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

Abstract

We present a fully unsupervised approach for the discovery of i) task relevant objects and ii) how these objects have been used. 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, the common modes of interaction for discovered objects are also found. We investigate using appearance, position, motion and attention, and present results using each and a combination of relevant features. Results show that the method is capable of discovering 95% of task relevant objects on a variety of daily tasks such as initialising a printer, preparing a coffee and setting up a gym machine. In addition, the approach enables the automatic generation of guidance video on how these objects have been used before.

1 Introduction

Humans learn how to deal with their surroundings through several means, one of which is observing others. An intelligent agent that aims to learn objects in an environment, and more importantly how these objects have been used, is of importance in robotics and assistive systems. This is intrinsically distinct from learning a specific task or an activity, as the same object can be used in many tasks, while the ways in which one object can be interacted with are usually limited to a finite set of possible interactions.

This work attempts, to **fully unsupervised**, discover objects and how they have been used from observing several operators. As opposed to discovering all objects in the environment, we focus on discovering task relevant objects. A **Task Relevant Object (TRO)** is an object, or part of an object, with which a person interacts during task performance. For example, a

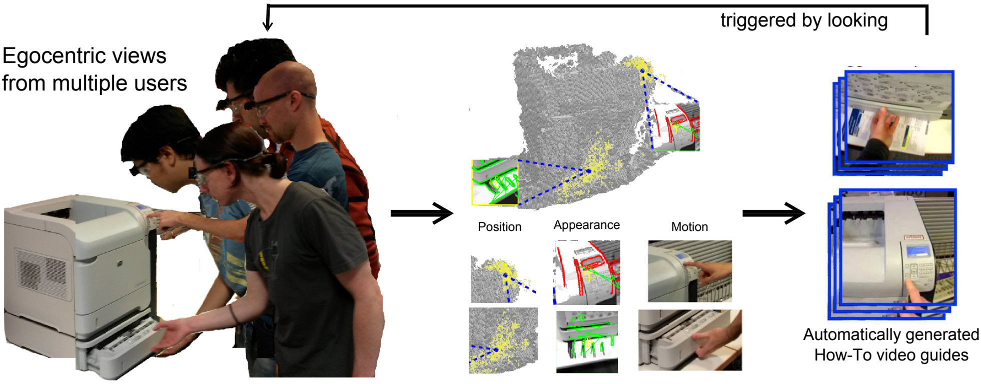

triggered by looking
Egocentric views from multiple users
Position
Appearance
Motion
Automatically generated How-To video guides

Figure 1: Given egocentric videos from multiple users, three strands of information (position, appearance and motion) are used to discover task relevant objects **fully unsupervised**. Harvested video snippets show how a discovered object is used for an assistive mode.

person using a printer may interact with the paper drawer and/or the keypad while operating it. A system that aims to discover TROs would attempt to discover these objects (drawer and keypad) as opposed to the full machine or all its parts (Fig. 1). Importantly, we also introduce the term **Modes of Interaction (MOI)** to refer to the different ways in which TROs are used. Say, a cup can be lifted, washed, or poured into. All these are different MOIs associated to the cup. When harvesting interactions with the same object from multiple operators, common MOIs can be discovered.

In attempting the discovery of TROs and their MOIs, here we particularly focus on an egocentric view of the world. This first-person view of the world offers a unique perspective on object-level interactions. With the introduction of wearable systems, egocentric videos from multiple individuals around a common environment will become easily and widely available. But crucially we note that most available systems to analyse object-interactions from first-person view [5, 11, 23, 26, 28, 29, 34] expect the objects involved to be known in advance. The better way is to extract relevant information automatically. Recently, a shift towards automatic discovery of objects for video summaries [25] or action recognition [9] has been witnessed, to which this work contributes. In this work, egocentric videos of multiple operators performing tasks around a common environment are recorded, along with the operators' gaze. We provide a fully unsupervised method to discover TROs, and identify common MOIs for each object. Additionally and importantly, we are able to harvest suitable video snippets representative of each MOI. These can be used, for example, to provide guidance to other users exploring the environment. Fully closing the cycle of identifying what objects are relevant and how these can be used has not been attempted before, to our knowledge, in an unsupervised manner.

2 Unsupervised Object Discovery - a Review

Unsupervised object discovery refers to grouping visual information into meaningful clusters that correspond to an entity worth discovering. We attempt to differentiate between the various ways in which entities can be discovered from egocentric video; appearance, position and motion. Figure 2 envisages what can be discovered if each, or a combination, of these information is used in the grouping. Previous works on unsupervised object discovery are reviewed here based on the information they use.

