Personal Object Discovery in First-Person Videos

Cewu Lu, *Member, IEEE*, Renjie Liao, and Jiaya Jia, *Senior Member, IEEE*

***Abstract*— People know and care for personal objects, which can be different for individuals. Automatically discovering personal objects is thus of great practical importance. We, in this paper, pursue this task with wearable cameras based on the common sense that personal objects generally company us in various scenes. With this clue, we exploit a new object-scene distribution for robust detection. Two technical challenges involved in estimating this distribution, i.e., scene extraction and unsupervised object discovery, are tackled. For scene extraction, we learn the latent representation instead of simply selecting a few frames from the videos. In object discovery, we build an interaction model to select frame-level objects and use nonparametric Bayesian clustering. Experiments verify the usefulness of our approach.**

***Index Terms*— Object discovery, object detection, scene understanding, first-person vision and wearable camera.**

I. Introduction

PERSONALLY used objects are individually characterized, important in our daily life. Their footage provides vital information about location, event, activity, and even individual personality and interests. In the era of social network, understanding personal objects and their relationship with different scenes avail a wide range of community service and even new business for individual-oriented commercials. For example, putting a medicine box to the personally-used object set could indicate the potential need of clinical service. Removal of a watch from the list, on the contrary, may suggest that it is lost or damaged. Therefore, personal object information could be a great feature for recommendation systems to establish efficacious links between companies and customers.

Albeit valuable, detecting and finding personal objects automatically from videos is still difficult, primarily due to its tight link to first-person view. In the intensively-studied third-person videos such as the surveillance ones, the definition of personal objects is vague and their detection is thus challenging. The difficulty also stems from distinguishing between general and personal objects.

Fig. 1. A hand-touching electronic lock is not a personal object.

The prevalence of wearable cameras brings forth the unique essentiality of wearable computing. In contrast to the conventional "third-person" video paradigm, wearable cameras record individual activities from another perspective. First-person perspective videos (simplified as first-person videos in what follows) contain egocentric cues. In this paper, we focus on personally and frequently used object discovery for camera wearers.

Even with a first-person video captured by a wearable camera, personal object discovery is still nontrivial. Personal objects vary from person to person and may have different appearance and structure properties. The hint that personal objects appear in the interaction process with human does not always hold. For example, in Fig. 1, the electronic locker interacted with the hands is not a personal object.

We study the nature of this problem and present an effective solution. It is based on the fact that personal objects, such as bags, wallets and mobile phones, company us in different events and scenes. So the personal-object probability can be high if an object appears in *sufficiently many* locations in a period of time. With this clue, we define a personal object following a few simple and conservative rules. That is, such an object should be with generally acceptable appearance, movable and, most importantly, be held or touched frequently in different scenes. Accordingly, we propose a *object-scene distribution* to measure interacted-object appearing frequencies in different scenes. The final result from first-person videos is a scored object list based on how close an object is to the camera recorder, based on a rule in information theory.

Our method boils down measuring object appearing frequencies in different scenes to two sub-problems, i.e., scene extraction and interacted-object discovery, which are

Manuscript received September 14, 2014; revised February 27, 2015, June 19, 2015, and July 27, 2015; accepted August 19, 2015. Date of publication October 7, 2015; date of current version October 27, 2015. This work was supported in part the Research Grants Council, Hong Kong, under Project 413113, and in part by the National Science Foundation, China, under Grant 61133009 and Grant 61472245. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jianfei Cai.

C. Lu is with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200030, China (e-mail: lu-cw@cs.sjtu.edu.cn).

R. Liao is with The Chinese University of Hong Kong, Hong Kong (e-mail: lrjconan@gmail.com).

J. Jia is with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong (e-mail: leojia@cse.cuhk.edu.hk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2015.2487868

tackled respectively. Scene extraction needs a collection of well-characterized data for training and comparison. However, this data cannot be found directly in video frames when the camera moves arbitrarily and is not deliberately path-controlled. We thus learn the latent representative scenes instead of simply selecting a few frames from the videos. Then in interacted-object discovery, supervised algorithms could be less optimal, as they require to include as many objects as possible in the training data. In our method, we build an interaction model to select the frame-level objects. Unsupervised object discovery is implemented using a nonparametric Bayesian clustering technique. Our scheme can discover and locate objects that appear in disjoint frame segments in the input video.

