Assessing Individual Dietary Intake In Food Sharing Scenarios with Food and Human Pose Detection*

Jiabao Lei[1,2], Jianing Qiu[1,3], Frank P.-W. Lo[1,2], and Benny Lo[1,2]

[1] The Hamlyn Centre, Imperial College London, London SW7 2AZ, U.K.
[2] Department of Surgery and Cancer, Imperial College London, London SW7 2AZ, U.K.
[3] Department of Computing, Imperial College London, London SW7 2AZ, U.K.
{j.lei19, jianing.qiu17, po.lo15, benny.lo}@imperial.ac.uk

Abstract. Food sharing and communal eating are very common in some countries. To assess individual dietary intake in food sharing scenarios, this work proposes a vision-based approach to first capturing the food sharing scenario with a 360-degree camera, and then using a neural network to infer different eating states of each individual based on their body pose and relative positions to the dishes. The number of bites each individual has taken of each dish is then deduced by analyzing the inferred eating states. A new dataset with 14 panoramic food sharing videos was constructed to validate our approach. The results show that our approach is able to reliably predict different eating states as well as individual's bite count with respect to each dish.

Keywords: Dietary intake assessment · 360-degree video · Food detection · Human pose estimation.

1 Introduction

With recent advances in deep learning and computer vision, a new range of methods have been proposed for food recognition [1, 19, 3, 11, 16, 13] and volume estimation [12, 9, 10], aiming to provide objective and accurate measurements of dietary intake. However, most approaches developed to date are mainly for western cultures, and not much research has been conducted to address the problem of dietary intake measurement in communal eating or shared plate scenarios, which are very common in typical Asian or African households. In nutritional epidemiological studies, analyzing individual intake in a communal eating or shared plate setting is one of the major challenges.

To estimate the individual dietary intake in a food sharing scenario, a system shall be able to monitor all movements of the individuals who are sharing the food. To capture such information, multiple cameras are often required, but the

* Supported by the Innovative Passive Dietary Monitoring Project funded by the Bill & Melinda Gates Foundation (Opportunity ID: OPP1171395)

needed set up, calibration and fusion greatly hinder the use of such approach for epidemiological studies. Through combining 2 cameras with wide angle lenses, 360 cameras can capture everything around in close proximity of the camera, and therefore, it is particularly suitable for capturing eating in shared plate and communal eating settings. This work proposes to use a 360 camera (Samsung's Gear 360) to capture the entire episode of a meal in which multiple people are sharing a number of dishes together, and a neural network is then applied to infer the eating states of each individual based on their body pose and position to dishes. The number of bites each person has taken of each dish is then deduced from the inferred eating states throughout the meal. From the number of bites taken, it is then possible to estimate the volume of food an individual has eaten.

2 Related Work

Most vision-based approaches proposed for dietary intake assessment are either constrained to laboratory environments or only targeted for a single subject having a meal alone [8, 20, 18, 4, 15]. Assessing dietary intake in food sharing scenarios have been rarely researched. A recent work in [14] proposes to use food detection, face recognition, and hand tracking to quantify individual dietary intake in food sharing scenarios pre-recorded by a 360-degree camera. In their recorded food sharing videos, 2-3 subjects were sharing food, such as sushi and pizza, together. In [17], a 360-degree camera was set up to record communal eating scenarios in which 4 people have meals together. However, in their settings, each individual eats their own meal on a plate placed in front of them. Therefore, by cropping out the subject and his/her plate, the dietary intake assessment process becomes similar to those common assessment scenarios (i.e., single subject having a meal alone). In this work, we also use a 360-degree camera to record the entire episode of multiple people eating together. Unlike [17], our work aims to quantify individual intake in shared food scenarios in a typical household setting and the data was collected in the subjects' own home. Our setting is closer to [14] (i.e., people share dishes from different plates), but our technical solutions differ from [14] in that we integrate human pose and dish detection, and use a neural network to infer the eating states of each individual to estimate which dishes each subject has eaten and how many bites of each dish they have taken.

