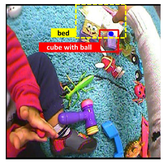
A Self-Validation Network for Object-Level Human Attention Estimation

Zehua Zhang, Chen Yu, David Crandall
Indiana University Bloomington
http://vision.soic.indiana.edu/mindreader/

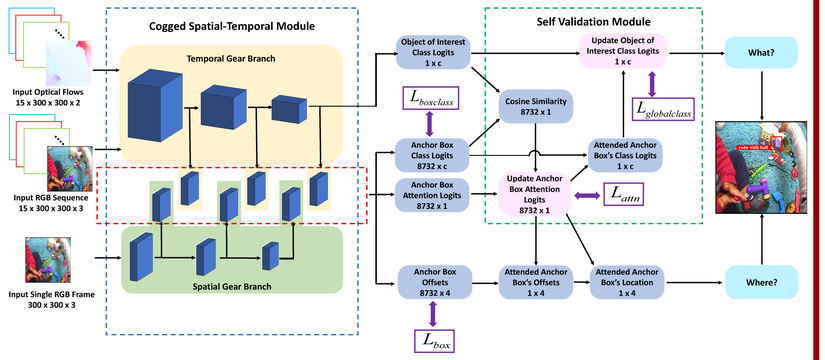
Introduction



- Due to the foveated nature of the human vision system, people can focus their visual attention on a small region of their visual field at a time.
- Our goal is to jointly identify and locate the attended object in first-person videos.
- Combining traditional eye gaze estimators and off-the-shelf object detectors fails because they solve the **where** and **what** problems separately.
- Directly performing anchor-level attention estimation by modifying existing object detection models don't give satisfying results.
- Various applications:
  - ❑ Applied in VR/AR devices, driver assistance systems
  - ❑ Auxiliary information for other computer vision tasks

Model

- We propose a novel unified model to enforce and leverage the visual consistency of our foveated visual system.



- Double self validation:
  - ❑ What→Where

$$A' = R(A) + V_{attn} = \frac{A - (max(A) + min(A))/2}{max(A) - (max(A) + min(A))/2} + V_{attn}, \quad = \frac{C_{global} C_{box_i}^T}{||C_{global}|| \times ||C_{box_i}||}.$$

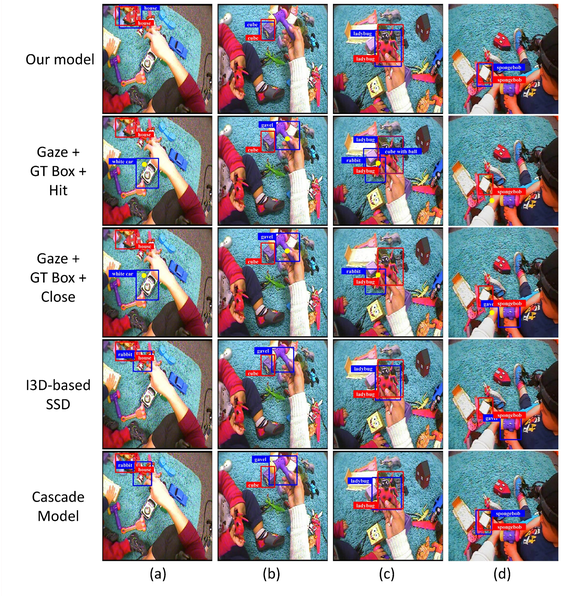
  - ❑ Where→What

$$V_{class} = C_{box_m} \quad C'_{global} = R(C_{global}) + R(V_{class}) \times$$

$$\hat{V}_{class} = \sum_{i=1}^{a} \widetilde{A}'_i C_{box_i} \quad \widetilde{A}'_i = \frac{e^{A'_i}}{\sum_{j=1}^{a} e^{A'_j}} \quad C'_{global} = R(C_{global}) + R(\hat{V}_{class}) \checkmark$$

Experiments

- Evaluate on ATT dataset and Epic-Kitchens dataset
- Report Acc_0.5, Acc_0.75 and mAcc

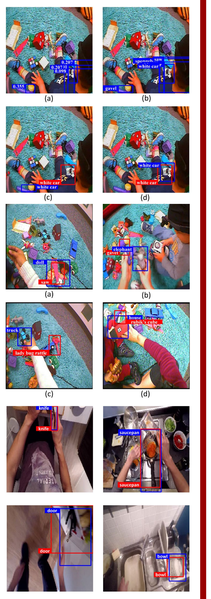


Method	$Acc_{0.5}$ ↑	$Acc_{0.75}$ ↑	$mAcc$ ↑
Our Mr. Net	**74.27**	**46.78**	**44.78**
Gaze [68] + GT Box + Hit	25.26	25.26	25.26
Gaze [68] + GT Box + Closest	35.86	35.86	35.86
I3D [9]-based SSD [40]	**70.11**	42.10	40.85
Cascade Model	66.97	**45.10**	41.93
OIH Detectors + WH Classifier	37.16	37.16	37.16
Left Handed Model	38.31	38.31	38.31
Right Handed Model	39.00	39.00	39.00
OIH GT + WH Classifier	40.83	40.83	40.83
Either Handed Model	42.94	42.94	**42.94**
Center GT Box	23.97	23.97	23.97

Table 1: *Accuracy of our method compared to others*, on the ATT dataset. OIH represents Object-in-Hand, while WH means Which-Hand.

Streams	Self validation? Training	Testing	$Acc_{0.5}$ ↑	$Acc_{0.75}$ ↑	$mAcc$ ↑
Two	yes	yes	**74.27**	**46.78**	**44.78**
Two	yes	half	—	—	43.88
Two	yes	no	68.19	42.83	41.18
Two	no	yes	67.18	40.06	39.48
Two	no	half	—	—	37.87
Two	no	no	62.33	38.31	37.18
RGB	yes	yes	74.59	43.15	42.48
Flow	yes	yes	64.30	38.63	37.60
Flow	no	yes	—	—	25.10
Flow	no	no	—	—	18.40

Table 2: *Ablation results.* Testing with half means that the model is tested with only what→where validation.



Model	$Acc_{0.5}$ ↑	$Acc_{0.75}$ ↑	$mAcc$ ↑
Mr. Net	**71.34**	**38.26**	**39.04**
Gaze [68] + GT Boxes Hit	26.46	26.46	26.46
Gaze [68] + GT Boxes Closest	36.81	36.81	36.81
I3D [9]-based SSD [40]	67.43	37.90	37.22
Cascade Model	65.96	38.01	37.93

Table 3: *Results of online detection.*

Method	$Acc_{0.5}$ ↑	$Acc_{0.75}$ ↑	$mAcc$ ↑
Our Mr. Net	**57.18**	**31.00**	**31.20**
I3D [9]-based SSD [40]	47.58	24.38	25.42
Cascade Model	51.20	28.18	28.36

Table 4: *Accuracies on the Epic-Kitchen dataset.*

INDIANA UNIVERSITY
SCHOOL OF INFORMATICS, COMPUTING, AND ENGINEERING