You-Do, I-Learn: Unsupervised Multi-User egocentric Approach Towards Video-Based Guidance

Dima Damen, Teesid Leelasawassuk, Walterio Mayol-Cuevas

Abstract

This paper presents an unsupervised approach towards automatically extracting video-based guidance on object usage, from egocentric video and wearable gaze tracking, collected from multiple users while performing tasks. The approach i) discovers task relevant objects, ii) builds a model for each, iii) distinguishes different ways in which each discovered object has been used and vi) discovers the dependencies between object interactions. The work investigates using appearance, position, motion and attention, and presents results using each and a combination of relevant features. Moreover, an online scalable approach is presented and is compared to offline results. In the assistive mode, the paper proposes a method for selecting a suitable video guide to be displayed to a novice user indicating how to use an object, purely triggered by the user’s gaze. The assistive mode also recommends an object to be used next based on the learnt sequence of object interactions. The approach was tested on a variety of daily tasks such as initialising a printer, preparing a coffee and setting up a gym machine.

Keywords: Video Guidance, Wearable Computing, Real-time Computer Vision, Assistive Computing, Object Discovery, Object Usage

1. Introduction

Increasingly, commercial interest in wearable devices, including cameras and head-mounted displays in miniature and in fully wearable form (e.g. Google’s Glass, Microsoft’s HoloLens, Sony’s SmartEyeglass) invited research into cognitive systems that take advantage of these platforms. Footage from wearable cameras has fuelled Internet-based video sharing sites. Interestingly, among the most sought after videos are *how to do* guides, accessed by people wishing to carry out tasks, from cooking to assembling furniture. Video is nowadays a natural teaching resource and wearable displays are a potentially ideal channel for
Figure 1: Given egocentric videos from multiple users (a), a map of the environment (b) and feature clustering are used to discover distinct task-relevant objects (TROs); e.g. paper drawer and keypad for the task of operating a printer (c,d). For each discovered TRO, a model is built to incorporate possible locations, appearance, usage and 3D structure, along with a probabilistic graph of object interactions (e). In the assistive mode, a TRO is recognised triggered by gaze (f), a usage guidance snippet is chosen (g) and displayed to the user to provide guidance on how to use the object (h) along with the most likely object to be used next (i).

Assistance in task performance (e.g. assembly, repair) using augmented reality or video guidance has been promised for a while. One of the key limitations to realise such systems is the time consuming and evidently limiting task of authoring the content by e.g. manually segmenting and annotating videos or creating three-dimensional models that represent meaningful guidance (e.g. [38],[1]). Approaches that can provide guidance without the need for manual intervention would enable a wider adoption of assistive wearable systems.

Figure 1 shows an overview of the You-Do, I-Learn approach, both the learning and the assistive modes. This work attempts, to fully unsupervised, discover
objects and their usage from multiple users in a common environment (Fig. 1h), then provides a complete automatic solution for object-based guidance. We particularly focus on an egocentric view of the world, taking advantage of wearable technology, as it offers a unique perspective on object-level interactions. 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 person using a printer may interact with the paper drawer (Fig. 1c) and/or the keypad (Fig. 1d) while operating it. A system that aims to discover TROs would attempt to discover these objects/parts as opposed to the full machine or all its parts.

For each discovered object, we build a location model, an appearance model, an approximate 3D model and collect usage snippets on how different users interacted with the same discovered object. 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. We distinguish between object-based guidance and task-based guidance. This is because 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. We also build a first Markovian model of object usages within a task (Fig. 1e).

In the assistive mode, we propose a method for object-based guidance, triggered simply by gaze (Fig. 1f), where the user is advised on how a TRO object can be used based on the object’s current state (Fig. 1h), as well as advise the user on the most-likely object to be used next (Fig. 1i).

Up to our knowledge, this is the first attempt to close the gap between object discovery and video-based guidance in a fully unsupervised way. The manuscript builds on our previous works towards offline [8] and online [7] discovery of task relevant objects, with an improved online discovery algorithm compared to the one proposed in [7]. It compares the results of both approaches and extends the guidance to sequences of object interactions. Section 2 presents an overview of previous attempts towards object discovery, ego-centric video analysis and video-based guidance. The learning and assistive modes are presented in Sections 3 and 4 respectively. A varied dataset from coffee preparation to operating a gym machine is presented, alongside results in Section 5. The paper concludes with future directions.
2. Video-Based Object Discovery and Guidance - a Review

In this review, we focus on unsupervised approaches to object discovery (Section 2.1), task-relevant object discovery (Section 2.2) as well as approaches that aim to link discovery to guidance (Section 2.3). The works presented here differ from attempts to recognise objects in egocentric video from supervised training (e.g. [14, 39, 41]) which are interesting in their own right.

2.1. Unsupervised Object Discovery

![Diagram](image)

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

Object discovery refers to grouping feature descriptors into meaningful clusters that correspond to entities worth discovering. We attempt to differentiate between the various ways in which entities can be discovered from egocentric video; appearance, position, motion. Figure 2 envisages what can be discovered if each, or a combination, of these information is used in the grouping.

**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 tap, or a temporary position of a moveable object. Position information is frequently used with a static camera, and has been used in [17] to discover objects, by aligning two point clouds and identifying changes that correspond to objects that have been placed or removed.