3 Method

Fig. 1 illustrates the framework of our proposed approach. A neural network is designed to infer three different eating states (i.e., grabbing the food, eating, and others). The network estimates each individual's state on the basis of their interaction with each dish on the table. Specifically, interactions are modeled using individual's body pose and the location of each dish on the table. We introduce techniques used in each module of our framework in the following.

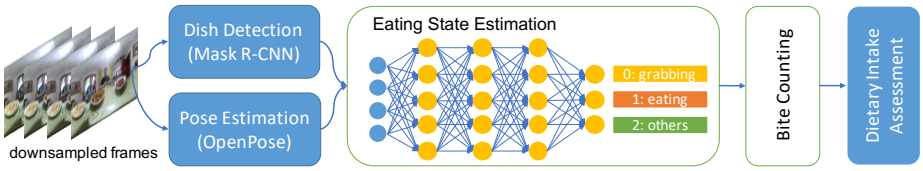


Fig. 1: The framework of our proposed approach, which includes dish detection, body pose estimation, a neural network for estimating the eating state of each individual (i.e., grabbing, eating and others), and a bite counting step, for assessing individual dietary intake in food sharing scenarios.

3.1 Dish Detection

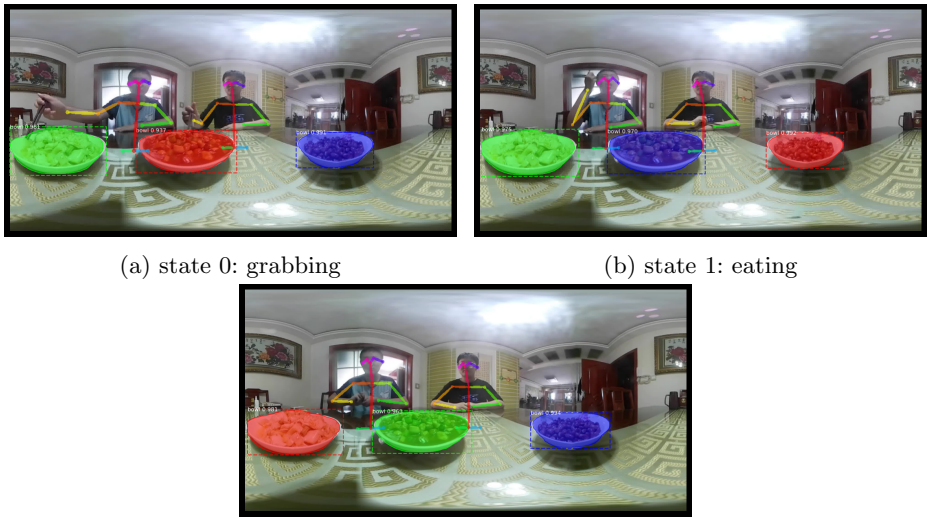
Mask R-CNN [5] is one of the most widely used method for instance segmentation and it is able to detect dishes accurately (i.e., localize and recognize dishes) on the dinning table once fine-tuned with a dish dataset. However, due to the wide variety of prepared dishes in this work, and the limited available data of each dish, we therefore did not resort to fine-tuning Mask R-CNN but instead we used the Mask R-CNN model pretrained on the COCO dataset [7] to detect plates and bowls as the proxy for the associated dishes. Recognizing dish types will be addressed in our future work as our dataset expands. At the current stage of this work, though the method does not recognize the type of each dish on the dinning table, it can detect the location of the dishes, which is sufficient to model subject-dish interactions. Every dish in each video is assigned with a pseudo dish type based on their location (e.g., dishes A, B, and C from left to right). Our approach then utilizes the pseudo dish types and the bounding boxes of the detected food containers for the subsequent dietary intake assessment for each individual. Specifically, for the location of each dish, the x and y coordinates of the upper-left and bottom-right corners of the plate/bowl bounding boxes are used.

