

Assessing Individual Dietary Intake in Food Sharing Scenarios with a 360 Camera and Deep Learning

Jianing Qiu

Department of Computing
Imperial College London

London, UK

jianing.qiu17@imperial.ac.uk

Frank P.-W. Lo

Department of Surgery and Cancer
Imperial College London

London, UK

po.lo15@imperial.ac.uk

Benny Lo

Department of Surgery and Cancer
Imperial College London

London, UK

benny.lo@imperial.ac.uk

Abstract—A novel vision-based approach for estimating individual dietary intake in food sharing scenarios is proposed in this paper, which incorporates food detection, face recognition and hand tracking techniques. The method is validated using panoramic videos which capture subjects’ eating episodes. The results demonstrate that the proposed approach is able to reliably estimate food intake of each individual as well as the food eating sequence. To identify the food items ingested by the subject, a transfer learning approach is designed. 4,200 food images with segmentation masks, among which 1,500 are newly annotated, are used to fine-tune the deep neural network for the targeted food intake application. In addition, a method for associating detected hands with subjects is developed and the outcomes of face recognition are refined to enable the quantification of individual dietary intake in communal eating settings.

Index Terms—dietary intake assessment, 360-degree video, object detection

I. INTRODUCTION

Dietary intake are usually assessed by techniques such as 24-hour dietary recall (24HR) and food frequency questionnaire (FFQ) [1]. These methods rely mainly on the subjects to recall and report their food intake which are subjective and highly inaccurate. This has been a major issue for nutritional epidemiological studies. In addition, as the collected information is often qualitative and abstract, it often requires the nutritionist to supervise the data collection, interpret the collected information, and provide an estimate of the nutrient intakes of the subject, which is a very labour intensive process. It is for these reasons that technological approaches are needed to enable objective and pervasive assessment of dietary intake. With the advances in computer vision and its applications in food-related areas such as food recognition [2], [3], volume estimation [4], [5] and food image-recipe retrieval [6], [7], vision-based approach is one of the major directions in addressing the unmet need for automatic objective assessment of dietary intake.

Although the recent development in deep learning has shown promising results in food recognition, the approaches are designed mainly for western cultures. In some cultures, such as those in Asian and African countries, communal eating and shared plates are very common in typical households. Identifying individual dietary intake in a communal eating or shared plate scenario is a very challenging problem in

This work is supported by the Innovative Passive Dietary Monitoring Project funded by the Bill & Melinda Gates Foundation (Opportunity ID: OPP1171395).

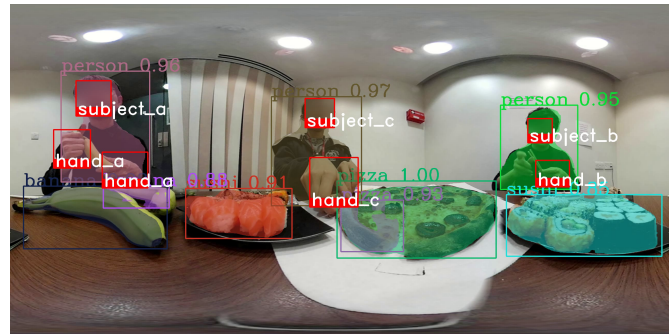


Fig. 1. Estimating dietary intake with food detection, face recognition and hand tracking based on a 360 video in which 3 subjects share a pizza, and a few banana and sushi. Subjects’ faces are blurred using Gaussian blur to protect their privacy.

nutritional studies. Thus far, no technological approach is able to tackle this challenge. With the aim of addressing this need for nutritional epidemiological studies, this paper proposes a novel approach of using a 360 camera (Samsung’s Gear 360), and applying deep learning techniques to identify individuals (i.e., recognizing the subjects’ faces and tracking their hands) and detect food items. In food sharing settings, using a 360 camera can capture all individuals’ faces, hands and food items. Based on this technology, the proposed approach infers the individual intake by associating each face recognized and hand detected with the food items being consumed. Fig. 1 illustrates the outcome of the deep learning methods in a food sharing scenario. To the best of our knowledge, this is the first work that combines these techniques together to tackle the food sharing problem.

The contributions of this work are threefold:

- A vision-based approach is proposed for estimating individual dietary intake in food sharing scenarios;
- A technique is developed to associate the detected hands with individuals;
- 1,500 images with segmentation masks of food items are newly annotated for food detection as well as food instance segmentation¹.