**Appearance:** Appearance, of object and context, are often used to discover categories (e.g. [23, 42, 50]) or instances of an object based on visual similarity (e.g. [22, 19, 20, 44]). When attempting to discover categories, works assume a
collection of images where ‘common’ features correspond to the category, with or without spatial consistency (e.g. \cite{23, 42}). In \cite{50}, Tuytelaars et al present a review of recent works in category-based discovery. In egocentric videos within a common environment, instance discovery is targeted. In \cite{22} an object segment is discovered from a single image using RGB-D data. Several measures assist segmentation: compactness, symmetry, local convexity, global convexity, smoothness and recurrence in different scenes. In \cite{19} a data-driven objectness measure is proposed where a segment is compared to a database of general object segments. In \cite{20}, colour, texture and shape-based features are used to construct a network of finely-segmented regions. Segments are then iteratively grouped and refined until the algorithm converges to discovered objects. While a very interesting approach with promising results, \cite{20} 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. Many objects of daily living tasks such as a coffee machine or an electric socket remain fixed to their surroundings. The approaches in \cite{24, 42, 22, 19, 20} 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. Second, these approaches can only work offline after the whole dataset is collected. This assumption is also considered by other works \cite{44}.

**Motion:** Motion in egocentric video is a result of the wearer’s self-motion or 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 \cite{30} or dense features \cite{51} 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 \cite{25}) or object interactions \cite{47, 48}.

**Combinations:** When combining appearance with position, one can separate two instances of a mug when viewed in different locations. Using multiple cues has been recently attempted to improve recognition results. RGB-D images are also used in \cite{45} where sensor tracking is used to combine segments from multiple views to enable constructing a model of the object. A database of models is used to refine the reconstruction. Highly-textured objects (SIFT) In \cite{35}, position and appearance are combined where the disappearance of features in the same position is used to discover objects. An object is discovered if SIFT features seem to disappear when observing the same location again. Similarly in \cite{17}, table-top segments are compared and those that move place or change pose are labelled discovered as objects of interest. Experiments are very staged, In Collet et al. \cite{4},
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.

2.2. Unsupervised Discovery of Task-Relevant Objects from Egocentric Video

Combining appearance with motion has also been attempted \cite{12, 52}.

\cite{31, 34, 12} In \cite{12}, an action is identified by the change in appearance of the object before and after the action is performed. In \cite{31}, 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).

In a recent work, Bolanos proposed a semi-supervised approach to discover objects from wearable sensors \cite{2}. Given sparse sampling of images from a wearable camera (\(\frac{1}{60}\)fps), from a partially labelled dataset, objectness measures along with Convolutional Neural Networks (CNN) are used as features with iterative clustering. Clustering is evaluated using silhouette coefficients, to evaluate the discovered clusters.

In \cite{53}, clustering unlabelled video snippets representing activities is formulated as a linear program and a solver is used to find the optimal clustering. The number of tasks (i.e. number of clusters) is known a priori. This is based on the knowledge that the people perform the same set of tasks (e.g. in the morning). Segments of 15 seconds are used in the first dataset and pre-clipped sequences in the other. The multi-task clustering is formulated using earth mover’s distance and linear optimisation. Results seem to outperform previous multi-tasking approaches used in non-first-person data. The approach compares K-means, Kernel K-means as well as convex and semi-nonnegative matrix factorisation.

In \cite{37} tracked gaze along with motion information are used to classify actions such as ‘read’, ‘write’, browse, where they confirm previous results that gaze information is helpful in performing this task.

As opposed to discovering all objects in the scene, several approaches attempted to discover objects ‘of interest’, based on their similarity to a database of objects \cite{19} or more relevant to our work, discovering usable/used objects. Common approaches to discovering TROs in egocentric vision include i) segmenting the area surrounding the user’s hand \cite{13, 12, 51}, ii) extracting foreground regions through frame stabilisation or scene planarity assumptions \cite{40, 48} or iii) detecting ‘object-like’ regions \cite{34}. 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 [29, 28]. Gaze has been successfully used in to assist action recognition [11, 33] or supervised object recognition [46, 10].

2.3. Unsupervised Video-Based Guidance

Unsupervised generation of video snippets from a continuous egocentric video has mostly targeted video summarisation [34, 31]. The earliest example we could trace of determining video snippets that correspond to tasks or object manipulations is the work of Kang and Ikeuchi [21] that uses stereo visual data and other sensors, and focuses on the determination of hand motion in the video, for guidance of robotic arms. For human user assistance, Hashimoto et al. [16] proposed view sharing of video from wearable cameras to guide novice users. Their work does not require unsupervised segmentation of video guides but focus on live sharing of views of experts and beginners. In [15], instructional videos are projected onto an AR display for task guidance. While manually divided instruction video is employed, the system is able to pace the steps with the current step in performing the task. In [38], automatic extraction of snippets is performed using novelty detection. As the distance between consecutive images increases above a certain threshold, video clips are extracted. The work’s novelty focuses on overlaying the segmentation videos onto the task.

3. You-Do, I-Learn: Learning Mode

During learning, Task-Relevant Objects (TRO) need to be discovered (Sec. 3.1) and a model to be built for each object (Sec. 3.2). For each discovered TRO, usage snippets are automatically collected showing multiple people interacting with the same object. These usage snippets can then be analysed to discover the various Modes of Interaction (MOI) in an unsupervised manner (Sec. 3.3). Sequences of object interactions can also be discovered, highlighting strong dependencies (Sec. 3.4).

3.1. Discovering Task Relevant Objects (TRO)

Given a sequence of egocentric images \{I_1, ..., I_T\} collected from multiple operators around a common environment, we aim to automatically discover all
task-relevant objects. Assume $\Omega(I_t)$ is a part of the image $I_t$ (e.g., a segmentation or a bounding box within $I_t$), we aim to extract $K$ TROs, where each $O_k$ is a set of image parts from the sequence.

$$O_k = \{\Omega(I_t); 1 \leq t \leq T\}$$

(1)

TRO discovery is then the process of finding $K$ TROs, $\{O_k; 1 \leq k \leq K\}$, where the number of objects $K$ is not known a priori.

