Public Health Nutrition: 22(7), 1168–1179 doi:10.1017/S1368980018000538

Automatic food detection in egocentric images using artificial intelligence technology

Wenyan Jia[1,*], Yuecheng Li[1], Ruowei Qu[1,2], Thomas Baranowski[3], Lora E Burke[4], Hong Zhang[5], Yicheng Bai[1], Juliet M Mancino[4], Guizhi Xu[2], Zhi-Hong Mao[6] and Mingui Sun[1,6,7]

[1]Department of Neurosurgery, University of Pittsburgh, 3520 Forbes Avenue, Suite 202, Pittsburgh, PA 15213, USA: [2]Department of Biomedical Engineering, Hebei University of Technology, Tianjin, People's Republic of China: [3]Department of Pediatrics, Baylor College of Medicine, Houston, TX, USA: [4]School of Nursing, University of Pittsburgh, Pittsburgh, PA, USA: [5]Image Processing Center, Beihang University, Beijing, People's Republic of China: [6]Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA, USA: [7]Department of Bioengineering, University of Pittsburgh, Pittsburgh, PA, USA

Submitted 18 April 2017: Final revision received 29 December 2017: Accepted 9 February 2018: First published online 26 March 2018

Abstract

Objective: To develop an artificial intelligence (AI)-based algorithm which can automatically detect food items from images acquired by an egocentric wearable camera for dietary assessment.

Design: To study human diet and lifestyle, large sets of egocentric images were acquired using a wearable device, called eButton, from free-living individuals. Three thousand nine hundred images containing real-world activities, which formed eButton data set 1, were manually selected from thirty subjects. eButton data set 2 contained 29 515 images acquired from a research participant in a week-long unrestricted recording. They included both food- and non-food-related real-life activities, such as dining at both home and restaurants, cooking, shopping, gardening, housekeeping chores, taking classes, gym exercise, etc. All images in these data sets were classified as food/non-food images based on their tags generated by a convolutional neural network.

Results: A cross data-set test was conducted on eButton data set 1. The overall accuracy of food detection was 91·5 and 86·4 %, respectively, when one-half of data set 1 was used for training and the other half for testing. For eButton data set 2, 74·0 % sensitivity and 87·0 % specificity were obtained if both 'food' and 'drink' were considered as food images. Alternatively, if only 'food' items were considered, the sensitivity and specificity reached 85·0 and 85·8 %, respectively.

Conclusions: The AI technology can automatically detect foods from low-quality, wearable camera-acquired real-world egocentric images with reasonable accuracy, reducing both the burden of data processing and privacy concerns.

Keywords
Food detection
Artificial intelligence
Wearable device
Deep learning
Egocentric image
Technology-assisted dietary intake estimation

In recent years, meal pictures taken by smartphones have been investigated as a novel tool for dietary assessment[1–7]. This method reduces the burden of collecting dietary data and the food pictures are helpful in refreshing people's memory about their food intake. However, picture taking must be volitionally initiated for each plate of food before and after consumption, which is inconvenient and impractical. This approach may also disrupt normal eating habits because of the picture-taking process. Thereby it is difficult to use this method to conduct long-term dietary assessment.

Since a wearable camera can record the scenes in front of the wearer continuously and automatically, its potential application in objective dietary studies has been explored[8–25]. For example, the SenseCam (originally developed by Microsoft[26]), the eButton (developed by our team)[13–15], an ear-worn micro-camera[8,21] and a camera-enabled cell phone worn around the neck[19] have been used to conduct dietary assessment and/or compared with 24 h recalls[8–12,16,17,23]. These studies demonstrated that, with the help of a wearable camera, not only can food intake be evaluated, but also the eating environment/behaviour can be studied. However, researchers, research subjects or 'Mechanical Turk Workers' (i.e. people who perform tasks that computers are currently unable to do)[27] must manually observe and

*Corresponding author: Email jiawenyan@gmail.com

identify eating episodes from large image data sets, which is both tedious and time-consuming[16–18]. Furthermore, privacy becomes a concern when the images of the wearer are observed by another person[28–32]. The current strategy is to ask the wearer to review all the images before sending to researchers and delete some images if necessary, or to control the on/off switch of the camera. This approach reduces privacy concerns but is very time-consuming or inconvenient.

To minimize the burden of image review and reduce privacy concerns caused by manual processing, additional sensor(s) can be attached to the body of the research subject which initiate picture taking when an eating activity is detected by the sensor(s). Piezoelectric sensors and microphones have been used to measure chewing and swallowing during eating[33–39]. Motion sensors, magnetic proximity sensors and infrared sensors have also been tested to detect hand movements during eating[40–45]. A number of other sensors, such as an audio sensor within an ear and a motion sensor on the head/wrist, have also been used to detect eating[21,24,46,47]. However, these sensors are often intrusive, uncomfortable and/or cosmetically unpleasant for long-term wear. A better solution would be if foods could be detected directly from the acquired images without human involvement and without adding extra sensors. With food detected and portion size estimated for each food, information about the nutrient and energy content of the food can be retrieved from a dietary database[23,24,48].

Artificial intelligence (AI) is a field aiming to develop tools/algorithms that allow machines to function intelligently (e.g. learn, reason, solve problems). A recent technological breakthrough in AI has enabled image/scene understanding to become increasingly accurate[49–54]. In the field of computer vision, automatic human or object detection and recognition from images has become an active research topic, for example differentiating food/non-food images or recognizing different types of food from high-quality images[55–66]. However, automatic detection of foods from the data acquired by a wearable camera is challenging because the quality of wearable camera-acquired images is usually low due to practical factors such as improper illumination, low image resolution, motion-caused blur, etc. In addition, the scenes in these images are not intentionally selected. As a result, the images may obtain numerous irrelevant objects but incomplete or even missing objects of interest. In the present study, we conducted an initial investigation on automatic food detection from real-world images acquired by a wearable device, called eButton, during daily life. eButton, which looks like a round chest badge, contains a miniature camera and a motion sensor[14,15]. It automatically takes images of the view in front of the wearer at a pre-set rate after being turned on. Due to the limit on the size and weight of the eButton, the performance of the camera is not as good as that of a smartphone (see Figs 1 and 5 for examples of eButton images). In the current paper we introduce our food detection algorithm and validate its performance using eButton data sets.

Methods

The deep-learning technology for AI has attracted great attention as it constantly breaks records in a variety of common benchmark tests[49,51,67–69]. Deep learning is a subfield of machine learning allowing multi-layered computational models (e.g. neural network with multiple hidden layers) to learn representations from tremendous amounts of raw data. Deep leaning has been successfully applied to a wide range of fields, such as automatic machine translation, computer vision and speech recognition. As one type of state-of-the-art deep-learning structure, the convolutional neural network (CNN) has shown exceptional performance in object detection/classification from images[49–53]. CNN is a feed-forward artificial neural network, typically consisting of a series of stacked convolutional layers and one or more fully connected layers. The output of this network (i.e. a layer of neurons) determines the class of the object in the input image. Training with annotated images is required for this network to learn how to differentiate objects. For example, a deep CNN trained using 1·2 million images is able to recognize 1000 different objects or 'concepts' with high accuracy[51]. However, large amounts of annotated images are required for training the deep network, and the training process is both system-demanding (e.g. special hardware required) and time-consuming. Thus, it is usually not very practical for nutrition scientists to implement this technology without substantial computational support.

