Imperial College
London

# Assessing Individual Dietary Intake in Food Sharing Scenarios with Food and Human Pose Detection

Jiabao Lei, Jianing Qiu, Frank Lo, and Benny Lo

The Hamlyn Centre
Institute of Global Health and Innovation

# An Overview of This Work

- A novel food sharing dataset has been constructed (14 videos)
- Human pose estimation and dish detection are integrated. Neural network is used to infer different eating states of a subject
- The number of bites a subject has taken of each dish on the dinning table is predicted

# Method - framework

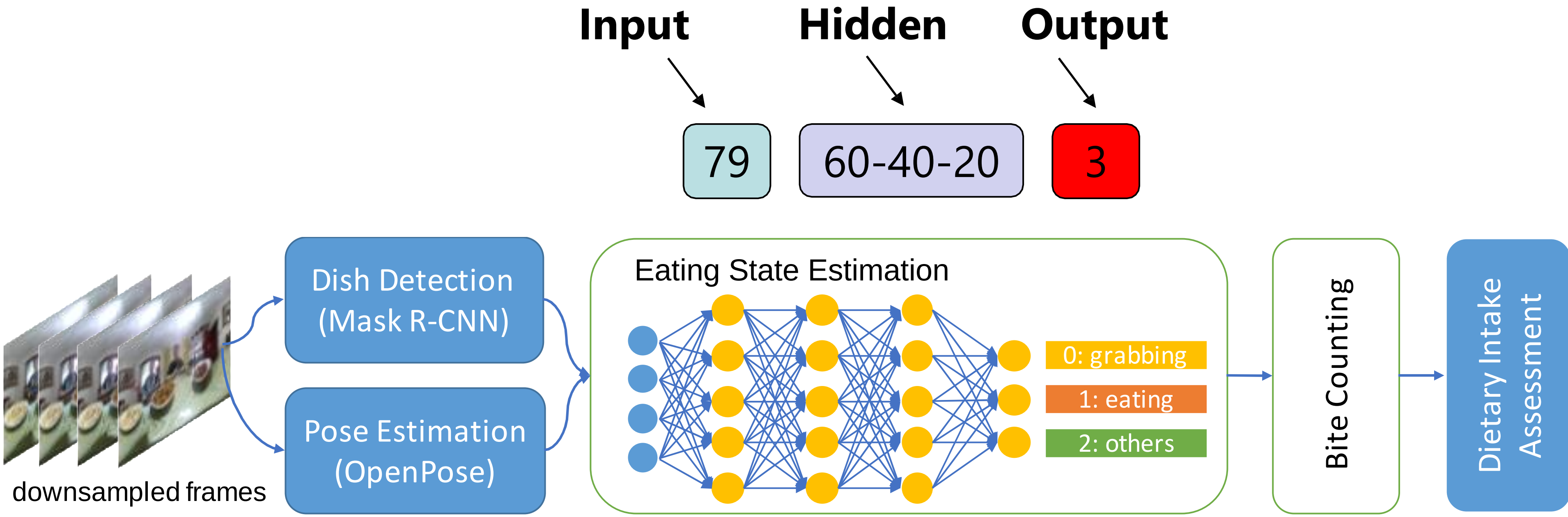
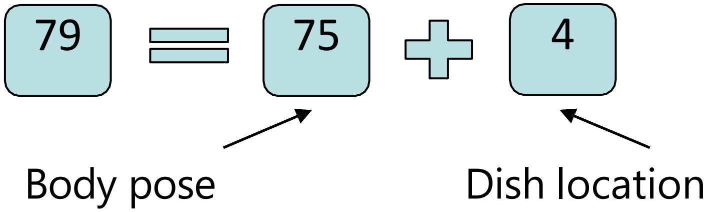


Fig. 1: The framework of our proposed approach, which includes dish detection, body pose estimation, and a neural network for estimating the eating state of each individual



# Method – dish detection

Dish location

4

Upper-left corner (x, y)

Bottom-right corner (x, y)