In this work, we make the assumption that at most one task-relevant image part is present within each image $I_t$. While multiple objects can be visible in $I_t$, only one object, and thus one image part, would be the focus of attention, and is task-relevant at each frame. The person could be interacting with multiple objects, for example placing one object on top of another, yet the attention is believed to shift between these objects [13]. This assumption simplifies the discovery of TROs without much loss in generality. Accordingly, the sets of image parts representing discovered TROs $\{O_k\}$ are believed to be disjoint and form a subset of all images in the sequence. Figure 3 shows a visual representation of TRO discovery.

Figure 3: Given a sequence of images from egocentric views, the objective is to group image parts into TROs, based on the assumption that one image part at most is task-relevant in each image. In this example, two TROs are shown (yellow, green)

We propose two approaches to TRO discovery, one is offline assuming all sequences from multiple users are collected prior to the discovery. The second approach is online, and thus is scalable to multiple users and different TROs.
We also compare two approaches to image parts $\Omega(I_t)$. The first $\Omega_c(I_t)$ crops the image around the centre. Given a glass-mounted camera, it is expected to have the object of interest at the centre of the frame during interaction. We compare this approach to gaze fixation $\Omega_g(I_t)$ where the image is cropped around a known gaze fixation. Given a wearable gaze tracker, we filter saccades using the velocity-based approach [43], where the average angular velocity over a sliding temporal window is considered a saccade if it is greater than $100^{\circ}/sec$, and is thus discarded. Alternative approaches to image parts could be explored like super-pixels but are left for future work.

To describe image parts, we use position, appearance features and their combination following Fig. 2:

- **Position:** The Image $I_t$ is positioned relative to the scene using sparse Simultaneous Localisation and Mapping (SLAM) [26] and a triangular tessellation of tracked interest points is built (similar to [49]). Given the 6D pose of the scene camera, a 3D ray links the centre of the image part $\Omega(I_t)$ to a point in the scene. Using the tessellation, the 3D position of the intersection point is calculated using linear interpolation.

- **Appearance:** To represent appearance, Histogram of Oriented Gradients (HOG) [5] is calculated over sub-windows within the image patch $\Omega(I_t)$. In offline TRO discovery, bag of Words (BoW) are used to represent appearance information. BoWs cannot be used in online TRO discovery, and thus HOG features are used as appearance information directly.

- **Combining Position and Appearance:** When combining position and appearance features, the normalised affinity matrices are summed with equal weighting in offline TRO discovery. The features are simply combined for online TRO discovery.

We also compare to results that accumulate features over a sliding window $w$ centred around each image ($\Omega(I_{t-w}), \ldots, \Omega(I_t), \ldots, \Omega(I_{t+w})$). In the experiments section, we test features that use position, appearance and their combination, over a sliding window and the two image part methods. We use the term $f_t$ next to refer to the feature vector representing an image patch where,

$$f_t = (F(\Omega(I_{t-w})), \ldots, F(\Omega(I_t)), \ldots, F(\Omega(I_{t+w}))$$

and $F(\Omega(I_t))$ is the feature descriptor for the image part $\Omega(I_t)$. 

[9]
3.1.1. Offline TRO Discovery

Offline TRO discovery refers to the attempt to discover all TROs after the dataset is fully collected. The sequencing of images is thus discarded and a data point $x_i = f_i$ refers to the descriptor of an image part in the dataset. We compare k-means clustering to spectral clustering from Ng et al. [36]. These approaches were compared in [50] for a known number of object categories.

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 \emph{a priori} [23, 50] to avoid the complexity. We propose estimating the optimal number of clusters $\hat{K}$ using the standard Davies-Bouldin (DB) index [9]. For an object $O_k$ with $n_k$ data points $\{x_i; i = 1..n_k\}$ assigned to this cluster, and $\mu_k$ is the mean of these data points, the intra-cluster distance $S_k$ can be measured as (Euclidean distance used):

$$S_k = \sqrt{\frac{1}{n_k} \sum_{i=1}^{n_k} ||x_i - \mu_k||_2^2} \quad (3)$$

The inter-cluster distances between two objects $O_k$ and $O_j$ is measured as $M_{kj} = ||\mu_k - \mu_j||_2$. The cluster similarity measure $R_{kj} = \frac{S_k + S_j}{M_{kj}}$ is used to calculate DB index,

$$V_{DB}(K) = \frac{1}{K} \sum_{k=1}^{K} \max_{j \neq k} R_{kj} \quad (4)$$

The optimal number of clusters is calculated to be

$$\hat{K} = \arg \max_K V_{DB}(K) \quad (5)$$

Recall that some images do not contain a TRO (Fig[3]), while clustering assigns a label for each image part. We assign a probability to each cluster being a TRO as the ratio of the number of points in the cluster to the total number of points,

$$p(S_k) = \frac{n_k}{\hat{K} \sum_{j=1}^{n_j}} \quad (6)$$

Clusters are also refined by removing the furthest $\beta\%$ of points in the cluster from the mean $\mu_k$. The refinement threshold, $\beta$, was set to 75 in all experimental results for the offline processing.
3.1.2. Online TRO Discovery

To discover objects in an online manner, image parts are clustered as they are collected and clusters are incrementally updated. An approach for online TRO discovery should enable clustering image parts of the same object as the object is used by multiple operators, whether in the same or a new location.

