A Study of Calorie Estimation in Pictures of Food

Jun Zhou[1], Dane Bell[2], Sabrina Nusrat[3], Melanie Hingle[4], Mihai Surdeanu[3], Stephen Kobourov[3]

[1]Department of Computer Science, Columbia University, New York, NY, United States
[2]Department of Linguistics, University of Arizona, Tucson, AZ, United States
[3]Department of Computer Science, University of Arizona, Tucson, AZ, United States
[4]Department of Nutritional Sciences, University of Arizona, Tucson, AZ, United States

Abstract

Background: Software designed to accurately estimate food calories from still images could help users and health professionals more efficiently identify dietary patterns and food choices associated with health and health risks. However, calorie estimation from images is difficult, and no publicly available software can do so accurately while minimizing the burden associated with data collection and analysis.

Objective: The aim of this study is to determine the accuracy of crowdsourced annotations of calorie content in food images, and to identify and quantify sources of bias and noise as a function of respondent characteristics and food qualities (e.g., energy density).

Methods: We invited adult social media users to provide calorie estimates for 20 food images (for which ground truth calorie data were known) using a custom-built webpage that administers an online quiz. The images were selected to provide a range of food types and energy density. Participants optionally provided age range, gender, and their height and weight. Additionally, five nutrition experts provided annotations for the same data to form a basis of comparison. We examined estimate accuracy on the basis of expertise, demographic data, and food qualities using linear mixed effects models with participant and image index as random variables. We also analyzed the advantage of aggregating nonexpert estimates.

Results: 2028 respondents agreed to participate in the study (males: 770 [38%], mean body mass index: 27.5). Average accuracy was 5 out of 20 correct guesses, where "correct" was defined as a number within 20% of the ground truth. Even a small crowd of 10 individuals achieved an accuracy of 7, exceeding the average individual's and expert annotator's accuracy of 5. Women were more accurate than men (P<.001), and younger people were more accurate than older people (P<.001). The calorie content of energy-dense foods was overestimated (P=.024). Participants performed worse when images contained reference objects, such as credit cards, for scale (P=.014).

Conclusions: Our findings provide new information about how calories are estimated from food images, which can inform the design of related software and analyses.

Keywords: Calorie estimation, image annotation, crowdsourcing, obesity, public health

Introduction

Estimating calories in pictures of food is an important task, providing data to inform nutrition research and practice, and helping individuals achieve optimal, balanced dietary intakes. Yet this task turns out to be difficult for both experts and nonexperts. We are using this study as an opportunity to enhance our understanding of whether and how calorie estimation works “in the wild”, i.e., in real-world scenarios. There are many applications of this understanding, ranging from improving the methodological rigor (and reducing the associated burden) of dietary assessment, a pervasive and unanswered question in nutrition science, as well as influencing the design of interventions focused on dietary behavior change.

The fact that individuals do not estimate calories well [23,65-67] has motivated the design of software applications ("apps") to help individuals better estimate different aspects of dietary intake (e.g., calories, energy density, nutrient density, portions) using machine learning (ML) and by harnessing the “wisdom of the crowd”. The latter phenomenon was first documented in a 1907 Nature paper [62] and has been successfully used in many domains, ranging from gene network inference [63] to computational problems [64]. Apps in this space remain quite difficult to use, requiring burdensome manual logging of what one eats, or, when ML is used to classify pictures of foods, explicit weight values to be entered manually. To a large extent, the identification of calorie content from images of food, either through crowd sourcing or machine learning, remains an open research question.

This work is a necessary step towards the automated identification of calorie content from images of food.

The aim of this study is to determine the accuracy of crowdsourced annotations of calorie content in food images, and to identify and quantify sources of bias and noise as a function of respondent characteristics and food qualities (e.g., energy density).

Methods

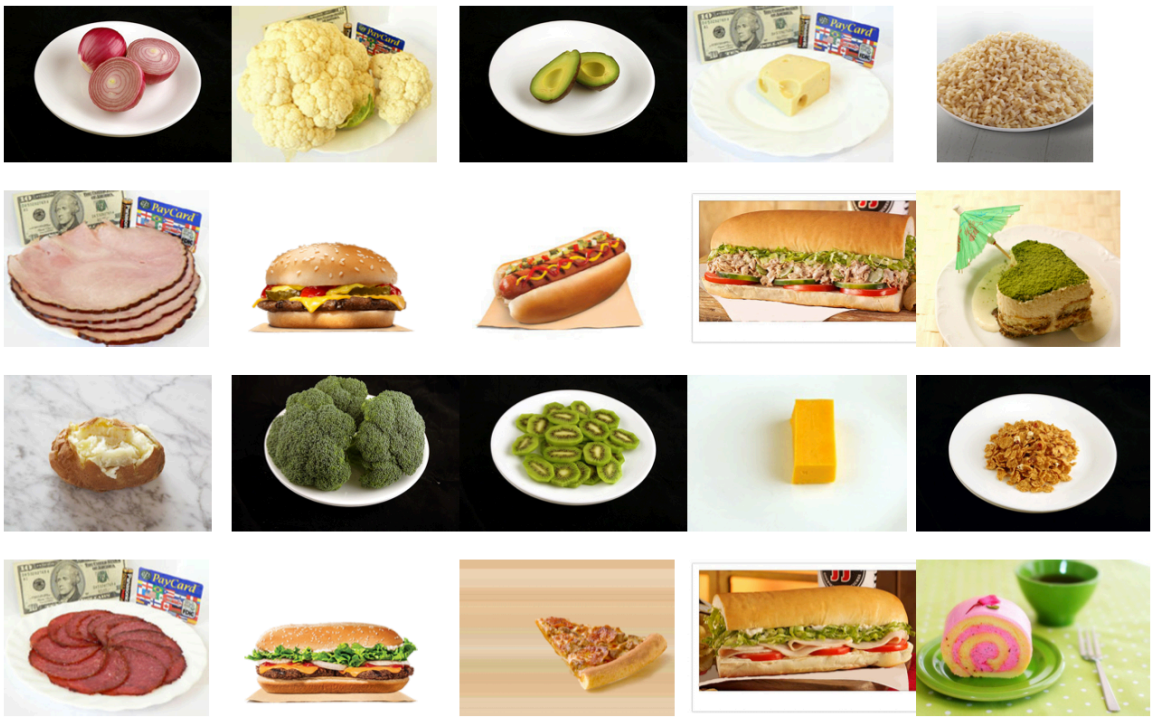
The proposed task is essentially a combination of two tests individuals must engage in when estimating calories. The first test relates to the relative energy density of the food pictured, whereas the second test discerns the portion size. Thus, we contend that the ecological validity of our approach is high, despite the task's complexity. The study protocol described herein was reviewed by an Institutional Review Board at the University of Arizona and met the criteria for exemption under 45 CFR 46.101(b).

We designed a simple online quiz administered by a custom-built webpage to measure the accuracy of calorie estimation in pictures of food, verify the existence of collective wisdom, and analyze data and find patterns and trends that can be useful in the design of calorie tracking apps.

We posted the quiz to SampleSize [59], a subreddit (i.e., a forum on reddit) dedicated to posting surveys and survey results. This choice was made on the basis of having a large, active user base that reflects the demographics likely to make large-scale food annotations for reasons of personal interest in self-quantification.

The quiz began with a short introduction "We would like to see whether you have a good understanding about calories. We will show you several pictures of food and your task, should you choose to accept it, is to guess how many calories are in the food. We will not share any identifying information about you. All of the data is anonymous."

Figure 1. Untrained participants estimated the food calories in these twenty images.



The quiz included 20 questions. Each question consisted of a picture of some food item (see Fig. 1) and the prompt, "How many calories are in the food pictured here? (Type a number in the box between 50 and 800)." Implausible dietary data, from (un)intentional under-reporting or over-reporting, are a pervasive problem in nutrition research and can introduce bias or lead to erroneous interpretations of diet-weight or diet-disease relationships. A common way of handling this issue is to exclude extreme values after the fact based on the distribution of the data (e.g., removing data more than 2 standard deviations from the mean) or by subjective assessment [35]. In contrast, we provided the upper and lower limits on the guesses, based on the ground truth data, in order to ease an already difficult task and thereby reduce the amount of data it would later be necessary to remove. The numbers also helped clarify that we were referring to kilocalories and helped reduce outliers.