Our main contribution in this paper is as follows.

1) We tackle the personal-object-discovery problem based on useful clues.
2) We propose the simple and yet effective object-scene distribution to differentiate between personal and other objects.
3) Latent scene learning is used for scene extraction in first-person videos.
4) We provide a reliable unsupervised interacted-object discovery framework for first-person videos.

We clarify due to the complicated nature of first-person videos, our solution inevitably has several limitations. We describe them later along with concluding remarks.

II. Related Work

We review image/video object detection and discovering work. Research about first-person video processing and understanding is also discussed.

A. Object-Related Methods

Since "object" is fundamental for many tasks, object-related topics were widely discussed, such as inpainting [1], segmentation [2], [3], tracking [4], video coding [5], and saliency detection [6]. We mainly investigate object discovery and detection.

Early research [7], [8] tends to achieve object detection by filtering. Luo and Crandall [9] developed a color object detection algorithm by building a spatial-color joint probability function incorporating the color edge co-occurrence histogram and perceptual color naming. In [10], binary partition trees (BPTs) were used that involve hierarchical region-based representation for object detection. It reduces search space by tree construction, similarity criterion development, and node extension. Sun and Lam [11] introduced a forward feature-selection technique combining a coarse-to-fine learning scheme to construct an efficient classifier while yielding good performance.

Guo and Wang [12] tackled the domain adaptation problem of image object detection by employing kernel analysis. The domain adaptive input-output kernel learning (DA-IOKL) algorithm addresses the feature distribution change issue. In [13], an effective deformable part model was proposed making use of latent SVM. Rather than working on images, detecting or discovering objects in videos is also common. Li *et al*. [14] proposed a Bayesian framework incorporating spectral, spatial, and temporal features to describe the background appearance. A learning method adaptive to gradual and sudden "once-off" background change benefits foreground object detection. In [15], abandoned object detection in video was studied. In [16], a method for detecting primary objects was proposed. The key contribution is to design a self-adaptive saliency map fusion method by learning the reliability of saliency maps from training data.

For video object discovery and video segmentation, a topic model by incorporating a word co-occurrence prior into LDA was proposed [17] for efficient discovery of topical video objects. In [18], video object discovery was achieved using a co-segmentation scheme where an optimization-based method identifies relevant frames containing the target object. In [19], Faktor and Michal solved the foreground/background segmentation of unconstrained videos. Saliency is taken to initialize seed and segmentation is iteratively corrected by consensus voting. Thematic objects in videos were studied in [20]. A bottom-up approach gradually prunes uncommon local visual primitives for finally locating the thematic objects.

These methods are for general object detection and discovery. As personal objects are special, our problem definition and solution are inherently different.

B. First-Person Perspective Methods

Pioneer research based on first-person videos captured by wearable cameras was presented in [21] and [22]. In [23], an automatic approach was proposed to structure and produce video summarization. Goto and Tanaka [24] presented a text tracking and detection system to facilitate blind people. Han *et al*. [25] developed a real-time face detection, recognition, and identification system. Based on a normalized optical flow field, a bottom-up segmentation algorithm was proposed [26] to separate foreground objects in ego-centric videos. In [27], camera wearer activity recognition was discussed. Combining optical flow features and several classifiers, this method can achieve good recognition results. In [28], a random-walk based influence metric, which reflects how objects contribute to the progression of events, was exploited to define a new objective for egocentric video summarization.

In [29], Lee et al. learned a regressor to predict important objects. They also use human interaction cues such as gaze. This method is unlike ours in several ways – the goal is not personal object discovery and important objects were detected regarding a short time frame. To find personal objects, more frames across scenes need to be processed.