Because large image data sets for object detection/recognition (i.e. ImageNet, Pascal VOC)[48,70,71] are already available, several CNN structures have been successfully trained and applied to these tasks. We thus utilized the object recognition results from an existing CNN for food detection, instead of building our own CNN. Figure 1 shows two examples in which objects in eButton-acquired real-world images are recognized and output is obtained in keywords (or tags)[72]. Although not all descriptions are correct, the potential of machine intelligence for detecting food images is clear. Combining the CNN-produced tags of a class of images, we can build histograms of these tags demonstrating the occurrence frequency of each tag. Two histograms computed separately for fifty images containing at least one food (we call them 'food images') and fifty images containing no food (we call them 'non-food images') are each visualized in Fig. 2 as a 'word cloud' using a web-provided software tool[73]. The frequently presented tags in these two classes are quite different. This observation inspired us to use tags to construct a classifier to detect whether an image contains edible objects or an eating activity, based on the contents of the image. Individual food-related tags (e.g. tag 'food' or 'dish') can be used

https://doi.org/10.1017/S1368980018000538 Published online by Cambridge University Press

Tags:

food, dish, dinner, bowl, meal, no person, plate, vegetable, restaurant, cooking, table

Tags:

group, adult, meeting, indoors, flatware, wood, grow, table, dairy product, spoon

Fig. 1 (colour online) Examples of the tags generated by Clarifai for eButton-acquired images. Some descriptions are correct, although not all of them. Red tags appear to be questionable ones

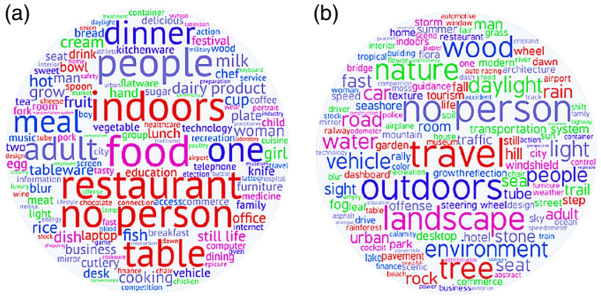

Fig. 2 (colour online) Word clouds of a set of tag histograms from fifty food images (a) and fifty non-food images (b)

as an indicator to detect food images, but it is not very reliable especially when the food covers only a small part of the image. For our study, we selected Clarifai CNN[72] which provides an easy-to-use Application Program Interface to produce the tag output from the input image. In addition, this particular CNN has been one of the top performers in the computer vision community. For example, it won first place in the image classification task at the ImageNet Large Scale Visual Recognition competition in 2013. According to the company's report[72], Clarifai can identify more than 11 000 general concepts including objects, ideas and emotions.

The actual computational algorithm is mathematical in nature and is described in the online supplementary material for interested readers. Here we only highlight the concepts involved. Since the CNN-generated tags are English words, we first construct a dictionary (which is essentially a subset of the ordinary English dictionary) containing possible words that the Clarifai CNN could produce. After the tag words of a specific image are obtained from this CNN, we look up the dictionary and record the locations of these tag words (analogous to 'page numbers') in the dictionary. Note that each location (analogous to each dictionary 'page') is pre-marked with 'food-related' or 'not food-related' (imagine that each dictionary page is coloured either red (denoting 'food-related') or yellow (denoting 'not food-related')), determined mathematically according to a probability calculation in a 'training' process. Now, the classification of the input image into the 'food' or 'non-food' class becomes simply counting the numbers of the red and yellow pages. Specifically, if the number of red pages is larger than a pre-set threshold, the image is determined to be 'food-related'; otherwise it is 'not food-related'.

We finally comment here that, in the past, determining whether an image contains at least one edible product was an extremely difficult task when only traditional image processing techniques were available. Now, with the availability of advanced AI technology and the algorithm described above, this task becomes straightforward.

Experimental results

We validated the performance of our AI-based image classification method using three data sets, including one public data set and two eButton data sets. The public data set is called Food-5K, a recently published benchmark for image classification tasks[55]. We used this data set to compare the performance of our algorithm with that from an existing study[55]. In this data set, the images were acquired by smartphones and handheld cameras rather than wearable devices, and most food images contain only a food with or without a container without other objects. Therefore, the classification of this data set was relatively easy. The other two data sets were constructed using our eButton images. eButton data set 1 contains 3900 egocentric images acquired from thirty wearers (research participants) who wore an eButton for different durations in their real living environment, divided equally as training and testing sets. Snack, food preparation and shopping images were also included as food images in this data set. eButton data set 2 was collected from only one female participant who wore an eButton for one entire week during daytime, containing 29 515 images. The performance of our AI system was evaluated using measures of sensitivity (recall), specificity, precision and overall accuracy, defined as follows:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}},$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}},$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}$$

and

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FN}+\text{FP}},$$

where TP (true positive), TN (true negative), FP (false positive) and FN (false negative) are clarified in Table 1. Sensitivity and specificity are also known as the 'true

https://doi.org/10.1017/S1368980018000538 Published online by Cambridge University Press

positive rate' and the 'true negative rate', respectively. Precision is also called 'positive predictive value', which is the ratio of true positives to combined true and false positives.

Food-5K data set

Food-5K contains 2500 food images and 2500 non-food images (see examples in Fig. 3). These images were selected from publicly available data sets, such as Food-101[74], UEC-FOOD-100[75], Caltech256[76], the Images of Groups of People[77] and Emotion6[78]. A human observer checked all images to guarantee that they were distinguishable. To make the classification task more challenging, the food item in the image may take up only a small part of the image (see Fig. 4 for examples). Three thousand images were selected as the training set, and two sets, each containing 1000 images, were used as the validation and evaluation set, respectively. The number of food and non-food images in each data set was equal to guarantee balanced classes. In the present study, we used only the evaluation set as the testing set to evaluate our algorithm and compare with the published result[55].

