute error (%)	RMSE (%)	*P* value
Milk	240·00	59·66	63·22	66·43	91·92	−6·19	31·56	20·42	29·90	<0·01
Peach	151·67	26·76	45·58	39·81	57·00	−8·35	21·92	17·56	23·72	<0·01
Broccoli	119·93	6·68	64·62	36·91	53·83	−1·89	23·89	15·38	21·02	<0·05
Potato	112·67	13·55	61·55	40·48	68·29	−4·03	23·58	20·30	29·66	<0·01
Bread	106·00	43·70	79·37	71·83	103·19	−14·06	29·85	18·64	23·50	<0·01
Angel cake	93·67	54·76	69·30	65·32	85·26	−10·85	11·85	14·91	21·08	<0·01
Ice cream	79·67	14·34	28·62	28·53	40·78	−3·99	16·32	14·79	21·10	<0·01
Cottage cheese	64·03	19·45	99·65	51·32	70·12	−16·66	30·73	21·17	29·42	<0·01
Chicken thigh	64·00	42·33	53·37	60·05	70·48	0·00	13·69	12·18	20·14	<0·01
Brown rice	52·33	124·83	81·28	118·71	140·22	1·96	18·75	11·18	15·28	<0·01
Carrot	25·97	194·52	226·99	181·07	219·26	5·46	20·88	30·66	58·27	<0·01
Egg	20·67	46·22	105·96	97·74	154·41	19·98	68·95	41·10	59·42	>0·05
Ketchup	11·37	289·82	176·37	285·40	336·49	17·20	93·74	50·62	83·32	<0·01
Peanut butter	11·07	172·97	117·90	157·14	181·19	−3·09	91·72	46·14	72·10	<0·01
Strawberry	8	209·94	231·04	286·27	545·12	7·37	100·16	69·47	127·94	<0·01
Total	–	87·97	100·32	105·80	147·84	−1·14	39·84	26·97	42·39	–

IQR, interquartile range; RMSE, root mean-square error.

few items. These items, which are shown in the last row of Fig. 7(a), have the smallest volumes among the fifteen foods tested, as indicated in the second column in Table 3. The average absolute error is 16·65 % for the first ten foods (calculated from the ninth column in Table 3), indicating that, for these large-volume items, the 20 % error bound was well achieved. On the other hand, for the last five foods, the average absolute error is 47·60 %, indicating that, for these small-volume items, the 20 % error bound was not achieved.

Discussion

One of the desired goals in portion size estimation is to gauge and control, in a statistical sense, the average estimation error. Our pre-training method is, in theory, capable of reaching this goal if the statistical features of food items for training and testing are the same. Our experimental results indicated that this goal was achieved for large-volume foods (average absolute error 16·65 % < 20 %). However, it failed for small-volume foods (average absolute error 47·60 % > 20 %). Several factors may have contributed to this failure. First, we found that it was difficult to visually judge the VE when the display of the food and the cube was small. Thus, we believe that the excessive error was mainly due to this visual effect. This effect could be corrected by allowing the estimator to manually zoom the display window, so the food and cube appear in normal sizes. Second, our training samples (Fig. 7(b)) lacked small objects with similar shapes to the small foods utilized in the test (last row in Fig. 7(a)). As a result, participants may not be trained adequately for the types of foods tested. This problem can be solved by enlarging the training set properly. Third, in Study 2, we asked the participants to decide by themselves when to stop training. This self-judgement may have also contributed to the large error because some participants may be over-confident in their skills and stopped training too early. A more appropriate stop criterion could be to monitor the participant's skills by averaging the absolute estimation errors in the past *N* trials, where *N* is to be determined empirically, and stopping the training when the average error is below the pre-set error bound.

