cTADA: The Design of a Crowdsourcing Tool for Online Food Image Identification and Segmentation

Shaobo Fang, Chang Liu, Khalid Tahboub, Fengqing Zhu and Edward J. Delp
School of Electrical and Computer Engineering
Purdue University
West Lafayette, Indiana, USA

Carol J. Boushey
Cancer Epidemiology Program
University of Hawaii Cancer Center
Honolulu, Hawaii, USA

***Abstract*—Measuring accurate dietary intake, the process of determining what someone eats during the course of the day is considered to be an open research problem in the nutrition and health fields. We have developed image-based tools to automatically obtain accurate estimates of what foods and how much energy/nutrients a user consumes. In this work, we present a crowdsourcing tool we designed and implemented to collect large sets of relevant online food images. This tool can be used to locate food items and obtaining groundtruth segmentation masks associated with all the foods presented in an image. We present a systematic design for a crowdsourcing tool aiming specifically for the task of online food image collection and annotations with a detailed description. The crowdsoucing tool we designed is tailored to meet the needs of building a large image dataset for developing automatic dietary assessment tools in the nutrition and health fields.**

***Keywords*-Dietary Assessment; Crowdsourcing; Food Image Analysis; Groundtruth Segmentation**

I. Introduction

Six of the ten leading causes of death in the United States, including cancer, diabetes and heart diseases can be directly related to diet. Due to the growing concern of chronic diseases and other health problems related to diet, there is a need to develop accurate methods to estimate individual's food and energy intake. Dietary assessment, the process of determining what someone eats during the course of the day, provides valuable insights for mounting intervention programs for prevention of many chronic diseases. Measuring accurate dietary intake is considered to be an open research problem in the nutrition and health fields. Traditional dietary assessment techniques, such as the dietary record, requires individuals to keep detailed written reports for 3-7 days of all food or drink consumed [1], hence it is a time consuming and tedious process.

With the smartphone quickly gaining popularity in recent years, the use of a smartphone can provide a unique mechanism for collecting dietary information of individuals. By February 2016, 72% of American adults were smartphone owners and there has been a noticeable rise in mobile phone and internet usage in the past few years in the emerging and developing nations [2]. We have been investigating the use of images that users take of their meal before and after eating occasions to assess dietary intake. We have developed a system, known as the Technology Assisted Dietary Assessment System (TADA), to acquire and analyze food images [3], [4], [5]. The TADA system and the associated mobile Food Record (mFR), a mobile application, allows users to acquire food images using a mobile telephones [3], [4], [5]. The TADA system has been used in more than 14 scientifically implemented user studies, including environments in the wild, by more than 800 users who have taken more than 60,000 food images. Image processing and computer vision methods are then used to determine the food type, volume, energy (kilocalories) and nutrients of the food [5], [6], [7] present in the images. Other mobile dietary assessment systems have also been developed such as FoodLog [8], FoodCam [9], DietCam [10] and Im2Calories [11]. However, these systems have not been tested under rigorous experimental conditions as has the TADA system.

Training-based techniques have been widely used in recent years for developing automatic dietary assessment systems [5], [7], [11]. For training-based techniques, increasing the training data size would in general improve the accuracy of the system, thus a larger image dataset is always preferred. To date we have a food image dataset with more than 60,000 food images all collected from scientific studies that can be possibly used as training data for our system. We have groundtruth food labels, segmentation masks and portion sizes information for thousands of the food images. In addition to the food images we have collected, a few other food image datasets are available, namely the PFID: Pittsburgh fast-food image dataset [12], UEC-Food 100/256 [13] and Food-101 [14]. The images in [12] are collected under laboratory set-up and only with fast food. Thus the categories and the appearances of the eating scenes do not best suit our use to examine realistic, diverse eating occasions. Furthermore, although both [14] and [13] contain a large amount of food images and a decent range of food types, we feel a detailed description for systematic design of food images collection and annotation is not revealed. Without a well-designed user interface, removing the noisy images from candidate sets and generating the groundtruth segmentation masks are inefficient and not feasible. In addition, many food tags in [14] and [13] are dish names instead of individual foods (in [13] many are Asian style cuisine), we feel the datasets do not meet all of our needs. As our goal is not only to identify the food items but also to estimate the energy/nutrient information from the food images, we are interested in food items that have nutrient information made available by standard food nutrient databases, such as the United States Department of Agriculture (USDA) Food and Nutrient Database for Dietary Studies (FNDDS) [15].

