You-Do, I-Learn: Discovering Task Relevant Objects and their Modes of Interaction from Multi-User Egocentric Video

Dima Damen
Dima.Damen@bristol.ac.uk

Teesid Leelasawassuk
Csztl@bristol.ac.uk

Osian Haines
Osian.Haines@bristol.ac.uk

Andrew Calway
Andrew.Calway@bristol.ac.uk

Walterio Mayol-Cuevas
Walterio.Mayol-Cuevas@bristol.ac.uk

Computer Science Department
University of Bristol
Bristol, UK

We present a **fully unsupervised** approach for the discovery of i) task relevant objects and ii) how these objects have been used. A **Task Relevant Object (TRO)** is an object, or part of an object, with which a person interacts during task performance. Given egocentric video from multiple operators, the approach can discover objects with which the users interact, both static objects such as a coffee machine as well as movable ones such as a cup. Importantly, we also introduce the term **Mode of Interaction (MOI)** to refer to the different ways in which TROs are used. Say, a cup can be lifted, washed, or poured into. When harvesting interactions with the same object from multiple operators, common MOIs can be found.

Setup and Dataset: Using a wearable camera and gaze tracker (Mobile Eye-XG from ASL), egocentric video is collected of users performing tasks, along with their gaze in pixel coordinates. Six locations were chosen: kitchen, workspace, laser printer, corridor with a locked door, cardiac gym and weight-lifting machine. The Bristol Egocentric Object Interactions Dataset is publically available [1].

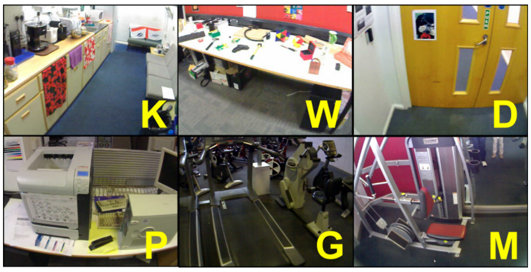

Figure 1: An overview of the locations in the dataset.

Discovering TROs: Given a sequence of images $\{I_1,..,I_T\}$ collected from multiple operators around a common environment, we aim to extract K TROs, where each object TRO_k is represented by the images from the sequence that feature the object of interest . We investigate using appearance, position and attention, and present results using each and a combination of relevant features. For attention, we exploit the high quality and predictive nature of eye gaze fixations.

Results compare k-means clustering to spectral clustering, and propose estimating the optimal number of clusters using the standard Davies-Bouldin (DB) index. Figure 2 shows the best performance for discovering TROs by combining position (relative to a map of the scene) and appearance (HOG features within BoW) over a sliding window $w = 25$, using gaze fixations for attention, spectral clustering and estimating the number of clusters using the Davies-Bouldin (DB) index.

Finding MOIs: Given consecutive images $(I_t, I_{t+1}, I_{t+\rho})$ clustered into the same TRO, a video snippet u_i^k for TRO k is defined as

$$u_i^k = \{\Psi(I_j, \Delta(j), \omega); \quad I_j \in TRO_k; \quad j = t..t+\rho; \quad \rho \geq \xi\} \qquad (1)$$

where Ψ crops a window of size ω from image I_j around $\Delta(j)$, and $\Delta(j)$ is the interpolated gaze at frame j as gaze information is missing in some frames. The collection of all video snippets $U_k = \{u_i^k\}$ shows different ways in which TRO_k was used.

On average, 16.6 video snippets are extracted for each TRO ($\sigma = 7.4$). We cluster U_j, and represent each cluster by the video snippet $\hat{u_j}$ closest to the centre of the cluster μ_j (i.e. mean snippet), as well as the percentage of snippets within that cluster $p(MOI_j)$. We vary the threshold λ to accept $p(MOI_j)$ to produce recall-precision curves. Figure 3 shows an example of the method successfully discovering two MOIs for the 'socket'.

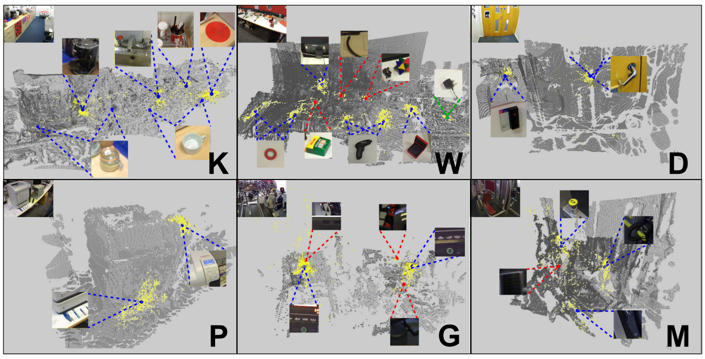

Figure 2: Discovered TROs. An overview of the locations is shown at the top. Blue dots represent true-positive (19 objs), red dots represent false positive (7 objs) and green dots represent false negative (1 obj).

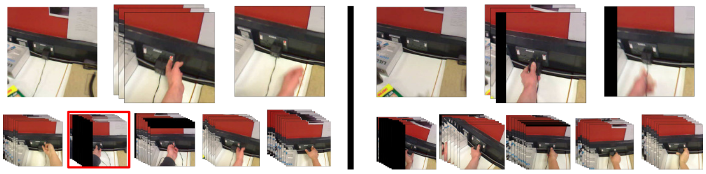

Figure 3: For the 'socket', the two common MOIs ('switching', 'plugging') are found (left & right). The representative *video snippet* is shown (up) with the other snippets in the same cluster (below) - only one snippet is incorrectly clustered (shown in red).

Video Guides: In addition, the approach enables the automatic generation of *help snippets* on how objects have been used before. We showcase video help guides using inserts on a pre-recorded video. A suitable video insert (i.e. MOI snippet) is chosen every time a gazed-at object is first recognised. In this assistive mode, we use the real-time texture-minimal scalable detector [2] due to its light-weight computational load that makes it amendable to wearable systems. Figure 4 shows frames from the help videos and a full sequence is available [3]. Recall that these inserts are *extracted, selected and displayed* fully automatically.

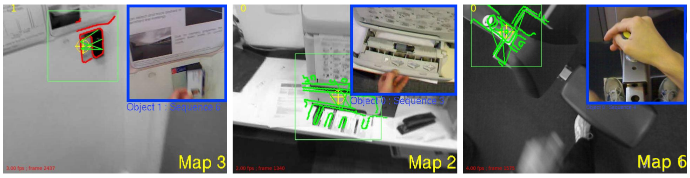

Figure 4: In the assistive mode, when a TRO is detected, video snippet is inserted showing the most relevant common MOI based on the object's current appearance.

[1] http://www.cs.bris.ac.uk/~damen/BEOID/

[2] http://www.cs.bris.ac.uk/~damen/MultiObjDetector.htm

[3] http://www.cs.bris.ac.uk/~damen/You-Do-I-Learn