In our method, the IFU™ played an important role in the algorithm development. We used the cubic-shaped virtual dimensional reference to provide a better visual effect in estimating food volume, especially providing a reference in the height of the food which was not provided previously when using the actual two-dimensional fiducial markers. In addition, the scalable virtual cube allows the use of this new VE concept. We believe that, for most people, the mental task involved in assessing VE is easier than coming up with a numerical number for the volume. However, this assertion is still subject to a solid proof, perhaps by a psychological or experimental study.

https://doi.org/10.1017/S136898001800054X Published online by Cambridge University Press

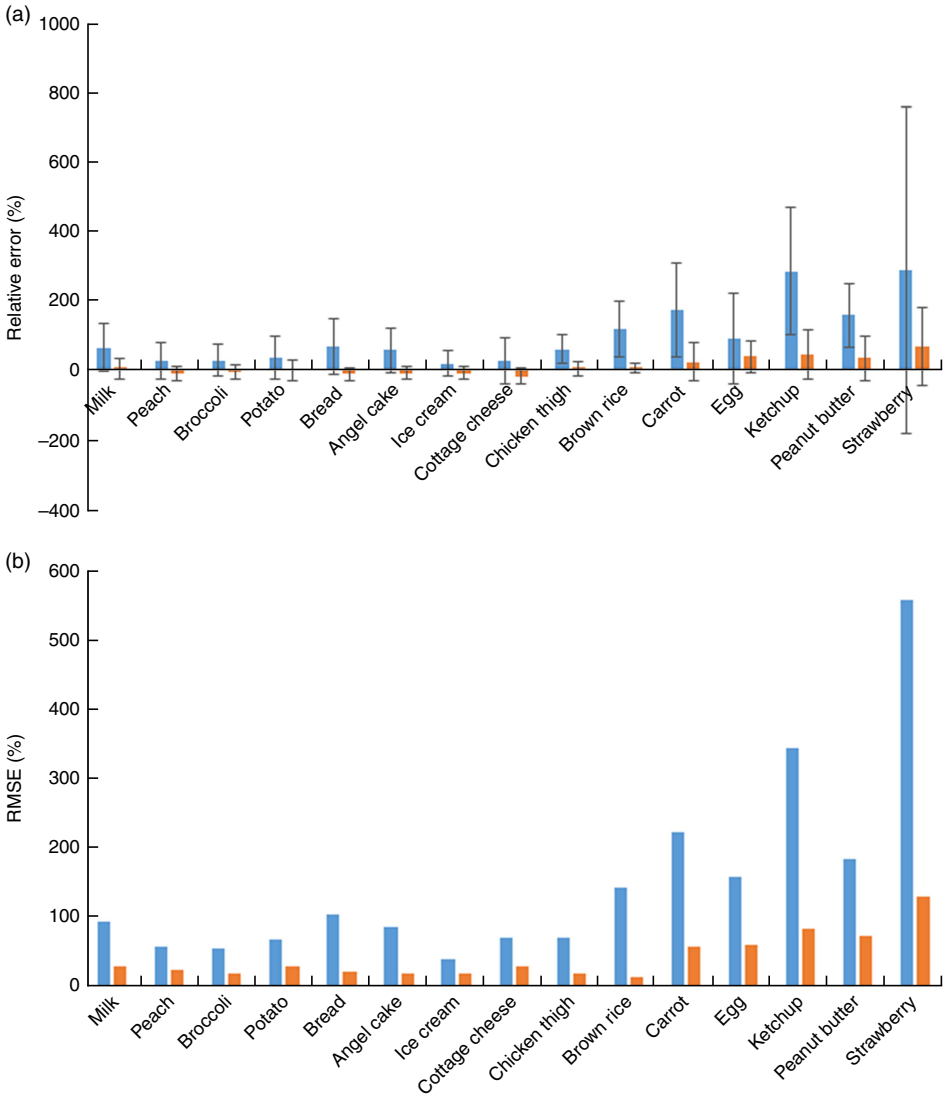

Fig. 8 (colour online) (a) Mean relative error ((estimate – ground truth)/ground truth × 100 %) with standard deviation indicated by vertical rules and (b) root mean-square error (RMSE) for each of the fifteen foods tested using the new approach of fiducial-marker-free image-based food portion size estimation using a smartphone: ■, Study 1, without training; ■, Study 2, with training

Currently, the VE is established manually. However, it is possible to develop an algorithm to initiate this equivalence using a computer and display the computed result for a person to validate or modify. This computational approach could make the table-side food volume estimation both quicker and more accurate.

Clearly, our purpose of computing food volume is to use it for the estimation of energy and nutrients. However, this final step of estimation is not straightforward. There are numerous unsolved problems, such as food/ingredient recognition and food database development/improvement. Despite solving these problems is important, the present work focuses on food volume estimation.