Online image sharing is quickly gaining popularity in recent years (for example, through social networks such as Facebook and review orientated websites such as Yelp), and there are hundred-thousands of food images uploaded by smartphone users. We believe online food images can be used as part of our training data developing automatic dietary assessment techniques and provide valuable contextual information such as users' dietary patterns and food co-occurrence patterns. We define the contextual information as the data that is not directly produced by the visual appearance

of an object in the image, but yields information about users' diet pattern or can be used for diet planning [7]. Collecting food images with proper annotations in a systematic way is a challenging task and requires systematic designs [16]. "Crowdsourcing", as defined in [17], also referred to as the collective intelligence, the wisdom of the crowd or human computation, is often considered as an effective solution to problems that involve cognitive tasks. Amazon Mechanical Turk (AMT) has been used in the past for food image collection and annotation tasks [13], [18] however the AMT is not tailored for the needs emerged from our research of building a large food image dataset efficiently with food items labeled, localized, and segmented.

In this work we present a crowdsourcing tool, namely the crowdsourcing TADA (cTADA), that is tailored to address our needs of online food image collection and annotation. In addition to label and localize the target objects in the images [16], the cTADA is also capable of generating accurate segmentation masks for food objects based on users' input. To generate the segmentation masks, both the user input and automatic segmentation technique [19] are required. Furthermore, the categorical labels (such as "meats" and "beverages") that are assigned to food items in segmentation step are food attributes (similar to the "biometrics" obtained in annotation tasks in [20], [21]). We used a programming interface to collect a large amount of online food images. We designed criteria for the removal of noise from images. Similar to [16], we are able to label and localize the food objects in images. In addition, the cTADA tool allows us to identify all the food items in an image (located by bounding boxes) and generate associated segmentation masks for each food item.

II. The Design of cTADA Crowdsourcing Tool

Various food websites (such as `foodspotting.com`, `foodgawker.com`) contain large amounts of food images. Many food images are uploaded by users on reviews-oriented websites (such as Yelp) and image sharing/social networks (such as Flickr, Instagram, Pinterest, Facebook). We believe many of those food images can be used as the training data in our TADA image analysis system. We define a set of criteria for a food image to be included in our dataset. In addition, the crowdsourcing tool must be efficient and effective as each of the crowd members will go through thousands of food images.

A. Obtaining Online Food Images

Manually downloading thousands of online food images is not feasible. We use Application Programming Interface (API) made available by image website or the search engine for image collection. The APIs we used were Flickr API [22] and Google Custom Search Engine (CSE) API [23]. The APIs allow us to obtain the food images based on the search terms (food tags) we are interested in. Existing datasets frequently use dish names as food tags. The disadvantage of using dish name is that the same type of dish posts very large variation by the look, ingredients and layouts as they were prepared by different people/restaurants. We use the food categories that are frequently present in our existing food image dataset collected from users in nutrition/health studies. The advantage of using such food categories is the energy and nutrient information is made available by the FNDDS database [15].

Figure 1. Examples of food images we collected for the nutrition scientific studies (left) v.s. food images collected online with aesthetic appearances (right).

Figure 2. Defining the foreground (green) and background (red).

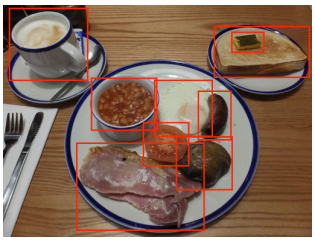

Figure 3. An example of online food image that contains multiple food items.

