Multi-User Egocentric Online System for Unsupervised Assistance on Object Usage

Dima Damen, Osian Haines, Teesid Leelasawassuk,
Andrew Calway and Walterio Mayol-Cuevas

Department of Computer Science, University of Bristol, Bristol, UK

Abstract. We present an online fully unsupervised approach for automatically extracting video guides of how objects are used from wearable gaze trackers worn by multiple users. Given egocentric video and eye gaze from multiple users performing tasks, the system discovers task-relevant objects and automatically extracts guidance videos on how these objects have been used. 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 approach is tested on a variety of daily tasks ranging from opening a door, to preparing coffee and operating a gym machine.

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

1 Introduction

With the advent of wearable devices, systems able to provide guidance to users remain a possibility and a challenge. In particular in industrial settings (e.g. assembly, repair), operations using augmented reality or video-based manuals have been promised for a while. One of the key limitations to realize such systems is the need for authoring the content by e.g. manually segmenting and annotating videos or creating three-dimensional models that represent meaningful guidance [16],[1]. Authoring is time consuming and evidently limiting. Approaches that can provide guidance without the need for any manual intervention would enable a wider adoption of assistive wearable systems.

In this paper we present a fully automated, online and real-time approach for providing video-based guidance on object usage from egocentric video and eye gaze. The system has two modes, a *learning mode* where video snippets are automatically extracted from videos of multiple users performing tasks around a shared environment, and an *assistive mode* where a 'suitable' video snippet from the automatically collected video guides is selected, triggered by gaze. In strong contrast to most previous work on assistive egocentric guidance, we require no pre-training of the objects involved in tasks, nor knowledge of the tasks' scripts or the knowledge of how many objects will be used or interacted with. The approach is able to harvest video snippets for objects of interest as a precursor for cognitive assistance. The system selects a short assistive snippet or *video*

guide to be shown when a gazed-at object is recognised, to illustrate how the object was used before. This paper presents a prototype for the system and concentrates on evaluating the extraction of objects and their use. We illustrate the annotation of test videos with the automatically extracted video guides[1], and leave the evaluation of the effectiveness of the *assistive mode* with real users for future work.

The setup uses a single wearable gaze-tracker eyepiece which features a camera that looks out towards the scene and a pupil tracker that indicates where the eye is looking.

2 Related Work

Related systems to the problem we are aiming to address expect the objects to have visual markers (e.g. [18]), use model-based tracking(e.g. [16]) or be specified in advance of task performance (e.g. [1]). This review focuses on the ability to find objects of interest, i.e. task-relevant objects (TRO), from egocentric video during task performance. Common approaches include i) segmenting the area surrounding the user's hand [7],[6],[12], ii) extracting foreground regions through frame stabilisation or scene planarity assumptions [17],[21] or iii) detecting 'object-like' regions [15].

One uniquely rich source of information in egocentric sensing is eye gaze. Eye gaze has been studied for hundreds of years and more intensively since the 19th century [23]. 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 [11],[8].

However, eye gaze has been rarely considered as part of wearable systems, perhaps due to the scarcity of *mobile* gaze tracking hardware. Exceptions include [5] which exemplifies how gaze can assist in predicting the current action, and how the predicted action can be used to estimate the forthcoming gaze position. In [20], a wearable gaze-controlled camera provides a cropped image dictated by eye gaze locations to enhance object tracking and in [4], interest points are extracted around the gaze point and matched to pre-learnt highly textured objects. None of these approaches discover objects using gaze. In [14], object segmentation using gaze is attempted from annotated short clips containing action, though the work focuses on gaze estimation.

Our recent work [3] has compared the influence of gaze, position, appearance and motion *using offline processing* on the extraction of objects in egocentric video. Results prove that 80% of objects were correctly extracted by localising gaze within a 3D map. In this work, we use the same dataset but propose an *online* incremental algorithm that learns objects and extracts video help guides incrementally from multiple operators. An online approach would scale with more users without the need for re-training, and data can be processed on

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

the fly avoiding the need to store lengthy hours of egocentric video collected from multiple users. To enable real-time processing, we learn objects using the shape-based real-time object learning and detection method [2], which is capable of accommodating multiple objects in a scalable manner using constellations of edgelets. The algorithm is also capable of detecting 'moveable' objects and distinguishing them from 'static' objects that remain fixed in the 3D map.