Appearance: Appearance, of object and context, are often used to discover *categories* - multiple instances are grouped based on visual similarity (ref. recent survey [38]). Applied mostly to datasets of images, the aim is to group images (e.g. [16]) or segments of images (e.g. [30]) into object categories such as cars or giraffes. In [16], for example, Harris-affine interest points with SIFT descriptors are used to construct a visual similarity network. Edges between features are constrained by the geometric consistency of matching pairs of images, and weights are calculated to minimise the deformation cost in image matching. The network is used to find image groupings based on nodes' structural similarity and a PageRank algorithm. Similar approaches were applied to instance discovery [15] - colour, texture and shape-based features are used to construct a network of finely-segmented regions. While a very interesting approach with promising results, [15] assumes that objects of daily living are moveable. A computer screen, for example, needs to be moved to a different background to enable its discovery. This assumption is also made by other works [32, 33]. Many objects of daily living tasks such as a coffee machine or an electric socket remain fixed to their surroundings. Moreover, all these approaches [15, 16, 38] assume the dataset contains a single instance of an object of interest per image. When using video as input, a significant number of frames might not contain TROs as the user roams around an environment.

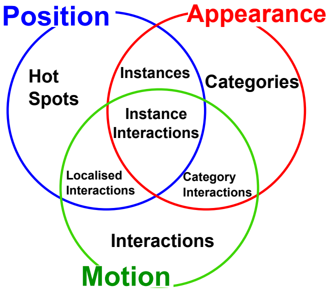


Figure 2: Using appearance, position, motion and combinations for object discovery

Position: The position, relative to an environment, can be grouped into hot-spots. A *hot spot* is a position at which object interaction takes place. It can refer to a fixed object in the scene such as a kitchen sink, or a temporary position of a moveable object. Position has been used in [13] to discover objects, by aligning two point clouds and identifying changes that correspond to objects that have been placed or removed.

Motion: Motion in egocentric video is a result of the wearer's self-motion or that of objects in an environment. Motion features can be grouped into actions, such as putting, drinking or stirring. The bag of quantised features approach for sparse spatio-temporal interest points [22] or dense features [39] has produced state-of-the-art results in action recognition. In egocentric video, motion descriptors have also been used to recognise actions, either full-body action (such as in sports [17]) or object interactions [9, 23, 25, 35, 36].

Combinations: Using multiple cues has been recently attempted to improve recognition results. When combining *appearance with position*, one can separate two instances of a mug when viewed in different locations. In Collet *et al.* [2], RGB-D images collected from a robot in a common environment were first separated into discrete locations (rooms, in their case), then appearance and depth data are clustered to extract instances. The approach assumes that all objects are placed on a planar surface (e.g. table-top) and employs a prior on the object's shape and size. Combining *appearance with motion* has also been attempted [9, 40]. In [9], an action is identified by the change in appearance of the object before and after the action is performed. In [23], objects of 'importance' are segmented from egocentric video sequences using appearance and motion features. The approach learns 'objects of importance' from a manually labelled training set (collected via crowd sourcing).

Attention: Attention is an important aspect to support relevance in tasks. Common approaches in egocentric vision include i) segmenting the area surrounding the user's hand [9, 10, 23], ii) extracting foreground regions through frame stabilisation or scene planarity assumptions [29, 36] or iii) detecting 'object-like' regions [25]. The first two approaches are only able to segment objects being manipulated, during which objects could be heavily occluded by the hand. In the second approach, fixed objects like a sink tap or a coffee machine, which can be quite crucial to a task, are ignored. In the third approach, 'object-like' regions can focus on salient rather than used objects. Very few systems exploit the high quality and predictive nature of eye gaze fixation. Its anticipatory nature allows estimating which object will be used next [20, 21]. Gaze has been successfully used in to assist action recognition [11, 24] or supervised object recognition [8, 34].

Contributions - Our approach

I. **Discovers Task Relevant Objects from Multiple Users** - using position and appearance, along with attention (particularly gaze fixations).

II. **Finds Common Modes of Interaction** - as the object is used multiple times by the same or multiple operators, the system automatically harvests various ways of usage, and identifies the most frequent ones.

III. **Provides Video Help Guides** - we showcase automatically extracted video guides; a suitable video insert is triggered when a gazed-at object is recognised, to illustrate how the object was used before.

Up to our knowledge, *all the three tasks above have not been achieved before in a fully unsupervised manner*. The approach is tested on 6 different tasks involving 20 task relevant objects and 3-5 operators.

3 The Method

Using a wearable camera and gaze tracker [19], egocentric video is collected of users performing tasks, along with their gaze in pixel coordinates. There are two principal eye behaviours: fast motion transitions (aka saccades) and eye fixations. Importantly, studies of eye fixations during everyday tasks show substantial similarities in the locations and number of fixations by different operators, that gaze rarely visits irrelevant objects and that fixations precede actions [12, 14, 20]. To filter saccades, we follow the velocity-based approach [31], where the average angular velocity v_t over a sliding temporal window is calculated

$$v_t = \frac{1}{N}\sum_i \theta(g_i, g_{i-1}) \tag{1}$$

The function θ calculates the angular velocity between two consecutive gaze rays g_i and g_{i-1}, and N is the number of samples within the temporal window. As in [31], When $v_t \geq 100°/sec$, the gaze sample is classified as a saccade is thus discarded.