Neither the correct calorie amounts nor other participants' answers were visible to a participant during the estimation portion of the experiment, although it is possible that some might have read the reddit comments prior to participation, which revealed some calorie values. We decided to not add additional information to the pictures (e.g., *does the sandwich contain mayonnaise?*) in order to keep the task closer to a realistic image annotation task.

Following the food-related questions, the participants were asked to provide their age group, gender, and BMI. An option to calculate BMI via height and weight information was also available. We deliberately chose not to ask for additional demographic questions (e.g., location, income, education) in order to protect participant privacy.

We reported the accuracy of the individual participant who just completed the quiz, as well as the average accuracy of all prior participants, using a breakdown showing the performance of each question.

We used two categories of food for the pictures: single-ingredient (e.g., broccoli, cheese) and mixed-ingredients (e.g., sandwich, pizza). There were 20 pictures of food items in total (Figure 1), 12 single-ingredient and 8 mixed. The food shown in the pictures ranged from 100 to 720 kcal. Importantly, we chose these food items according to the USDA's MyPlate model [9] that captures the building blocks for a healthy diet, and which includes five types of food (vegetables, fruits, protein, dairy, and grain), as well as mixed foods containing these ingredients. Our selection aimed to follow this model, to include realistic foods that appear in daily consumption, and to be concise so participants engage with the quiz.

The food portions selected are summarized in Table 1. The images were ordered so that each food type was maximally separated from other instances of its type, and the order was the same for each participant. We collected nutrition information about some food items from official restaurant websites. While the calorie content of the foods pictured was not directly measured, US federal statute requires the published calorie values of restaurant food items to be within 20% of actual calorie value [31].

Table 1. Foods were chosen for the quiz to attain maximum coverage of food types encountered in daily life by likely participants. *Scaling* refers to the presence of reference objects such as credit cards, which could indicate food volume.

Food	Type	Energy (kcal)	Mass (g)	Scaling?	Source
Cheddar cheese	dairy	200	51	no	[11]
Gouda cheese	dairy	300	84	yes	[1]

avocado	fruit	200	125	no	[11]
kiwi	fruit	200	328	no	[11]
brown rice	grain	420	297.7	no	[2]
cereal	grain	200	55	no	[11]
ham	meat	300	185.1	yes	[1]
salami	meat	300	72.9	yes	[1]
red onion	vegetable	200	475	no	[11]
potato	vegetable	100	141.7	no	[6]
broccoli	vegetable	200	588	no	[11]
cauliflower	vegetable	300	1200	yes	[1]
cheeseburger	mixed	270	104	no	[3]
hot dog	mixed	310	123	no	[3]
green tea cake	mixed	136	40	no	[8]
long cheeseburger	mixed	590	213	no	[3]
pepperoni and sausage pizza	mixed	240	97	no	[10]
Swiss roll	mixed	251	96	no	[5]
tuna sandwich	mixed	720	420	yes	[7]
turkey sandwich	mixed	510	254	yes	[7]

We chose not to inform the participants of the sources of the images, in order to reduce the potential that they would search the web for “ground truth” data, e.g., by going to the actual Burger King website. Likewise, the participants were not explicitly told that some images were from fast food restaurants, and thus more likely to be subject to food engineering, e.g., replacing sugar or using sweetness enhancers, or adding water or protein to enhance food properties and palatability [47, 50]. The fact that this was not explicitly mentioned to the participants raises the possibility that participants might have considered these foods as "homemade", which may have influenced perceived energy density and calories. However, since a majority of hamburgers are eaten at restaurants rather than homemade, judgments

about engineered foods are as or more relevant than home-cooked foods for both naturalistic and app-related purposes.

Patterns and Analysis

Three measures were relevant to our analysis:

1. Error, e, is estimated kilocalories ($\hat{c}$) minus ground truth kilocalories (c), $e = \hat{c} - c$, and percent error, η, is error as a percentage of the ground truth kilocalories, $\eta = \frac{\hat{c}-c}{c}$, both of which are positive in overestimation and negative in underestimation. Because of the variation in the ground truth kilocalories of the foods, the latter is a more reliable indicator of the scale of response bias.
2. Absolute error, $|e|$, measures accuracy irrespective of the direction of estimation bias ($|e| = |\hat{c} - c|$).
3. Discrete accuracy, D, is the number of estimates that were within 20% of the true calorie value (out of twenty estimates).

$$d_i = \begin{cases} 1, if\, 0.8c_i \leq \hat{c}_i \leq 1.2c_i \\ 0, otherwise \end{cases}$$

$$D = \sum_{i=1}^{20} d_i$$

Discrete accuracy was the measure reported to quiz participants.

Prior to this analysis, we removed participants who reported a BMI less than 15 or more than 50 kg/m^2 (which are unlikely to be correct), and participants who did not report their gender. Additionally, we eliminated responses of less than 50 kcal or greater than 800 kcal, and kept all remaining ones.

We analyzed the results of the survey using linear mixed-effects modeling in R [16, 43], allowing regression with random intercepts for both participants and foods simultaneously. The R^2 values are the proportion of the variance in the data that is described by the models' predicted values. For all analyses, a P value less than α=.05 was considered indicative of a statistically significant relation.

Results

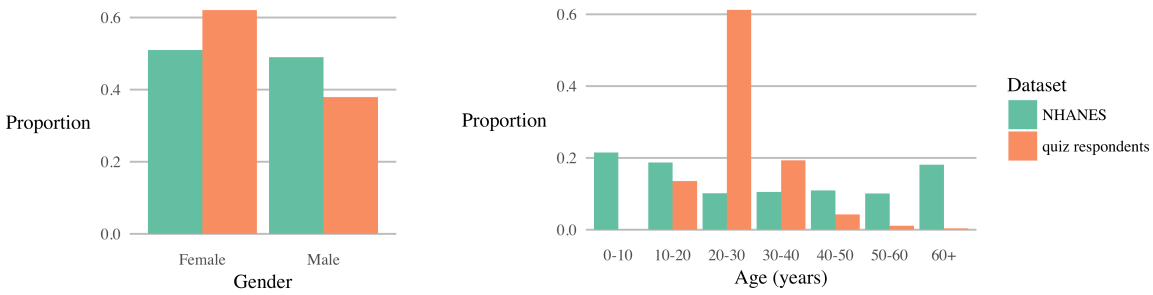
In total, 2,125 individuals participated in our reddit quiz. After removing 97 participants with missing or invalid demographic data, 2,028 individuals were included in the analysis.

Participant Demographics

The demographics of the participants are summarized in Figure 2. Although we collected no location data, an earlier study, again recruiting from the SampleSize

subreddit, found that 67% of participants reported a location within the United States [18, 51], a rate that is similar the 64% reported in another voluntary survey with participants from across reddit [51]. We also have a higher percentage of female participants than the US average, and a larger fraction of people with BMI around 25 kg/m2. It is possible that the participants in our quiz were more interested in this topic than the average person. However, in their self-selection, they are more demographically similar than the average person to likely crowdsourcing annotators for potential future app development.

Figure 2. Demographic data from our quiz is compared to data from NHANES



(National Health and Nutrition Examination Survey).

Participant Feedback

The participants volunteered their BMI and other demographic information, and 18 participants left 31 comments on the reddit thread. Table 2 summarizes the types of feedback comments we received, as well as some examples.

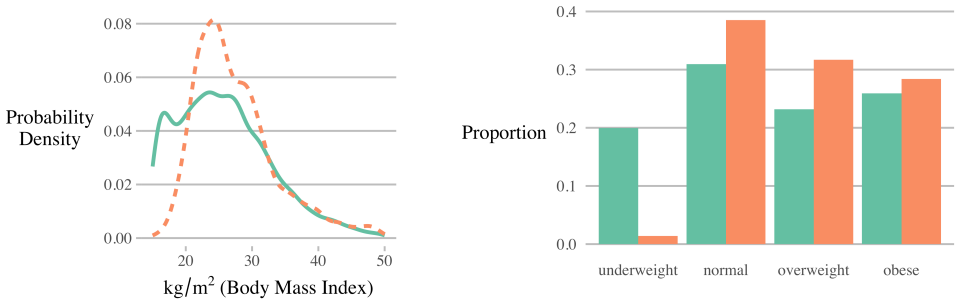


Table 2. Representative comments from the reddit post of the calorie estimation quiz.