3.2 Human Pose Estimation

OpenPose [2] is an accurate tool for estimating body pose from video images, and it is used in this work to estimate the body pose of each individual during the meal for capturing the subject-dish interactions. Specifically, we use it to detect 25 body keypoints, which returns 75 values (x, y, c where x and y indicate the location of a keypoint and c indicates the detection confidence).

3.3 Eating State Estimation

A 4-layer feed-forward neural network is designed to infer the eating state of each individual with regard to each dish on the basis of dish location and individual's body pose. The dimension of each layer of the network is as follows: `79-60-40-20-3 (input, 3 hidden, and output layers)`. The input of the



(a) state 0: grabbing (b) state 1: eating

(c) state 2: others (e.g., doing nothing)

Fig. 2: Illustration of the three states (left subject in each image) during a meal. The estimated body poses and detected dish bounding boxes and masks are overlaid on top of the images.

network is a 79 dimensional vector composed of 4 elements of the location of one target dish and 75 elements of body keypoints. Therefore, given a frame with $\mathcal{N}$ subjects and $\mathcal{M}$ dishes, for each subject, $\mathcal{M}$ number of 79-D vectors are constructed to infer the subject's interactions with all $\mathcal{M}$ dishes. Each 79-D vector has a ground truth label of one of the three eating states. We define these three states as follows: 1) `state 0` (grabbing): the subject reaches out to grab the food and then gradually moves his/her hand close to the face; 2) `state 1` (eating): following grabbing, the subject's hand is near his/her face area until the subject puts down his/her hand; 3) `state 2` (others): the rest situations are treated as others.

3.4 Bite Counting

Based on the inferred states of each individual throughout the meal, we then count how many bites each individual has taken of each dish. Starting from the first predicted `eating` state (`state 1`) of that dish, which is counted as the first bite, the second bite is the next predicted `eating` state that has at least 4 predicted `non-eating` states in between it and the first predicted `eating` state. This then continues to count the bites based on the condition that at least 4 `non-eating` states exist in between 2 predicted `eating` states.

4 Experiments

4.1 Dataset

Data Collection and Pre-processing 14 videos of a household in north-western China eating shared food were collected using a 360-degree camera. 4 subjects participated and consented to be recorded while having their self-cooked meals with their family members. They were asked to eat as they normally do and were free to converse with each other during the meal. Among 14 recorded videos, the longest one lasts for 13 m 58 s, and the shortest one lasts for 5 m 36 s, with the average length being 10 m 49 s. Each video has 2-3 subjects sitting around the table sharing 3-4 dishes. All videos were then downsampled to 2 frames per second for later dietary intake assessment. For each frame, one annotator labelled each subject's states with regard to all dishes present.

Dataset Preparation Leave-one-out cross-validation (LOOCV) is chosen to evaluate eating state estimation and the bite counting method. As we observe the number of 79-D vectors labelled as `state 2` (i.e., others) is far more than that of the other two states, In each fold, the training set is therefore balanced to have its 50% of data as `state 2` and 25% each as `state 0` and `1`. For the test set, we test the network with both the balanced set (i.e., `state 0, 1`, and `2` account for 25%, 25%, and 50% respectively), and the full unbalanced set (i.e., all vectors from the downsampled frames of the test video).

4.2 Results

4.3 Eating State Estimation

Table 1 summarizes the results of the eating state estimation. The network was trained for 20 epochs with cross entropy loss. Adam optimization [6] was used with a learning rate of 0.001. The average top-1 testing accuracy on the balanced test sets is 87.7%. The average accuracy on the unbalanced sets is lower than that of the balanced sets. This is because each unbalanced set has more `state 2` samples than its balanced counterpart, which increases the likelihood of `state 2` samples being mis-recognized, as some of them appear to be similar to grabbing or eating.