3.1 Discovering Task Relevant Objects (TRO)

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 . As we are targeting instances rather than categories, Fig. 2 suggests combining position and appearance information.
Position: The Image I_t is positioned relative to the scene using sparse Simultaneous Localisation and Mapping (SLAM) [18]. Given the 6D pose of the scene camera, a 3D ray links either the gaze point or the centre of the image[1] to a point in the scene. A dense depth map is estimated, using a triangular tessellation on the tracked interest points that are visible to the scene camera (similar to [37]). A 3D point in space f_t is thus calculated from the triangle $\triangle Y_0 Y_1 Y_2$ of tracked interest points that includes the intersection point.

$$f_t = Y_0 + u(Y_1 - Y_0) + v(Y_2 - Y_0) \tag{2}$$

where u and v are projections of $\overrightarrow{Y_0 f_t}$ onto $\overrightarrow{Y_0 Y_1}$ and $\overrightarrow{Y_0 Y_2}$, respectively.
Appearance: To represent appearance, images are cropped around the gaze point or the centre of the image to a window of size ω. The Histogram of Oriented Gradients (HOG) [3], calculated on patches of size p, is chosen as it achieves reasonable results for both highly-textured and minimal-texture objects - the latter being frequent in the dataset used. Bag of Words (BoW) is then used to represent each image. When combining position and appearance, the normalised affinity matrices are summed with equal weighting. We also compare to results that accumulate features over a sliding window w centred around each image $(I_{t-\frac{w-1}{2}}, .., I_t, .., I_{t+\frac{w-1}{2}})$.

For discovery, we compare k-means clustering to spectral clustering from Ng *et al.* [27]. Unsupervised discovery, like other grouping problems, suffers from the dilemma of model selection (i.e. the optimal number of groups). Most previous approaches assume the number of groupings is known apriori [16, 38] to avoid the complexity. We propose estimating the optimal number of clusters $\hat{\kappa}$ using the standard Davies-Bouldin (DB) index [7]. For a cluster TRO_i with data points $\{x_j; j = 1..n_i\}$ assigned to this cluster, and μ_i is the mean of these data points, the intra-cluster distance S_i can be measured as (Euclidean distance used):

$$S_i = \sqrt{\frac{1}{n_i} \sum_{j=1}^{n_i} ||X_j - \mu_i||_2} \tag{3}$$

The inter-cluster distance between two clusters TRO_i and TRO_j is measured as $M_{ij} = ||\mu_i - \mu_j||_2$. The cluster similarity measure $R_{ij} = \frac{S_i + S_j}{M_{ij}}$ is used to calculate the DB index,

$$V_{DB}(\kappa) = \frac{1}{\kappa} \sum_{i=1}^{\kappa} \max_{j \neq i} R_{ij} \tag{4}$$

The optimal number of clusters is calculated to be $\hat{\kappa} = \arg\max_{\kappa} V_{DB}(\kappa)$.

3.2 Finding Modes of Interaction (MOI)

For discovered TROs, we then aim to find common MOIs for each discovered object by clustering **video snippets**, each representing one usage of a TRO. Given consecutive images $(I_t, I_{t+1}, I_{t+\rho})$ clustered into the same TRO[2], a video snippet u_i^k for TRO k is defined as

$$u_i^k = \{\Psi(I_j, \Delta(j), \omega); \quad I_j \in TRO_k; \quad j = t..t+\rho; \quad \rho \geq \xi\} \tag{5}$$

[1] the gaze point is used in experiments where attention is considered, otherwise the centre of the image is used
[2] the sequence may also include images that have not been clustered due to the lack of gaze information

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.

Position and appearance information of all frames in u_i (superscript k removed for simplicity) are the same features used for discovering objects. These are augmented with motion information collected using the Histogram of Optical Flow (HOF) descriptors around 3D Harris points [22].

In addition to the BoW representation, we also use a *temporal pyramid* to encode the descriptors. At each level $l = \{1..L\}$, the snippet is split into l equally-sized temporal segments, and the descriptor is calculated for each segment. The temporal pyramid could potentially separate MOIs that differ in their temporal ordering, such as opening and closing. A one-dimensional representation of the temporal pyramid formulates the descriptor $d(u_i)$. Clustering then follows (as in 3.1) to find the MOIs.