Our classification results on the evaluation set are shown in Table 2 using different similarity measures (Jaccard and Dice, explained in the online supplementary material). The tag dictionary was constructed from the training data set with n 761. It can be observed from Table 2 that when threshold was set to 3 (i.e. $k=3$), the overall accuracy, sensitivity, specificity and precision were 98·7, 98·2, 99·2 and 99·2 %, respectively, with the Dice similarity measure. We noticed that the misclassified images were mostly difficult cases that were challenging even for a human to perform the task (examples in Fig. 4). The overall accuracy reported by another group using the non-tag-based approach was 99·2 %[55], which was only slightly better than our result.

eButton data set 1: food/non-food data sets

A total of 3900 real-life images were selected from the images acquired by eButton. The resolution of these images is 640 pixels × 480 pixels. Half of the images, including 950 food images and 1000 non-food images, were acquired by twelve participants in two field studies conducted at the University of Pittsburgh. The other half were acquired mainly by eighteen lab members and collaborators during their daily life or travel. Since eButton recorded image sequences with the speed of one image every 2–5 s, adjacent images in the sequences were usually similar. Therefore, we down-sampled the image sequences by a factor of 10 so that the resulting images were separated by 20–50 s. Even after this down-sampling, some images were still quite similar. We deleted similar ones further manually to keep the number of images recorded from the same event to less than fifteen. The images containing likely private information, or that were too dark or too blurred, were also removed. All images were annotated to provide the ground truth (food or non-food) and some detailed information, such as 'eating a meal', 'having a drink', 'eating a snack', 'food preparation' and 'food shopping', which were all considered food images (see Table 3 and Fig. 5). Since, in the egocentric images, no eating behaviour of the wearer can be seen directly from the images, an image with a mug on the table or a cup in hand was considered as 'having a drink'. Comparing Fig. 5 and Fig. 3, it can be seen that the eButton image data set was more difficult to process due to uncontrolled imaging, low-resolution, unfavourable illumination, small objects within complex background and motion artifacts.

In this experiment, we first built a tag dictionary from 6000 randomly selected eButton images including different daily activities. Only a very small portion of these images contained food-related contents, so only a few food-related tags were included in this dictionary. To make the tag dictionary better fit the study, we also included 2560 food images (ten images from each food category) from a food image database[79]. In total, we got a dictionary with 1253 tags. Since the images in these two sets of data were acquired exclusively by different people, we conducted a cross data-set evaluation. The results are shown in Fig. 6. The Dice measure was used and ε was set to 0·05. It can be observed that threshold k is an important factor determining the classification result. When $k=2$, both sensitivity and specificity are high. The overall accuracy measures are 91·5 and 86·4 %, respectively, for these two cases (see Fig. 7). A more detailed analysis (in Table 4) shows that the classification accuracies in eating and food shopping category are higher than the accuracies in other categories.

Table 1 Definitions of true positive (TP), true negative (TN), false positive (FP) and false negative (FN)

	Predicted	
True label	Food	Non-food
Food	TP	FN
Non-food	FP	TN

Fig. 3 (colour online) Typical images in the Food-5K data set. Left four images are labelled as food images; the right four are non-food images

https://doi.org/10.1017/S1368980018000538 Published online by Cambridge University Press

Table 2 Classification results on the Food-5K data set

	k	TP	FN	TN	FP	Sensitivity (%)	Specificity (%)	Precision (%)	Accuracy (%)
Jaccard*,	1	496	4	454	46	99·2	90·8	91·5	95·0
$\varepsilon = 0{\cdot}05$	2	493	7	490	10	98·6	98·0	98·0	98·3
	3	490	10	496	4	98·0	99·2	99·2	98·6
	4	484	16	498	2	96·8	99·6	99·6	98·2
Dice*,	1	496	4	453	47	99·2	90·6	91·3	94·9
$\varepsilon = 0{\cdot}05$	2	493	7	490	10	98·6	98·0	98·0	98·3
	3	491	9	496	4	98·2	99·2	99·2	98·7
	4	487	13	497	3	97·4	99·4	99·4	98·4

TP, true positive; FN, false negative; TN, true negative; FP, false positive.
*The calculations of Jaccard and Dice are explained in the online supplementary material.

Fig. 4 (colour online) Misclassified images in the Food-5K data set. Left four were misclassified as food images; the right nine were misclassified as non-food images

Table 3 Categories and number of images in the eButton food/non-food data set

	Food images					
	Eating a meal	Having a drink	Eating a snack	Food shopping	Food preparation	Non-food images
Session 1	855	10	34	8	43	1000
Session 2	795	84	25	15	31	1000

eButton data set 2: one-week data set

With institutional review board approval, we recorded whole-day images using eButton. A female participant wore an eButton continuously during daytime for one week. As a result, a variety of real-life activities were recorded in these images, including eating/drinking/cooking, attending parties, garden work, taking classes, gym exercise, etc. All seven image sequences (one per day) were down-sampled first (resulting in a picture-taking rate of one image every 10 s). Human faces were detected using the Picasa software and blocked before sending to an observer, who further filtered out images with possible privacy concerns and annotated images in the same way as described in the last case. After these procedures, 29 515 images were obtained, exemplified in Table 5. All these images were then processed by the AI software (Clarifai CNN) to generate twenty tags for each image. Based on these tags, the number of food-related tags for each image, defined as the evidence index, was calculated, as described in the online supplementary material. If this number was higher than the pre-set threshold k, we determined it as a food image. The tag dictionary was obtained using the same approach as described for eButton data set 1 and food vector $\boldsymbol{y}$ was trained using eButton data set 1.

Table 5 shows that the numbers of real food images, including 'food' and 'drink' images, were much lower than the number of the 'other' (i.e. non-food) images. In this case, the overall precision and accuracy are not good performance measures because of the imbalance between the sample sizes. Since images classified as food images will be further studied to obtain dietary information, which is a time-consuming procedure, it is desirable to minimize the total number of positive images (being classified as food images, including true positives and false positives) while limiting the impact on both sensitivity and specificity measures. Since the evidence index represents how many tags of each image belong to food images, a higher value of threshold k results in fewer positive images.

https://doi.org/10.1017/S1368980018000538 Published online by Cambridge University Press