3 Proposed Method

Our method is based on four principles:

- Spatio-temporally consistent gaze fixations indicate an observation of a task-relevant object (TRO).
- Each observation represents a candidate video snippet for assistive guidance.
- Spatially consistent observations correspond to a fixed TRO (i.e. an object with a fixed location in the scene).
- Appearance-consistent observations, observed in different locations, correspond to a moveable TRO.

The input to the system is real-time egocentric video with 2D gaze fixations. In the *learning mode* (Sec 3.1), the system aims to learn objects of interest as well as extract video snippets on how these objects are used. In the *assistive mode* (Sec 3.2), the system aims to recognise gazed-at objects and select a suitable video snippet for guidance from the automatically extracted snippets in the learning mode. The approach is completely unsupervised, and details of both modes are discussed next.

3.1 Learning Mode

First, we follow the velocity-based approach from [19] to distinguish saccades from fixations, and position the 2D fixation relative to the scene using sparse Simultaneous Localisation and Mapping (SLAM) [9]. Given the 6D pose of the scene camera, a 3D gaze ray links the direction of the gaze to a point in the scene. A dense depth map is estimated, using a triangular tessellation on the tracked interest points that are visible on the scene camera (similar to [22]). To distinguish between the 2D fixation at time t and its corresponding 3D position within the map, we refer to these as f_t^2 and f_t^3 respectively.

Next, objects are discovered using online clustering, as explained below and in Algo. 1. We define a *gaze cluster (GC)* as a collection of 'at least' ξ spatially-close consecutive gaze fixations, and use this to learn objects. Two consecutive fixations, f_t^3 and f_{t-1}^3 belong to the same GC if $||f_t^3 - f_{t-1}^3|| < \epsilon$, where ϵ is the distance threshold selected to accept clustering consecutive fixations and $||.||$ is the Euclidean distance. Notice that the temporal difference between t and $t-1$ might not correspond to one frame, as some frames have missing gaze information, or the gaze might have been discarded as a saccade. If and only

input : fixations $\{(f_t^2, f_t^3)\}$, images $\{I_t\}; t = 1..T$
output: TROs $\{(A_k, U_k, m_k, \nu_k); 1 \leq k \leq K\}$
A_k learnt view-based appearance model for TRO i
U_k video snippets for TRO i
$\nu_k \in$ {fixed, moveable} type of TRO
m_k segmented 3D model for TRO k

$K = previousK = 0$
$stableGC = 0$
for $t = 1..T$ **do**
  find closest gaze cluster k: $\min \arg_k ||f_t^3 - \mu_k||_{\Sigma_k}$
  Extract window w_t centred around f_t^2 from I_t
  // Object Discovery
  if $||f_t^3 - \mu_k||_{\Sigma_k} \leq 1$ **then**
    Update $\mu_k(Eq\ 1), \Sigma_k(Eq\ 2)$
  else
    if $||f_t^3 - f_{t-1}^3|| < \epsilon$ **then**
      $stableGC = stableGC + 1$
      if $stableGC \geq \xi$ **then**
        $K = K + 1$
        Add a new gaze cluster $k = K$
        Learn the first view of a new object
        $\nu_k =$ 'fixed'
    else
      $stableGC = 0$
  // Learn Appearance
  Detect an object within the window w_t
  if *recognised as TRO* j **then**
    if $j \neq k$ **then**
      if *confirmed from several detections* **then**
        $\nu_k =$ 'moveable'
  else
    if *Object was not detected in last* δ *frames* **then**
      Learn a new view for object k
  // Video Snippets and Model
  if $k \neq previousK$ **then**
    add *video snippet* u_i^k to U_k (Eq. 3)
    build 3D model m_k
  // Keep track of current GC
  $previousK = k$