In proposing an algorithm for online TRO discovery, we rely on the assumption that consecutive similar image parts (\(\Omega(I_{t-\xi+1}), \ldots, \Omega(I_t)\)) indicate an observation of a task-relevant object (TRO). We thus define an object (\(O_k\)) as a collection of ‘at least’ \(\xi\) consecutive and similar image parts. The notion of similarity relies on the features used. For example, when \(f_t\) is the 3D position of image part \(\Omega(I_t)\), then at least \(\xi\) spatially-close consecutive image parts are labelled as a TRO. Alternatively when \(f_t\) is the appearance of image part \(\Omega(I_t)\), then at least \(\xi\) consecutive image parts of similar appearance enable discovering a TRO.

Two consecutive image parts, \(\Omega(I_t)\) and \(\Omega(I_{t-1})\) belong to the same object if \(||f_t - f_{t-1}|| < \epsilon_1\), where \(\epsilon_1\) is the threshold selected to accept clustering consecutive image parts and ||.|| is the Euclidean distance (Algo. 1 L. 9-14). The strict consecutive constraint between \(t\) and \(t-1\) can be relaxed to allow proximity within a sliding window. The mean and covariance of \(O_k\) are updated incrementally as further image parts are located within the threshold \(\epsilon_1\). Equations 7 and 8 show the incremental update for the mean and covariance of a \(O_k\).

\[
||f_t - f_{t-1}|| < \epsilon_1 \rightarrow \mu^k_t = \frac{\mu^k_{t-1} \times (n^k_{t-1} - 1) + f_t}{n^k_t} \tag{7}
\]

\[
\rightarrow \Sigma^k_t = \frac{n^k_t - 2}{n^k_t - 1} \Sigma^k_{t-1} + \frac{1}{n^k_t} (f_t - \mu^k_{n^k_t-1})^T (f_t - \mu^k_{n^k_t-1}) \tag{8}
\]

where \(\mu^k_t\) is the mean, \(\Sigma^k_t\) is the covariance matrix and \(n^k_t\) is the number of image parts within \(O_k\) at time \(t\).

Attention is believed to have moved to another object when \(||f_t - f_{t-1}|| \geq \epsilon_1\). At a future point in time \(t + \rho\), further image patches \(\Omega(I_{t+\rho})\) can belong to the same TRO \(O_k\) if it is within \(\epsilon_2\) standard deviations from the TRO \(k\) according to the Mahalanobis distance (Algo. 1 L. 4-5). This clustering method does not predefine the size of the clusters. When using position as a feature, it enables both small-sized and large TROs to be discovered.

The algorithm enables merging two clusters based on a measure of similarity (Algo. 1 L. 17). This could simply be appearance-based similarity to ensure moveable objects are grouped together. Two clusters \((\mu^j_t, \Sigma^j_t)\) and \((\mu^k_t, \Sigma^k_t)\) are
Input: Image parts and feature vectors \(\{(\Omega(I_t), f_t)\}; t = 1..T\)

Output: TROs \(\{O_k; 1 \leq k \leq K\}\) where \(O_k = (\{\Omega(I_t)\}, \Phi_k)\) and 
\[\Phi_k = \{({\theta_{ki}}, {\mu_{ki}}, {\Sigma_{ki}}); i = 1..L_k\}\]

1. \(K = 0\)
2. \(candidate = 0\)
3. for \(t = 1..T\) do
   4. find closest cluster \(k\): \(\text{min arg} \sum_{i=1}^{L_k} \theta_{ki}||f_t - \mu_{ki}||\Sigma_{ki}\)
   5. if \(\sum_{i=1}^{L_k} \theta_{ki}||f_t - \mu_{ki}||\Sigma_{ki} \leq \epsilon_2\) then
      6. \(l = \text{min arg} ||f_t - \mu_l||\Sigma_{kl}; 1 \leq l \leq L_k\)
      7. Update \(\theta_{ki}, \mu_{kl}(Eq.7), \Sigma_{kl}(Eq.8)\)
   8. else
      9. if \(||f_t - f_{t-1}|| < \epsilon_1\) then
         10. \(candidate = candidate + 1\)
         11. if \(candidate \geq \xi\) then
            12. \(K = K + 1\)
            13. \(L_K = 1\)
            14. Calculate \(\mu_K\) and \(\Sigma_K\)
         else
            15. \(candidate = 0\)
      16. end if
      17. end if
      18. if \(\min_{j \neq k} d_B(O_k, O_j) < \epsilon_3\) then
         19. \(L_j = L_j + 1\)
         20. \(\mu_{jL_j} = \mu_k\)
         21. \(\Sigma_{jL_j} = \Sigma_k\)
         22. Calculate mixture components \(\theta_j\)
         23. Delete \(O_k\) (objects merged)
      24. K = K - 1
     end if
   end if
end for

Algorithm 1: Proposed algorithm for online learning mode

merged if the distance measure \(d_B\) is below a threshold \(\epsilon_3\). We investigate using Bhattacharyya distance for merging clusters. The location model of the moveable object \(O_k\) is then a Gaussian Mixture Model (GMM) \(\{({\theta_i}, {\mu_i}, {\Sigma_i}), i = 1..L_k\}\) where \(\theta_i\) is the mixture component of the Gaussian \(i\) and \(L_k\) is the number of Gaussians in the GMM (Algo. 1 L. 18-23). When using appearance similarity, a new Gaussian is added to the GMM every time an object of similar appearance is found in a new position.
3.2. Building Models of TROs

For each discovered object $O_k$, we build four models that encapsulate the object’s location $\Phi_k$, appearance $A_k$, three-dimensional shape $M_k$ as well as its usage $U_k$. As the models are built from multiple operators with different heights and interaction behaviours, they give a good representation of the object (e.g. Fig. 4).