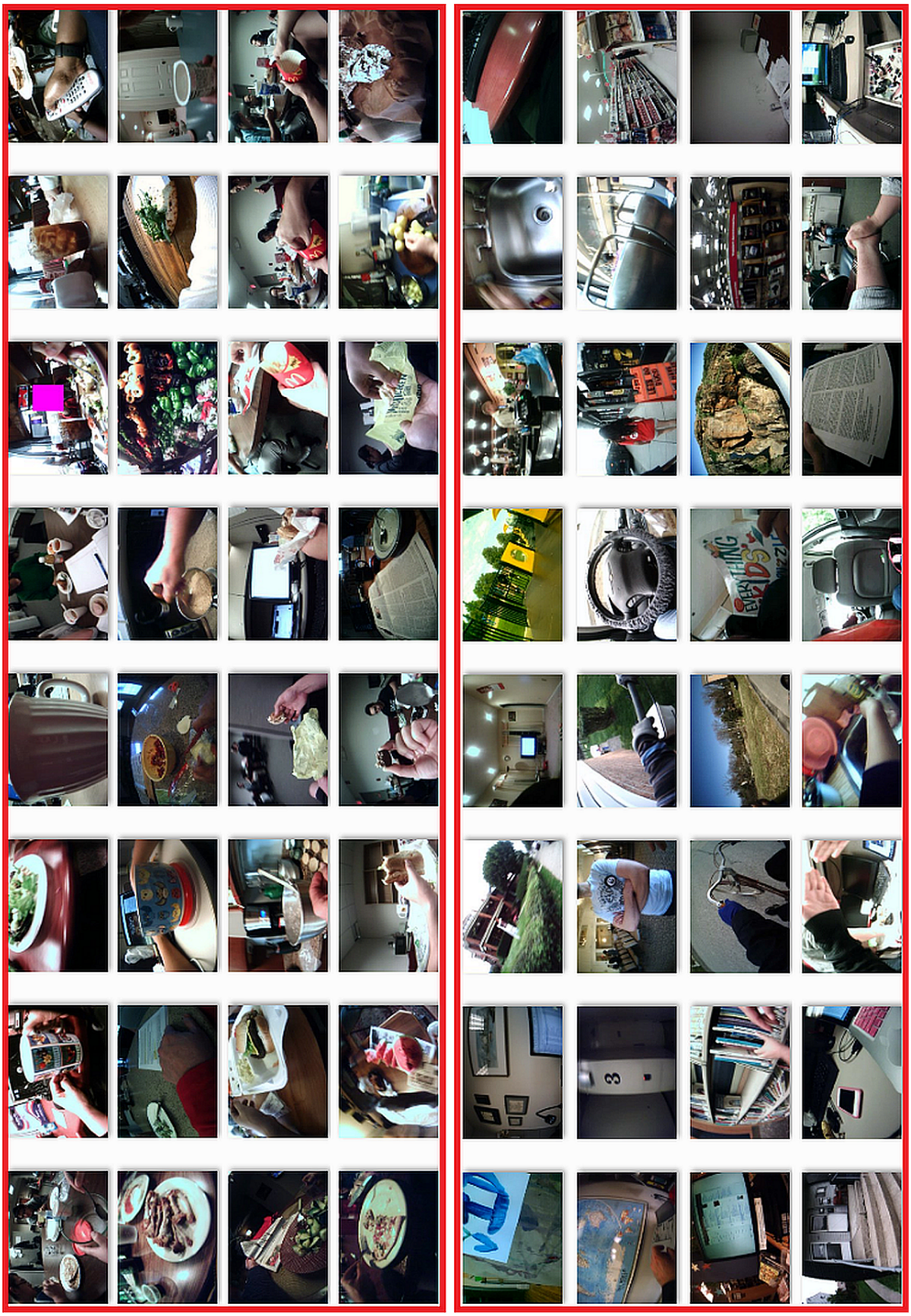

Fig. 5 (colour online) Examples in eButton data set 1. Images in the top four rows are labelled as food images, and those in the bottom four rows are non-food images. Compared with the images in Fig. 3, it can be seen that the egocentric images were more difficult to classify

https://doi.org/10.1017/S1368980018000538 Published online by Cambridge University Press

In our experiment, we defined a burden index to represent the ratio between the number of total positive images and the number of all images (see Fig. 8). When $k=1$, 37·8 % of the total images needed a further check, which may be too much burden for researchers although the overall sensitivity was as high as 89·5 %. Threshold k is an adjustable parameter in this algorithm. Smaller k results in higher sensitivity, but also higher burden. When $k=2$, the overall burden decreased to 18 % with 74·0 % sensitivity and 87·0 % specificity (shown in Fig. 8). The sensitivity of day 1 and day 6 was significantly lower than on the other days. After extracting all of the false negative images, we found on the first day that 206 out of 354 'drink' images (similar to the first two images in Fig. 9) were missed due to the dark environment and the small coffee cup, while on sixth day, seventy-eight out of 154 eating images were not found probably because of the small food item and the overexposed images (similar to the last three images in Fig. 10).

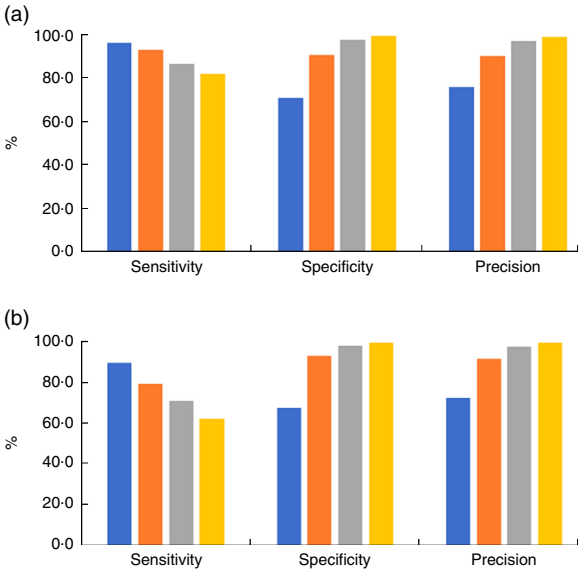

Fig. 6 (colour online) The effect of threshold k (■, $k=1$; ■, $k=2$; ■, $k=3$; ■, $k=4$) on the sensitivity, specificity and precision in the cross data-set evaluation of the eButton data set 1: (a) case 1 (training: session 2, testing: session 1); (b) case 2 (training: session 1, testing: session 2)

Table 5 Durations of recording and numbers of images in the seven-day study

	Total no. of images	No. of 'food' images	No. of 'drink' images	No. of 'other' images	Special food-related events
Day 1	4869	250	354	4265	A cup on table lasting about 75 min
Day 2	3400	183	50	3167	
Day 3	4513	265	185	4063	
Day 4	4254	233	92	3929	Attend a baby shower event, about 30 min
Day 5	4213	288	0	3925	Activities at other's house, about 5 h
Day 6	4367	154	3	4210	Make muffin, about 70 min in total, then go to party, about 3 h
Day 7	3899	170	146	3583	
Total	29 515	1543	830	27 142	

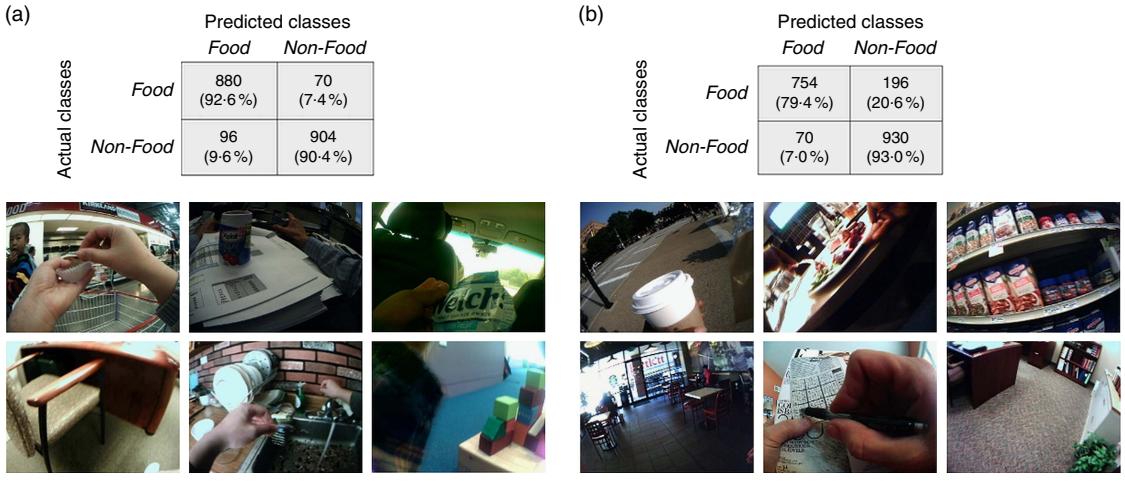

Fig. 7 (colour online) Tag-based classification results in the cross data-set evaluation of the eButton data set 1 and misclassified examples (top images were misclassified as non-food images and bottom images were misclassified as food images): (a) case 1 (training: session 2, testing: session 1); (b) case 2 (training: session 1, testing: session 2)