Type	Example

fun	"That was fun! I think the folks in 'loseIt' [another subreddit] and on various MFP [MyFitnessPal] forums would enjoy taking this, too."
surprise	"I'm really really doubtful that burger is only 270 cal." "[N]o way are two red onions 200 calories."
units	"[C]ountries other than the US use the actual unit of energy- Joules"
scale	"It would have been great to have a ruler next to the food." "[I]f you show me a plate of rice, I can't guess how much rice are on the plate because I don't know how big the plate is."
difficulty	"Shoot, got 1 right out of 20 LOL. No wonder my BMI is 29." "I dont know if there was mayo on [the submarine sandwiches] or not, which changes things a lot."

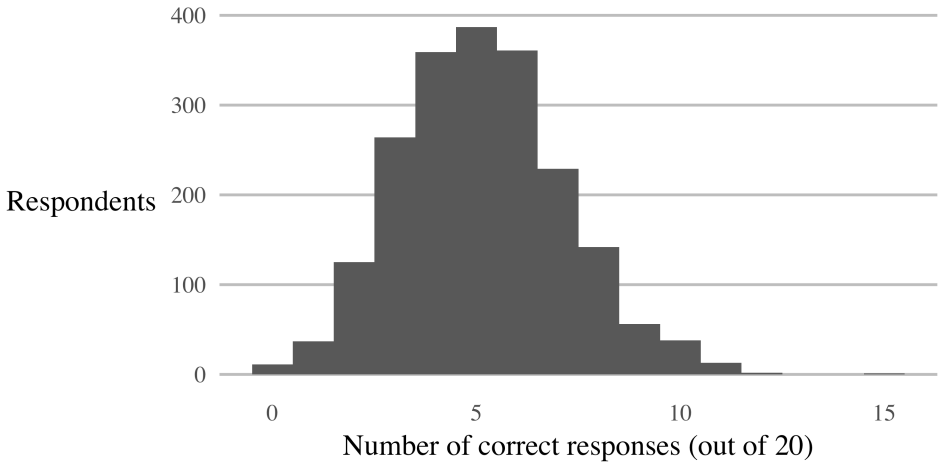
The feedback from the participants demonstrates engagement, interest, and curiosity. This implies that such tasks could be legitimately gamified (applying game mechanics and game design techniques to engage and motivate people to achieve their goals). It also shows that unlike Mechanical Turk participants, the participants in our study were engaged and motivated by intrinsic interest.

Note that our work addresses some of the requests shown in Table 2. For example, we found no increased accuracy from the presence of reference objects for scale in the pictures.

How good are people at estimating calories?

The participants' estimates had a mean absolute error ($|e|$) of 58% (136 kcal). In terms of discrete accuracy (D), the mean participant answered 5.15 questions correctly out of 20. Figure 3 shows the distribution of correct responses. Absolute error varied considerably by item from the most accurate item—a turkey sandwich, with a mean absolute error of 23.0% (39 kcal)—to the least—green tea cake, at 241% (327 kcal) absolute error. Figure 4 illustrates the variety of estimates and percent error (η) distributions for different items. Together, these facts show that human calorie estimates are both inaccurate overall and inconsistent in their inaccuracies.

Figure 3. A histogram of the number of correct estimates each participant made. See



Patterns and Analysis for the definition of this measure, *D*.

Does the wisdom of the crowd phenomenon apply here?

A consensus formed rapidly for each food, as shown in Figure 5 (see the dashed orange line in the figure), so that ten responses gave a very good estimate of the next 1,000 responses. In fact, a bootstrap significance test shows that the average of 10 randomly selected participants' guesses is no more (or less) accurate than the average of those of 1,000 random participants (P=.358). Moreover, the consensus responses had greater discrete accuracy (D) than that of the individual participants, achieving 7 correct responses out of 20, a 36% relative improvement over the 5.15 correct among individual participants. This result is consistent with previous studies demonstrating the wisdom of the crowd, in which the accuracy of consensus judgments exceeds that of individual judgments (see Comparison with Prior Work).

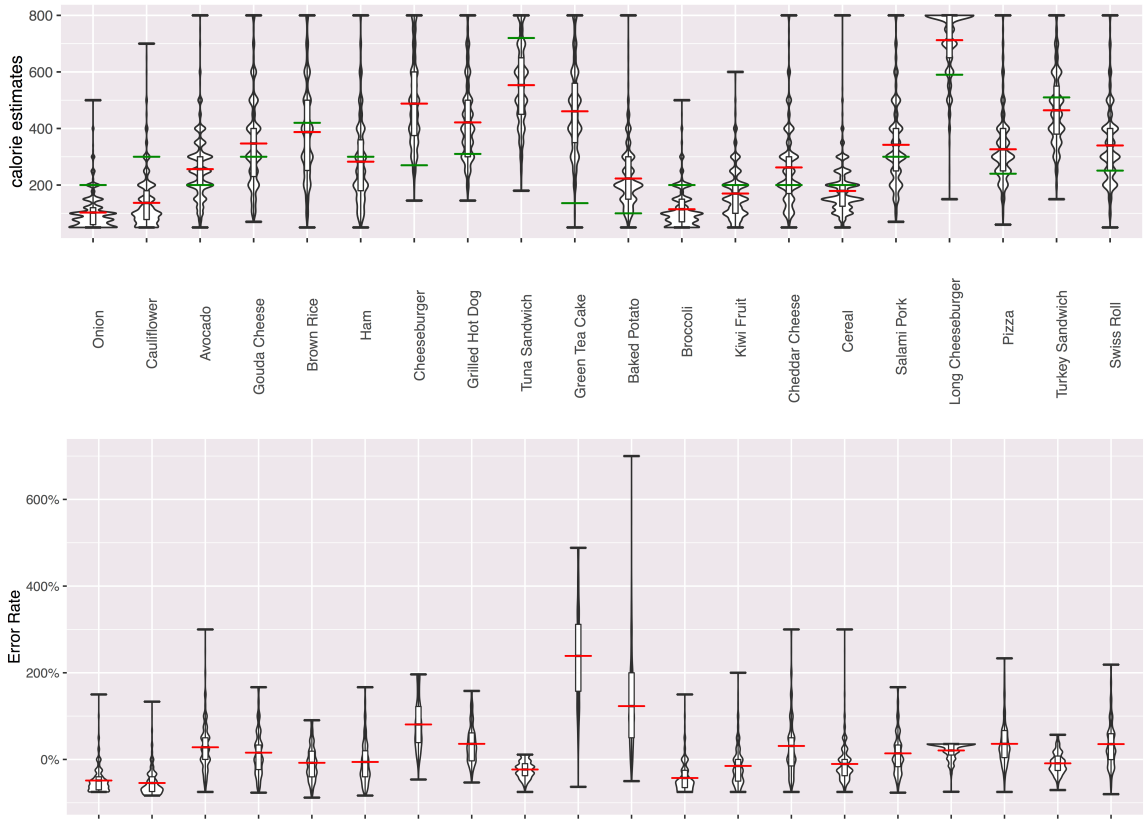
Another important observation is that, although error was high for individual responses and individual foods, the bias in the errors was low overall across all questions, such that the median of the error across items and participants is 0 (when using crowdsourcing over 2,028 participants). While this result is not actionable in itself, since it is averaged across all questions, it does demonstrate the power of crowds to converge toward high-accuracy judgments.

Do the nutritional experts outperform the crowd?

In addition to redditors, we solicited participation from five nutritional experts. We recruited faculty on a voluntary basis from the Department of Nutritional Science at the University of Arizona and the School of Nutrition and Health Promotion at Arizona State University. Somewhat surprisingly, neither the absolute error of their responses nor their discrete accuracy was statistically different from those of the average nonexpert participant (P=.195). In fact, a small crowd of only 2 randomly

selected nonexperts was required to outperform the highest performing expert, achieving an average absolute error ($|e|$) of 119.3 (52.34%) compared with the expert's 130.2 (55.34%). Expert performance is shown in comparison with nonexpert performance in Figure 5. This result is consistent with the hypothesis that the sources of error (e.g., erroneous volume estimation due to a notion of typical portion size) apply equally to experts and nonexperts. Prior work in many domains of estimation has supported the notion that a relatively small group of nonexperts can estimate just as well as a single expert [54, 57] (see also Comparison with Prior Work).