II. METHODOLOGY

To estimate individual dietary intake using a 360 camera, the proposed method measures the hand-face distance and verifies the hand-food interaction of each subject during their

¹https://www.doc.ic.ac.uk/~jq916/food_anno_1500

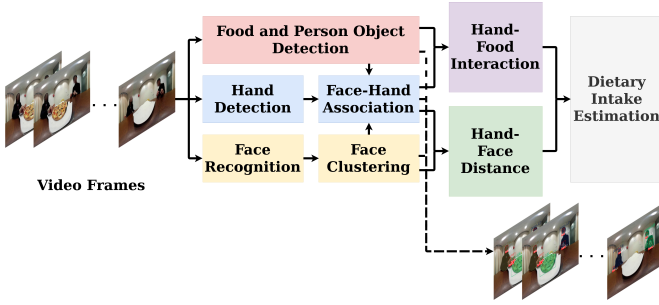


Fig. 2. Framework of the proposed vision-based approach for assessing individual dietary intake in food sharing scenarios.

meals. Fig. 2 illustrates the framework of the method. The components of the framework are described in the following.

A. Food and Person Detection

To detect the food items being consumed, Mask R-CNN [8] is adopted, which outputs the food class, the bounding box indicating the location, and the segmentation mask of each detected food item. The predicted food masks can facilitate the volume estimation but which is beyond the scope of this paper. Mask R-CNN can also detect the *person* object but without knowing the identity, which can be solved through face recognition.

Fine-tune Mask R-CNN. The Mask R-CNN used in this paper is implemented by [9], which was pre-trained on the COCO dataset [10]. However, due to the limited number of food categories in the COCO dataset, the Mask R-CNN can only detect a very small number of food items and which is not sufficient for our application for shared plate scenarios. Therefore, another 5 common food classes are selected from the Food-101 dataset [2], and annotated using VGG Image Annotator [11]. The resulting images are then merged with a subset of the COCO which contains 8 food classes as well as the *person* class to construct a new dataset with 4, 200 images (300 images for each class). Table I shows the selected classes. 90% of the dataset is used for fine-tuning Mask R-CNN, and the rest is retained for validation. The best trained model is then applied on 360 videos to detect food and *person* objects.

B. Face Recognition

For face recognition, a library called *face_recognition* [12] is adopted in which the pre-trained convolutional neural network (CNN) model is enabled to identify individuals, and output the bounding boxes and labels of the faces detected.

Refine Face Recognition. In our experiments, mis-recognition of subjects' faces occurs in some frames. To rectify this mis-identification issue, *KMeans* algorithm is utilized to cluster all faces after the whole video is processed. Faces of the same cluster are then assigned with the label that is the most frequent in that cluster based on the assumption that subjects do not change their seats during the eating episodes.

C. Hand Tracking

Accurate capturing of hand-food interaction could provide essential information to improve the determination of food consumption of each subject. The subjects' hands are first detected using the pre-trained model from [13]. Each hand

TABLE I
SELECTED CLASSES FOR FINE-TUNING MASK R-CNN

Dataset	Class					
COCO	person	apple	banana	sandwich	hot dog	donut
Food-101	fish and chips	fried rice	hamburger	pizza	spaghetti bolognese	sushi

is localized by a bounding box. Detected hands are then associated with subjects through the following method (recall Mask R-CNN and *face_recognition* return the bounding boxes of *person* instances and faces, respectively): 1) For each face and *person* instance of a frame, if the bounding box of a face intersects with that of a *person* instance, then the *person* instance is assigned with the identity (i.e., label) of that face. If a face intersects multiple *person* instances (e.g., if a subject leans forward to grab food away from him/her, its enlarged bounding box may overlap with other subject's face), then the one with the smallest bounding box is associated with the face; 2) For each hand instance, if its bounding box intersects with that of a *person* instance which has been assigned with an identity, then the same identity is passed to the hand instance. Once a hand is labelled with an identity, no more intersections between it and the remaining *person* instances will be measured. A threshold ϵ determining the extent to which 2 bounding boxes must intersect is defined.

D. Eating Events Detection

Based on the hand-face distance and hand-food interaction determined, an *Eating Events Detection Algorithm* (see Algorithm 1) is developed to estimate individual food consumption in shared plate scenarios.