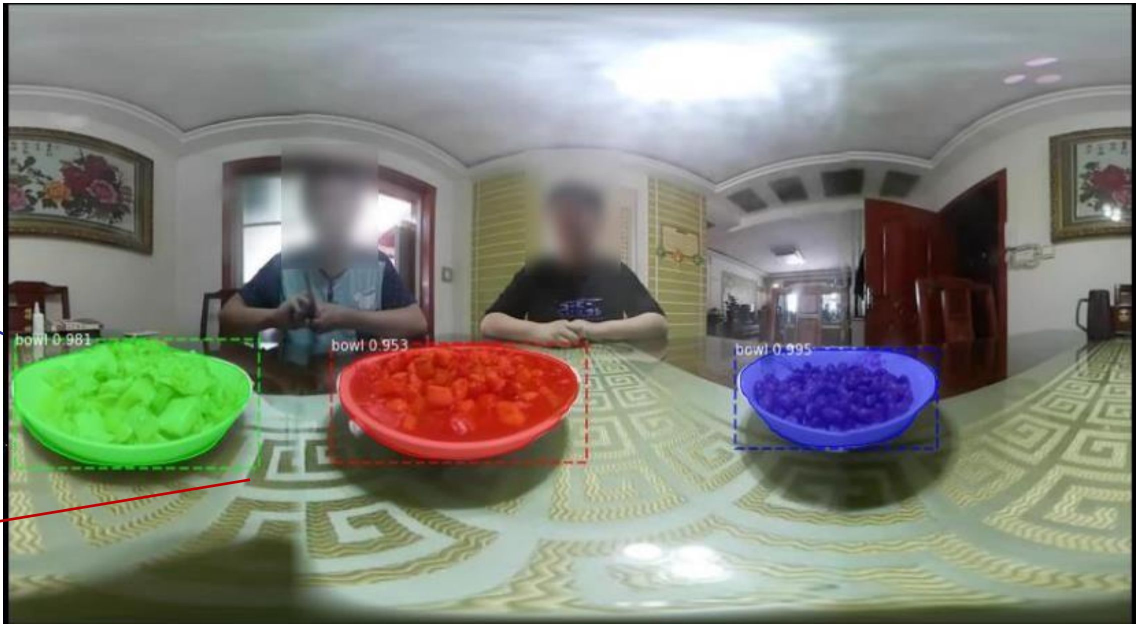
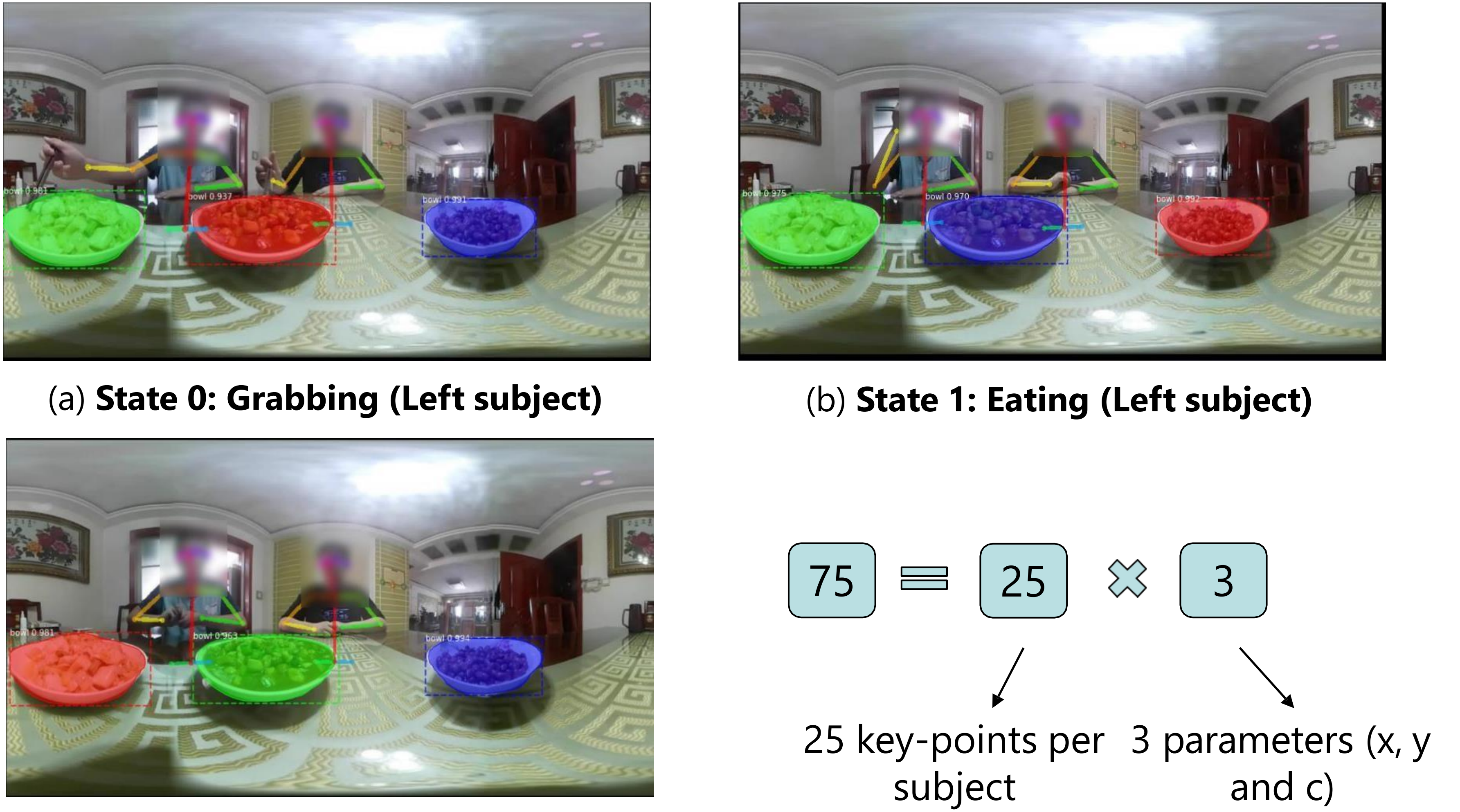