Location Model $\Phi_k$: The location model represents the position and extent of the object using a Gaussian Mixture Model $\Phi_k$; a single Gaussian for a fixed object and a multi-variate Gaussian for moveable objects. The likelihood of the object’s
position is indicated by

$$P(f_t|O_k) = \sum_{l=1}^{L_k} \theta_{kl} e^{-\frac{1}{2} (f_t - \mu_{kl})^T \sum_{kl}^{-1} (f_t - \mu_{kl})}$$

(9)

**Appearance Model** $\Phi_k$: For a view-based appearance model, we use the real-time method from [6] for learning views of the object. This method is particularly helpful for online learning as it is scalable and works in real-time. It is shape-based and thus particularly suitable for texture-minimal objects, many of which are present in indoor environments.

**Three-dimensional Model** $M_k$: To build a three dimensional representation of the object $M_k$, we adapt the work of [32] so it does not require the detection of keyframes and uses input from multiple users. Given a sparse map of the environment, the 3D points-of-regards are found by back-projecting the rays connecting the camera to the image part. These are used as seeds for super-pixel segmentation. The method uses outlier removal to reduce the error in volume estimation. In this work, we exploit 3D position information to generate textured three-dimensional models of the TROs as a byproduct of the process. Despite not being perfect models, due to the fact that they are created during task performance, the resulting models are useful visualisations of what objects the system has discovered. Ultimately, having a 3D model could facilitate applications such as augmented reality guidance.

**Usage Model** $U_k$: The image parts can also be combined to represent video snippets indicating how a TRO was used by multiple operators. Given consecutive image parts $\{\Omega(I_t), \Omega(I_{t+1}), \Omega(I_{t+\rho})\}; \rho \geq \xi$ belonging to the same TRO, a usage snippet $u^k_i$ can be formed from the sequence of image parts. Notice that when using gaze fixations $\Omega_s$, interpolation is needed when gaze information is missing. The collection of all usage snippets $U_k = \{u^k_i\}$ shows different ways in which $O_k$ was used or interacted with.

### 3.3. Finding Modes of Interaction (MOI) for TROs

For discovered TROs, we aim to find common MOIs for each TRO by analysing usage snippets, each representing a sample usage. The collection of all usage snippets $U_k = \{u^k_i\}$ shows different ways in which $O_k$ 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 to encode the interaction with the object [30].
We also use a temporal pyramid to encode the descriptors. At each level \( l \), 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 3.1.1) to find the MOIs.

Each cluster is represented by the snippet \( \hat{u}_j \) closest to the centre of the cluster \( \mu_j \) (i.e. mean snippet),

\[
\hat{u}_j = \arg \min_{u_l \in \text{MOI}_j} ||d(u_l) - \mu_j||; \quad \mu_j = \frac{1}{|\text{MOI}_j|} \sum_{u_l \in \text{MOI}_j} d(u_l)
\]  

and the confidence in a cluster being a common mode of interaction is represented by the percentage of snippets within that cluster \( p(\text{MOI}_j) \),

\[
p(\text{MOI}_j) = \frac{|\text{MOI}_j|}{|U_k|}
\]

A threshold \( \lambda \) can be used to select common MOIs such that \( p(\text{MOI}_j) \geq \lambda \).

3.4. Graphs of Object Interactions

Following the discovery of TROs, it is also possible to model, in an unsupervised way, the sequence of object interactions towards modelling tasks or simply discovering strong links between object interactions. For example, after using the tap, the user is likely to follow that by interacting with the drainer or with a towel. These strong links between objects can be automatically discovered from sequences of multiple users. We model these interaction sequences by a graph-based representation as follows.

For all discovered objects \( \{O_k; k = 1..K\} \), a complete directed graph \( G \) is constructed so each TRO is represented by a node and the weight \( W_{O_k \rightarrow O_j} \) of the directed edge \( O_k \rightarrow O_j \) represents the probability of interacting with object \( O_j \) directly after having interacted with object \( O_k \). Note that we loosely define interaction as attending or looking at an object. Algorithm 2 details how the graph is constructed. The edge weights are initialised with a small value \( \alpha \). The temporal transitions from one discovered object to another are accounted for, followed by edge-weight normalisation.
input : Image parts and feature vectors \(\{(\Omega(I_t), f_t)\}; t = 1..T\) 
Discovered TROs \(\{O_k; k = 1..K\}\)
output: Graph representing the interaction \(\{G_{K \times K}\}\)

\[
G = [\alpha]_{K \times K}, \text{CurrentIndex} = \text{PreviousIndex} = -1
\]

for \(t = 1..T\) do
  for \(k = 1..K\) do
    if \(\Omega(I_t) \in O_k\) then
      CurrentIndex = k
    end
  end
  if CurrentIndex \(\neq\) PreviousIndex then
    \(G[\text{CurrentIndex}][\text{PreviousIndex}]++\)
    PreviousIndex = CurrentIndex
  end
end

Normalise the rows of the graph \(G\)

Algorithm 2: Algorithm for constructing the graph-based representation of TRO interaction sequences

4. You-Do, I-Learn: Assistive Mode

In the assistive mode, the location models \(\{\Phi_k\}\), the appearance models \(\{A_k\}\), the usage snippets \(\{U_k\}\), the various modes of interaction \(\{MOI_k\}\) as well as the graph of object interactions \(G_{K \times K}\) are used to provide a recommendation of how the object can be used, as well as what object to use next. For each test image, the image \(\Omega(I_t)\) is compared to the discovered TROs. Upon recognition of a TRO \(O_k\), help can be provided showing one of the possible MOIs, that is most relevant to the task or the object status. The help snippet \(h_t = u_k \in U_k\) is displayed to show how this object was previously used. From the possibly many usage snippets featuring the TRO, we chose the help snippet \(h_t\) as a video guide at time \(t\) such that the appearance of the first frame in the snippet, is closest to the recognised view.