Hand-Face Distance. In each frame for each subject, if there exist bounding boxes of his/her face and hand, then the hand-face distance is measured as the distance between the centers of their bounding boxes. The hand-face distance data is downsampled after all frames have been processed, and outliers are then removed using *Local Outlier Factor* (LOF), and *Exponentially Weighted Moving Average* (EWMA) is then applied in both forward and backward directions to smooth the distance data. The list of distances after EWMA is represented as D . The mean of the 200 shortest distances as well as that of the 200 longest distances are denoted as d_{min} and d_{max} . An upper bound distance of eating is calculated using (1), which is denoted as τ :

$$\tau = d_{min} + \frac{(d_{max} - d_{min})}{\lambda} \quad (1)$$

where λ is a scale factor.

From the experiments, it is found that the hand-face distance increases when the subject is about to grab a food item. When the subject is eating the food, the hand-face distance is relatively short. An intuitive pattern of the variation of hand-face distance is hypothesized as a wave-like signal, with peaks being grabbing food and the interval between 2 peaks being eating. Peaks inside D are found by comparing the adjacent values with the *prominence* set to half of $d_{max} - d_{min}$. The list D is then split into intervals using the peaks detected. Considering when the subject's hands are close to the first food

item he/she is about to eat at the start of recording, then the hand-face distance normally decreases immediately at the first few seconds, which results in a *complete* peak that indicates the start of eating unlikely to form. Therefore the first value in D as well as the last one are also appended to the list of found peaks at the front and the rear, respectively (i.e., treated as peaks as well). To restrict the condition of having eaten a food item, a ratio α is introduced as the number of distances below τ between 2 peaks to all of distances between them.

Note that when the subject is eating the food, the food that is being consumed is very difficult to be recognized due to the deformation and very small size. Whereas at the grabbing stage, since hands and food items are close to the camera, the interactions are easier to be captured. The determination of which food items are consumed by the subject is therefore based on the hand-food interaction at the grabbing stages.

Hand-Food Interaction. To capture the hand-food interaction, whether the bounding box of a detected hand of the subject intersects with those of detected food items is measured. A different threshold v is introduced to control the extent of intersections. After all of the frames are processed, those food items whose interactions with the subject’s hands are less than 10% of all of the subject’s interactions are then filtered out. This is mainly designed to minimize the affect of mis-recognition due to noise or image distortion. The list of remaining interactions are denoted as I .

III. EXPERIMENTS AND RESULTS

The proposed approach for assessing dietary intake is validated with videos of food sharing captured by a 360 camera. Implementation details including the settings of parameters mentioned in Section II are presented in this section followed by experimental results and analysis.

A. Implementation Details

Due to the high resolution and geometric distortion of the 360 videos, each frame is rescaled to 1024×1024 pixels *without* keeping the original aspect ratio before being input into the Mask R-CNN. For face recognition and hand tracking, no scaling or preprocessing is performed. The threshold ϵ controls the intersection between the bounding boxes of a face and a *person*, and those of a hand and a *person* is set to -10 pixels (i.e. 2 bounding boxes must overlap at least 10 pixels in 2 perpendicular sides) whereas the threshold v defined for the bounding boxes of a hand and a food item is set to 50 pixels, which allows a 50-pixel margin between the 2 boxes. The scale factor λ that adjusts the upper bound distance of eating is set to 3.0. The margin δ introduced in Algorithm 1 is set to 150 frames and η is 15. The ratio α controls the confidence of detecting food consumption is set to 0.1.

B. Dietary Intake Assessment

To evaluate the proposed approach, 3 healthy subjects were recruited and 2 video sequences were captured using the 360 camera, where one sequence with 2 subjects sharing a *pizza*, and the other with 3 subjects sharing a *pizza*, and a few *banana* and *sushi*, which are denoted as *Scenario - 1* (refer to Fig. 2) and *Scenario - 2* (refer to Fig. 1), respectively. Given that

Algorithm 1: Proposed Eating Events Detection

Data: N : number of peaks; P : a list of peaks; D : a list of distances; I : a list of interactions; τ : upper bound distance of eating; δ : margin in frame coordinate; η : number of interactions expected in 2δ consecutive frames; α : ratio adjusts the confidence of detecting food consumption.