In short, different from previous methods, we solve a fundamental problem in personal object discovery, which relies on the object-scene distribution. We provide new features for effectively representing both scene and objects in egocentric videos. An unsupervised learning method is also proposed to find personal objects, which saves time on collecting large labeled data.

Fig. 2. Six representative frames in a first-person video. Background changes rapidly.

III. Our Approach

We start with the introduction of the object-scene distribution (OSD), which involves methods to tackle scene extraction and wearer-interacted object discovery. Personal object scores derived from OSD in information theory are also discussed. In what follows, we abbreviate "wearable camera wearer" to "wearer".

A. Object-Scene Distribution

We denote by $\{I_1, \ldots, I_T\}$ the frame set of a first-person video, $\{S_1, \ldots, S_m\}$ the m different scenes, and by $\{O_1, \ldots, O_n\}$ the n objects that have interaction with the wearer in the video, which appear in multiple scenes. We also define x and y as the scene and object appeared respectively in videos. Thus, the object-scene distribution of object O_i is $P(x|y = O_i) = \{P(x = S_1|y = O_i), \ldots, P(x = S_j|y = O_i)\}$, where $P(x = S_j|y = O_i)$ is the probability of object O_i appearing in S_j. It can be further expressed as

$$P(x = S_j|y = O_i) = \frac{P(x = S_j, y = O_i)}{P(y = O_i)}. \tag{1}$$

We introduce three indicator functions $\mathbf{1}[x_t = S_j, y_t = O_i]$, $\mathbf{1}[y_t = O_i]$, and $\mathbf{1}[x_t = S_j]$, which represent three events: "the t^{th} frame falls into scene S_j and contains O_i", "the t^{th} frame contains O_i", and "the t^{th} frame falls into scene S_j" respectively. When the event happens, the indicator function outputs 1; otherwise it outputs 0.

An event's probability is the limit of its relative frequency in a large number of trials. With sufficient large frame number T in first-person videos, $P(x = S_j, y = O_i)$ and $P(y = O_i)$ can be written as

$$\begin{aligned} P(x = S_j, y = O_i) &= \frac{1}{T}\sum_{t=1}^{T}\mathbf{1}[x_t = S_j, y_t = O_i] \\ P(y = O_i) &= \frac{1}{T}\sum_{t=1}^{T}\mathbf{1}[y_t = O_i]. \end{aligned} \tag{2}$$

We denote the frame set that contains object O_i as Ω_i to build the link between object O_i and the scene. $\sum_{t=1}^{T}\mathbf{1}[y_t = O_i]$ thus can be $|\Omega_i|$, where $|\Omega_i|$ is the size of Ω_i. $P(x = S_j, y = O_i)$ is simplified to

$$P(x = S_j, y = O_i) = \frac{1}{T}\sum_{t\in\Omega_i}\mathbf{1}[x_t = S_j]. \tag{3}$$

In our model, considering the special properties of first-person videos, $\mathbf{1}[x_t = S_j]$ is replaced by $\mathbf{h}[x_t = S_j]$ indicating the likelihood whether frame I_t falls into scene S_j or not. This is because it is difficult to say a frame is exactly a particular scene or not. As the camera moves freely, most frames describe transformation and fall into different scenes partly. Therefore, we employ $\mathbf{h}[x_t = S_j]$ instead of $\mathbf{1}[x_t = S_j]$ to measure how significant frame I_t falls into scene S_i. Eq. (1) is rewritten as

$$P(x = S_j|y = O_i) \propto \frac{1}{|\Omega_i|}\sum_{t\in\Omega_i}\mathbf{h}[x_t = S_j]. \tag{4}$$

In this framework, two challenges are computing $\mathbf{h}[x_t = S_j]$ and set Ω_i respectively. The second challenge involves discovering wearer-interacted objects and obtaining objects' temporal and spatial location. In what follows, we address these challenges.

B. Latent Scene Extraction

We estimate $\mathbf{h}[x_t = S_j]$ by considering special properties of first-person videos. As the camera can move freely, it is over-simplified to define some frames as semantic "scenes". Even if one manually labels a few frames, view transformation could exist, as illustrated in Fig. 2. This brings special difficulty to identify which frames are "scenes".