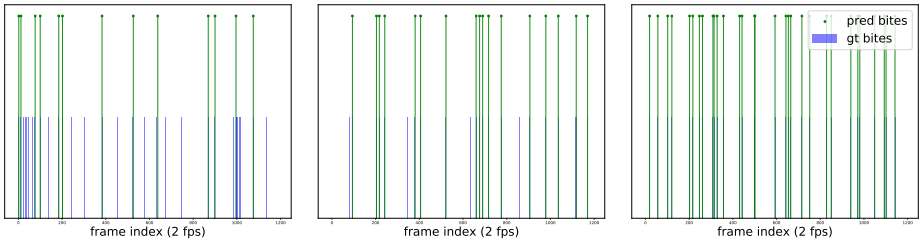
Table 1: The results of eating state estimation (Top-1 Accuracy). V1 to V14 are the recorded video sequences, each used as a test set during LOOCV.

Dataset	V01	V02	V03	V04	V05	V06	V07	V08	V09	V10	V11	V12	V13	V14	**Avg.**
Balanced	93.3	94.0	87.3	70.2	90.8	89.3	46.7	92.2	94.5	94.8	93.5	93.3	94.3	93.2	**87.7**
Unbalanced	59.0	47.1	42.9	60.8	47.8	48.9	52.1	54.4	70.7	51.7	54.2	54.4	59.3	52.2	**54.0**

4.4 Bite Counting

Table 2: The number of bites all subjects in a video have taken. G.T. bites are the ground truth data of each video sequence, Pred. bites are the predicted number of bites in each respective video sequence, △ bites are the difference between the ground truth and the predicted bite counts, and Bite err. % is the error percentage calculated as △ bites / G.T. bites.

	V01	V02	V03	V04	V05	V06	V07	V08	V09	V10	V11	V12	V13	V14	**Avg.**
G.T. bites	168	333	354	104	107	197	124	84	89	134	87	69	84	107	**145.8**
Pred. bites	130	279	195	49	94	162	98	61	78	97	74	56	29	93	**106.8**
△ bites	38	54	159	55	13	35	26	23	11	37	13	13	55	14	**39.0**
Bite err. %	22.6	16.2	44.9	52.9	12.1	17.8	21.0	27.4	12.4	27.6	14.9	18.8	65.5	13.1	**26.2**



(a) subject I (b) subject II (c) subject III

Fig. 3: The predicted and ground truth bites of 3 different subjects in video 8.

Based on the results of eating state estimation on the unbalanced sets, we then count bites for each individual. We first report the sum of predicted bites of all individuals in each test video, and the difference between it and ground truth. Table 2 shows the results. On average, the predicted number has 39.0 less bites than ground truth (106.8 vs 145.8 bites). The bite error percentage after averaging across 14 videos is 26.2%. An example is shown in Fig. 3, which illustrates the predicted and ground truth bites of three different subjects in video 8. It can be seen that in this particular example, most predicted bites are exactly at the same time point as their ground truth bite occurs. While there are some bites not predicted as occurring in the cases of subjects I (Fig. 3a) and II (Fig. 3b), the predicted number of bites and their occurring time points are exactly the same as ground truth in the case of subject III (Fig. 3c).

We then show the bite error percentage of each subject with respect to each dish in each video. Table 3 summarizes the results. In 12 videos, 3 subjects