Algorithm 1: Proposed algorithm for *learning mode*

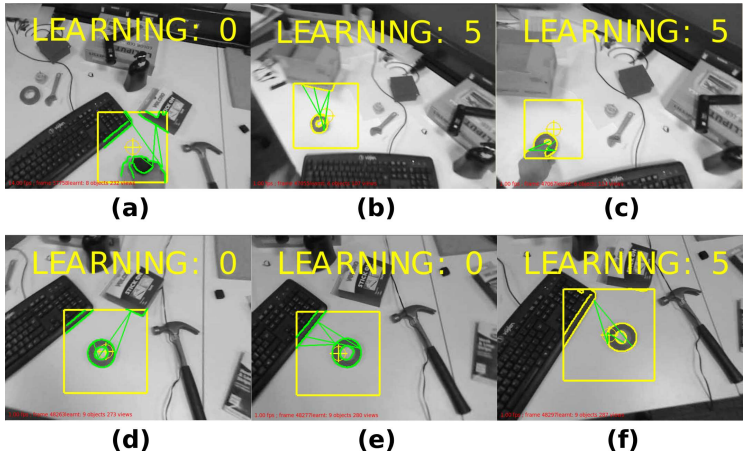

Fig. 1. Two TROs were discovered (a,b). Later, the tape was moved (c). A new fixation is spatially close to TRO '0' (d). Initially, further views were collected for TRO '0' (d,e). A few frames later, the object is consistently recognised as TRO '5' by appearance matching. Both TRO '0' and '5' are marked as 'moveable' (f).

if ξ consecutive fixations are within the same GC, an observation of a TRO k is discovered (Algo. 1 L. 9-15). The mean and covariance of GC are updated incrementally as further fixations are located within the threshold ϵ. Equations 1 and 2 show the incremental update for the mean and covariance of a GC.

$$||f_t^3 - f_{t-1}^3|| < \epsilon \rightarrow \mu_t^k = \frac{\mu_{t-1}^k \times (n-1) + f_t^3}{n} \tag{1}$$

$$\rightarrow \Sigma_t^k = \frac{n-2}{n-1}\Sigma_{t-1}^k + \frac{1}{n}(f_t^3 - \mu_{n-1})^T(f_t^3 - \mu_{n-1}) \tag{2}$$

where μ_t^k is the mean, Σ_t^k is the covariance matrix and n is the number of clustered fixations at time t.

Attention is believed to have moved to another location when $||f_t^3 - f_{t-1}^3|| \geq \epsilon$. At a future point in time $t+\rho$, further fixations can belong to the same TRO k if it is within one standard deviation from the mean of the TRO k according to the Mahalanobis distance (Algo. 1 L. 6-7). This clustering enables both small-sized and large TROs to be discovered, as it does not limit or pre-define the size of the GC. However, it assumes that the object is fixed, i.e. remains within the same 3D location.

To accommodate for moveable objects, appearance matching is considered. For every TRO k, views around the object are learnt using the real-time method from [2]. Only novel views are added to the appearance model - a view is added if the object fails to be recognised in the past δ frames (Algo. 1 L. 24-25). The gazed-at object is compared to the previously learnt K objects $\{A_k; k = 1..K\}$. If the appearance matches a learnt TRO, at a different location, the object is

believed to have moved, and is thus identified as a 'moveable' object (Algo. 1 L. 19-22). To avoid incorrect detections, multiple consecutive matching appearances are required before an object is identified as 'moveable'. Figure 1 shows an example of identifying a 'moveable' object.

Notice that identifying an object as 'moveable' could result from multiple instances of the same object. A limitation of the approach arises when a new object replaces a learnt TRO. The object is then incorrectly learnt as novel views of the previously learnt TRO. This does not affect the assistive nature of the method, as we use the current object's appearance to select a suitable *help snippet* as will be explained next.

As we position gaze in 3D space, we can exploit this information to generate visualisations of the TROs as a byproduct of the process (Algo. 1 L. 28). This step adapts [13] so it does not require the detection of keyframes form the user's motion and does not assume a single user is providing input. Despite not being perfect models, due to the fact that they are created during an action, the resulting models are useful visualisations of what objects the system has discovered. Ultimately, having a 3D model facilitates applications such as augmented reality guidance which we leave for future work.