Fig. 2: Dish detection example

- ❖ Due to various dish types used in our dataset. We used the detected food container as the proxy for the associated dish
- ❖ Mask-RCNN pretrained on the COCO dataset was used to detect plates and bowls.

# Method – body pose and eating states



(a) **State 0: Grabbing (Left subject)**

(b) **State 1: Eating (Left subject)**

75 = 25 × 3

25 key-points per subject

3 parameters (x, y and c)

(c) **State 2: Others (Left subject)**

Fig. 3: Illustration of body pose estimation and 3 different eating states

# Method – bite counting

- **First ‘State 1 (eating)**’ frame = first bite
- More than **4 non-eating** frames between **2 eating** frames = next bite
- And so on …

# Experiments - dataset

- **14** videos (lunch and dinner)
- 2 or 3 subjects and 3 or 4 dishes in each video
- Subjects grab and eat food in their normal ways
- Average time: 10 min 43 s
- Down sampled frequency: 2 frames / s

# Experiments – data pre-processing

## Training set

- Too many 'State 2' frames. In order to make dataset balanced, the state distribution is balanced as follows:

  50% 'State 2', 25% 'State 1' and 25% 'State 0'

## Testing set

- Balanced set: 50% 'State 2', 25% 'State 1' and 25% 'State 0'
- Unbalanced set: **All samples** from the down sampled frames of the test video