Each cluster is represented 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 $p(MOI_j)$.

$$\hat{u_j} = \arg\min_{u_l \in MOI_j} ||d(u_l) - \mu_j||; \quad \mu_j = \frac{1}{|MOI_j|} \sum_{u_l \in MOI_j} d(u_l); \quad p(MOI_j) = \frac{|MOI_j|}{|U_k|} \tag{6}$$

A threshold λ can be used to select common MOIs such that $p(MOI_j) \geq \lambda$.

3.3 Providing Video Help Guides

We present a possible application for unsupervised discovery of TROs and their MOIs. In the assistive mode, when a discovered TRO is recognised, a *help snippet* is displayed to show how this object was previously used. Notice that the assistive mode does not require tracking of the camera relative to an environment, and objects are recognised within a 2D patch around the gaze point. From the possibly many MOIs, we choose the *help snippet* h_t such as,

$$h_t = \arg\min_{u_j} ||A^{1st}(u_j) - A^{1st}(\Psi(I_t, f_t, \omega))|| \tag{7}$$

where A^{1st} is the appearance of the first frame in the snippet, and Ψ is the cropped image as in Eq. 5. If the object changes state, the initial appearance is a good indicator of which video snippet to show.

4 Experiments and Results

Setup & Dataset The wearable gaze tracker hardware (ASL Mobile Eye XG [19]) consists of two cameras, one looking at the scene and another looking at the eye. After calibration, the scene images are synchronised with, if available, 2D gaze points. Six locations were chosen: kitchen (K), workspace (W), laser printer (P), corridor with a locked door (D), cardiac gym (G) and weight-lifting machine (M) (Fig. 3). For the first four locations (K, W, P, D), sequences from five different operators were recorded, and from three operators for the last two locations (G, M)[3]. Following the gaze tracker calibration, the operator moved freely between the locations performing verbally-communicated tasks (Tab. 1). Two sequences were recorded for each operator.

[3] Dataset available at: http://www.cs.bris.ac.uk/~damen/BEOID

	Number of sequences	**Sequence length**		**Tracked (%)**		**Gaze Fixations (%)**	
		μ	σ	μ	σ	μ	σ
K	10	1905	386	69.4	9.1	58.9	11.1
	Prepare coffee using the machine, place the cup on the mat and add sugar [tap, coffee machine, heat mat, cutlery drainer], (cup, sugar jar)						
W	10	1221	194	78.3	12.4	61.9	18.1
	Plug the screwdriver for charging and place the tape in the red box [Socket, Box], (screwdriver, charger, tape)						
P	10	596	77	75.8	13.3	70.5	14.1
	Check the printer is loaded with paper manually and using the keypad [drawer, keypad]						
D	10	303	83	71.8	15.8	56.2	14.7
	Go through the locked door [door lock, door handle]						
G	6	5183	482	76.4	9.0	66.7	11.0
	Use the treadmill and the bicycle next to it [treadmill panel, bicycle panel]						
M	6	2059	624	24.5	16.2	14.6	15.2
	Adjust the seat, chest pad and weight then use the machine [seat adjuster, pad adjuster, weight adjuster]						

Table 1: For the six locations, the number of sequences, average number of frames, percentage of tracked frames, percentage of gaze fixations, as well as the verbally communicated tasks, fixed "[]" and movable "()" ground-truth TROs.

The operators were then asked to watch the videos, and write down a narration of what they have performed. Narrations were stemmed manually to unify nouns and verbs which are semantically identical (e.g. adapter vs. charger, pick vs. retrieve). Nouns narrated by more than 50% of the operators represent the twenty ground-truth TROs. Narrated verb-noun combinations are labelled as MOIs. Objects varied between having a single MOI (e.g. door handle: open) and up to three different usage methods (e.g. sugar jar: pick, put, get sugar). For each location, a map is built using Parallel Tracking and Mapping (PTAM) [18]. A 3D bounding box around each object is manually labelled for evaluation. For moveable objects, their different locations are ground-truthed.

Fixed Parameters The temporal sliding window for discarding saccades N (Eq. 1) was set to 9 frames. On average, 3D fixations were found at 66% of the sequences' frames (80% of localised frames) (Tab. 1). The appearance and motion descriptors are calculated in a window of size $\omega = 200 \times 200$, divided into 10×10 non-overlapping patches for calculating HOG descriptors. This corresponds to 19.3° visual angles in the scene camera. The number of words in BoW representation is set to 200. In calculating the BD index, $\kappa = [2..2No_{gt}]$ (Eq. 4) where No_{gt} is the number of ground-truth objects.