Table 4 Accuracy of food image detection in different categories of the eButton food/non-food data set

	Eating a meal	Having a drink	Eating a snack	Food shopping	Food preparation
Case 1	(855 − 41)/855 = 95·2 %	(10 − 5)/10 = 50·0 %	(8 − 4)/8 = 50·0 %	(43 − 2)/43 = 95·4 %	(34 − 18)/34 = 47·1 %
Case 2	(795 − 145)/795 = 81·8 %	(84 − 31)/84 = 63·1 %	(15 − 6)/15 = 60·0 %	(31 − 3)/31 = 90·3 %	(25 − 11)/25 = 56·0 %

https://doi.org/10.1017/S1368980018000538 Published online by Cambridge University Press

The results of the one-week data set are summarized in Fig. 11(a) where the red bars represent true 'food' images, the green bars represent true 'drink' images and grey bars represent the detected food images. The calculated evidence index for each image was averaged over a 1 min window first and then binarized by threshold *k*. It can be observed that most of the eating/drinking events (i.e. a cluster of red or green bars) were detected by our AI approach.

From Fig. 11(a), scattered false positives can be found which were caused mainly by unattended drinks in the scene of the eButton. The research participant brewed or purchased a cup of coffee for breakfast and often left it on the table or held it by hand while performing other activities, such as taking a class or reading a newspaper (Fig. 9). If we consider the 'drink' images as non-food images, the sensitivity can reach 85·0 %, while specificity remained at 85·8 % (Fig. 12).

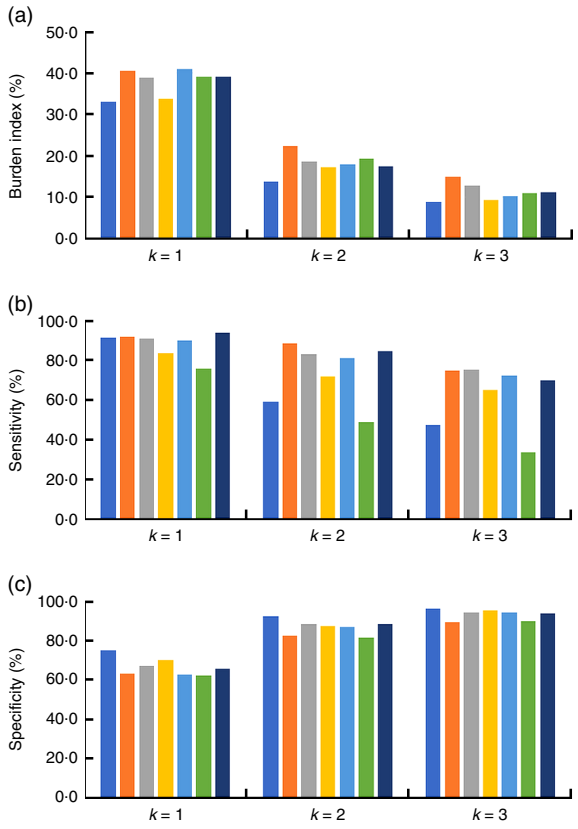

Fig. 8 (colour online) (a) Burden index, (b) sensitivity and (c) specificity in the eButton one-week data set with changing *k* (■, day 1; ■, day 2; ■, day 3; ■, day 4; ■, day 5; ■, day 6; ■, day 7)

Discussion

Our experimental results indicate that the AI algorithm performed well on both the Food-5K and eButton data sets. However, the performance on the Food-5K data set was better. This difference may be due to three reasons. First, blurred pictures cannot be avoided if images are recorded when the wearer of the eButton is moving. When applying a blur detection algorithm to the one-week eButton data set, we found that 5233 images (17·7 %) were blurred[80]. Several blurred images are shown in Fig. 10. If these images were removed before processing with the proposed method, the overall sensitivity increased only by about 1 %, which shows that the proposed algorithm is robust to blur. Second, in some pictures, the food covered only a very small part of the image, especially when drinking. In the one-week data set, the wearer sometimes eats/drinks while reading or doing computer work, so the food/plate is located in the corner of the images, see Figs 9 and 10 for examples. Third, compared with the Food-5K data set, more objects were included in the egocentric images due to the use of a wide-angle camera in eButton (120° field of view) and the passive image capture. It makes the detection task more challenging.

Fig. 9 (colour online) Several images in the eButton one-week data set annotated as 'drink' because a cup can be seen on the table or in hand

https://doi.org/10.1017/S1368980018000538 Published online by Cambridge University Press

Fig. 10 (colour online) Examples of blurred images (top row) and images with a small portion of food (bottom row)

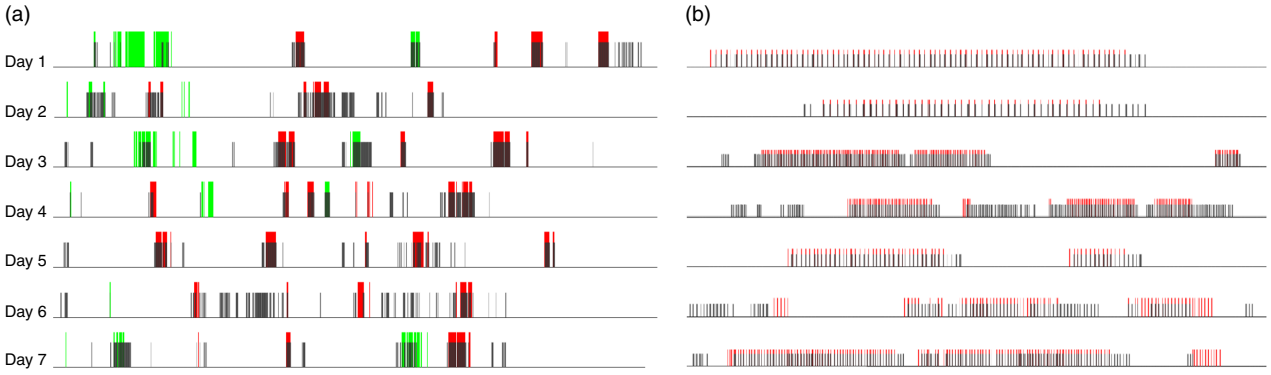

Fig. 11 (a) (colour online) (a) Results of food detection for eButton one-week data set. The red bars represent true 'food' images, the green bars represent true 'drink' images and grey bars represent the detected food images. (b) A zoomed-in picture of the last meal in each day