The food images obtained based on the tags will inevitably contain noisy images that we can not use. We define the noisy images as those that either contain irrelevant content, or have significant different appearances compared to our existing food images collected from scientific studies. A crowdsourcing process is required to remove the noisy images from the candidate food images collected.

B. cTADA: Noisy Image Removal

We first remove images that contain irrelevant contents. The irrelevant content means no food item in the image, images with logos/watermarks/texts and images containing faces. As our goal is to incorporate the online food images collected as part of the training dataset, we want to only include the images that are taken by actual users and exclude those images with aesthetic appearances (a comparison as shown in Figure 1). Food images with aesthetic appearances are likely captured and/or retouched by professional photographers and have fundamental differences compared to the images taken by average users regarding textures, colors, angles and layouts. To guide the crowds to successfully remove images with such aesthetic appearances, we define clear criteria for crowd members with image examples that show different lightings (e.g. professional lighting versus environment light), colors (e.g. vivid and saturated color versus natural color), textures (e.g. very smooth and reflective surface versus regular surface), angles (e.g. close-up or other creative angles versus common camera poses).

We do not exclude the food images that contain multiple food items. In fact, we believe food images that contains multiple food items will help us better understand the users' diet patterns and food co-occurrence patterns. Such patterns can provides us with important insights that can help dietary assessment.

C. cTADA: Food Item Localization and Segmentation

In addition to removing noisy images, we also want to be able to efficiently have crowds locating and obtaining the segmentation masks associated with the food items in an image. We only assign food images that passed the noisy image removal step to crowds for food item localization and segmentation. Users can still discard noisy images in case of a false positive (where the image should be neglected in a noisy image removal step).

To locate the food item, we ask the users to first draw a bounding box around one food item. This task can be performed easily and efficiently by click-and-drag using a computer mouse on our web interface. The bounding box drawn is then cropped out of the original image as preparation for generating the segmentation mask. Users can then select a food tag associated with the bounding box from the hierarchical drop-down food list. The hierarchical drop-down list is designed to best incorporate users' intuitions, for example, we use "meats", "beverages", "green vegetables", "red and orange vegetables" as top level entries where more food categories are available once a top level entry is selected.

To segment the food items, we implemented a stroke tool for users to define foreground and background. Foreground is the area that is associated with the food item, otherwise it will be defined as background. Users do not need to cover all areas of foreground nor background. Drawing lines (the traces of the stroke) across the foreground and background (shown in Figure 2) is sufficient. Similar to many drawing softwares, users can select the linewidth of the stroke tool. With foregrounds and backgrounds defined within the bounding boxes, we use automatic segmentation technique to generate the segmentation mask within the bounding box using the grab cut technique [24], [19]. For the food images that contains multiple food items (shown in Figure 3), the above procedures are repeated till bounding boxes associated with all food tags are located and a segmentation mask is generated for each food item in the images.

III. Experimental Results

For the initial crowdsourcing experiment we recruited the crowds from graduate school students pool all with engineering background in the field of image processing and computer vision. Our crowds are able to give valuable feedback on improving the cTADA crowdsourcing tool at initial design stage. The crowd users can only use our web interface, and were not involved in any of the programming tasks.

For noisy image removal, we implemented a one-click confirmation and short-cut keys on the keyboard, so users can even skip the point-and-click using the computer mouse. The confirmation is then saved in our database and the next image will be automatically present to users to minimize a user's effort. We provide a tutorial on the criteria of noisy image removal to the users. In tutorials, we provide side-to-side comparisons of images and a descriptions for the criteria we designed. We found that users can easily adapt to our set of noisy image removal criteria. With the tutorials, identifying aesthetic appearances is no longer a challenging task even for the crowd members lacking experiences in photography. Based on our observation, we find examining one image takes one second on average for the user, and a maximum of a few seconds. The cTADA system has shown great efficiency in the task of noise image removal and we were able to obtain almost 40,000 food images that can be added our dataset.