Results for discovering TROs To calculate precision and recall, the smallest bounding box encompassing 75% of the points in each cluster is computed, to avoid outliers. This is compared to the ground-truth bounding boxes, and the PASCAL overlap criteria (in 3D) of

w	clustering		Davies-Bouldin (DB) index						Known Number of Objects (Known K)					
			Without Attention			With Attention			Without Attention			With Attention		
			app	pos	both	app	pos	both	app	pos	both	app	pos	both
1	k-means	Recall	35.0	40.0	40.0	55.0	65.0	**65.0**	50.5	55.0	60.0	55.0	**80.0**	**80.0**
		Precision	50.0	40.0	44.4	40.7	59.1	**61.9**	52.6	61.1	66.7	61.1	**84.2**	**84.2**
	Spectral	Recall	50.0	65.0	60.0	65.0	85.0	**90.0**	45.0	60.0	50.0	60.0	80.0	**90.0**
		Precision	41.7	54.2	52.2	41.9	68.0	**75.0**	47.4	66.7	58.8	60.0	80.8	**90.0**
25	k-means	Recall	60.0	40.0	45.0	60.0	65.0	**70.0**	50.0	60.0	55.0	60.0	**85.0**	**85.0**
		Precision	44.4	42.1	52.9	42.9	59.1	**63.6**	52.6	70.6	64.7	60.0	**89.5**	**89.5**
	Spectral	Recall	70.0	75.0	60.0	70.0	80.0	**95.0**	50.0	60.0	55.0	70.0	90.0	**90.0**
		Precision	45.2	51.7	50.0	48.3	59.3	**73.0**	55.6	66.7	57.9	73.7	90.0	**94.7**

Table 2: Recall and precision results for discovering TROs using different features, clustering methods, with/without attention and sliding window.

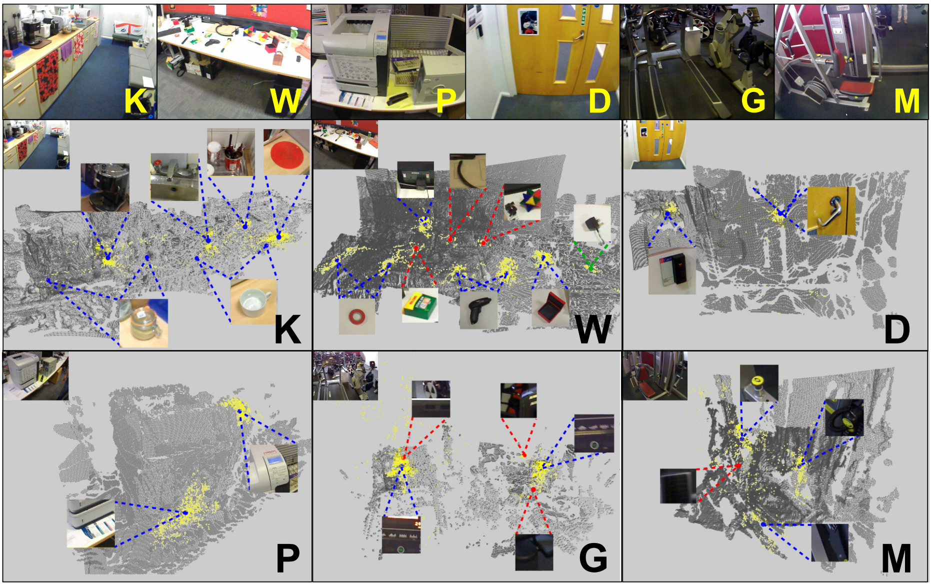
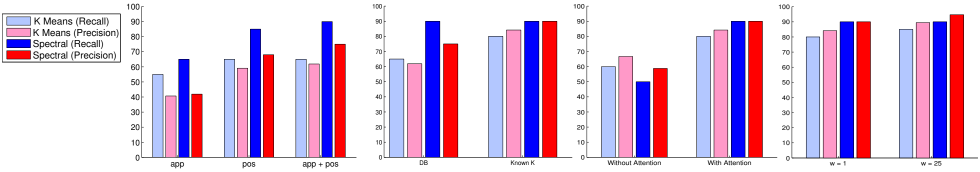


Figure 3: Discovered TROs (appearance, position, attention, spectral clustering, $w = 25$ and DB index (i.e. number of objects is unknown)). An overview of the locations is shown at the top. Blue dots represent true-positive (19 objs), red dots represent false positive (7 objs) and green dots represent false negative (1 obj).

20% indicates a true-positive. This is because the viewed positions don't typically cover the full extent of the object. Table 2 shows the complete set of results for discovering TROs. Two clustering methods are compared - spectral clustering and k-means. Appearance and position features are used individually or combined, either for a single frame ($w = 1$) or a sliding window ($w = 25$). The importance of gaze fixations as an attention mechanism is compared - results 'without attention' consider the centre of the image instead. Estimating the number of clusters using the Davies-Bouldin (DB) index is compared to knowing the number of clusters apriori (ref. *Known K*).