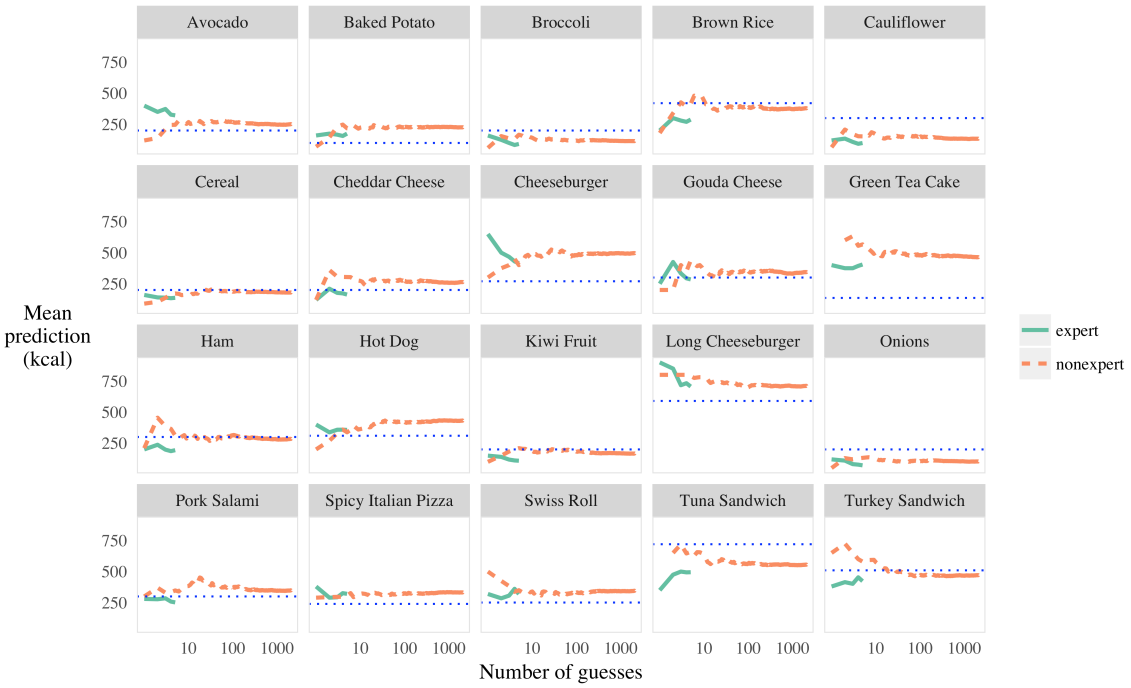
Figure 4. (Top) Calorie estimates for 20 food items. For each food item, the violin plot represents the distribution of the calorie estimates by the participants. The bottom and top of the boxes and the red band represent the first and the third quartile, and the mean of the calorie estimates, respectively. The green band represents the actual calorie value for each food item. The actual calories are shown in parentheses next to each food item. (Bottom) Percent error (η) for each food item. The bottom and top of the boxes and the red band represent the first and the third quartile, and the mean of the error rates, respectively.



Does having an object for scale in the picture help?

Several comments in the reddit thread expressed the hypothesis that pictures featuring standard-sized reference object (such as a credit card) were easier to answer. The results showed that reference objects, far from aiding estimation, increased absolute error ($|e|$) by a mean 4.6 kcal (P=.014, R^2=.31). Our hypothesis is that participants used background knowledge about the typical size of foods to scale foods, but were not able to profit from comparison against the reference objects. This is statistically significant evidence for the notion that scale information does not aid calorie estimation in digital images (compare [68]). However, it is important to note that this was a post-hoc analysis only; the experiment was not designed to analyze this hypothesis. For example, we included objects that come in many different sizes, e.g., forks, as reference objects, which may have confused the quiz takers. We leave a more careful evaluation of this particular observation as future work.

Figure 5. Mean estimates for each food as more participants are added show that a consensus forms rapidly. The dotted blue lines show the true calorie value for each food. The *x*-axis uses a logarithmic scale. The orange dashed line indicates the estimates of non-experts. The green continuous line represents the estimates of nutrition science experts. Note that the range of acceptable calorie estimates was 50–800 calories, for each food item.

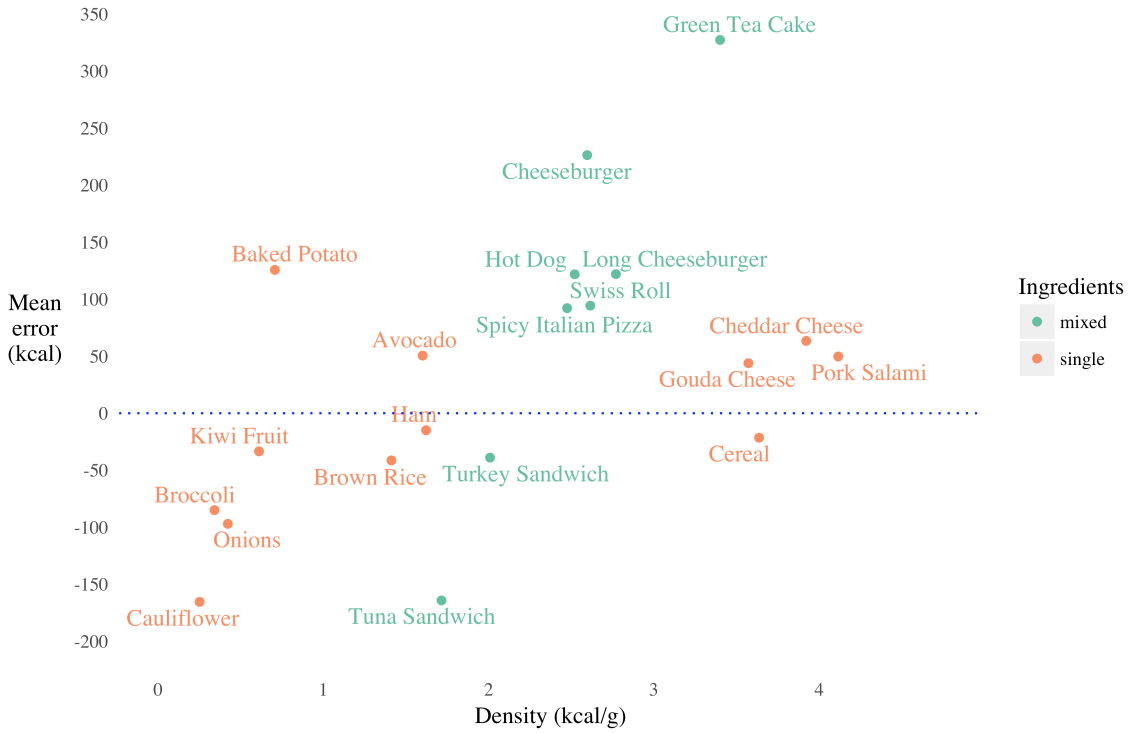


Does energy density of foods predict estimation error?

As shown in Figure 6, the caloric content of energy-dense foods was systematically overestimated, and that of energy-sparse foods underestimated, as measured by

error (*e*, P=.024, R^2=.57). This bias is similar to one found by Almiron-Roig et al. [13] in estimating in-person portion sizes, and could reflect two non-exclusive sources. First, it could result from the perceived healthiness of the food items [30]. For example, broccoli is a prototypically healthy food, but is not devoid of calories; conversely, prototypically unhealthy foods such as cheeseburgers have often been "engineered" for low calories [50]. This explanation aligns with the results of Carels et al. [23], who found that college students overestimated the caloric content of foods considered to be unhealthy while they underestimated the amount of calories in healthy foods. Second, the bias could result from an assumption that the items would have a similar weight to one another, when in fact there was an inverse relationship between the energy density and weight of the items (Pearson correlation: ρ=–0.70). We hypothesize that inelastic adjustment of portion size according to energy density could contribute to obesity.

Figure 6. Participants underestimated the calorie content of calorie-sparse foods and overestimated that of calorie-rich foods.



Does BMI predict estimation errors?