Result: F : a list of ids of consumed food.

```

/*  $x$  in  $P, D, I$  represents frame number.
    $y$  in  $P, D$  is the distance and in  $I$ 
   the food id. */
1  $F \leftarrow []$ ;
2 for  $n \leftarrow 0$  to  $N - 2$  do
3   if  $P_y[n] < \tau$  then
4     | continue;
5   end
6    $\text{dBetw} \leftarrow \text{findDistsBetwPeaks}(D, (P_x[n],$ 
7      $P_x[n + 1]))$ ;
8    $\text{dBelow} \leftarrow \text{findDistsBelowUpBound}(\text{dBetw})$ ;
9   if  $\text{ratio}(\text{dBelow}, \text{dBetw}) > \alpha$  then
10     $\text{foodIDs} \leftarrow []$ ;
11    foreach  $i$  in  $I$  do
12      if  $P_x[n] - \delta \leq i_x$  and  $i_x \leq P_x[n] + \delta$  then
13        |  $\text{foodIDs} \leftarrow \text{append}(i_y)$ ;
14      end
15    if  $\text{theNumberOf}(\text{foodIDs}) > \eta$  then
16      |  $F \leftarrow \text{append}(\text{theMost}(\text{foodIDs}))$ ;
17    end
18  end
19 end

```

the food detection is not fine-grained and there is no volume estimation involved, the dietary intake is calculated based on the abstracted information (e.g., a pizza consumption event detected means a *slice* of pizza is consumed).

Table II shows the detailed results of each subject’s food intake in the 2 scenarios. The actual amount of food each subject consumed as well as the predicted amount are reported. To better evaluate the approach, the predicted food eating sequence is also compared with the ground truth.

In *Scenario - 1* where 2 subjects were sharing a pizza, the proposed method is able to recognize all of the pizza consumption activities of each subject. In *Scenario - 2* in which 3 subjects were sharing multiple food items, the approach also achieves satisfactory results. Food items consumed by Subject C are all correctly estimated as well as the sequence of the consumption. For Subject A, the method mis-identifies one episode of pizza consumption as eating a banana, and for Subject B, it misses the last sushi consumption, and wrongly recognizes the first sushi consumption as eating pizza.

The variation of hand-face distance and hand-food interaction during the eating procedure of Subject C in *Scenario - 1* and Subject A in *Scenario - 2* are shown in Fig. 3. For a typical distance changing pattern (see Fig. 3(a)) where the peaks can

TABLE II
RESULTS OF DIETARY INTAKE ESTIMATION USING THE PROPOSED APPROACH. P, S, AND B, STAND FOR PIZZA, SUSHI AND BANANA, RESPECTIVELY

		Banana (piece)		Pizza (slice)		Sushi (piece)		Eating Sequence						
		Ground Truth	Prediction	Ground Truth	Prediction	Ground Truth	Prediction	Ground Truth			Prediction			
Scenario - 1	Subject A			3	3			P.	P.	P.	P.	P.	P.	
	Subject C			3	3			P.	P.	P.	P.	P.	P.	
Scenario - 2	Subject A	2	3	1	0			B.	P.	B.	B.	B.	B.	
	Subject B			1	2	2	0	S.	P.	S.	P.	P.		
	Subject C			2	2	2	2	P.	P.	S.	S.	P.	P.	S.