Limitations

Our fiducial-marker-free method for image-based food portion size estimation requires taking the food picture with the bottom of the smartphone sitting on the tabletop. This may become too restrictive in cases where the food is tall, and the single image may not cover the food adequately. In this case, the second image can be taken at an arbitrary position where the view is more desirable. Since some shape and dimensional features of the food have been available in the first image, this image can be used as an important reference while volume estimation is performed using the first image. Alternatively, computational algorithms could be developed to automatically transfer the calibration information from the first image to the second image so that volume estimation can be performed in the same way as that for the first image. Finally, we remark that there are limitations in our experiments. It should be noted that the number of human foods utilized in our study (i.e. sample size) was small, and the sample population (college students) was more educated than the average population. Thus, our subjects were digitally

https://doi.org/10.1017/S136898001800054X Published online by Cambridge University Press

knowledgeable, which may have helped them learn more quickly than the general population in using the computational tools. In addition, we have not yet been able to fully delineate the mechanisms of the large performance difference before and after training. It was likely that different individuals visualized the sizes of objects (including foods) differently, which affected their initial estimates. Further, the colour and contrast of the virtual IFU™ and grid displayed on the computer screen could also be factors affecting the initial estimates, even for experienced computer users. However, we believe that, after repeated training with feedbacks of true food volumes, the individual differences with respect to subject population, visualization of object size and response to computer display will all decrease and likely disappear, in a similar trend as that observed in our experiments.

Conclusion

The present work contributes to the field of smartphone-based food portion size estimation from images by: (i) elimination of the fiducial marker; (ii) achievement of a greater robustness and practical utility by the introduction of new concepts and application of advanced technologies, including the IFU™, virtual tablecloth, scalable cube as a reference and VE; and (iii) assessment and control of the estimation error using an automated training process. Our methods are suitable for implementation in an app, allowing the smartphone owner to perform table-side estimation of the energy and nutrients of his/her food before making a dietary decision. In the future, the presented technology may be combined with methods for automated food recognition[43] to quickly search a nutrient database and determine and monitor energy and nutrient intakes.

Acknowledgements

Acknowledgements: The authors acknowledge T.J. Carabuena, A.E. Richards, T.A. Pecarchik, S. Chow and S.A. Duran for their contributions to experimental studies. *Financial support:* Y.Y. and H.Z. acknowledge a research grant from the National Natural Science Foundation of China (NSFC) (grant number 61571026) and the Ministry of Science and Technology of China (MOST) (grant number 2016YFE0108100); W.J. and M.S. acknowledge the US National Institutes of Health (NIH) (grant numbers R01CA165255 and R21CA172864); T.B. acknowledges financial support from the School of Health Sciences and the Faculty of Health and Medicine of the University of Newcastle, Australia. The NSFC, MOST and NIH had no role in the design, analysis or writing of this article. *Conflict of interest:* The authors declare no conflict of interest. *Authorship:* Y.Y., M.S. and W.J. were responsible for research method design, mathematical analysis and experiments. T.B. contributed to experimental design and data analysis. H.Z. contributed to computational development. Y.Y., M.S., W.J. and T.B. contributed to drafting and editing of the manuscript. *Ethics of human subject participation*: The study was conducted according to the guidelines in the Declaration of Helsinki. Ethics of this study were authorized by the University of Pittsburgh Institutional Review Board.