Given consecutive fixations $(f_t^2, f_{t+1}^2, ..., f_{t+\rho}^2); \rho \geq \xi$ belonging to the same TRO, a **video snippet** u_i^k for TRO k is defined as

$$u_i^k = \{\Psi(I_j, \Delta(j), \omega) \tag{3}$$

where Ψ crops a window of size ω from Image I_j around the interpolated fixation $\Delta(j)$ as gaze information is missing in some frames (Algo. 1 L. 27). The collection of all video snippets U_k shows different ways in which the object k was used or interacted with.

As multiple operators with different heights and interaction behaviours use the same object, the method is capable of expanding the learnt views, the 3D model m_k and gather further interaction video snippets U_k. Figure 2 shows the advantages of learning from multiple users.

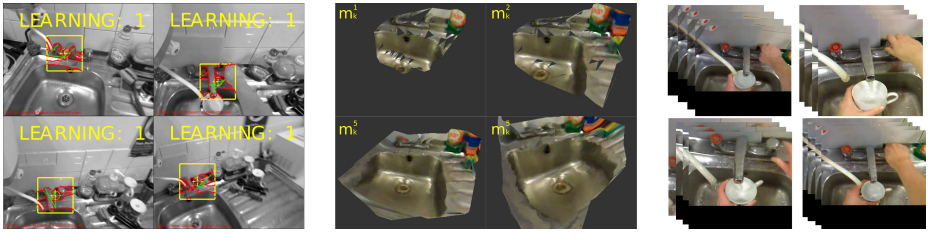

Fig. 2. For the same discovered object (sink): multiple users enable learning varying views in the appearance model A_k (left); the 3D model m_k (middle) is refined (m_k^1) shows the model for one user, two users (m_k^2) as well as five users (m_k^5); different video snippets U_k show multiple interactions with the same object (right).

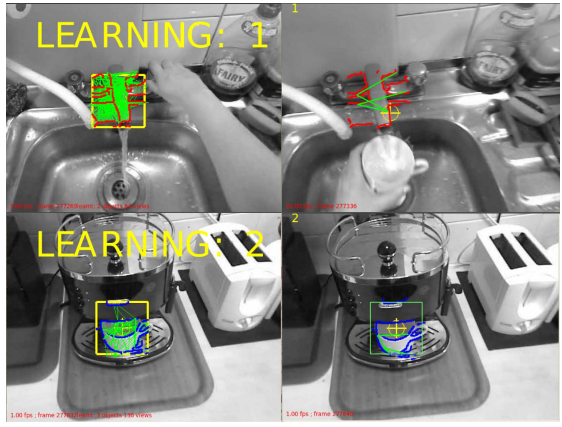

Fig. 3. During discovery (left), edges within a window around the gaze are captured as object views, and represented using affine-invariant descriptors. These are used to detect objects around the gaze point in real-time (right).

3.2 Assistive Mode:

In the assistive mode, *video snippets* $\{U_k; k = 1..K\}$ can be used to provide automatic assistance for novice operators. First, the system needs to identify which object the person intends to use next, then the system would select a *video snippet*, from the potentially many snippets collected from multiple operators using the object one or more times, to be displayed to the novice operator.

We recognise objects based on the learnt views in an image patch around the gaze point using the scalable real-time texture-minimal object detector from [2]. By using the combination of fixed paths and a hierarchical hash table, the method is scalable, and can reliably detect objects at frame rate. The descriptor is affine-invariant, and the method is tolerant to a level of occlusion but is also view-dependant. Figure 3 shows the method learning (left column) and subsequently recognising (right column) objects from our experiments. Notice that the assistive mode does not require 3D tracking, and objects are recognised around the 2D gaze point.