\[
h_t = \arg \min_{u_j} ||A^{1st}(u_j) - A(\Omega(I_t))||
\]

where \(A^{1st}\) is the appearance of the first frame in the snippet. If the object changes state, the initial appearance is a good indicator of which usage snippet to show. An additional advantage is to avoid showing a snippet observing the object from a different viewpoint, so the user can easily map what they see to what they could do. A help snippet is displayed each time a new object is detected, aiming to provide automatic assistance for novice operators.
The graph of object interactions $G_{K \times K}$ can be used to estimate the object to be next manipulated. The assistive mode would recommend the object to be used next, so that

$$\hat{j} = \arg \max_j p(O_j|O_k) = \arg \max_j W_{O_k \rightarrow O_j}$$  \hspace{1cm} (13)

Recommending a help snippet as well as the object to be used next is based on correctly recognising that the user is attending a TRO. We base the assistive mode on gaze fixations $\Omega_g(I_t)$ and investigate two approaches for recognising the TRO, 

1. **Using the location models** $\{\Phi_k\}$: a TRO is recognised based on Eq. [9] so that

$$k = \arg \max_k p(f_t|O_k); \quad p(f_t|O_k) \geq \lambda$$  \hspace{1cm} (14)

2. **Using the appearance models** $\{A_k\}$: this assistive mode does not require a map of the environment or tracking of the camera relative to the environment. Given the image part $\Omega_g(I_t)$, the appearance model is used to recognise the viewed object, from the set of appearance models. By using the combination of fixed paths and a hierarchical hash table, object recognition is scalable, and can reliably detect objects at frame rate [6]. The descriptor is affine-invariant, and the method is tolerant to a level of occlusion but is also view-dependant. Figure [5] shows the method learning (left column) and subsequently recognising (right column) objects from our experiments.

5. Experiments and Results

**Setup & Dataset:** The wearable gaze tracker hardware (ASL Mobile Eye XG [27]) consists of two cameras sharing a half-mirror, 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. [8]). 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) [1]. 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.

1 Dataset available at: [http://www.cs.bris.ac.uk/~damen/BEOID](http://www.cs.bris.ac.uk/~damen/BEOID)
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. adaptor 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). Figure 6 shows an example of how the narrations were used to generate the ground-truth TROs and MOIs. For each location, a map is built using Parallel Tracking and Mapping (PTAM) [26]. A 3D bounding box around each object is manually labelled for evaluation. For moveable objects, their different locations are ground-truthed.

**Parameters:** In all results, the image parts $\Omega(I_t)$ were fixed to a window size of $200 \times 200$ pixels, This corresponds to $19.3^\circ$ visual angles in the scene camera. To calculate appearance descriptors, $\Omega(I_t)$ is divided into $10 \times 10$ non-overlapping patches for calculating HOG descriptors. In offline processing, the number of words in BoW representation is set to 200. In calculating the BD index, $K = [2..2N_{ogt}]$ (Eq. 4) where $N_{ogt}$ is the number of ground-truth objects. In online TRO discovery $\xi$ was set to 40 frames, which corresponds to 1333ms of attention.

**Results for discovering TROs:** The results of offline and online TRO discovery are compared to the established ground-truth. The clusters’ bounding boxes are
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.

| Location | Number of sequences | Sequence length | Tracked (%) | Gaze Fixations (%) |
|----------|---------------------|-----------------|-------------|--------------------|
| K        | 10                  | 1905            | 69.4        | 58.9               |
| W        | 10                  | 1221            | 78.3        | 61.9               |
| P        | 10                  | 596             | 75.8        | 70.5               |
| D        | 10                  | 303             | 71.8        | 56.2               |
| G        | 6                   | 5183            | 76.4        | 66.7               |
| M        | 6                   | 2039            | 24.5        | 14.6               |

Prepare coffee using the machine, place the cup on the mat and add sugar [tap, coffee machine, heat mat, cutlery drainer], (cup, sugar jar)
Plug the screwdriver for charging and place the tape in the red box [Socket, Box], (screwdriver, charger, tape)
Check the printer is loaded with paper manually and using the keypad [drawer, keypad]
Go through the locked door [door lock, door handle]
Use the treadmill and the bicycle next to it [treadmill panel, bicycle panel]
Adjust the seat, chest pad and weight then use the machine [seat adjuster, pad adjuster, weight adjuster]

Table 2 shows the complete set of results. In offline TRO discovery, 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 image part mechanisms $\Omega_c$ and $\Omega_g$ are compared, where the latter crops an image around gaze fixations thus referred to as cropping ‘with attention’. Estimating the number of clusters using the Davies-Bouldin (DB) index is compared to knowing the number of clusters a priori (ref. Known $K$). For online results, the best

Figure 6: Example showing how the ground truth for TROs and MOIs was obtained from subject’s narrations. Ground-truth TROs narrated by more than 50% of subjects are framed in red, compared to less-frequent subjects (orange). Location names are ignored (blue). The verb-noun combinations are used to ground-truth MOIs (green). The narrations are released with the dataset.

compared to ground-truth bounding boxes and the PASCAL overlap criteria (in 3D) of 20% indicates a true-positive. This is because the viewed positions don’t typically cover the full extent of the object.
Table 2: Recall and precision results for discovering TROs using different features, clustering methods, with/without attention and sliding window.

Table 2 shows that the best offline 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. 8. If the number of clusters was known a priori 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. In online TRO discovery, attention significantly improves the results as the chance of $\xi$ consecutive similar image parts increases. Interestingly, when combining appearance and position, 85% of the objects were retrieved with 77.3% precision showing the potentials of the scalable algorithm.