Table 2 shows that the best results are obtained using spectral clustering, combining appearance and position, with attention and over a sliding window. Using Davies-Bouldin (DB) index, 95% of the TROs were retrieved with 73% precision. These discovered TROs are shown in Fig. 3. If the number of clusters was known apriori 90% of TROs would be discovered with 94% precision. This is because the optimal number of clusters using DB index was higher than ground-truth K, resulting in one more correct object and several false positive clusters.

Fig. 4 highlights several conclusions from the results: (a) shows that for [DB, attention, $w = 1$] position achieves better than appearance when used solely. This is because most of the objects in our dataset (15/20) are fixed objects. As expected, adding appearance information increases the precision as this clusters instances of moveable objects into a single cluster. Fig. 4 (b) shows that DB index achieves the same recall as Known K when using spectral clustering [app+pos, attention, $w = 1$]. Precision increases when K is known - i.e. smaller discarded clusters actually do not represent TROs. Fig. 4 (c) shows the importance of within-image attention [app+pos, Known K, $w = 1$]. A significant drop in recall is observed when



(a) Appearance vs. Position (b) Known K vs. BDI (c) without vs. with attention (d) single frame vs. sliding window

Figure 4: **(a)** appearance (app) vs position (pos) and their combination (app+pos) using spectral vs. k-means clustering using DB index. **(b)** Using app+pos, DB index vs. known number of clusters. **(c)** For app+pos and known K, patches around centre of image vs. gaze fixations. **(d)** Single-frame vs. sliding window representations.

the information is gathered around the image centre rather than gaze fixations. Fig. 4 (d) shows that a sliding window gives a slight improvement in performance.

Results for discovering MOIs For each discovered object, the video snippets longer than $\xi = 1s$ (Eq. 5) are used to discover MOIs. On average, 16.6 video snippets are extracted for each TRO ($\sigma = 7.4$). We vary the threshold λ to accept $p(MOI_j)$ (Eq. 6) to produce recall-precision curves. A cluster is true-positive if its representative snippet matches one ground-truth MOI; a duplicate match for the same ground-truth MOI is a false-positive. We compare using position, appearance and motion features with a temporal pyramid (Fig. 6). We then compare the features at their best temporal pyramid level, as well as their combination (Fig. 7). Using the combination of features and $\lambda = 0.2$, the approach is able to discover meaningful MOIs. Figure 8 shows an example of the method successfully discovering two MOIs for the 'socket'. Similarly, Fig. 5 shows further discovered MOIs for the sugar jar and the door handle.

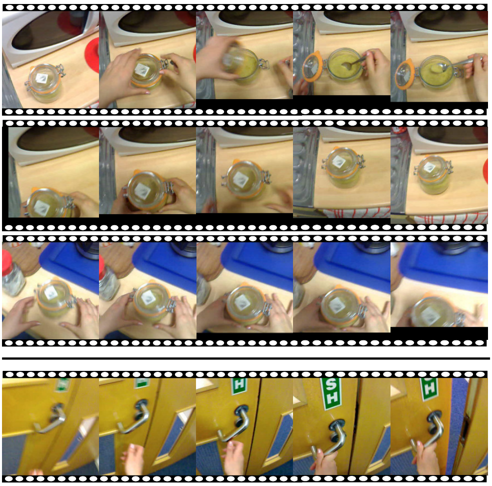


Figure 5: For TRO 'jar', 3 MOIs are discovered ('get sugar', 'put', 'pick'). For the handle, one MOI is discovered. Frames from the representative snippets are shown.

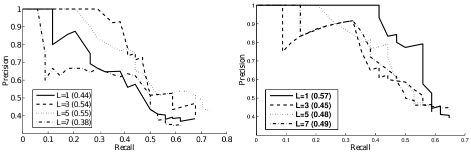


Figure 6: For position (left), temporal pyramid (L=5) performed best, while motion (right) performed best on L=1.

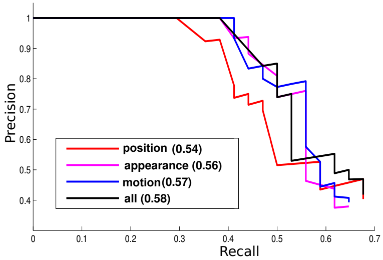


Figure 7: Motion features achieved the highest AUC (shown in brackets), with a slight improvement when features are combined.