Upon recognition, a *help snippet* is displayed to show how this object was previously used. From the possibly many *video snippets* featuring the TRO, collected in learning mode, 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. If the object changes state, the initial appearance is a good indicator of which video 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. As some objects can be gazed-at multiple times during the task, we employ temporal

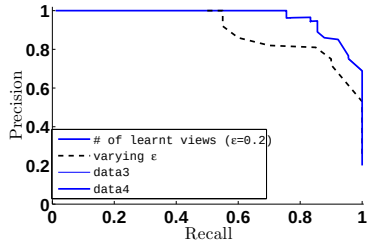

Fig. 4. Precision-Recall curve for discovering TROs as ϵ (Eq. 1) varies. For ϵ=0.2 metres, discovered objects are filtered based on the number of learnt views - at 76% recall, 100% precision was achieved.

ordering in choosing the *help snippet*. That is, for a given object we choose its snippets in order, starting first from all the first encounters of that object in all training sequences. When the same object is gazed-at again, a snippet from the set from all the second encounters in the training sequences is displayed and so on.

4 Experiments and Results

Setup & Dataset We use the dataset from [3] which was recorded using the wearable gaze tracker hardware [10]. After calibration, the scene images are synchronised with, if available, 2D gaze points. Twenty objects were ground-truthed, of which 5 are moveable objects.

To evaluate the ability of online clustering to find TROs, a 3D bounding box around fixations from one discovered TRO is compared to the manually labelled 3D bounding box on the map's point cloud. The PASCAL overlap criterion (adapted for 3D) of 20% is used for a true positive discovery, using the algorithm detailed above (parameter choices in Tab. 1). The main parameter for clustering is the threshold for 3D distances (ϵ). As ϵ (Eq. 1) varies, the number of discovered objects changes. The recall-precision results are shown in Fig. 4.

For $\epsilon = 0.2$, Tab. 2 shows the mean and standard deviation for the number of discovered, merged and split objects in one and all sequences. Since the clustering is online, different runs would result in a different set of discovered objects depending on the ordering of sequences. We run the experiments multiple times (5 times), starting from a different sequence, and record the results. As the table

Eq. 1		Algo. 1	
k	10 frames	δ	5 frames
ϵ	20 cm	w	150 × 150

Table 1. Parameter choices for object discovery

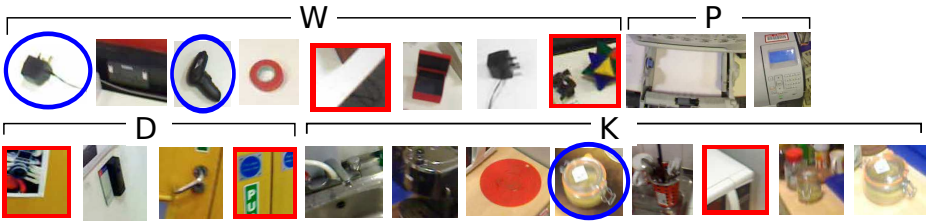

Fig. 5. [**Best seen in colour**] 22 objects were discovered within the four maps (W, D, P and K) listed from left to right by order of discovery. Out of these, 14 ground-truth objects are found (with 3 splits), and 5 are task-irrelevant (red rectangles). The 'cup' was missed at this iteration. Three objects were classified as 'moveable' (blue ellipses), out of possible 4. The charger was discovered twice and the sugar jar was discovered as three different objects.

shows, when trained using a single operator, precision of 79% is achieved alongside 86% of recall. When training on all operators, on average, 97% recall was achieved, with an increase in the total number of discovered objects. The number of false positives can be dropped by filtering for the number of learnt views, as operators observe TROs for longer than other objects in the scene (Fig. 4). The approach also separates fixed from moveable TROs. Recall that a TRO is 'moveable' if it is detected in different locations, using appearance matching. On average, 77% of TROs were correctly classified. The set of discovered objects from a single run is shown in Fig 5. Examples of learnt views for the discovered objects can be found in Fig 6.