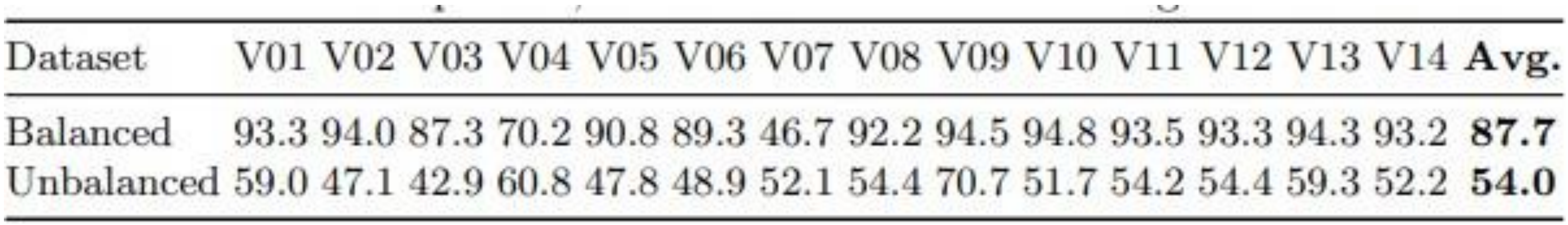
# Experiments – implementation details

- **Leave-one-out cross-validation (LOOCV) was used**

- **The network was trained using cross entropy loss with 20 epochs. Adam optimization was used. Learning rate was set to 0.001**

# Experiments – eating state estimation

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.

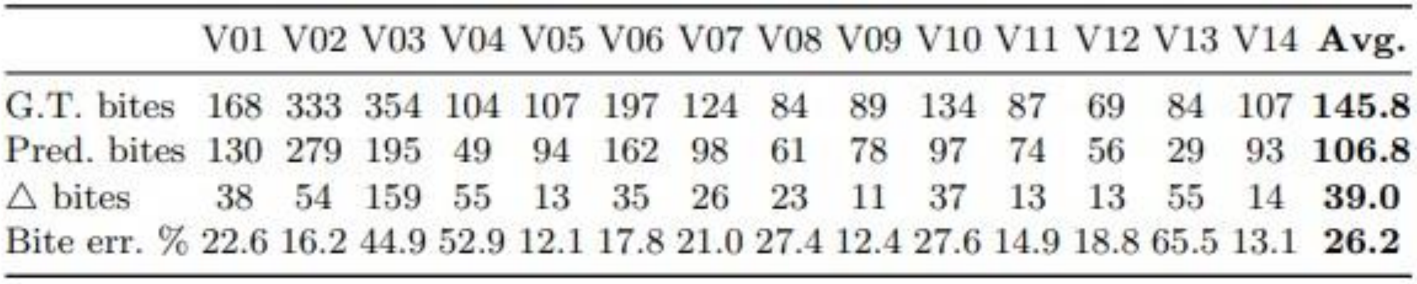


| Dataset | V01 | V02 | V03 | V04 | V05 | V06 | V07 | V08 | V09 | V10 | V11 | V12 | V13 | V14 | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Balanced | 93.3 | 94.0 | 87.3 | 70.2 | 90.8 | 89.3 | 46.7 | 92.2 | 94.5 | 94.8 | 93.5 | 93.3 | 94.3 | 93.2 | **87.7** |
| Unbalanced | 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** |