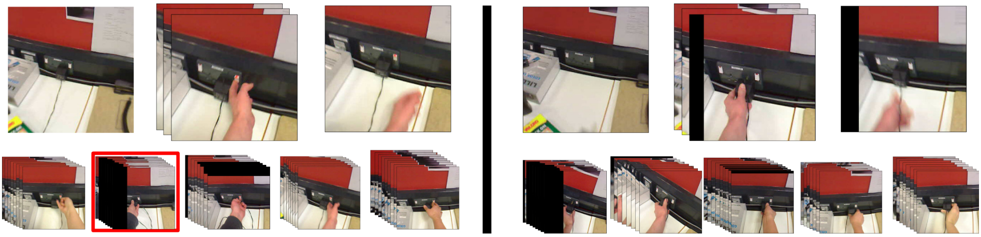


Figure 8: For the 'socket', the two common MOIs ('switching', 'plugging') are found (left & right). The representative *video snippet* is shown (top) with the other snippets in the same cluster (bottom) - only one snippet is incorrectly clustered (shown in red).

Discovering TROs could be attempted online, clustering and refining features during task performance as data becomes available. In [6], we propose an online system, using features suitable for real-time performance, to discover TROs and present results on the same dataset.

Video Help Guides To assess the ability of the approach to provide video guides, the method is run using leave-one-out. For every operator, TROs are discovered and common MOIs are found from sequences of other operators. In the assistive mode, when a discovered TRO is detected, an insert is shown indicating a suggestive way of how the object can be used. In this mode, we use the real-time texture-minimal scalable detector code from [4] due to its light-weight computational load that makes it amendable to wearable systems [1]. A *help snippet* is displayed each time a new object is recognised. We showcase video help guides using inserts on a pre-recorded video. These could in principle be shown on a head-mounted display, but is not considered in this study. Figure 9 shows frames from the help videos and a full sequence is available[4]. Recall that these inserts are *extracted, selected and displayed* fully automatically.

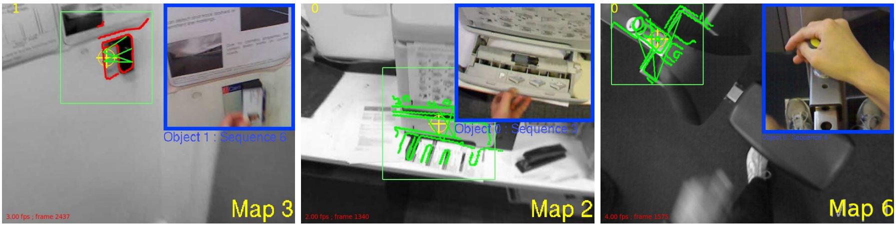


Figure 9: In the assistive mode, when a TRO is detected, video snippet is inserted showing the most relevant common MOI based on the initial appearance.

5 Conclusion and Future Work

In this work, we investigate discovering task relevant objects and their common modes of interaction from multi-user egocentric video, *fully automatically*. We compare appearance, position and motion features, along with gaze fixations to indicate attention, for the discovery. The method is able to produce high levels of precision and recall for task relevant objects as well as meaningful modes of interaction. Video guides on how objects have been used can also be automatically provided. We next aim to assess the usefulness of video guides for human operators, and compare gaze to other relevance cues.

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

References

[1] P Bunnun, D Damen, A Calway, and W Mayol-Cuevas. Integrating 3D object detection, modelling and tracking on a mobile phone. In *Int. Symp. on Mixed and Augmented Reality (ISMAR)*, 2012.

[2] A Collet, B Xiong, C Gurau, M Hebert, and S Srinivasa. Exploiting domain knowledge for object discovery. In *Int. Conf. on Robotics and Automation (ICRA)*, 2013.

[3] N Dalal and B Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition (CVPR)*, 2005.

[4] D Damen, P Bunnun, A Calway, and W Mayol-Cuevas. Real-time learning and detection of 3D texture-less objects: A scalable approach. In *British Machine Vision Conference (BMVC)*, 2012.

[5] D Damen, A Gee, W Mayol-Cuevas, and A Calway. Egocentric real-time workspace monitoring using an RGB-D camera. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2012.

[6] D Damen, O Haines, T Leelasawassuk, A Calway, and W Mayol-Cuevas. Multi-user egocentric online system for unsupervised assistance on object usage. In *ECCV Workshop on Assistive Computer Vision and Robotics (ACVR)*, 2014.

[7] D Davies and D Bouldin. A cluster separation measure. *Pattern Analysis and Machine Intelligence (PAMI)*, 1(2):224–227, 1979.

[8] S De Beugher, Y Ichiche, G Brone, and T Geodeme. Automatic analysis of eye-tracking data using object detection algorithms. In *Workshop on Perasive Eye Traking and Mobile Eye-based Interaction (PETMEI)*, 2012.

[9] A Fathi and J Rehg. Modeling actions through state changes. In *Computer Vision and Pattern Recognition (CVPR)*, 2013.

[10] A Fathi, X Ren, and J Rehg. Learning to recognise objects in egocentric activities. In *Computer Vision and Pattern Recognition (CVPR)*, 2011.