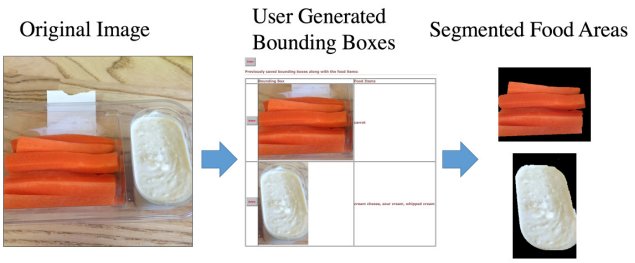

Figure 4. Locating the food items and obtaining the segmentation masks.

The process of localizing and obtaining the segmentation masks associated with all the food items in an image is shown in Figure 4. Users work on one food item at a time. For example, a user will first obtain the bounding box associated with one food item, then identify the food type and define the foreground and background using a 'stroke' tool and 'save' the action performed using the user interface. If there is more than one food item present, an 'add' button can be clicked to repeat the above procedures till all food items are done. The procedure is straight forward and minimizes users' efforts. We do not require users to manually crop out the segmentation masks as it is time consuming and not feasible when working with a large image dataset. Instead, the automatic segmentation tool [24], [19] we implemented on our server will generated very accurate segmentation masks from the bounding boxes and foregrounds and backgrounds defined, as shown in Figure 4.

IV. Conclusion and Future Work

We have designed and implemented the cTADA crowdsourcing tool tailored for the task of incorporating online food images into our food image dataset. We show that cTADA is efficient and effective in removing noisy images, locating the bounding boxes containing the food items and obtaining segmentation masks associated with all the food items in the image. However, we have noticed some mistakes are made unwillingly by the users, especially for the noisy image removal step as each task is done on the scale of a few seconds. In the future, we would like to address the issues of minimizing/avoiding mistakes made unwillingly by the users.

We have gained valuable insights from our experiments on the design of cTADA crowdsourcing tool. For example, which food tags to use as search entries and common appearances of food images taken by the users. Online food images introduce new perspectives as how we can collect and work on food images that are captured by users with no specific instructions. With the cTADA tool, we are capable of expanding our food image dataset with online food images based on the food tags. We no longer have the issue of lacking training images for new food categories in our TADA image analysis system. We are investigating the use of contextual information for the refinement of food identification and portion size estimation. In the future we are also interested in relating texts (e.g. recipes/comments on the same webpage) to food images as more nutrient or contextual information can be revealed and used. It still remains a challenging task to estimate valuable

information from the large amount of image data generated by numerous users which can potentially contribute to research in the health and nutrient fields.

ACKNOWLEDGMENT

This work was partially sponsored by the US National Institutes of Health under grant NIH/NCI 1U01CA130784-01 and NIH/NIDDK 1R01-DK073711-01A1, 2R56DK073711-04 and a Healthway Health Promotion Research Grant and from the Department of Health, Western Australia. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the US National Institutes of Health. Address all correspondence to Edward J. Delp, ace@ecn.purdue.edu or see www.tadaproject.org.

REFERENCES

[1] B. Six, T. Schap, F. Zhu, A. Mariappan, M. Bosch, E. Delp, D. Ebert, D. Kerr, and C. Boushey, "Evidence-based development of a mobile telephone food record," *Journal of the American Dietetic Association*, vol. 110, no. 1, pp. 74–79, January 2010.

[2] "Smartphone ownership and internet usage continues to climb in emerging economies," Pew Research Center. [Online]. Available: http://www.pewglobal.org/2016/02/22/smartphone-ownership-and-internet-usage-continues-to-climb-in-emerging-economies/

[3] C. Boushey, D. Kerr, J. Wright, K. Lutes, D. Ebert, and E. Delp, "Use of technology in children's dietary assessment," *European Journal of Clinical Nutrition*, vol. 63, pp. S50–S57, February 2009.

[4] F. Zhu, M. Bosch, I. Woo, S. Kim, C. Boushey, D. Ebert, and E. J. Delp, "The use of mobile devices in aiding dietary assessment and evaluation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 4, pp. 756 –766, August 2010.