# Experiments – bite counting

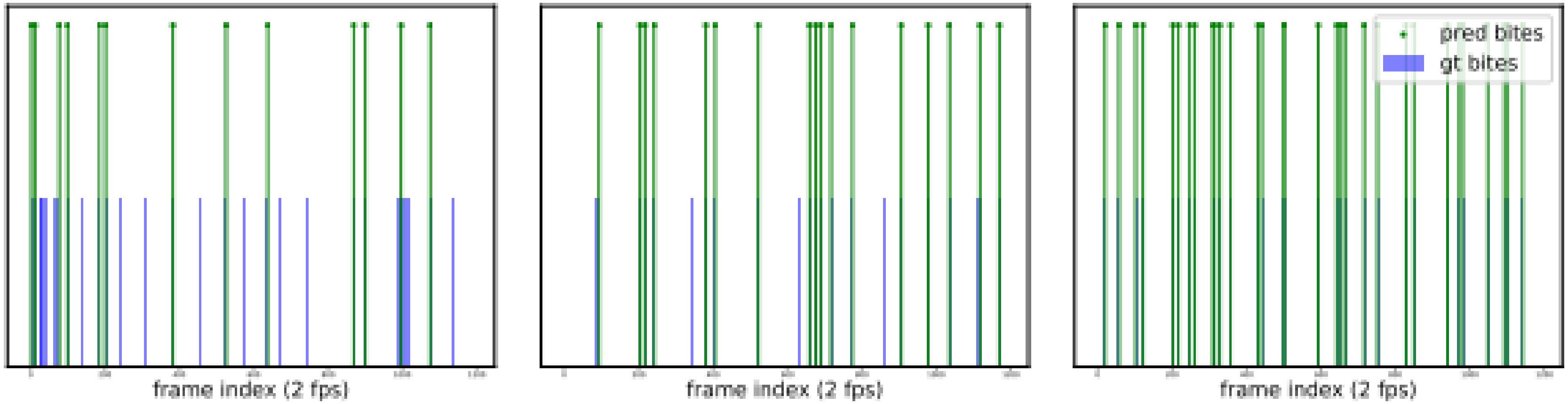
- **G.T. bites:** ground truth data
- **Pred. bites:** prediction
- **Δ bites:** difference between G.T. and Pred.
- **Bite err. %:** Δ bites / G.T. bites

Table 2: The number of bites all subjects in a video have taken



| | 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** |

# Experiments – bite counting



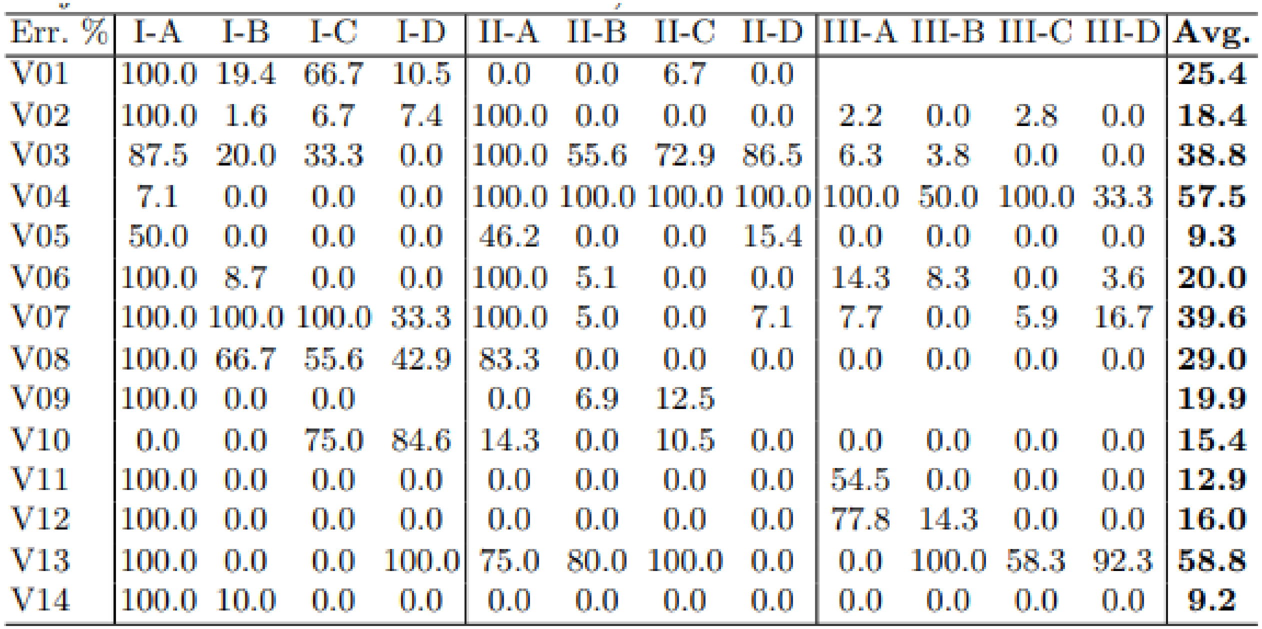
(a) subject I (b) subject II (c) subject III

Fig. 4: The predicted and ground truth bites of **3** different subjects in **video 8**.