As such, we do not select frames as representative, but instead regard each frame as a mixture of a small number of scenes. We denote the features respectively for frame I_t and m scenes as x_t and $\{\mathbf{d}_1, \ldots, \mathbf{d}_m\}$. Feature extraction will be detailed later. Modeling the frame $\mathbf{z}_t$ as a linear combination of features from a few scenes thus yields

$$\mathbf{z}_t = \mathbf{D}\boldsymbol{\beta}_t \text{ s.t. } \|\boldsymbol{\beta}_t\|_1 < \xi, \tag{5}$$

where $\mathbf{D} \triangleq \{\mathbf{d}_1, \ldots, \mathbf{d}_m\}$. The constraint $\|\boldsymbol{\beta}_t\|_1 < \xi$ models sparse representation on coefficients $\boldsymbol{\beta}$. It stems from the fact that one frame I_t can at most be a mixture of a small number of scenes in general.

We now discuss how to obtain the m scene features $\{\mathbf{d}_1, \ldots, \mathbf{d}_m\}$. We take $\{\mathbf{d}_1, \ldots, \mathbf{d}_m\}$ as *latent scene variables*. We learn them by solving

$$\begin{aligned} &\min_{\mathbf{D}\in\mathcal{D},\mathcal{B}} \frac{1}{T}\sum_{t=1}^{T}\{\|\mathbf{D}\boldsymbol{\beta}_t - \mathbf{z}_t\|_2^2 + \lambda\|\boldsymbol{\beta}_t\|_1\} \\ &\text{s.t. } \mathcal{D} \triangleq \{\mathbf{d} \in \mathcal{D} |\ \|\mathbf{d}\|_2^2 = 1\} \end{aligned} \tag{6}$$

where $\mathcal{B} \triangleq \{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_T\}$ and $\|\cdot\|_1$ is the sparsity inducing norm. Eq. (6) models sparse dictionary learning. Its solver updates $\mathcal{B}$ and $\mathbf{D}$ iteratively in two steps. We generate a random matrix as the initial dictionary with L_2 unit column. Because each atom of the dictionary is an abstract feature representation of a latent scene, our method is effective to represent complex video structures.

(a) (b)

Fig. 3. (a) Two similar-scene images yield alike LLC codes. (b) Two different scenes produce dissimilar LLC codes.

Given a learned latent scene dictionary, we sparsely code each frame by locality-constrained linear coding (LLC) [30] with the criterion

$$\min_{\boldsymbol{\beta}} \{\|\mathbf{z}-\mathbf{D}\boldsymbol{\beta}\|_2^2+\lambda\sum_{j=1}^{m}(\rho_j\beta_j)^2\}$$
$$s.t\ \sum_{j=1}^{m}\beta_j=1 \tag{7}$$

where β_j is the j^{th} element of vector $\boldsymbol{\beta}$ and ρ_j models similarity between each basis vector and the input descriptor $\mathbf{z}$. It is written as

$$\rho_j=\exp(\frac{1}{\sigma}dist(\mathbf{d}_j,\mathbf{z})), \tag{8}$$

where $dist(\mathbf{d}_j,\mathbf{z})$ is the Euclidean distance between $\mathbf{z}$ and $\mathbf{d}_j$. Compared to the ℓ_1 sparse coding (SC) that might select quite different basis vectors for similar inputs to favor sparsity, LLC guarantees that these inputs obtain similar codes, more desirable in representation.

We normalize the LLC results for each frame and take $\boldsymbol{\beta}_{t,j}$ – the j^{th} element of $\boldsymbol{\beta}_t$ – as the likelihood that I_t falls in latent scene S_j. That is, $\mathbf{h}[x_t=S_j]=\boldsymbol{\beta}_{t,j}$. LLC ensures that similar frames represent corresponding latent scenes (see Fig. 3). We illustrate the representation in Fig. 4.