Fig. 7 highlights several conclusions from the results of offline TRO discovery: (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. 7(b) shows...
Figure 7: (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+knownK, patches around centre of image vs. gaze fixations. (d) Single-frame vs. sliding window representations.

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. 7(c) shows the importance of within-image attention [app+pos, KnownK, \( w = 1 \)]. A significant drop in recall is observed when the information is gathered around the image centre rather than gaze fixations. Fig. 7(d) shows that a sliding window gives a slight improvement in performance.

Examples of learnt views for the discovered objects can be found in Fig 9. The accuracy of the model relies on whether it has been viewed from multiple views by the users. The importance of sequences from multiple users is particularly noticed when attempting to build these approximate 3D models. Figure 10 shows the models extracted from one, two and five operators. Figure 11 presents further three-dimensional models for eight discovered objects. Note that the method is capable of discovering and representing small-sized (a,d,e,h) as well as larger objects (b,f,g).

For each discovered object, the usage snippets longer than \( \xi = 1s \) are used
Figure 8: 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).
Figure 9: Learnt views from training sequences of multiple users for a variety of objects: coffee machine, tap, seat adjuster and screwdriver.

Figure 10: Textured three-dimensional models are built for two TROs (tap - top and door handle - bottom) from one operator (a), two operators (b) and two views from five operators (c,d). As sequences from multiple users with various heights and viewpoints are incorporated, the segmented 3D models are further refined.

to build a usage model. On average, 16.6 usage snippets are extracted for each TRO ($\sigma = 7.4$). Figure [12] presents a collection of usage snippets for the same
Figure 11: Textured three-dimensional models (two views each) for eight discovered TROs.

discovered TRO. Notice that these snippets are extracted automatically based on the discovered TRO. The example shown here is from the online discovery of TROs for the object tap.

Figure 12: Sample frames, at equal sampling rates, from five usage snippets cropped around gaze fixation ($\Omega_g(I_t)$) from five different operators for the discovered TRO tap. Note that the snippets vary in length and show various ways of using the same object. In these examples, the tap is used to wash (a), rinse (b,c) and fill (d,e) the cup.
Results for discovering MOIs  We vary the threshold $\lambda$ to accept $p(MOI_j)$ (Eq. 11) 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 as well as their combination (Fig. 13). As anticipated, motion information solely is capable of distinguishing the various modes of interaction with the same object. Using the combination of features and $\lambda = 0.2$, the approach is able to discover meaningful MOIs. Figure 14 shows an example of the method successfully discovering two MOIs for the ‘socket’. Similarly, Fig. 15 shows further discovered MOIs for the sugar jar and the door handle.

Figure 13: For position (left), temporal pyramid (L=5) performed best, while motion (right) performed best on L=1. When using motion only versus combining all features at their best temporal pyramid level, a minor improvement is observed.

Figure 14: For the ‘socket’, the two common MOIs (‘switching’, ‘plugging’) are found (left & right). The representative usage snippet is shown (up) with the other snippets in the same cluster (below) - only one snippet is incorrectly clustered (shown in red).

Results for Graph of Object Interactions:  With the discovered TROs, we trained the graphs representing the interactions using Algorithm 2. The initial link $\alpha$ was set as 0.05. The generated graphs for the Kitchen (K) and weight Machine (M) sequences are presented in Figure 16. The Notice the strong causal links between coffee machine→heat mat, tap→coffee machine, sugar jar→heat
mat, seat adjuster→pad adjuster all being meaningful strong links between interactions with these objects in the dataset.

Results for Assistive Mode: While we do not test the assistive mode with users to evaluate the ‘usefulness’ of the provided usage snippets or the recommendation for the object to use next, we qualitatively assess the ability of the assistive mode to provide meaningful help snippets. In running the assistive mode, we use a 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 and what object to use next. We show results from the two recognition methods, first employing the position models to predict the object being used, then employing the appearance models.

Fig[17] shows a sequence of object interactions in the assistive mode. When the
user fixates at a discovered TRO, a usage snippet indicating how to use that object is recommended along with the object to be used next. The figure also shows links (coloured using the heat map in Fig. 16) to indicate weight of the edges in the object interactions graph. When the heat map is recognised (Fig 17a), a usage snippet is shown recommending a cup to be placed on the mat. The next object to be used is thought to be the drainer. The operator indeed moves towards the drainer (Fig 17b), and the drainer is recognised (Fig 17c). The recommended usage is to pick a cutlery and the suggested next object is the heat mat. The harvested view of the heat mat is that with the cutlery being used. Though this is automatically chosen, it is extracted from the set of usage snippets that follow using the drainer. The attention is indeed shifted to the heat mat (Fig 17d).

Next, we use the real-time texture-minimal scalable detector code from [6] due to its light-weight computational load that makes it amendable to wearable systems [3] to recognise TROs. 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 18 shows frames from the help videos and a full sequence is available [7]. Recall that these inserts are extracted, selected and displayed fully automatically. This assistive mode presents a possible application for unsupervised discovery of TROs and their MOIs. We believe other potential applications could be explored.

https://youtu.be/vUeRJmwm7DA
Figure 17: For a pre-built map of the environment (left) and egocentric images with tracked gaze (right), the position (green-dot) is used to recognise TRO, a usage snippet is inserted (yellow-framed insert) along with the object to be used next (blue-framed insert). The recommendations are inserted everytime a different TRO is recognised.

6. 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. An offline approach is compared to an online approach that enables scalability. Usage snippets are also automatically extracted along with sequences of object interactions.

While this paper provides a complete framework that bridges the gap between unsupervised object discovery and video-based guidance with promising preliminary results, it aims to trigger further research and discussions, particularly related to the usefulness of automatically extracted video guides for human operators and/or autonomous systems, the importance of attention information in egocentric video analysis and more advanced techniques towards discovering modes of interactions for everyday objects.