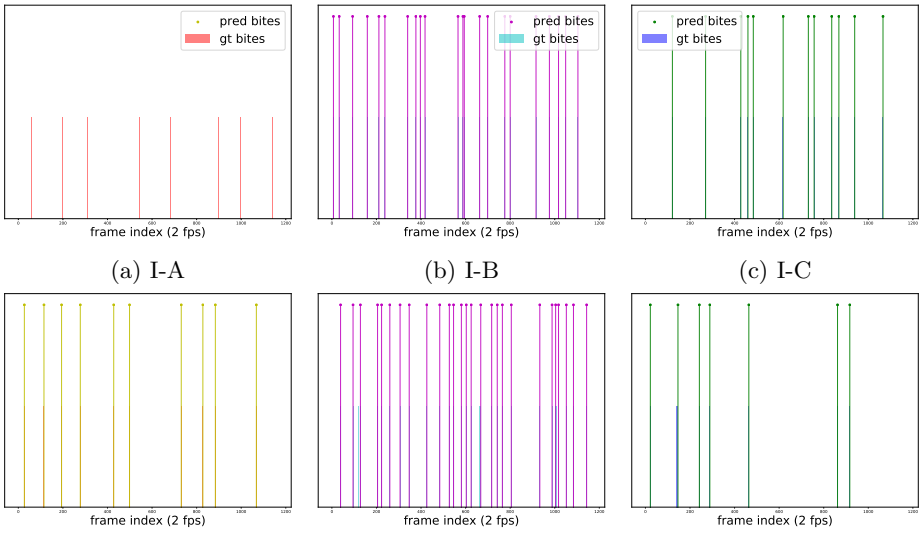
shared 4 dishes. In 81 out of 158 possible subject-dish associations, the predicted number of bites matches exactly with the number of ground truth bites (bite error percentage shown as 0.0 in Table 3). An example of bite prediction of each individual with regard to each dish is shown in Fig. 4. Despite in the case of subject I having dish A, where no bite is successfully predicted, most predicted bites in other cases match their associated true bites.

Table 3: Bite error percentage (each subject with respect to each dish in a video sequence), calculated as △ bites / G.T. bites where △ bites are the difference between the predicted number of bites a subject has taken of a specific dish and the respective ground truth (G.T. bites). Subjects are represented as capital roman numerals, and dishes are indicated using capital letters (i.e., I-A refers to subject I with respect to dish A in a video sequence. I and A can be different subjects and dishes in different videos.)

Err. %	I-A	I-B	I-C	I-D	II-A	II-B	II-C	II-D	III-A	III-B	III-C	III-D	**Avg.**
V01	100.0	19.4	66.7	10.5	0.0	0.0	6.7	0.0					**25.4**
V02	100.0	1.6	6.7	7.4	100.0	0.0	0.0	0.0	2.2	0.0	2.8	0.0	**18.4**
V03	87.5	20.0	33.3	0.0	100.0	55.6	72.9	86.5	6.3	3.8	0.0	0.0	**38.8**
V04	7.1	0.0	0.0	0.0	100.0	100.0	100.0	100.0	100.0	50.0	100.0	33.3	**57.5**
V05	50.0	0.0	0.0	0.0	46.2	0.0	0.0	15.4	0.0	0.0	0.0	0.0	**9.3**
V06	100.0	8.7	0.0	0.0	100.0	5.1	0.0	0.0	14.3	8.3	0.0	3.6	**20.0**
V07	100.0	100.0	100.0	33.3	100.0	5.0	0.0	7.1	7.7	0.0	5.9	16.7	**39.6**
V08	100.0	66.7	55.6	42.9	83.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**29.0**
V09	100.0	0.0	0.0		0.0	6.9	12.5						**19.9**
V10	0.0	0.0	75.0	84.6	14.3	0.0	10.5	0.0	0.0	0.0	0.0	0.0	**15.4**
V11	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	54.5	0.0	0.0	0.0	**12.9**
V12	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	77.8	14.3	0.0	0.0	**16.0**
V13	100.0	0.0	0.0	100.0	75.0	80.0	100.0	0.0	0.0	100.0	58.3	92.3	**58.8**
V14	100.0	10.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**9.2**

5 Conclusion

A novel vision and neural network-based approach for assessing individual dietary intake in food sharing scenarios has been proposed. Based on an individual's body pose and their position to each dish, the network estimates the eating states of each individual throughout the meal. The resulting estimated states are then used to count the number bites each individual has taken of each dish. Experiments have shown that the proposed approach is promising in assessing individual dietary intake in Chinese food sharing scenarios. Evaluating the performance of the proposed approach on more diverse communal eating and food sharing scenarios is planned in future work.