Assistive Mode: To assess the ability of the approach to provide video guides, the approach is run using leave-one-out. For every operator, the *learning mode* is run on the remainder sequences to discover TROs and collect video guides. The appearance models of discovered TROs are then used to recognise objects in the 'left' sequence (i.e. not used for discovery), within patches around the 2D gaze. When an object is recognised, an insert is added indicating a suggestive way of how the object can be used. A *help snippet* h_t is displayed each time a new object is recognised. We showcase video help guides using inserts on a pre-

Op		total	gt TROs	merged	split	type
1	μ	21.6	17.2	0.7	1.5	-
	σ	1.5	0.9	0.5	0.8	-
All	μ	33.2	19.3	0.2	3.8	15.5
	σ	2.0	0.8	0.4	1.6	0.9

Table 2. At $\epsilon = 0.2$ metres, from one and all operators, the avg. (μ) and std dev. (σ) of the # of discovered TROs, the # of true TROs (ground-truth=20), the # of merged objects (ground-truthed as two separate objects), the # of split objects. For distinguishing moveable from fixed objects, the # of correctly classified objects.

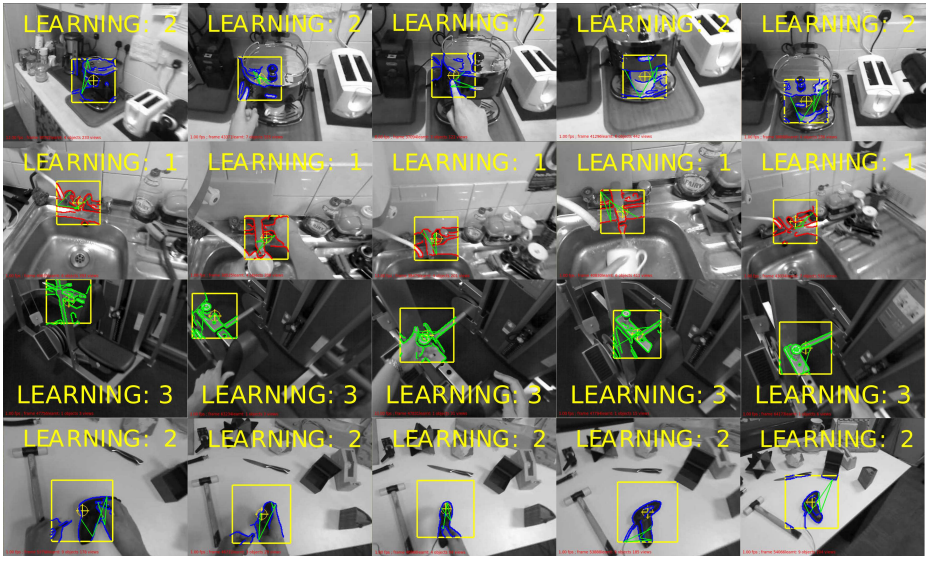

Fig. 6. Learnt views from training sequences of multiple users for a variety of objects: coffee machine, tap, seat adjustor and screwdriver.

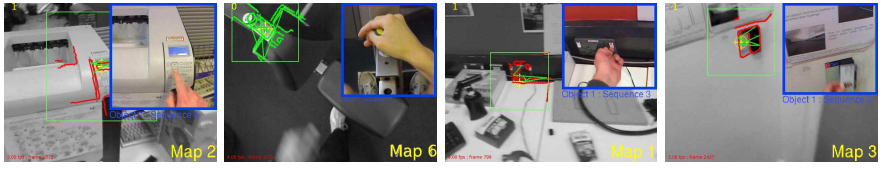

Fig. 7. In the assistive mode, when a TRO is detected, a video snippet is inserted showing the most relevant video guide based on the initial appearance.

recorded video. Figure 7 shows frames from the help videos and a full sequence is provided[2]. Recall that these inserts are *extracted, selected and shown* fully automatically. These could in principle be shown on a head-mounted display, but is not considered in this study. We believe this highlights the success and potentials of the work in this paper.

5 Conclusions and Future Work

In this paper we develop an online real-time system based on egocentric video with gaze. In its *learning mode*, the system discovers task-relevant objects and automatically collects video snippets from multiple users on how they used the discovered object. In the *assistive mode*, video guides are shown on how objects have been used before, triggered by recognising the gazed-at object. This could

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

be useful to novice users exploring the same environment and objects. This paper explains a complete online prototype, and future work aims to evaluate the benefits of the assistive mode on the performance of novice users.

