HowToEat: Exploring Human Object Interaction and Eating Action in Eating Scenarios

Yingcheng Wang, Junwen Chen, Keiji Yanai

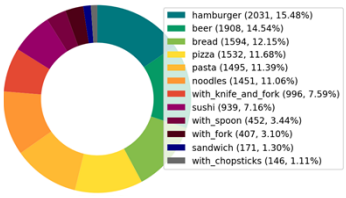
Department of Informatics
The University of Electro-Communications, Tokyo, Japan


MADiMa 2023

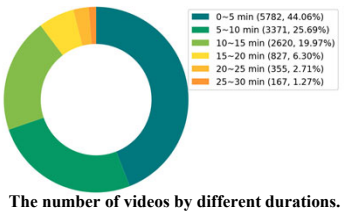
Contributions

- A new dataset **HowToEat**, which is specifically constructed for eating scenarios. Multi-task annotations on both eating action and hand-object interaction.
- Adapting an **HOI detection method to detect hand-object interaction** and achieve notable performance.
- A novel dataset and corresponding detection methods for **eating action of the face**.
- An **eating analysis system**, which is capable of simultaneously detecting hand-object interactions and eating actions.

HowToEat



The number of videos in each eating scenario.



The number of videos by different durations.

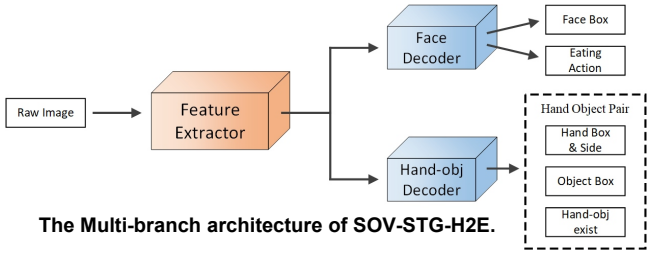
- Videos collected from YouTube in five languages (English, Japanese, French, Chinese, German)
- The total videos is 66 days. 70% of videos are 0~10 minutes.

Hand-Object Annotation with SOV-STG

The results of SOV-STG-Hand on 100DOH with redefined categories.

Model	Left Hand		Right Hand		Hand-Object
	No Contact	Portable Object	No Contact	Portable Object	mAP
SOV-STG-Hand-S	64.61	73.04	56.99	72.76	66.85
SOV-STG-Hand-Swin-L	70.16	80.50	65.35	77.05	73.26

- PPDM used for automatic image extraction is high efficiency, but not the best choice to generate high-quantity annotations.
- We implement SOV-STG-Hand for the hand-object interaction detection task and reannotate HowToEat.

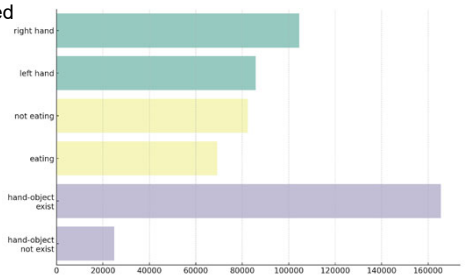


The Multi-branch architecture of SOV-STG-H2E.

The results of the baseline model on HowToEat.

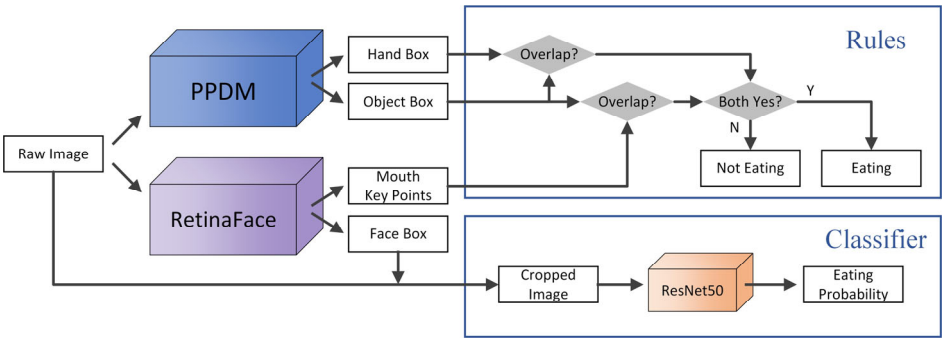
Model	Left Hand		Right Hand		Hand-object	Face		Face
	No Contact	Portable Object	No Contact	Portable Object	mAP	Not Eating	Eating	mAP
SOV-STG-H2E-S	61.91	87.79	47.98	88.56	71.56	57.43	73.89	65.66

- Our SOV-STG-H2E trained end-to-end on HowToEat.



Statistics of HowToEat.

Automatic Image Extraction and Face Box Annotation

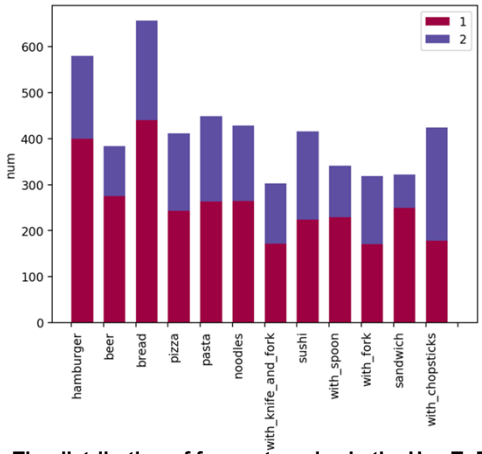


The pipeline of automatic image extraction and labeling. The top half is a rule-based face-eating action detection pipeline, and the bottom half is face-eating action detection based on classifiers.

1. An HOI detection model PPDM trained on 100DOH with a face detection model to **detect eating action by rules**. The frames containing eating actions are extracted, and face boxes are annotated in this step.
2. We crop the face bounding box and manually annotate a HowToEat face dataset, and **train a model for eating action recognition**.

HOI detection evaluation result on the hand-object interaction detection dataset, 100DOH.

75866 Train 8547 Test (instance num)	Self-Contact (1521)	Another Person (163)	Portable Object (11521)	Stationary Object (830)	mAP (14035)
AP	43.56	19.64	65.80	19.86	37.22
max Recall	71.33	61.35	80.61	74.82	72.03



The distribution of face categories in the HowToEat face dataset. [1] and [2] represent [eating] and [not eating].

Qualitative Results

- Our SOV-STG-H2E model can effectively detect hand-object pair bounding boxes and categories, while relatively reliably recognizing and classifying facial regions.
- This aims to advance dietary behavior research and contribute to broader applications.