(a) I-A (b) I-B (c) I-C

(d) II-A (e) II-B (f) II-C

Fig. 4: The predicted and ground truth bites subjects I and II have taken of dishes A, B, and C in video 9.

References

1. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101–mining discriminative components with random forests. In: European conference on computer vision. pp. 446–461. Springer (2014)
2. Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., Sheikh, Y.A.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)
3. Chen, J., Ngo, C.W.: Deep-based ingredient recognition for cooking recipe retrieval. In: Proceedings of the 24th ACM international conference on Multimedia. pp. 32–41 (2016)
4. Doulah, A., Ghosh, T., Hossain, D., Imtiaz, M.H., Sazonov, E.: "automatic ingestion monitor version 2"—a novel wearable device for automatic food intake detection and passive capture of food images. IEEE Journal of Biomedical and Health Informatics (2020)
5. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
6. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
7. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
8. Liu, J., Johns, E., Atallah, L., Pettitt, C., Lo, B., Frost, G., Yang, G.Z.: An intelligent food-intake monitoring system using wearable sensors. In: 2012 ninth international conference on wearable and implantable body sensor networks. pp. 154–160. IEEE (2012)

9. Lo, F.P.W., Sun, Y., Qiu, J., Lo, B.: Food volume estimation based on deep learning view synthesis from a single depth map. Nutrients **10**(12), 2005 (2018)
10. Lo, F.P.W., Sun, Y., Qiu, J., Lo, B.P.: Point2volume: A vision-based dietary assessment approach using view synthesis. IEEE Transactions on Industrial Informatics **16**(1), 577–586 (2019)
11. Martinel, N., Foresti, G.L., Micheloni, C.: Wide-slice residual networks for food recognition. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 567–576. IEEE (2018)
12. Meyers, A., Johnston, N., Rathod, V., Korattikara, A., Gorban, A., Silberman, N., Guadarrama, S., Papandreou, G., Huang, J., Murphy, K.P.: Im2calories: towards an automated mobile vision food diary. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1233–1241 (2015)
13. Min, W., Liu, L., Luo, Z., Jiang, S.: Ingredient-guided cascaded multi-attention network for food recognition. In: Proceedings of the 27th ACM International Conference on Multimedia. pp. 1331–1339 (2019)
14. Qiu, J., Lo, F.P.W., Lo, B.: Assessing individual dietary intake in food sharing scenarios with a 360 camera and deep learning. In: 2019 IEEE 16th International Conference on Wearable and Implantable Body Sensor Networks (BSN). pp. 1–4. IEEE (2019)
15. Qiu, J., Lo, F.P.W., Jiang, S., Tsai, C., Sun, Y., Lo, B.: Counting bites and recognizing consumed food from videos for passive dietary monitoring. IEEE Journal of Biomedical and Health Informatics (2020)
16. Qiu, J., Lo, F.P.W., Sun, Y., Wang, S., Lo, B.: Mining discriminative food regions for accurate food recognition. In: British Machine Vision Conference (2019)
17. Rouast, P.V., Adam, M.T.: Learning deep representations for video-based intake gesture detection. IEEE Journal of Biomedical and Health Informatics **24**(6), 1727–1737 (2019)
18. Sun, M., Burke, L.E., Baranowski, T., Fernstrom, J.D., Zhang, H., Chen, H.C., Bai, Y., Li, Y., Li, C., Yue, Y., et al.: An exploratory study on a chest-worn computer for evaluation of diet, physical activity and lifestyle. Journal of healthcare engineering **6**
19. Yanai, K., Kawano, Y.: Food image recognition using deep convolutional network with pre-training and fine-tuning. In: 2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). pp. 1–6. IEEE (2015)
20. Zhu, F., Bosch, M., Khanna, N., Boushey, C.J., Delp, E.J.: Multiple hypotheses image segmentation and classification with application to dietary assessment. IEEE journal of biomedical and health informatics **19**(1), 377–388 (2014)