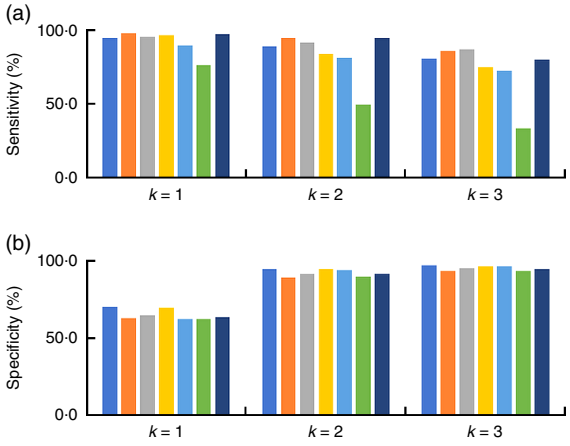

Fig. 12 (colour online) Classification results in the eButton one-week data set when only 'food' images were considered as food images (■, day 1; ■, day 2; ■, day 3; ■, day 4; ■, day 5; ■, day 6; ■, day 7)

In our previous work, all of the images were observed by researchers and dietary-related images were chosen manually for nutritional analysis[17]. This is the most time-consuming part in the whole data analysis procedure. If research burden is a significant concern, we can adjust the threshold k to compromise between research burden and sensitivity, as shown in Fig. 8. In addition, because a sequence of images can be acquired with eButton, whole eating/drinking events can be recorded. Even if some food images were not detected using our algorithm, missing a whole meal seldom happened. When we zoom in on the last meal of each day in Fig. 11(a), we can see most of the red bars are covered by the grey bars, which means these images are detected (see Fig. 11(b)). The uncovered red bars do not provide much new information since they were acquired from the same meal. We also found that some shopping or food preparation images were misclassified as food images. Checking these images will provide supplementary information in the actual dietary assessment. If, in a specific study, only food/drink episodes are of interest, we could use the motion data also acquired from eButton, which contains a three-axis accelerometer and a three-axis gyroscope, to exclude these scenarios.

Acknowledgements

Acknowledgements: The authors would like to acknowledge all participants for their significant contributions to this research study, as well as Clarifai for providing online service. *Financial support:* This work was supported by the

https://doi.org/10.1017/S1368980018000538 Published online by Cambridge University Press

National Institutes of Health (NIH) (W.J., M.S., Z.-H.M., grant number U01HL91736), (W.J., M.S., Z.-H.M., L.E.B., grant number R01CA165255), (W.J., M.S., T.B., grant number R21CA172864); the National Natural Science Foundation of China (H.Z., grant number 61571026); and the National Key Research and Development Program of China (H.Z., M.S., W.J., grant number 2016YFE0108100). The NIH and other agencies had no role in the design, analysis or writing of this article. *Conflict of interest:* None to declare. *Authorship:* W.J., Y.L. and R.Q. were responsible for image collection/annotation/analysis. G.X., H.Z., Z.-H.M. and M.S. contributed to the algorithm for data analysis. Y.L. and Y.B. designed and constructed the prototype of the eButtons used in this study. T.B., L.E.B. and J.M.M. conducted the field studies for acquiring images. W.J., T.B., L.E.B. and M.S. contributed to final drafting and editing of the manuscript. *Ethics of human subject participation:* This study was approved by the University of Pittsburgh Institutional Review Board.

Supplementary material

To view supplementary material for this article, please visit https://doi.org/10.1017/S1368980018000538

References

1. 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.
2. Martin CK, Nicklas T, Gunturk B *et al.* (2014) Measuring food intake with digital photography. *J Hum Nutr Diet* **27**, Suppl. 1, 72–81.
3. Stumbo PJ (2013) New technology in dietary assessment: a review of digital methods in improving food record accuracy. *Proc Nutr Soc* **72**, 70–76.
4. 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.
5. 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.
6. Boushey C, Spoden M, Zhu F *et al.* (2017) New mobile methods for dietary assessment: review of image-assisted and image-based dietary assessment methods. *Proc Nutr Soc* **76**, 283–294.
7. Illner AK, Freisling H, Boeing H *et al.* (2012) Review and evaluation of innovative technologies for measuring diet in nutritional epidemiology. *Int J Epidemiol* **41**, 1187–1203.
8. Pettitt C, Liu J, Kwasnicki RM *et al.* (2016) A pilot study to determine whether using a lightweight, wearable micro-camera improves dietary assessment accuracy and offers information on macronutrients and eating rate. *Br J Nutr* **115**, 160–167.
9. O'Loughlin G, Cullen SJ, McGoldrick A *et al.* (2013) Using a wearable camera to increase the accuracy of dietary analysis. *Am J Prev Med* **44**, 297–301.
10. Gemming L, Rush E, Maddison R *et al.* (2015) Wearable cameras can reduce dietary under-reporting: doubly labelled water validation of a camera-assisted 24 h recall. *Br J Nutr* **113**, 284–291.
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. Arab L, Estrin D, Kim DH *et al.* (2011) Feasibility testing of an automated image-capture method to aid dietary recall. *Eur J Clin Nutr* **65**, 1156–1162.
13. Sun M, Fernstrom JD, Jia W *et al.* (2010) A wearable electronic system for objective dietary assessment. *J Am Diet Assoc* **110**, 45–47.
14. Sun M, Burke LE, Mao Z-H *et al.* (2014) eButton: a wearable computer for health monitoring and personal assistance. *Proc Des Autom Conf* **2014**, 1–6.
15. 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.
16. Beltran A, Dadhaboy H, Lin C *et al.* (2015) Minimizing memory errors in child dietary assessment with a wearable camera: formative research. *J Acad Nutr Diet* **115**, Suppl. A86.
17. 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).
18. Thomaz E, Parnami A, Essa I *et al.* (2013) Feasibility of identifying eating moments from first-person images leveraging human computation. In *SenseCam '13. Proceedings of the 4th International SenseCam & Pervasive Imaging Conference*, San Diego, CA, USA, 18–19 November 2013, pp. 26–33. New York: Association for Computing Machinery; available at https://dl.acm.org/citation.cfm?id=2526672&dl=ACM&coll=DL
19. Arab L & Winter A (2010) Automated camera-phone experience with the frequency of imaging necessary to capture diet. *J Am Diet Assoc* **110**, 1238–1241.
20. Gemming L, Doherty A, Utter J *et al.* (2015) The use of a wearable camera to capture and categorise the environmental and social context of self-identified eating episodes. *Appetite* **92**, 118–125.
21. Liu J, Johns E, Atallah L *et al.* (2012) An intelligent food-intake monitoring system using wearable sensors. In *Proceedings of the 2012 Ninth International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, London, 9–12 May 2012, pp. 154–160. New York: Institute of Electrical and Electronics Engineers; available at http://ieeexplore.ieee.org/document/6200559/
22. Illner AK, Slimani N & Lachat C (2014) Assessing food intake through a chest-worn camera device. *Public Health Nutr* **17**, 1669–1670.
23. Jia W, Chen H-C, 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.
24. Vu T, Lin F, Alshurafa N *et al.* (2017) Wearable food intake monitoring technologies: a comprehensive review. *Computers* **6**, 4.
25. Vieira Resende Silva B & Cui J (2017) A survey on automated food monitoring and dietary management systems. *J Health Med Inform* **8**, 3.
26. Microsoft Research (2004) SenseCam. https://www.microsoft.com/en-us/research/project/sensecam/ (accessed March 2018).
27. Chen H-C, 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.
28. Doherty AR, Hodges SE, King AC *et al.* (2013) Wearable cameras in health: the state of the art and future possibilities. *Am J Prev Med* **44**, 320–323.