- Subject III: high accuracy
- Subject I: low accuracy

# Experiments – bite counting

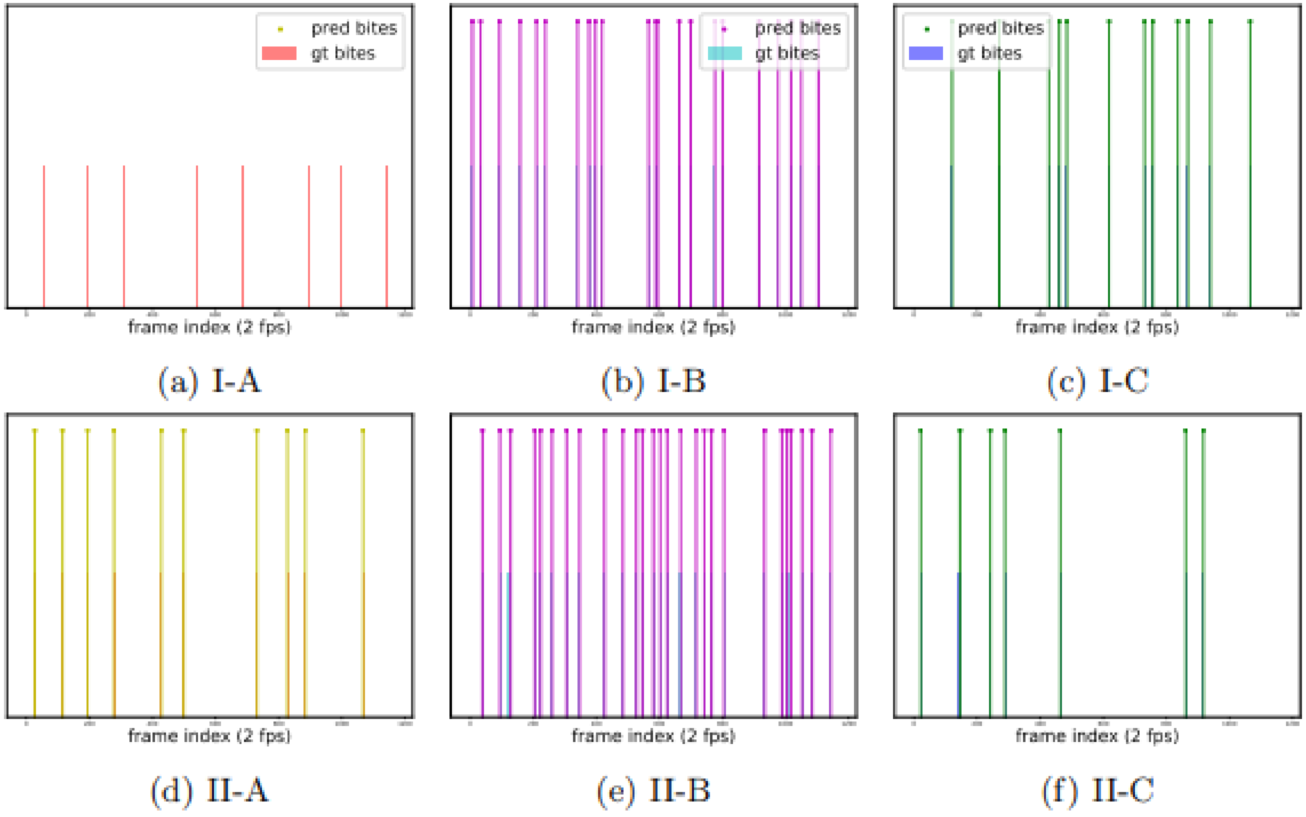
Table 3: Bite error percentage (each subject with respect to each dish in a video sequence)



| 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** |

- I, II, III represent subjects
- A, B, C, D represent dishes

# Experiments – bite counting



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

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

Fig. 5: The predicted and ground truth bites **subjects 1 and 2** have taken of **dishes A, B, and C** in **video 9**

# Thank You!