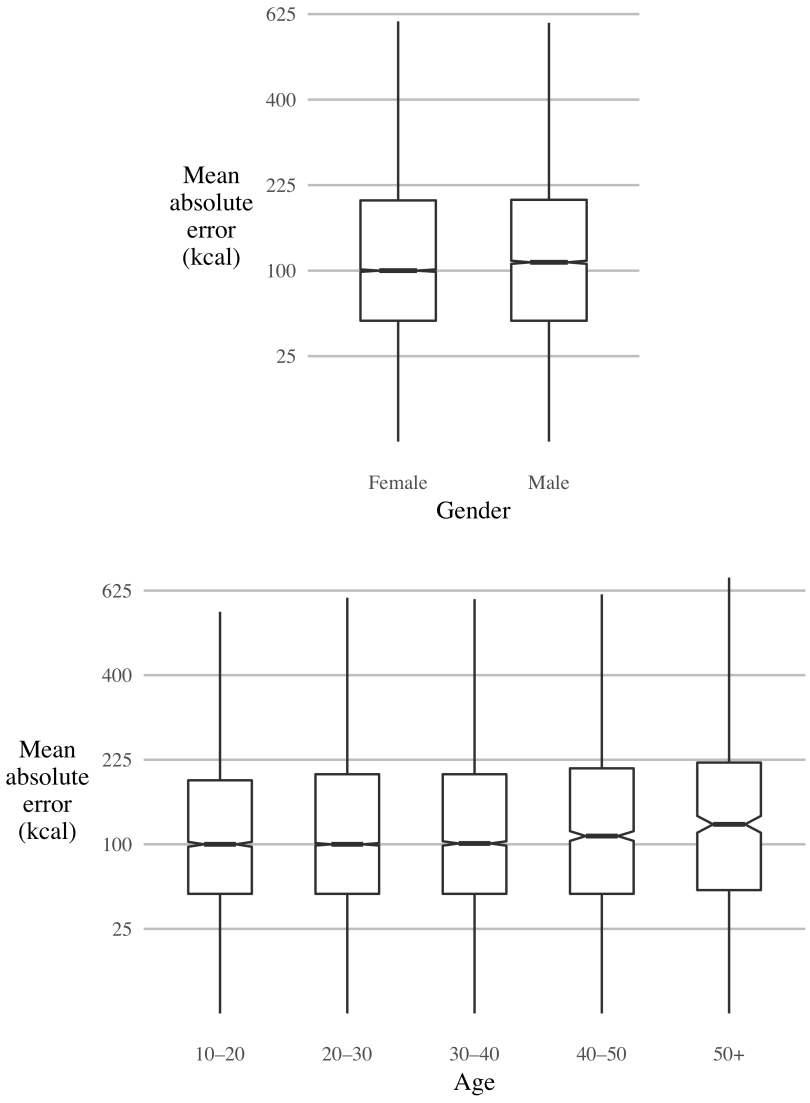
BMI itself does not predict accuracy or bias in this data, similar to Blake et al. [21] and Chandon and Wansink [25]. Other studies show that overweight and obese individuals consistently under-report calorie intake to a greater degree than non-overweight individuals [14, 42]. However, BMI does significantly interact with energy density in predicting percent error (η, P=.002, R^2=.57), such that the higher a participant's BMI, the more they exaggerated the calorie content of calorie-rich

foods. We hypothesize that overweight individuals are more sensitive to perceptions of food.

Do gender and age predict estimation errors?

No biasing effect (toward underestimation, for example) was found, but absolute error (|*e*|) was greater for men than for women ($P<.001$, $R^2=.31$), similar to the portion judgment result by Almiron-Roig et al. [13]. Additionally, the absolute error was greater for older participants ($P<.001$, $R^2=.31$), but these effects did not interact. Figure 7 summarizes these differences. We hypothesize that the primary reason for these differences is cultural, reflecting gender norms and the relatively recent cultural emphasis on calories as a measure of healthiness.

Figure 7. The absolute error (|*e*|) of participants differs by gender (top) and age (bottom). Box edges show the first and third quartiles and are split by the median. The boxes' whiskers extend to the farthest point within 1.5 times the interquartile range from the box ends. The notches denote the 95% confidence interval of the

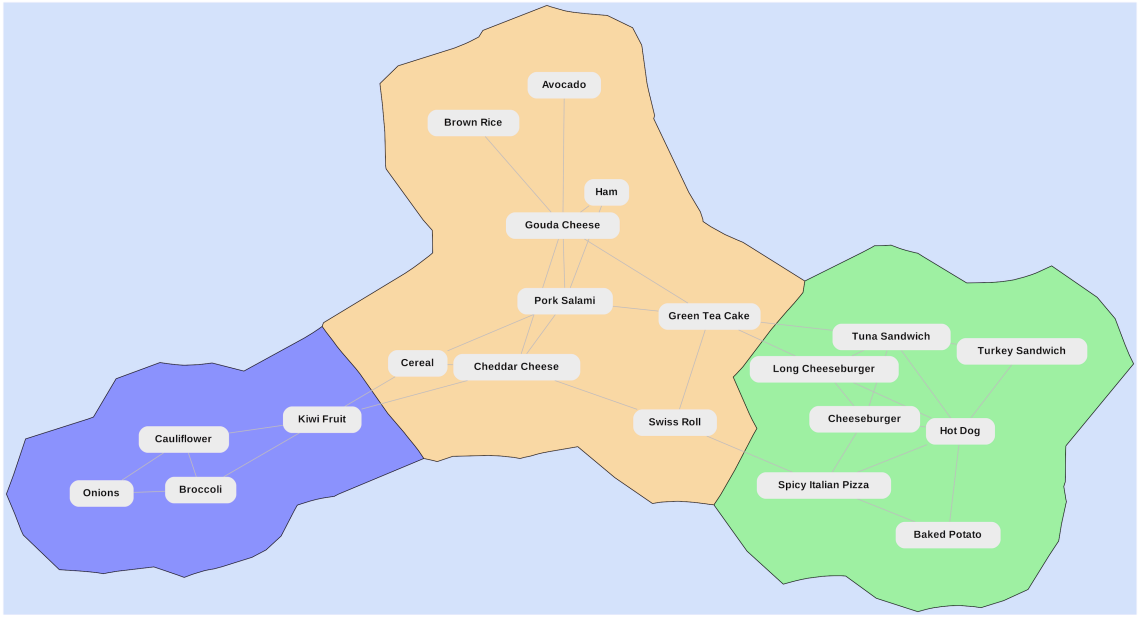


median. The *y*-axis is on a square-root scale.

Do estimation errors cluster by food type?

The over- and underestimation errors for some foods correlate with those for others. For example, a participant who underestimates the calories in broccoli is likely to do so for cauliflower as well. Figure 8 shows an automatically generated map [41] illustrating these correlations, with clusters showing similar subnetworks. A larger map with more food items would be a strong basis for predicting human bias on clusters of food types (e.g., vegetables).

Figure 8. This map shows a network of food items in our survey based on correlation of estimation errors. Pairs of strongly correlated foods are connected by edges. The stronger the correlation, the closer the two (distances are inverse to correlation) are. Clusters show groups of similar subnetworks.



Discussion

Principal Results

The above analysis identifies several patterns that are important for the design of calorie estimation apps.

First and foremost, our study demonstrates that individuals are poor judges of calorie content in images, and prior work has shown that they are poor judges of portion size in real-life situations (see next subsection). This suggests the utility of a machine-learning approach to calorie estimation to facilitate meal planning. Keeping

track of calories by describing foods and guessing quantities and values is a tedious and inaccurate strategy, yet it is the one most commonly used in apps today. Given that "a picture is worth a thousand words," our initial hypothesis was that using images (rather than descriptions of foods) should lead to better estimates. Our results, however, do not support this hypothesis: on average, participants performed poorly at estimating the amount of calories in pictures of food, answering 5 of 20 questions correctly on average. Our analysis indicates that participants in our dataset tended to exaggerate common dietary knowledge: they underestimated the amount of calories in energy-sparse foods, and overestimated them in energy-dense ones.

Our related work discussion (Comparison with Prior Work, below) highlights that estimating calories using machine learning remains an open research problem. However, our work suggests that such apps could take advantage of the wisdom of the crowd for estimation. We showed that the crowd performs better than experts, on average, even when the crowd is small. This suggests that this annotation could be implemented accurately and at low cost.

The results suggest that for apps that focus on calorie monitoring (including self-reporting), it might be a good idea to characterize users' demographic data (age, gender, and BMI) shown to influence the accuracy of calories estimates, either directly or when combined with other factors such as energy density.

We identified additional patterns that simplify the design and implementation of calorie-tracking apps. The first such pattern is that scale information does not improve estimation accuracy. The second is that estimation errors cluster by food types, which indicates that the app may extrapolate user patterns between foods in the same group.