29. Safavi S & Shukur Z (2014) Conceptual privacy framework for health information on wearable device. *PLoS One* **9**, e114306.
30. McCarthy M (2016) Federal privacy rules offer scant protection for users of health apps and wearable devices. *BMJ* **354**, i4115.
31. Roberto H, Robert T, Steven A *et al.* (2014) Privacy behaviors of lifeloggers using wearable cameras. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, Seattle, WA, USA, 13–17 September 2014, pp. 571–582. New York: Association for Computing Machinery; available at https://dl.acm.org/citation.cfm?doid=2632048.2632079
32. Kelly P, Marshall SJ, Badland H *et al.* (2013) An ethical framework for automated, wearable cameras in health behavior research. *Am J Prev Med* **44**, 314–319.
33. Kalantarian H & Sarrafzadeh M (2015) Audio-based detection and evaluation of eating behavior using the smartwatch platform. *Comput Biol Med* **65**, 1–9.
34. Fukuike C, Kodama N, Manda Y *et al.* (2015) A novel automated detection system for swallowing sounds during eating and speech under everyday conditions. *J Oral Rehabil* **42**, 340–347.
35. Gao Y, Zhang N, Wang H *et al.* (2016) iHear food: eating detection using commodity bluetooth headsets. In *Proceedings of the 2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, Washington, DC, 27–29 June 2016, pp. 163–172. New York: Institute of Electrical and Electronics Engineers; available at http://ieeexplore.ieee.org/document/7545830/
36. Xu Y, Guanling C & Yu C (2015) Automatic eating detection using head-mount and wrist-worn accelerometers. In *Proceedings of the 2015 17th International Conference on E-health Networking, Application & Services (HealthCom)*, Boston, MA, USA, 13–17 October 2015, pp. 578–581. New York: Institute of Electrical and Electronics Engineers; available at http://ieeexplore.ieee.org/document/7454568/
37. Amft O, Junker H & Troster G (2005) Detection of eating and drinking arm gestures using inertial body-worn sensors. In *Proceedings of the Ninth IEEE International Symposium on Wearable Computers (ISWC'05)*, Washington, DC, 18–21 October 2005, pp. 160–163. New York: Institute of Electrical and Electronics Engineers; available at http://ieeexplore.ieee.org/document/1550801/
38. Dong Y, Scisco J, Wilson M *et al.* (2013) Detecting periods of eating during free-living by tracking wrist motion. *IEEE J Biomed Health Inform* **18**, 1253–1260.
39. Päßler S & Fischer W-J (2014) Food intake monitoring: automated chew event detection in chewing sounds. *IEEE J Biomed Health Inform* **18**, 278–289.
40. Ramos Garcia R (2014) Using hidden Markov models to segment and classify wrist motions related to eating activities. PhD Thesis, Clemson University.
41. Zhang S, Ang MH Jr, Xiao W *et al.* (2009) Detection of activities by wireless sensors for daily life surveillance: eating and drinking. *Sensors (Basel)* **9**, 1499–1517.
42. Thomaz E, Essa I & Abowd GD (2015) A practical approach for recognizing eating moments with wrist-mounted inertial sensing. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, Osaka, Japan, 7–11 September 2015, pp. 1029–1040. New York: Association for Computing Machinery; available at https://dl.acm.org/citation.cfm?id=2807545
43. Ye X, Chen G, Gao Y *et al.* (2016) Assisting food journaling with automatic eating detection. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, San Jose, CA, USA, 7–12 May 2016, pp. 3255–3262. New York: Association for Computing Machinery; available at https://dl.acm.org/citation.cfm?doid=2851581.2892426
44. Li C, Bai Y, Jia W *et al.* (2013) Eating event detection by magnetic proximity sensing. In *Proceedings of the 2013 39th Annual Northeast Bioengineering Conference (NEBEC)*, Syracuse, NY, USA, 5–7 April 2013, pp. 15–16. New York: Institute of Electrical and Electronics Engineers; available at http://ieeexplore.ieee.org/document/6574334/
45. Bai Y, Jia W, Mao Z-H *et al.* (2014) Automatic eating detection using a proximity sensor. In *Proceedings of the 2014 40th Annual Northeast Bioengineering Conference (NEBEC)*, Boston, MA, USA, 25–27 April 2014, pp. 1–2. New York: Institute of Electrical and Electronics Engineers; available at http://ieeexplore.ieee.org/document/6972716/
46. Mirtchouk M, Merck C & Kleinberg S (2016) Automated estimation of food type and amount consumed from body-worn audio and motion sensors. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, Heidelberg, Germany, 12–16 September 2016, pp. 451–462. New York: Association for Computing Machinery; available at https://dl.acm.org/citation.cfm?id=2971677
47. Rahman T, Czerwinski M, Gilad-Bachrach R *et al.* (2016) Predicting 'about-to-eat' moments for just-in-time eating intervention. In *Proceedings of the 6th International Conference on Digital Health*, Montréal, QC, Canada, 11–13 April 2016, pp. 141–150. New York: Association for Computing Machinery; available at https://dl.acm.org/citation.cfm?id=2896359
48. Chang X, Ye H, Albert P *et al.* (2013) Image-based food volume estimation. *CEA13 (2013)* **2013**, 75–80.
49. LeCun Y, Bengio Y & Hinton G (2015) Deep learning. *Nature* **521**, 436–444.
50. Farabet C, Couprie C, Najman L *et al.* (2013) Learning hierarchical features for scene labeling. *IEEE Trans Pattern Anal Mach Intell* **35**, 1915–1929.
51. Krizhevsky A, Sutskever I & Hinton GE (2012) ImageNet classification with deep convolutional neural networks. In *NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems – Volume 1*, Lake Tahoe, NV, USA, 3–6 December 2012, pp. 1097–1105. New York: Association for Computing Machinery; available at https://dl.acm.org/citation.cfm?id=2999257
52. Karpathy A & Li F-F (2015) Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 7–12 June 2015, pp. 3128–3137. New York: Institute of Electrical and Electronics Engineers; available at http://ieeexplore.ieee.org/document/7534740/
53. Johnson J, Karpathy A & Li F-F (2016) DenseCap: fully convolutional localization networks for dense captioning. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 27–30 June 2016, pp. 4565–4574. New York: Institute of Electrical and Electronics Engineers; available at http://ieeexplore.ieee.org/document/7780863/
54. Johnson J, Krishna R, Stark M *et al.* (2015) Image retrieval using scene graphs. *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 7–12 June 2015, pp. 3668–3678. New York: Institute of Electrical and Electronics Engineers; available at http://ieeexplore.ieee.org/document/7298990/
55. Singla A, Yuan L & Ebrahimi T (2016) Food/non-food image classification and food categorization using pre-trained GoogLeNet model. In *MADiMa '16 Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management*, Amsterdam, 16 October 2016, pp. 3–11. New York: Association for Computing Machinery; available at https://dl.acm.org/citation.cfm?id=2986039