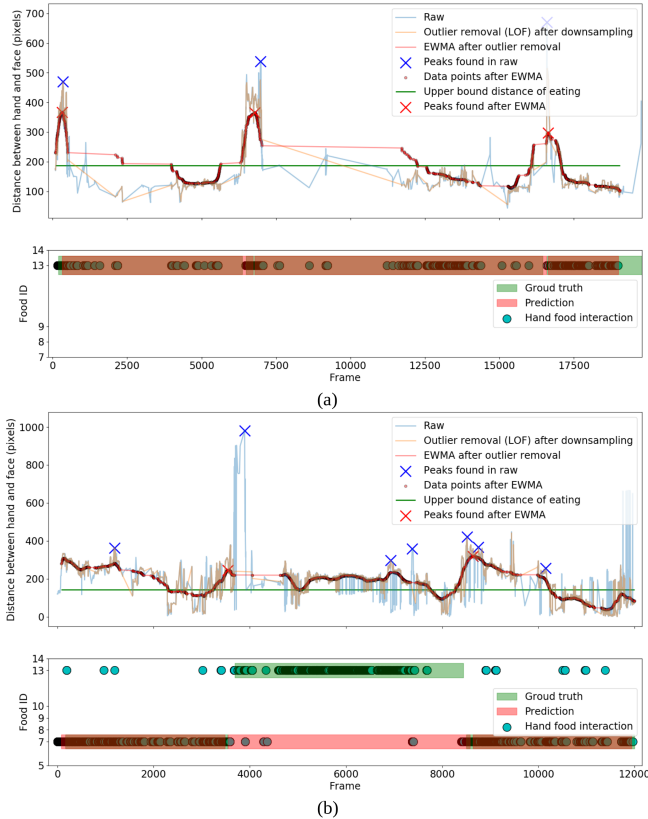


Fig. 3. The variation of hand-face distance and hand-food interaction throughout the eating procedure. (a) Subject C in Scenario - 1; (b) Subject A in Scenario - 2.

be clearly identified and the hand-face distance is relatively short while the subject was eating, the approach can robustly detect each eating episode and predicts the consumed food item correctly based on the hand-food interaction captured at the grabbing stage. For patterns of distance variation that are less intuitive like the one in Fig. 3(b), the approach is still able to find decisive peaks that indicate the start of eating. It is also worth noting that after EWMA, noisy peaks found in raw data are significantly reduced. Though in these 2 examples, food items like *pizza* and *banana* can still be recognized while being eaten, this does not apply for other items. The hands of Subject A in Scenario - 2 was close to the first item (*banana*) he was about to consume at the start of recording. A complete peak therefore is not able to be detected for this consumption. By treating the first value of the list D as a peak as mentioned in Section II-D, the consumption activity is able

to be recovered. In addition, the eating *pizza* event of Subject A is mis-recognized as eating *banana* because the number of hand-banana interaction detected in 2δ -frame interval around the peak (refer to the first red cross in Fig. 3(b)) is slightly larger than that of hand-pizza interaction.

IV. CONCLUSION

A vision-based approach for estimating individual dietary intake in food sharing settings is proposed. Based on transfer learning, Mask R-CNN was fine-tuned for the targeted food intake application. By integrating the food detection with face recognition and hand tracking, the proposed approach is able to infer the individual food consumption. This is the first attempt to use a 360 camera and deep learning to tackle the challenge in quantifying individual food intake in shared plate scenarios. The results from the preliminary study have shown the robustness of the proposed method for individual dietary assessment. In future work, we are investigating the integration of vision techniques with wearable sensors, and utilizing more generic and stable patterns to better evaluate dietary intake.

REFERENCES

- [1] J.-S. Shim, K. Oh, and H. C. Kim, "Dietary assessment methods in epidemiologic studies," *Epidemiology and health*, vol. 36, 2014.
- [2] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101 – mining discriminative components with random forests," in *ECCV*, 2014.
- [3] K. Yanai and Y. Kawano, "Food image recognition using deep convolutional network with pre-training and fine-tuning," in *ICMEW*, 2015.
- [4] A. Meyers, N. Johnston, V. Rathod, A. Korattikara, A. Gorban, N. Silberman, S. Guadarrama, G. Papandreou, J. Huang, and K. P. Murphy, "Im2calories: Towards an automated mobile vision food diary," in *ICCV*, 2015.
- [5] F. Lo, Y. Sun, J. Qiu, and B. Lo, "Food volume estimation based on deep learning view synthesis from a single depth map," *Nutrients*, vol. 10, no. 12, p. 2005, 2018.
- [6] A. Salvador, N. Hynes, Y. Aytar, J. Marin, F. Ofl, I. Weber, and A. Torralba, "Learning cross-modal embeddings for cooking recipes and food images," in *CVPR*, 2017.
- [7] J.-j. Chen, C.-W. Ngo, and T.-S. Chua, "Cross-modal recipe retrieval with rich food attributes," in *ACM Multimedia*, 2017.
- [8] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *ICCV*, 2017.
- [9] W. Abdulla, "Mask r-cnn for object detection and instance segmentation on keras and tensorflow," https://github.com/matterport/Mask_RCNN, 2017.
- [10] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014.
- [11] A. Dutta, A. Gupta, and A. Zissermann, "VGG image annotator (VIA)," <http://www.robots.ox.ac.uk/~vgg/software/via/>, 2016, version: 2.0.2, Accessed: 28-11-2018.
- [12] ageitgey, "face_recognition," https://github.com/ageitgey/face_recognition, 2017.
- [13] D. Victor, "Real-time hand tracking using ssd on tensorflow," <https://github.com/victordibia/handtracking>, 2017.