References

1. Jeran S, Steinbrecher A & Pischon T (2016) Prediction of activity-related energy expenditure using accelerometer-derived physical activity under free-living conditions: a systematic review. *Int J Obes (Lond)* **40**, 1187–1197.
2. Huang Y, Xu J, Yu B *et al.* (2016) Validity of FitBit, Jawbone UP, Nike+ and other wearable devices for level and stair walking. *Gait Posture* **48**, 36–41.
3. Evenson KR, Goto MM & Furberg RD (2015) Systematic review of the validity and reliability of consumer-wearable activity trackers. *Int J Behav Nutr Phys Act* **12**, 159.
4. Taraldsen K, Chastin SF, Riphagen II *et al.* (2012) Physical activity monitoring by use of accelerometer-based body-worn sensors in older adults: a systematic literature review of current knowledge and applications. *Maturitas* **71**, 13–19.
5. Päßler S & Fischer W-J (2014) Food intake monitoring: automated chew event detection in chewing sounds. *IEEE J Biomed Health Inform* **18**, 278–289.
6. Bi Y, Lv M, Song C *et al.* (2016) AutoDietary: a wearable acoustic sensor system for food intake recognition in daily life. *IEEE Sensors J* **16**, 806–816.
7. Alshurafa N, Kalantarian H, Pourhomayoun M *et al.* (2015) Recognition of nutrition intake using time-frequency decomposition in a wearable necklace using a piezoelectric sensor. *IEEE Sensors J* **15**, 3909–3916.
8. Farooq M & Sazonov E (2016) Linear regression models for chew count estimation from piezoelectric sensor signals. In *Proceedings of the 2016 10th International Conference on Sensing Technology (ICST)*, Nanjing, China, 11–13 November 2016, pp. 1–5. New York: Institute of Electrical and Electronics Engineers; available at http://ieeexplore.ieee.org/document/7796222/
9. Hoover A & Sazonov E (2016) Measuring human energy intake and ingestive behavior: challenges and opportunities. *IEEE Pulse* **7**, 6–7.
10. Nørnberg TR, Houlby L, Jørgensen LN *et al.* (2014) Do we know how much we put on the plate? Assessment of the accuracy of self-estimated versus weighed vegetables and whole grain portions using an intelligent buffet at the FoodScape lab. *Appetite* **81**, 162–167.
11. Gemming L, Doherty A, Kelly P *et al.* (2013) Feasibility of a SenseCam-assisted 24-h recall to reduce under-reporting of energy intake. *Eur J Clin Nutr* **67**, 1095–1099.
12. Sun M, Burke LE, Mao ZH *et al.* (2014) eButton: a wearable computer for health monitoring and personal assistance. *Proc Des Autom Conf* **2014**, 1–6.
13. Sun M, Burke LE, Baranowski T *et al.* (2015) An exploratory study on a chest-worn computer for evaluation of diet, physical activity and lifestyle. *J Healthc Eng* **6**, 1–22.
14. Mezgec S & Korousic Seljak B (2017) NutriNet: a deep learning food and drink image recognition system for dietary assessment. *Nutrients* **9**, E657.
15. Hassannejad H, Matrella G, Ciampolini P *et al.* (2016) Food image recognition using very deep convolutional networks. In *MADiMa '16 – Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management*, Amsterdam, 16 October 2016, pp. 41–49. New York: Association for Computing Machinery; available at https://dl.acm.org/citation.cfm?id=2986042