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.: Cognitive workflow capturing and rendering with on- body sensor networks (cognito). German Research Center for Artificial Intelligence, DFKI Research Reports (RR) (2013)
2. Damen, D., Bunnun, P., Calway, A., Mayol-Cuevas, W.: Real-time learning and detection of 3D texture-less objects: A scalable approach. In: British Machine Vision Conference (BMVC) (2012)
3. Damen, D., Leelasawassuk, T., Haines, O., Calway, A., Mayol-Cuevas, W.: You-do, i-learn: Discovering task relevant objects and their modes of interaction from multi-user egocentric video. In: British Machine Vision Conference (BMVC) (2014)
4. De Beugher, S., Ichiche, Y., Brone, G., Geodeme, T.: Automatic analysis of eye-tracking data using object detection algorithms. In: Workshop on Perasive Eye Traking and Mobile Eye-based Interaction (PETMEI) (2012)
5. Fathi, A., Li, Y., Rehg, J.: Learning to recognize daily actions using gaze. In: European Conference on Computer Vision (ECCV) (2012)
6. Fathi, A., Rehg, J.: Modeling actions through state changes. In: Computer Vision and Pattern Recognition (CVPR) (2013)
7. Fathi, A., Ren, X., Rehg, J.: Learning to recognise objects in egocentric activities. In: Computer Vision and Pattern Recognition (CVPR) (2011)
8. Henderson, J.: Human gaze control during real-world scene perception. Trends in Cognitive Sciences 7(11) (2003)
9. Klein, G., Murray, D.: Parallel Tracking and Mapping for Small AR Workspaces. In: Int. Sym. on Mixed and Augmented Reality (ISMAR) (2007)
10. Laboratories, A.S.: Mobile Eye-XG, `http://www.asleyetracking.com/`
11. Land, M.: Eye movements and the control of actions in everyday life. Progress in Retinal and Eye Research (2006)
12. Lee, Y., Ghosh, J., Grauman, K.: Discovering important people and objects for egocentric video summarization. In: Computer Vision and Pattern Recognition (CVPR) (2012)
13. Leelasawassuk, T., Mayol-Cuevas, W.: 3D from looking: Using wearable gaze tracking for hands-free and feedback-free object modelling. In: Int. Sym. on Wearable Computers (ISWC) (2013)
14. Li, Y., Fathi, A., Rehg, J.: Learning to predict gaze in egocentric video. In: Int. Conf. on Computer Vision (ICCV) (2013)
15. Lu, Z., Grauman, K.: Story-driven summarization for egocentric video. In: Computer Vision and Pattern Recognition (CVPR) (2013)
16. Petersen, N., Stricker, D.: Learning task structure from video examples for workflow tracking and authoring. In: International Symposium on Mixed and Augmented Reality (ISMAR) (2012)
17. Ren, X., Gu, C.: Figure-ground segmentation improves handled object recognition in egocentric video. In: Computer Vision and Pattern Recognition (CVPR) (2010)

18. Rosten, E., Reitmayr, G., Drummond, T.: Real-time video annotations for augmented reality. Advances in Visual Computing (2005)
19. Salvucci, D., Goldberg, J.: Identifying fixations and saccades in eye-tracking protocols. In: Sym. on Eye Tracking Research & Applications (2000)
20. Sun, L., Klank, U., Beetz, M.: EyeWatchMe - 3D hand and object tracking for inside out activity analysis. In: Computer Vision and Pattern Recognition Workshop (CVPRW) (2009)
21. Sundaram, S., Mayol-Cuevas, W.: What are we doing here? egocentric activity recognition on the move for contextual mapping. In: Int. Conf. on Robotics and Automation (ICRA) (2012)
22. Takemura, K., Kohashi, Y., Suenaga, T., Takamatsu, J., Ogasawara, T.: Estimating 3D point-of-regard and visualizing gaze trajectories under natural head movements. In: Sym. on Eye-Tracking Research & Applications (ETRA) (2010)
23. Wade, N., Tatler, B.: The moving tablet of the eye: the origins of modern eye movement research. Oxford University Press (2005)