RUTGERS
WINLAB | Wireless Information Network Laboratory

EgoUse

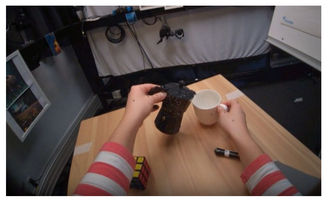
Learning Object use from egocentric video

Daniel Heller
Advisors: Zihao Ding, Yao Liu

Scan for Wiki Page

Overview

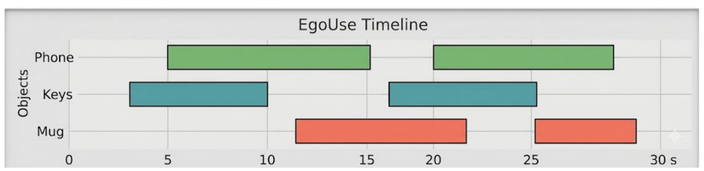
Problem



- **Egocentric video** is first-person video from a head-mounted camera.
- Detecting object use is difficult because of **motion, occlusion, and blur**.

Goal

- Produce a per-object timeline showing when the wearer is using each object.



Colored blocks indicate object-use intervals.

Data

- **HOT3D**: Project Aria recordings with 3D hand and object annotations.
- Used to build privileged teacher labels for training and evaluation.

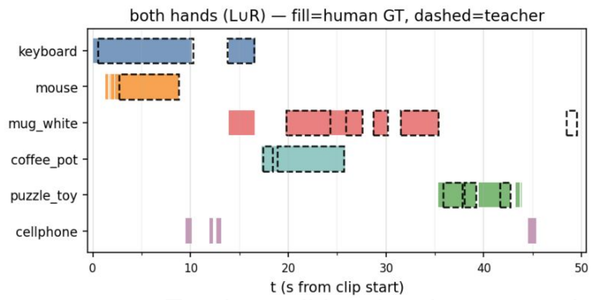
Approaches

Project path

1. 2D heuristics were fragile in egocentric video.
   ↓
2. 3D motion rules produced teacher labels.
   ↓
3. A GRU student learned from detector-derived features.

Teacher

- Generate pseudo-label interaction timelines from **HOT3D 3D hand and object tracks**.
- Interaction is defined using two rules for each hand–object pair:
  - **Co-motion**: hand and object move together in 3D
  - **Stationary interaction**: hand stays close with low relative motion
- These privileged 3D labels are used only to **supervise and evaluate** the student.

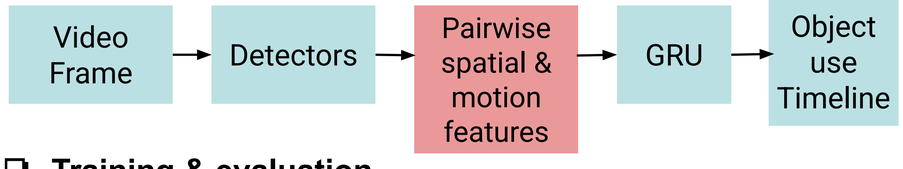


- Dashed = teacher
- Solid = human
- **Frame Precision:** 0.932
- **Frame F1:** 0.798
- **Event F1:** 0.667

Teacher validated on human-made segments

Student

- The student predicts object-use timelines from **detector-derived 2D spatial and motion features**.
- Inputs come from:
  - **100DOH** hand detections
  - **YOLOE** object detections
- For each hand–object pair, features include overlap, distance, confidence, and short motion history.
- A **GRU** predicts whether the object is in use over time.



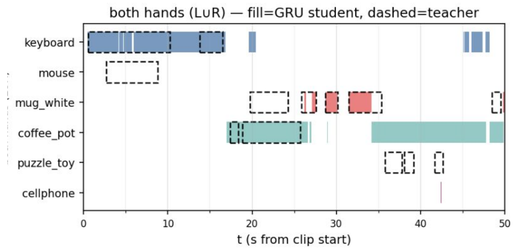
- **Training & evaluation**
  - **122 training sequences,10 validation sequences**
  - Metrics: **Frame f1, Event F1** (segment overlap)

Results

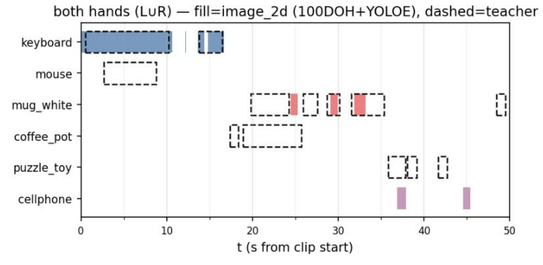
Method	Frame F1	Event F1 (tIoU ≥ 0.5)
2D baseline	0.146	0.061
GRU student	0.161	0.061

* Agreement with frozen teacher labels on the held-out proof set.

- On held-out clips, the GRU slightly improves Frame F1, while Event F1 remains low.



GRU vs teacher on held-out proof clip



2D baseline vs teacher on the same clip

Main findings

- The GRU shows **modest frame-level improvement** over the 2D baseline on held-out sequences.
- **Event-level quality** remains weak, especially for temporal boundaries and generalization.

Limitations & Future Work

- Limited to **one dataset/device**; detector quality affects results.
- Improve **generalization and event timing** before live deployment.

WINLAB