56. Kagaya H & Aizawa K (2015) Highly accurate food/non-food image classification based on a deep convolutional neural network. In *New Trends in Image Analysis and Processing – ICIAP 2015 Workshops: ICIAP 2015 International Workshops, BioFor, CTMR, RHEUMA, ISCA, MADiMa, SBMI, and QoEM, Genoa, Italy, September 7–8, 2015, Proceedings*, pp. 350–357 [V Murino, E Puppo, D Sona *et al.*, editors]. Cham: Springer International Publishing.
57. Ragusa F, Tomaselli V, Furnari A *et al.* (2016) Food vs non-food classification. In *MADiMa '16 Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management*, Amsterdam, 16 October 2016, pp. 77–81. New York: Association for Computing Machinery; available at https://dl.acm.org/citation.cfm?id=2986041
58. Farinella GM, Allegra D, Stanco F *et al.* (2015) On the exploitation of one class classification to distinguish food vs non-food images. In *New Trends in Image Analysis and Processing – ICIAP 2015 Workshops: ICIAP 2015 International Workshops, BioFor, CTMR, RHEUMA, ISCA, MADiMa, SBMI, and QoEM, Genoa, Italy, September 7–8, 2015, Proceedings*, pp. 375–383 [V Murino, E Puppo, D Sona *et al.*, editors]. Cham: Springer International Publishing.
59. 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=2654970
60. Kawano Y & Yanai K (2014) Food image recognition with deep convolutional features. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, Seattle, WA, USA, 13–17 September 2014, pp. 589–593. New York: Association for Computing Machinery; available at https://dl.acm.org/citation.cfm?doid=2638728.2641339
61. 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 International Workshops, BioFor, CTMR, RHEUMA, ISCA, MADiMa, SBMI, and QoEM, Genoa, Italy, September 7–8, 2015, Proceedings*, pp. 458–465 [V Murino, E Puppo, D Sona *et al.*, editors]. Cham: Springer International Publishing.
62. Yanai K & Kawano Y (2015) Food image recognition using deep convolutional network with pre-training and fine-tuning. In *Proceedings of the 2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, Turin, Italy, 29 June–3 July 2015, pp. 1–6. New York: Institute of Electrical and Electronics Engineers; available at http://ieeexplore.ieee.org/document/7169816/
63. 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 – Volume 9677*, Wuhun, China, 25–27 May 2016, pp. 37–48. New York: Springer-Verlag New York Inc.; available at https://dl.acm.org/citation.cfm?id=2960855
64. 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
65. Farinella GM, Allegra D, Moltisanti M *et al.* (2016) Retrieval and classification of food images. *Comput Biol Med* **77**, 23–39.
66. Mezgec S & Korousic Seljak B (2017) NutriNet: a deep learning food and drink image recognition system for dietary assessment. *Nutrients* **9**, E657.
67. Hinton G (2012) Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Process Mag* **29**, 82–97.
68. Mohamed AR, Dahl GE & Hinton G (2012) Acoustic modeling using deep belief networks. *IEEE Trans Audio Speech Lang Process* **20**, 14–22.
69. LeCun Y & Bengio Y (1995) Convolutional network for images, speech, and time series. In *The Handbook of Brain Theory and Neural Networks*, p. 3361 [MA Arbib, editor]. Cambridge, MA: MIT Press.
70. Stanford Vision Lab, Stanford University & Princeton University (2016) ImageNet: image database organized according to the WordNet hierarchy. http://www.image-net.org/index (accessed March 2018).
71. Everingham M, Gool L, Williams C *et al.* (2012) Pascal VOC: various object recognition challenges. http://host.robots.ox.ac.uk/pascal/VOC/ (accessed March 2018).
72. Clarifai, Inc. (2016) Clarifai. http://www.clarifai.com/ (accessed July 2016).
73. WordArt (2016) Tagul – WordCloud Art. https://wordart.com/tagul (accessed March 2018).
74. Bossard L, Guillaumin M & Van Gool L (2014) Food-101 – mining discriminative components with random forests. In *Proceedings of the 13th European Conference on Computer Vision (ECCV 2014), Part IV*, Zurich, Switzerland, 6–12 September 2014, pp. 446–461. Cham: Springer International Publishing.
75. Matsuda Y, Hoashi H & Yanai K (2012) Recognition of multiple-food images by detecting candidate regions. In *Proceedings of the 2012 IEEE International Conference on Multimedia and Expo (ICME)*, Melbourne, Australia, 9–12 July 2012, pp. 25–30. New York: Institute of Electrical and Electronics Engineers; available at http://ieeexplore.ieee.org/document/6298369/
76. Griffin G, Holub A & Perona P (2007) *Caltech-256 Object Category Dataset Technical Report 7694*. Pasadena, CA: California Institute of Technology.
77. Gallagher AC & Chen T (2009) Understanding images of groups of people. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Miami, FL, USA, 20–25 June 2009, pp. 256–263. New York: Institute of Electrical and Electronics Engineers; available at http://ieeexplore.ieee.org/document/5206828/
78. Peng K-C, Chen T, Sadovnik A *et al.* (2015) A mixed bag of emotions: model, predict, and transfer emotion distributions. In *Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 7–12 June 2015, pp. 860–868. New York: Institute of Electrical and Electronics Engineers; available at http://ieeexplore.ieee.org/document/7298687/
79. Kawano Y & Yanai K (2015) Automatic expansion of a food image dataset leveraging existing categories with domain adaptation. In *Computer Vision – ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part III*, pp. 3–17 [L Agapito, MM Bronstein and C Rother, editors]. Cham: Springer International Publishing.
80. Pech-Pacheco JL, Cristóbal G, Chamorro-Martinez J *et al.* (2000) Diatom autofocusing in brightfield microscopy: a comparative study. In *Proceedings of the 15th IEEE International Conference on Pattern Recognition*, Barcelona, Spain, 3–7 September 2000, pp. 314–317. New York: Institute of Electrical and Electronics Engineers; available at http://ieeexplore.ieee.org/document/903548/