Acknowledgement We would like to thank Pished Bunnun, Osian Haines and Andrew Calway for their input on previous iterations of parts of this work.

References

References

[1] Bleser, G., Almeida, L., Behera, A., Calway, A., Cohn, A., Damen, D., Domingues, H., Gee, A., Gorecky, D., Hogg, D., Kraly, M., Macaes, G., Marin, F., Mayol-Cuevas, W., Miezal, M., Mura, K., Petersen, N., Vignais, N., Santos, L., Spaas, G., Stricker, D., 2013. Cognitive workflow capturing and rendering with on-body sensor networks (cognito). German Research Center for Artificial Intelligence, DFKI Research Reports (RR).

[2] Bolanos, M., 2015. Ego-object discovery in lifelogging datasets. Master’s thesis, Universitat de Barcelona.
[3] Bunnun, P., Damen, D., Calway, A., Mayol-Cuevas, W., 2012. Integrating 3D object detection, modelling and tracking on a mobile phone. In: Int. Symp. on Mixed and Augmented Reality (ISMAR).

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

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

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

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

[8] Damen, D., Leelasawassuk, T., Haines, O., Calway, A., Mayol-Cuevas, W., 2014. You-do, i-learn: Discovering task relevant objects and their modes of interaction from multi-user egocentric video. In: British Machine Vision Conference (BMVC).

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

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

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

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

[13] Fathi, A., Ren, X., Rehg, J., 2011. Learning to recognise objects in egocentric activities. In: Computer Vision and Pattern Recognition (CVPR).
[14] Gonzalez-Diaz, I., Boujut, H., Buso, V., Benois-Pineau, J., Domenger, J., 2014. Saliency-based object recognition in video.

[15] Goto, M., Uematsu, Y., Saito, H., Senda, S., Iketani, A., 2010. Task support system by displaying instructional video onto ar workspace. In: Int. Symposium on Mixed and Augmented Reality (ISMAR).

[16] Hashimoto, Y., Kondo, D., Yonemur, T., Iizuka, H., Ando, H., Maeda, T., 2011. Improvement of wearable view sharing system for skill training. In: Int. Conference on Artificial Reality and Telexistence.

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

[18] Johansson, R., Westling, G., Backstrom, A., Flanagan, J., 2001. Eye-hand coordination in object manipulation. The Journal of Neuroscience 21 (17).

[19] Kang, H., Hebert, M., Efros, A., Kanade, T., 2015. Data-driven objectness. IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) 37.

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

[21] Kang, S., Ikeuchi, K., 1994. Determination of motion breakpoints in a task sequence from human hand motion. In: Int. Conference on Robotics and Automation (ICRA).

[22] Karpathy, A., Miller, S., Fei-Fei, L., 2013. Object discovery in 3d scenes via shape analysis. In: Int. Conf. on Robotics and Automation (ICRA).

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

[24] Kim, G., Hebert, M., Park, S., 2008. Preliminary development of a line feature-based object recognition system for textureless indoor objects 370, 255–268.

[25] Kitani, K., Okabe, T., Sato, Y., Sugimoto, A., 2011. Fast unsupervised ego-action learning for first-person sports videos. In: Computer Vision and Pattern Recognition (CVPR).
[26] Klein, G., Murray, D., 2007. Parallel Tracking and Mapping for Small AR Workspaces. In: Int. Sym. on Mixed and Augmented Reality (ISMAR).

[27] Laboratories, A. S., 2011. Mobile Eye-XG. URL http://www.asleyetracking.com/

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

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

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

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

[32] Leelasawassuk, T., Mayol-Cuevas, W., 2013. 3D from looking: Using wearable gaze tracking for hands-free and feedback-free object modelling. In: Int. Sym. on Wearable Computers (ISWC).

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

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

[35] Mason, J., MArthi, B., Parr, R., 2012. Object disappearance for object discovery. In: Int. Conf. or Intelligent Robots and Systems (IROS).

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

[37] Ogaki, K., Kitani, K., Sugano, Y., Sato, Y., 2012. Coupling eye-motion and ego-motion features for first-person activity recognition. In: Computer Vision and Pattern Recognition Workshops (CVPRW).

[38] Petersen, N., Stricker, D., 2012. Learning task structure from video examples for workflow tracking and authoring. In: International Symposium on Mixed and Augmented Reality (ISMAR).
[39] Pirsiavash, H., Ramanan, D., 2012. Detecting activities of daily living in first-person camera views. In: Computer Vision and Pattern Recognition (CVPR).

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

[41] Ren, X., Philipose, M., 2009. Egocentric recognition of handled objects: Benchmark and analysis. In: Computer Vision and Pattern Recognition Workshop (CVPRW).

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

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

[44] Sanders, B., Nelson, R., Sukthankar, R., 2002. A theory of the quasi-static world. In: Int. Conf. on Pattern Recognition (ICPR).

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

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

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

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

[49] Takemura, K., Kohashi, Y., Suenaga, T., Takamatsu, J., Ogasawara, T., 2010. Estimating 3D point-of-regard and visualizing gaze trajectories under natural head movements. In: Sym. on Eye-Tracking Research & Applications (ETRA).
[50] Tuytelaars, T., Lampert, C., Blaschko, M., Buntine, W., 2010. Unsupervised object discovery: A comparison. Int. J. on Computer Vision (IJCV).

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

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

[53] Yan, Y., Ricci, E., Liu, G., Sebe, N., 2014. Recognizing daily activities from first-person videos with multi-task clustering. In: Asian Conference on Computer Vision (ACCV).