It is important to note that the observations of this study are statistically significant and applicable to the population of interest to us (i.e., individuals likely to participate in crowdsourced annotations). This population is considerably younger than the US population ($\chi 2=3363$, $P<.001$), and contains more women proportionally ($\chi 2=80.98$, $P<.001$). In future work, we aim to repeat this study for a larger population that matches known demographics to verify the validity of our analysis on such populations.

Comparison with Prior Work

Related work includes prior work in nutritional sciences, machine learning, image processing, and crowdsourcing. We review a small but representative subset below.

Nutrition and diet

Bandini et al. [15] and Schoeller et al. [53] have reported that individuals tend to selectively under-report the energy intake when these data are manually logged. This seems to be especially true for overweight and obese individuals [14, 42], and could be associated with a failure to accurately estimate portions, although Blake et al. [21] and Chandon and Wansink [25] found that BMI does not correlate with the

ability to estimate calories when this task is conducted in person. Portion estimation of in-person food remains poor, whether in reference to images on computer screens or on printed images [34]. However, calorie estimation of large meals may be worse than that of small meals [56].

To monitor dietary intake more accurately, third-party automated food analysis systems have been proposed. Martin et al. [46] use the remote food photography method (RFPM), which requires individuals to upload three pictures when having a meal: the plate of the foods selected by an individual, standard portions of known quantities of the foods, and the leftovers. These pictures are sent to trained dietitians who verify portions with participants, and analyze this data using a standardized nutrient database. This approach relies on the judgment of trained nutrition professionals, and argues for the validity of RFPM. Providing all three pictures for each meal is a challenge, as indicated by Williamson et al. [58]. Beltran et al. [19] tested the reliability of the eButton system, in which a camera worn on the chest records images continuously. The images are captured passively while the participant goes about their day, but such a system still requires experts to identify foods in the images, and confirm them with participants. Similar to the RFPM employed by Martin et al. [46], the eButton system requires valid pictures before and after each meal, camera placement at a certain angle, and proper lighting. While promising, such systems are unlikely to scale to the millions of people who would like to accurately track their nutritional intake.

Machine learning and image processing

Given the challenges of the systems described above, a system that can automatically measure calories in pictures of food would be in great demand. Image processing techniques can be used to recognize food in images, and machine learning can be used to estimate the calories in the food.

Menu-Match [17] uses a database of restaurants and GPS locations, and attempts to guess what is in the picture, using image features such as color and scale-invariant feature transforms [44]. It has not been made available to the general public. Im2Calories [48] is built on the work of menu-match. A multi-label classifier is trained on a collection of images of food. The app locates the restaurant a user is dining in and, given an image from the user, the classifier (running on the user's phone) guesses which foods are present in the meal. Looking up the nutritional facts provided by the restaurant, using the resulting estimates, yields good results. Note, however, that Im2Calories has not been made available to the general public or even for research purposes.

Bettadapura et al. [20] show that food recognition using location data improves accuracy. Such systems, however, are inherently limited to the restaurants whose menus are in the database. These also assume that menus do not change often, and that the volume of food is the same from plate to plate. In reality, most meals are eaten either outside of restaurants or in restaurants whose menus are not included in some dataset. The “in the wild” problem is more natural, but also more difficult.

The web application and app Foodlog [39, 40] divides food images into 300 blocks each and extracts Discrete Cosine Transform (DCT) coefficients and color histogram from each block. Using this data, Foodlog classifies the food into five categories according to the USDA's My Pyramid system. Experimental results report 88% accuracy in the extraction of food and 73% accuracy in food balance estimation. The FoodCam system [38] segments the region of each food by GrabCut (an image segmentation approach based on iterative graph-cuts) [52], extracts image features of histogram of oriented gradients (HOG) [27] and color patches with the Fisher Vector (an image representation obtained by pooling local image features) [26] and finally classifies it into one of 100 food categories using linear support vector machines (SVM).

With the exception of Im2Calories, the systems above achieve relatively good food recognition, but without volume estimation. To estimate volume, Chae et al. [24] minimize the false-segmented regions, smooth the segmentation boundaries of food, and reconstruct 3D primitive shapes from a single food image. He et al. [33] estimate the weight of food given a single image using a shape template for regular-shape foods and area-based weight estimation for irregularly shaped food. The Im2Calories system [48] estimates the distance of every pixel from the camera by using a convolutional neural net (CNN) architecture, converts the depth map [4] into a voxel representation, and estimates the volume of the food. Although such approaches are effective, there is no application for estimation of food volumes that is available to the general public.

Crowdsourcing

Crowdsourcing sometimes makes it possible to use multiple nonexpert judgments to approach the high quality of expert annotation [22]. Surowiecki [54] argues that in many instances, the average nonexpert estimates can even outperform a single expert. Watson has shown that the average of the individual judgments can be equal or superior to the judgment of the best individual within the group [57]. Moreover, the validity of judgments increases with more judges [32]. The strength of the wisdom of the crowd over machine learning is well understood and exploited in industry. For example, CardMunch (now a service of Evernote [60]) uses crowdsourcing with Amazon Mechanical Turk to convert pictures of business cards into digital contact information. Eloquent Labs [61] uses a mix of crowdsourcing with an artificial intelligence to implement a conversational assistant for customer service.

In the nutrition domain, Mamykina et al. [45] show that crowdsourced ingredient annotations from food images are improved by expert annotation and by showing the annotators previous annotations of the images. The PlateMate [49] app leverages crowdsourcing to implement the first step in the Remote Food Photography Method. Rather than typing names of foods and estimating portions, users take photographs of their plates both at the beginning of the meal and at the end to accurately capture how much food was actually eaten. PlateMate uses annotations from nonexpert Amazon Mechanical Turk workers instead of expert dietitians to estimate the composition of foods in static images. PlateMate's results

are as accurate as the experts. Similarly, the Im2Calories [48] project uses crowdsourcing to annotate all the food terms that apply to an image. Manually merging synonymous terms, they create the Food201 multi-label dataset for training. Compared to the original Food101 classes, the new classes of Food201-MultiLabel do better according to mean average precision, since they often correspond to side dishes or small food items.

In sum, despite the abundance of interest in this and related topics, including calorie-tracking apps with manual entry, there exists no publicly available app that will accurately estimate calories from a single image. Likewise, while there are many studies of human bias in tracking calories and lack of skill in estimating portion sizes, no previous work establishes the accuracy and biases of crowdsourcing for calorie estimation, or what demographic factors might correlate with accuracy.

Learning from our study, we envision a very simple app, where the only action required from the user is to take a picture of her or his food. The estimation logic, driven by the wisdom of the crowd and machine learning, would be transparent to the user, i.e., it would be triggered automatically when the camera is used. The logic includes: (a) detecting if the picture is a picture of food using image classification [36, 37], and (b) routing the image for crowd annotations (similar to CardMunch, which routes the task of processing images of business cards to the crowd). We hope that this simplicity will yield wide adoption, which in turn will lead to measurable effects in dietary choices.

Conclusions

We described a study measuring the ability of over two thousands individuals to estimate calories in 20 pictures of food chosen to capture the building blocks of a healthy diet [9]. We believe this study should be read as an analysis that drives the design of future food-related apps, with additional impacts on crowdsourcing strategies and the design of human-computer interfaces.

Our analysis confirms some earlier observations (e.g., calorie estimation is a difficult task, even for the experts), and offers new insights:

1. Even a small crowd of two nonexperts achieves calorie estimation accuracy greater than that of the expert annotators. This suggests that semi-automated food labeling apps can be implemented at a low cost by harnessing the wisdom of the crowd, even when the crowd is small. Note that some prior approaches in this space, such as PlateMate [49], use crowdsourcing to provide calorie information to users. To the best of our knowledge, the crowdsourcing method had never been tested as a source of data for algorithmic calorie estimation before.
2. We found new type-of-food effects, with energy-dense foods (such as hamburgers) being consistently over-estimated and energy-sparse foods (such as broccoli) consistently underestimated. Future crowdsourcing (or machine learning) projects aiming to annotate food for calorie content will benefit from correction using these biases.

3. We found the absence of some expected correlations. For example, the presence of reference objects for scale does not improve accuracy but rather slightly decreases accuracy, and the Body Mass Index is not correlated with accuracy. These observations impact the design of interfaces for annotation apps, as well as data collection protocols.