[5] F. Zhu, M. Bosch, N. Khanna, C. Boushey, and E. Delp, "Multiple hypotheses image segmentation and classification with application to dietary assessment," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 1, pp. 377–388, January 2015.

[6] S. Fang, C. Liu, F. Zhu, E. Delp, and C. Boushey, "Single-view food portion estimation based on geometric models," *Proceedings of the IEEE International Symposium on Multimedia*, pp. 385–390, December 2015, Miami, FL.

[7] Y. Wang, Y. He, C. Boushey, F. Zhu, and E. Delp, "Context based image analysis with application in dietary assessment and evaluation," *Multimedia Tools and Applications*, pp. 1–26, November 2017.

[8] K. Kitamura, T. Yamasaki, and K. Aizawa, "Foodlog: Capture, analysis and retrieval of personal food images via web," *Proceedings of the ACM multimedia workshop on Multimedia for cooking and eating activities*, pp. 23–30, November 2009, Beijing, China.

[9] T. Joutou and K. Yanai, "A food image recognition system with multiple kernel learning," *Proceedings of the IEEE International Conference on Image Processing*, pp. 285–288, October 2009, cairo, Egypt.

[10] F. Kong and J. Tan, "Dietcam: Automatic dietary assessment with mobile camera phones," *Pervasive and Mobile Computing*, vol. 8, pp. 147–163, February 2012.

[11] A. Myers, N. Johnston, V. Rathod, A. Korattikara, A. Gorban, N. Silberman, S. Guadarrama, G. Papandreou, J. Huang, and K. Murphy, "Im2Calories: towards an automated mobile vision food diary," *Proceedings of the IEEE International Conference on Computer Vision*, December 2015, Santiago, Chile.

[12] M. Chen, K. Dhingra, W. Wu, L. Yang, R. Sukthankar, and J. Yang, "Pfid: Pittsburgh fast-food image dataset," *Proceedings of the IEEE International Conference on Image Processing*, pp. 289–292, November 2009, Cairo, Egypt.

[13] Y. Kawano and K. Yanai, "Automatic expansion of a food image dataset leveraging existing categories with domain adaptation," *Proceedings of European Conference on Computer Vision Workshops*, pp. 3–17, September 2014, Zurich, Switzerland.

[14] L. Bossard, M. Guillaumin, and L. V. Gool, "Food-101 – mining discriminative components with random forests," *Proceedings of European Conference on Computer Vision*, vol. 8694, pp. 446–461, September 2014.

[15] "USDA food and nutrient database for dietary studies, 1.0." Beltsville, MD: Agricultural Research Service, Food Surveys Research Group, 2004.

[16] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and F. Li, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211 – 252, 2015.

[17] J. Howe, "The rise of crowdsourcing," *Wired Magazine*, vol. 14, no. 6, pp. 1–4, 2006, Dorsey Press.

[18] M. Rabbi, J. Costa, F. Okeke, M. Schachere, M. Zhang, and T. Choudhury, "An intelligent crowd-worker selection approach for reliable content labeling of food images," *Proceedings of the conference on Wireless Health*, pp. 9:1–9:8, October 2015, bathesda, MD.

[19] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," *ACM Transactions on Graphics*, vol. 23, no. 3, pp. 309–314, 2004.

[20] E. Taborsky, K. Allen, A. Blanton, A. Jain, and B. Klare, "Annotating unconstrained face imagery: A scalable approach," *Proceedings of the International Conference on Biometrics*, pp. 264–271, May 2015, Phuket, Thailand.

[21] D. Martinho-Corbishley, M. Nixon, and J. Carter, "Analysing comparative soft biometrics from crowdsourced annotations," *IET Biometrics*, vol. 5, no. 4, pp. 276–283, November 2016.

[22] "Flickr: The app garden," [Online]. Available: https://www.flickr.com/services/api/.

[23] "Google: Custom search engine," [Online]. Available: https://cse.google.com/cse/.

[24] G. Bradski and A. Kaehler, *Learning OpenCV: Computer vision with the OpenCV library*. O'Reilly Media, Inc., 2008.