16. Liu C, Cao Y, Luo Y *et al.* (2016) DeepFood: deep learning-based food image recognition for computer-aided dietary assessment. In *ICOST 2016 Proceedings of the 14th International Conference on Inclusive Smart Cities and Digital Health*, Wuhan, China, 25–27 May 2016, pp. 37–48. New York: Springer; available at https://dl.acm.org/citation.cfm?id=2960855
17. Christodoulidis S, Anthimopoulos M & Mougiakakou S (2015) Food recognition for dietary assessment using deep convolutional neural networks. In *New Trends in Image Analysis and Processing – ICIAP 2015 Workshops. ICIAP 2015. Lecture Notes in Computer Science*, vol. 9281, pp. 458–465 [V Murino, E Puppo, D Sona *et al.*, editors]. Cham: Springer.
18. Kagaya H, Aizawa K & Ogawa M (2014) Food detection and recognition using convolutional neural network. In *Proceedings of the 22nd ACM International Conference on Multimedia*, Orlando, FL, USA, 3–7 November 2014, pp. 1085–1088. New York: Association for Computing Machinery; available at https://dl.acm.org/citation.cfm?id=2647868
19. Jia W, Li Y, Qu R *et al.* (2018) Automatic food detection in egocentric images using artificial intelligence technology. *Public Health Nutr* (In the Press).
20. Beltran A, Dadabhoy H, Chen TA *et al.* (2016) Adapting the eButton to the abilities of children for diet assessment. In *Proceedings of Measuring Behavior 2016 – 10th International Conference on Methods and Techniques in Behavioral Research*, pp. 72–81 [A Spink, G Riedel, L Zhou *et al.*, editors]. http://www.measuringbehavior.org/files/2016/MB2016_Proceedings.pdf (accessed February 2018).
21. Pew Research Center (2017) Mobile Fact Sheet. http://www.pewinternet.org/fact-sheet/mobile/ (accessed February 2018).
22. Gemming L, Utter J & Ni Mhurchu C (2015) Image-assisted dietary assessment: a systematic review of the evidence. *J Acad Nutr Diet* **115**, 64–77.
23. Martin CK, Nicklas T, Gunturk B *et al.* (2014) Measuring food intake with digital photography. *J Hum Nutr Diet* **27**, Suppl. 1, 72–81.
24. Stumbo PJ (2013) New technology in dietary assessment: a review of digital methods in improving food record accuracy. *Proc Nutr Soc* **72**, 70–76.
25. Boushey CJ, Kerr DA, Wright J *et al.* (2009) Use of technology in children's dietary assessment. *Eur J Clin Nutr* **63**, Suppl. 1, S50–S57.
26. Steele R (2015) An overview of the state of the art of automated capture of dietary intake information. *Crit Rev Food Sci Nutr* **55**, 1929–1938.
27. Puri M, Zhu Z, Lubin J *et al.* (2013) Food recognition using visual analysis and speech recognition. *US Patent 2013/0260345 A1*. Menlo Park, CA: SRI International.
28. Hardesty L (2016) Voice-controlled calorie counter. Spoken-language app makes meal logging easier, could aid weight loss. *MIT News*, 24 March. http://news.mit.edu/2016/voice-controlled-calorie-counter-0324 (accessed February 2018).
29. Khanna N, Boushey CJ, Kerr D *et al.* (2010) An overview of the technology assisted dietary assessment project at Purdue University. In *Proceedings of the 2010 IEEE International Symposium on Multimedia (ISM)*, Taichung, Taiwan, 13–15 December 2010, pp. 290–295. New York: Institute of Electrical and Electronics Engineers; available at http://ieeexplore.ieee.org/document/5693855/
30. Ashman AM, Collins CE, Brown LJ *et al.* (2016) A brief tool to assess image-based dietary records and guide nutrition counselling among pregnant women: an evaluation. *JMIR Mhealth Uhealth* **4**, e123.
31. Pendergast FJ, Ridgers ND, Worsley A *et al.* (2017) Evaluation of a smartphone food diary application using objectively measured energy expenditure. *Int J Behav Nutr Phys Act* **14**, 30.
32. Bucher T, Weltert M, Rollo ME *et al.* (2017) The international food unit: a new measurement aid that can improve portion size estimation. *Int J Behav Nutr Phys Act* **14**, 124.
33. Bucher T, Rollo M, Matthias W *et al.* (2017) The international food unit (IFU) can improve food volume estimation. In Abstract Book of ISBNPA Conference, Victoria, Canada, 7–10 June 2017, p. 100. https://www.isbnpa.org/files/articles/2018/01/30/70/attachments/5a70f8d2d3cc2.pdf (accessed March 2018).
34. Jia W, Yue Y, Fernstrom JD *et al.* (2012) Image-based estimation of food volume using circular referents in dietary assessment. *J Food Eng* **109**, 76–86.
35. Yao N (2010) Food dimension estimation from a single image using structured lights. PhD Thesis, University of Pittsburgh.
36. Langston J (2016) This smartphone technology 3-D maps your meal and counts its calories. *UW News*, 19 January. http://www.washington.edu/news/2016/01/19/this-smartphone-app-3-d-maps-your-meal-and-counts-its-calories/ (accessed February 2018).
37. Zhang Z (2010) Food volume estimation from a single image using virtual reality technology. Masters Thesis, University of Pittsburgh.
38. Jia W, Chen HC, Yue Y *et al.* (2014) Accuracy of food portion size estimation from digital pictures acquired by a chest-worn camera. *Public Health Nutr* **17**, 1671–1681.
39. Chae J, Woo I, Kim S *et al.* (2011) Volume estimation using food specific shape templates in mobile image-based dietary assessment. *Proc SPIE* **7873**, 78730K.
40. Chen HC, Jia W, Yue Y *et al.* (2013) Model-based measurement of food portion size for image-based dietary assessment using 3D/2D registration. *Meas Sci Technol* **24**, 105701.
41. Ma Y, Soatto S, Kosecka J *et al.* (2004) *Interdisciplinary Applied Mathematics*. vol. 26*: An Invitation to 3-D Vision: From Images to Geometric Models*. New York: Springer.
42. Valkenburg RJ & McIvor AM (1998) Accurate 3D measurement using a structured light system. *Image Vis Comput* **16**, 99–110.
43. Eftimov T, Korosec P & Korousic Seljak B (2017) StandFood: standardization of foods using a semi-automatic system for classifying and describing foods according to FoodEx2. *Nutrients* **9**, 542.