All in all, this work suggests that calorie-estimation apps are needed and can be built at low cost (e.g., using small annotator groups, and without the overhead of including reference objects in images, or controlling for the BMI of users).

Several interesting research questions remain. First, given the low calorie estimation accuracy (5 out of 20), and some clear patterns (underestimating "healthy" foods and overestimating "unhealthy" foods) it is natural to ask whether simple training with feedback can help improve accuracy for nonexperts. If so, how much training is required, what gains in accuracy can be obtained, and how much further can the crowd boost the results? Second, can we factor in biases (e.g., age, gender) to obtain better crowdsourced prediction? Third, can better (more consistent) reference objects lead to improvements in accuracy? Fourth, assuming the baseline accuracy for "simple" foods (e.g., fruits, vegetables, sandwiches) can be improved with some of the ideas above, can we hope to tackle more difficult challenges, such as amorphous foods (porridge, mashed potatoes) and liquids (soups, smoothies) in which ingredients and volume are less obvious? Lastly, but perhaps most importantly, we aim to apply the knowledge gained from this study beyond the understanding of how (or how well) people estimate calories, to include assessment of diet quality, which has become a dietary construct of interest in the past five years [55]. This change has occurred because dietary patterns and dietary quality (e.g., increased nutrient density, nutrient diversity, and nutrient adequacy) have been strongly associated with health and disease outcomes. This information provides potentially more meaningful metrics than amount of calories (which says nothing about the quality or "healthiness" of the food) when providing participants or patients with feedback.

We believe this study should be read as an analysis that informs the design of future food-related apps (in particular, apps that feature calorie estimation), with additional potential impacts on crowdsourcing strategies and the design of human-computer interfaces. Our future goal is to provide estimates about judging calories from images for the purpose of mass annotation (e.g., in support of a calorie-estimation app), which, in turn, is part of a larger system that analyzes text, images, and videos to estimate risk of diet-sensitive diseases such as type 2 diabetes mellitus [18, 69].

Conflicts of Interest

None declared.

Abbreviations

BMI: body mass index (kg/m^2)

ML: machine learning

USDA: United States Department of Agriculture

References

1. World Health Organization. 2017. Obesity and overweight. http://www.who.int/mediacentre/factsheets/fs311/en/. Archived at: http://www.webcitation.org/6tg2Yf7s5
2. HealthAssist. 2017. 300CalorieFoodPictureGallery. http://www.healthassist.net/food/300kcal/300.shtml. Archived at: http://www.webcitation.org/6tdAdUSkj
3. Panda Express. 2017. Brown Steamed Rice—Panda Express Chinese Restaurant. https://www.pandaexpress.com/menu/sides/brown-steamed-rice. Archived at: http://www.webcitation.org/6tbsukhFK
4. Burger King. 2017. Burger King. https://www.bk.com/. Archived at: http://www.webcitation.org/6tbqVFj2e
5. Slism. 2017. Food nutrients can be seen calorie calculation at a glance. http://calorie.slism.jp/200528/. Archived at: http://www.webcitation.org/6tbt9Z9Ym
6. Food Network. 2017. Foods with 100 Calories. https://www.foodnetwork.com/healthy/photos/foods-with-100-calories.html. Archived at: http://www.webcitation.org/6tbqxkJsp
7. Jimmy John's. 2017. Jimmy John's Gourmet Sandwiches. https://www.jimmyjohns.com/. Archived at: http://www.webcitation.org/6tbr3E3nW
8. Wit Co., Ltd. 2017. Matcha calorie calculation of cake. http://www.asken.jp/calculate/meal/94371. Archived at: http://www.webcitation.org/6tbs116BG
9. United States Department of Agriculture. 2017. MyPlate Model. https://www.choosemyplate.gov/MyPlate. Archived at: http://www.webcitation.org/6tbsR3lvJ
10. Papa John's Pizza. 2017. Papa John's Pizza. https://www.papajohns.com/. Archived at http://www.webcitation.org/6tbsSuImT
11. wiseGEEK. 2017. What does 200 Calories Look Like? http://www.wisegeek.com/what-does-200-calories-look-like.htm. Archived at: http://www.webcitation.org/6tbsWxbS1
12. Holleman F, Gale E. Diagnosis and Classification of diabetes mellitus (revision number 21). Diapedia. 2014. doi:10.14496/dia.11040851106.21.
13. Almiron-Roig E, Solis-Trapala I, Dodd J, Jebb SA. Estimating food portions. Influence of unit number, meal type and energy density. Appetite. 2013;71:95-103. doi:10.1016/j.appet.2013.07.012.
14. Bailey RL, Mitchell DC, Miller C, Smiciklas-Wright H. Assessing the effect of underreporting energy intake on dietary patterns and weight status. J Am Diet Assoc. 2007;107(1): 64-71. doi: 10.1016/j.jada.2006.10.009
15. Bandini, LG, Schoeller, DA, Cyr, HN, Dietz, WH. Validity of reported energy intake in obese and nonobese adolescents. Am J Clin Nutr. 1990;52(3):421-425.
16. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4. J Stat Softw. 2015;67(1). doi:10.18637/jss.v067.i01.

17. Beijbom, Oscar, et al. “Menu-Match: Restaurant-Specific Food Logging from Images.” 2015 IEEE Winter Conference on Applications of Computer Vision, 2015, doi:10.1109/wacv.2015.117.
18. Bell D, Fried D, Huangfu L, Surdeanu M, Kobourov S. Towards using social media to identify individuals at risk for preventable chronic illness. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016); 2016 May 23-28; Portorož, Slovenia; 2016.
19. Beltran A, Dadabhoy H, Chen TA, Lin C, Jia W, Baranowski J, Yan G, Sun M, Baranowski T. Adapting the eButton to the abilities of children for diet assessment. Proceedings of Measuring Behavior 2016; 2016 May 25-27; Dublin, Ireland; 2016.
20. Bettadapura V, Thomaz E, Parnami A, Abowd GD, Essa I. Leveraging Context to Support Automated Food Recognition in Restaurants. 2015 IEEE Winter Conference on Applications of Computer Vision. 2015. doi:10.1109/wacv.2015.83.
21. Blake, AJ, Guthrie, HA, Smiciklas-Wright, H. Accuracy of food portion estimation by overweight and normal-weight subjects. J Am Diet Assoc. 1989;89(7):962-964.
22. Brabham, DC. Crowdsourcing as a model for problem solving: An introduction and cases. Convergence. 2008;14(1):75-90.
23. Carels RA, Konrad K, Harper J. Individual differences in food perceptions and calorie estimation: An examination of dieting status, weight, and gender. Appetite. 2007;49(2):450-458. doi:10.1016/j.appet.2007.02.009.
24. Chae J, Woo I, Kim S, Maciejewski R, Zhu F, Delp EJ, Boushey CJ, Ebert DS. Volume estimation using food specific shape templates in mobile image-based dietary assessment. Proc SPIE Int Soc Opt Eng. 2011 Feb 7;7873:78730K. doi: 10.1117/12.876669.
25. Chandon P, Wansink B. Is Obesity Caused by Calorie Underestimation? A Psychophysical Model of Meal Size Estimation. J Mark Res. 2007;44(1):84-99. doi:10.1509/jmkr.44.1.84.
26. Csurka G, Perronnin F. Fisher vectors: Beyond bag-of-visual-words image representations. Communications in Computer and Information Science Computer Vision, Imaging and Computer Graphics Theory and Applications. 2011;28-42. doi:10.1007/978-3-642-25382-9_2.
27. Dalal N, Triggs B. Histograms of oriented gradients for human detection. Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05); 2005 Jun 20-26; 1:886-893.
28. Duffey KJ, Popkin BM. Energy density, portion size, and eating occasions: Contributions to increased energy intake in the United States, 1977–2006. PLoS Med. 2011;8(6): e1001050. doi:10.1371/journal.pmed.1001050.
29. Ello-Martin, JA, Ledikwe, JH, Rolls, BJ. The influence of food portion size and energy density on energy intake: Implications for weight management. Am J Clin Nutr 2005;82(1):236S-241S.
30. Faulkner GP, Pourshahidi LK, Wallace JMW, Kerr MA, Mccaffrey TA, Livingstone MBE. Perceived ‘healthiness’ of foods can influence consumers’ estimations of energy density and appropriate portion size. Int J Obes 2013;38(1):106-112. doi:10.1038/ijo.2013.69.
31. Food and Drug Administration, HHS. Food labeling; nutrition labeling of standard menu items in restaurants and similar retail food establishments. Final rule. Fed Reg 2014;79(230):71155.
32. Gordon K. Group judgments in the field of lifted weights. J Exp Psychol. 1924;7(5):398-400. doi:10.1037/h0074666.