[11] A Fathi, Y Li, and J Rehg. Learning to recognize daily actions using gaze. In *European Conference on Computer Vision (ECCV)*, 2012.

[12] J Henderson. Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*, 7(11), 2003.

[13] E Herbst, P Henry, Ren X, and D Fox. Toward object discovery and modeling via 3-D scene comparison. In *Int. Conf. on Robotics and Automation (ICRA)*, 2011.

[14] M Just and P Carpenter. A theory of reading: from eye fixations to comprehension. *Psychological Review*, 87, 1980.

[15] H Kang, M Hebert, and T Kanade. Discovering object instances from scenes of daily living. In *Int. Conference on Computer Vision (ICCV)*, 2011.

[16] G Kim, C Faloutsos, and M Herbert. Unsupervised modeling of object categories using link analysis techniques. In *Computer Vision and Pattern Recognition (CVPR)*, 2008.

[17] K Kitani, T Okabe, Y Sato, and A Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In *Computer Vision and Pattern Recognition (CVPR)*, 2011.

[18] G Klein and D Murray. Parallel Tracking and Mapping for Small AR Workspaces. In *Int. Sym. on Mixed and Augmented Reality (ISMAR)*, 2007.

[19] Applied Science Laboratories. Mobile Eye-XG. URL http://www.asleyetracking.com/.

[20] M Land. Eye movements and the control of actions in everyday life. *Progress in Retinal and Eye Research*, 2006.

[21] M Land, N Mennie, and J Rusted. The roles of vision and eye movements in the control of activities of daily living. *Perception*, 28, 1999.

[22] I Laptev. On space-time interest points. *Int. Journal of Computer Vision (IJCV)*, 64, 2005.

[23] Y Lee, J Ghosh, and K Grauman. Discovering important people and objects for egocentric video summarization. In *Computer Vision and Pattern Recognition (CVPR)*, 2012.

[24] Y Li, A Fathi, and J Rehg. Learning to predict gaze in egocentric video. In *Int. Conf. on Computer Vision (ICCV)*, 2013.

[25] Z Lu and K Grauman. Story-driven summarization for egocentric video. In *Computer Vision and Pattern Recognition (CVPR)*, 2013.

[26] W Mayol-Cuevas and D Murray. Wearable hand activity recognition for event summarization. 2005.

[27] A Ng, M Jordan, and Y Weiss. On spectral clustering: analysis and an algorithm. *Advances in Neural Information Processing Systems*, 2002.

[28] H Pirsiavash and D Ramanan. Detecting acitivites of daily living in first-person camera views. In *Computer Vision and Pattern Recognition (CVPR)*, 2012.

[29] X Ren and C Gu. Figure-ground segmentation improves handled object recognition in egocentric video. In *Computer Vision and Pattern Recognition (CVPR)*, 2010.

[30] B Russell, A Efros, J Sivic, W Freeman, and A Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *Computer Vision and Pattern Recognition (CVPR)*, 2006.

[31] D Salvucci and J Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Sym. on Eye Tracking Research & Applications*, 2000.

[32] B Sanders, R Nelson, and R Sukthankar. A theory of the quasi-static world. In *Int. Conf. on Pattern Recongition (ICPR)*, 2002.

[33] G Somanath, R Mv, D Metaxas, and C Kambhamett. D - clutter: Building object model library from unsupervised segmentation of cluttered scenes. In *Computer Vision and Pattern Recognition (CVPR)*, 2009.

[34] L Sun, U Klank, and M Beetz. EyeWatchMe - 3D hand and object tracking for inside out activity analysis. In *Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2009.

[35] S Sundaram and W Mayol-Cuevas. High level activity recognition using low resolution wearable vision. In *First Workshop on Egocentric Vision, Computer Vision and Pattern Recognition (CVPRW)*, 2009.

[36] S Sundaram and W Mayol-Cuevas. What are we doing here? egocentric activity recognition on the move for contextual mapping. In *Int. Conf. on Robotics and Automation (ICRA)*, 2012.

[37] K Takemura, Y Kohashi, T Suenaga, J Takamatsu, and T Ogasawara. Estimating 3D point-of-regard and visualizing gaze trajectories under natural head movements. In *Sym. on Eye-Tracking Research & Applications (ETRA)*, 2010.

[38] T Tuytelaars, C Lampert, M Blaschko, and W Buntine. Unsupervised object discovery: A comparison. *Int. J. on Computer Vision (IJCV)*, 2010.

[39] H Wang, A Kläser, C Schmid, and C Liu. Action Recognition by Dense Trajectories. In *Computer Vision and Pattern Recognition (CVPR)*, 2011.

[40] J Wu, A Osuntogun, T Choudhury, M Philipose, and J Rehg. A scalable approach to activity recognition based on object use. In *Int. Conf. on Computer Vision (ICCV)*, 2007.