Appendix

In this section, we explain the derivation for Eq. (10). Let $Z = 0$ which represents the tabletop. According to Eq. (1), we have:

$$\begin{bmatrix} U \\ V \\ W \end{bmatrix} = \boldsymbol{R} \begin{bmatrix} X \\ Y \\ 0 \end{bmatrix} + \boldsymbol{T}, \tag{13}$$

where

$$\boldsymbol{R} = \begin{bmatrix} R_{11} & R_{12} & R_{13} \\ R_{21} & R_{22} & R_{23} \\ R_{31} & R_{32} & R_{33} \end{bmatrix} \tag{14}$$

https://doi.org/10.1017/S136898001800054X Published online by Cambridge University Press

and

$$\boldsymbol{T} = \begin{bmatrix} T_1 \\ T_2 \\ T_3 \end{bmatrix}. \tag{15}$$

By substitution, Eqs (13) to (15) yield:

$$\begin{bmatrix} U \\ V \end{bmatrix} = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} + \begin{bmatrix} T_1 \\ T_2 \end{bmatrix} \tag{16}$$

and

$$W = \begin{bmatrix} R_{31} & R_{32} \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} + T_3. \tag{17}$$

Since $\boldsymbol{R}$ is a rotation matrix, its inverse,

$$\begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}^{-1},$$

exists:

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}^{-1} \left\{ \begin{bmatrix} U \\ V \end{bmatrix} - \begin{bmatrix} T_1 \\ T_2 \end{bmatrix} \right\}. \tag{18}$$

Substituting Eq. (18) into Eq. (17), we have:

$$W = \begin{bmatrix} R_{31} & R_{32} \end{bmatrix} \cdot \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}^{-1} \left\{ \begin{bmatrix} U \\ V \end{bmatrix} - \begin{bmatrix} T_1 \\ T_2 \end{bmatrix} \right\} + T_3. \tag{19}$$

Because

$$\begin{bmatrix} u \\ v \end{bmatrix} = \frac{f_c}{W} \begin{bmatrix} U \\ V \end{bmatrix}$$

as expressed in Eq. (3), we have:

$$\begin{bmatrix} U \\ V \end{bmatrix} = \frac{W}{f_c} \begin{bmatrix} u \\ v \end{bmatrix}. \tag{20}$$

Substituting Eq. (20) into Eq. (19), we obtain:

$$W = \begin{bmatrix} R_{31} & R_{32} \end{bmatrix} \cdot \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}^{-1} \left\{ \frac{W}{f_c} \begin{bmatrix} u \\ v \end{bmatrix} - \begin{bmatrix} T_1 \\ T_2 \end{bmatrix} \right\} + T_3. \tag{21}$$

Extracting W from both sides of Eq. (21) gives:

$$W = \frac{-\begin{bmatrix} R_{31} & R_{32} \end{bmatrix} \cdot \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}^{-1} \begin{bmatrix} T_1 \\ T_2 \end{bmatrix} + T_3}{1 - \begin{bmatrix} R_{31} & R_{32} \end{bmatrix} \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}^{-1} \begin{bmatrix} u \\ v \end{bmatrix} / f_c}. \tag{22}$$

For the specific setup in Fig. 3,

$$\boldsymbol{T} = \begin{bmatrix} 0 \\ -h \\ 0 \end{bmatrix}$$

and

$$\begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1/\cos\theta \end{bmatrix}.$$

Substituting them into Eq. (22), the relationship between W and v can be simplified as follows:

$$\begin{aligned} W &= \frac{-\begin{bmatrix} 0 & \sin\theta \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \\ 0 & 1/\cos\theta \end{bmatrix} \begin{bmatrix} 0 \\ -h \end{bmatrix} + 0}{1 - \begin{bmatrix} 0 & -h \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1/\cos\theta \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} / f_c} \\ &= \frac{h \cdot (\sin\theta / \cos\theta)}{1 - (v \cdot \sin\theta / f_c \cdot \cos\theta)} \\ &= \frac{h \cdot \sin\theta}{\cos\theta - (v / f_c) \cdot \sin\theta}, \end{aligned} \tag{23}$$

which is Eq. (10).

https://doi.org/10.1017/S136898001800054X Published online by Cambridge University Press