33. He Y, Xu C, Khanna N, Boushey CJ, Delp EJ. Food image analysis: Segmentation, identification and weight estimation. 2013 IEEE International Conference on Multimedia and Expo (ICME). 2013. doi:10.1109/icme.2013.6607548.
34. Hernández T, Wilder L, Kuehn D, Rubotzky K, Moser-Veillon P, Godwin S, Thompson C, Wang C. Portion size estimation and expectation of accuracy. J Food Comp Anal 2006;19:S14-S21.
35. Huang TT-K, Roberts SB, Howarth NC, Mccrory MA. Effect of screening out implausible energy intake reports on relationships between diet and BMI. Obes Res 2005;13(7):1205-1217. doi:10.1038/oby.2005.143.
36. Kagaya H, Aizawa K. Highly accurate food/non-food image classification based on a deep convolutional neural network. International Conference on Image Analysis and Processing; 2015 Sep 7; Genova, Italy. 2015:350-357. doi:10.1007/978-3-319-23222-5_43.
37. Kagaya H, Aizawa K, Ogawa M. Food Detection and Recognition Using Convolutional Neural Network. Proceedings of the ACM International Conference on Multimedia - MM 14. 2014. doi:10.1145/2647868.2654970.
38. Kawano Y, Yanai K. FoodCam: A real-time food recognition system on a smartphone. Multimed Tools Appl 2014;74(14):5263-5287. doi:10.1007/s11042-014-2000-8
39. Kitamura K, Yamasaki T, Aizawa K. Food log by analyzing food images. Proceeding of the 16th ACM international conference on Multimedia - MM 08. 2008. doi:10.1145/1459359.1459548.
40. Kitamura K, Yamasaki T, Aizawa K. FoodLog. Proceedings of the ACM multimedia 2009 workshop on Multimedia for cooking and eating activities - CEA 09. 2009. doi:10.1145/1630995.1631001.
41. Kobourov SG, Pupyrev S, Simonetto P. Visualizing graphs as maps with contiguous regions. Eurographics Conference on Visualization (EuroVis). 2014.
42. Kretsch MJ, Fong AK, Green MW. Behavioral and Body Size Correlates of Energy Intake Underreporting by Obese and Normal-weight Women. J Amer Diet Assoc 1999;99(3):300-306. doi:10.1016/s0002-8223(99)00078-4.
43. Kuznetsova A, Brockho PB, Christensen RHB. lmerTest: Tests in linear mix effects models. 2016.
44. Lowe D. Object recognition from local scale-invariant features. Proceedings of the Seventh IEEE International Conference on Computer Vision. 1999. doi:10.1109/iccv.1999.790410.
45. Mamykina L, Smyth TN, Dimond JP, Gajos KZ. Learning from the crowd: Observational learning in crowdsourcing communities. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI 16. 2016. doi:10.1145/2858036.2858560.
46. Martin CK, Han H, Coulon SM, Allen HR, Champagne CM, Anton SD. A novel method to remotely measure food intake of free-living individuals in real time: the remote food photography method. Br J Nutr 2008;101(03):446. doi:10.1017/s0007114508027438.
47. Meister K, Doyle ME. 2009. Obesity and Food Technology. American Council on Science and Health. https://www.scribd.com/document/37170221/. Archived at http://www.webcitation.org/6tfsQIMLD.
48. Myers A, Johnston N, Rathod V, et al. Im2Calories: Towards an Automated Mobile Vision Food Diary. 2015 IEEE International Conference on Computer Vision (ICCV). 2015. doi:10.1109/iccv.2015.146.

49. Noronha J, Hysen E, Zhang H, Gajos KZ. Platemate. Proceedings of the 24th Annual ACM symposium on User interface software and technology - UIST 11. 2011. doi:10.1145/2047196.2047198.
50. Pérez-Escamilla R, Obbagy JE, Altman JM, Essery EV, McGrane MM, Wong YP, Spahn JM, Williams CL. Dietary energy density and body weight in adults and children: a systematic review. J Acad Nutr Diet 2012;112(5):671-84.
51. Reddit. 2011. Who in the World is reddit? Results are in.... https://redditblog.com/2011/09/12/who-in-the-world-is-reddit-results-are-in/. Archived at: http://www.webcitation.org/6tbvF5LJE.
52. Rother C, Kolmogorov V, Blake A. Grabcut: Interactive foreground extraction using iterated graph cuts. ACM SIGGRAPH 2004 Papers on - SIGGRAPH 04. 2004. doi:10.1145/1186562.1015720.
53. Schoeller DA, Bandini LG, Dietz WH. Inaccuracies in self-reported intake identified by comparison with the doubly labelled water method. Can J Phys Pharmacol 1990;68(7):941-949. doi:10.1139/y90-143.
54. Surowiecki J. The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations. 1st ed. New York: Doubleday; 2004. ISBN:978-0-385-50386-0.
55. Tapsell LC, Neale EP, Satija A, Hu FB. Foods, nutrients, and dietary patterns: Interconnections and implications for dietary guidelines. Adv Nutr 2016;7(3):445-454. doi:10.3945/an.115.011718.
56. Wansink B, Chandon P. Meal size, not body size, explains errors in estimating the calorie content of meals. Ann Intern Med 2006;145(5):326. doi:10.7326/0003-4819-145-5-200609050-00005.
57. Watson GB. Do groups think more efficiently than individuals? J Abnorm Soc Psychol. 1928;23(3):328-336. doi:10.1037/h0072661.
58. Williamson DA, Allen HR, Martin PD, Alfonso A, Gerald B, Hunt A. Digital photography: A new method for estimating food intake in cafeteria settings. Eat Weight Disord. 2004;9(1):24-28. doi:10.1007/bf03325041.
59. Reddit. 2017. /r/SampleSize: Where your opinions actually matter! https://www.reddit.com/r/SampleSize/. Archived at: https://www.webcitation.org/6tg5V8wTB
60. Evernote. 2017. Get organized. Work smarter. Remember everything. http://evernote.com. Archived at: http://www.webcitation.org/6udUVuy9M
61. Eloquent Labs. 2017. Eloquent Labs. http://eloquent.ai. Archived athttp://www.webcitation.org/6uhsMCtST
62. Galton F. Vox populi (The wisdom of crowds). Nature 1907;75(7):450-451. doi:10.1038/075450a0.
63. Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, Aderhold A, Bonneau R, Chen Y, Collins JJ. Wisdom of crowds for robust gene network inference. Nat methods 2002;9(8):796-806. doi:10.1038/nmeth.2016.
64. Yi SKM, Steyvers M, Lee MD, Dry MJ. The wisdom of the crowd in combinatorial problems. Cogn sci 2012;36(3):452-470. doi: 10.1111/j.1551-6709.2011.01223.x.
65. Carels RA, Harper J, Konrad K. Qualitative perceptions and caloric estimations of healthy and unhealthy foods by behavioral weight loss participants. Appetite 2006;46(2)199-206. doi:10.1016/j.appet.2005.12.002.
66. Block JP, Condon SK, Kleinman K, Mullen J, Linakis S, Rifas-Shiman S, Gillman MW. Consumers' estimation of calorie content at fast food restaurants: cross sectional observational study. BMJ 2013;346:f2907. doi:10.1136/bmj.f2907.

67. Brown RE, Canning KL, Fung M, Jiandani D, Riddell MC, Macpherson AK, Kuk JL. Calorie Estimation in Adults Differing in Body Weight Class and Weight Loss Status. Med Sci Sports Exerc 2016;48(3):521-526. doi:10.1249/MSS.0000000000000796
68. Hernández T, Wilder L, Kuehn D, Rubotzky K, Moser-Veillon P, Godwin S, Thompson C, Wang C. Portion size estimation and expectation of accuracy. J Food Compos Anal 2006;19:S14-S21. doi:10.1016/j.jfca.2006.02.010
69. Rains SA, Hingle MD, Surdeanu M, Bell D, Kobourov S. A Test of The Risk Perception Attitude Framework as a Message Tailoring Strategy to Promote Diabetes Screening. Health Comm 2018;0(0):1-8. doi:10.1080/10410236.2018.1431024