1) Scene Features: We use several low-level features to represent each frame. They include pyramid of histograms of oriented gradients (PHOG) [31] which produces a 680-dimension feature vector, a bag-of-features (BoF) descriptor and a color histogram. BoF is created using dense SIFT and a vocabulary of dimension 1200. The vocabulary was computed using k-means. The color histogram is made by quantizing each color channel in 8 levels, resulting in a 512-element feature vector. We have used the code available online for BoF and PHOG. Before combining these feature vectors, each of them is normalized. The dimension of the final feature vector is reduced to 256 using PCA. We solve Eq. (6) using standard dictionary learning [32]. The dictionary is with scale 256×512.

C. Wearer-Interacted Object Discovery

In this section, we describe the procedure to detect and locate candidates of wearer-interacted objects $\{O_1,\ldots,O_n\}$. Since personal objects cannot be fully described by appearance, we turn to object discovery in an unsupervised way.

In the first place, we generate object-like regions by CPMC [33]. Given an image, CPMC produces a large number

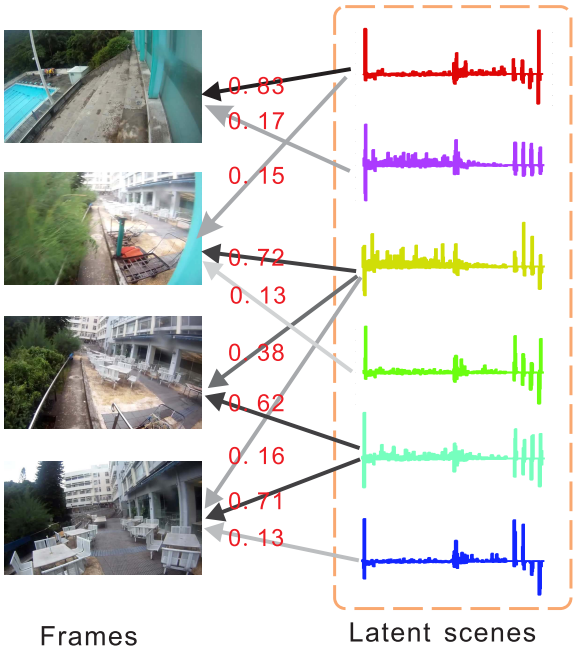


Fig. 4. Latent scene combination in frames. We visualize latent scenes using their corresponding dictionary atoms (obtained by solving Eq. (6)). The intensity of arrows indicates coding coefficients (Coding value 1 makes the arrow black).

of regions with object scores. A set of object-like regions produced by CPMC are illustrated in Fig. 5. We select 100 regions with the highest scores per frame, resulting in $100T$ object candidates in total, where T is the number of frames. We found empirically most personal objects are included in these many regions.

Since the regions are clutter, it is necessary to find correct wearer-interacted objects from them. We build an interaction model to reject a large number of regions without interaction with the wearer, which greatly reduces the size of regions. Then we extract mid-level features from remaining candidates and cluster them.

1) Interaction Model: Our interaction model is implemented in a two-step rejection using hand and Gaze priors.

a) Hand prior rejection: An important feature we make use of is frequent hand-interaction with objects. We resort to the statistical color model in [34] to detect human skin based on an EM algorithm. We take super-pixels with the mean of likelihood (also called hand scores) larger than 0.6 as hands. Euclidean distance of each region's centroid to closest hand is computed in each frame. If it is larger than 30% of the region size, we reject it. Note it is only a rejection process and does not matter if some pixels are wrongly detected as hand skin.

b) Gaze prior rejection: In general, human eye focus is more likely to be on personal objects than other things (e.g. trash bin). This condition helps reject a few unimportant objects. We track and backtrack each region using the method of [35] until the region centroid moves more than 5% of the frame size. The track and backtrack time is recorded as t_1 and t_2. If t_1+t_2 is reasonably long in our experiments, we keep the region.

It is noted that thresholds used in these rejection conditions are both loose, in order to avoid rejecting correct objects.

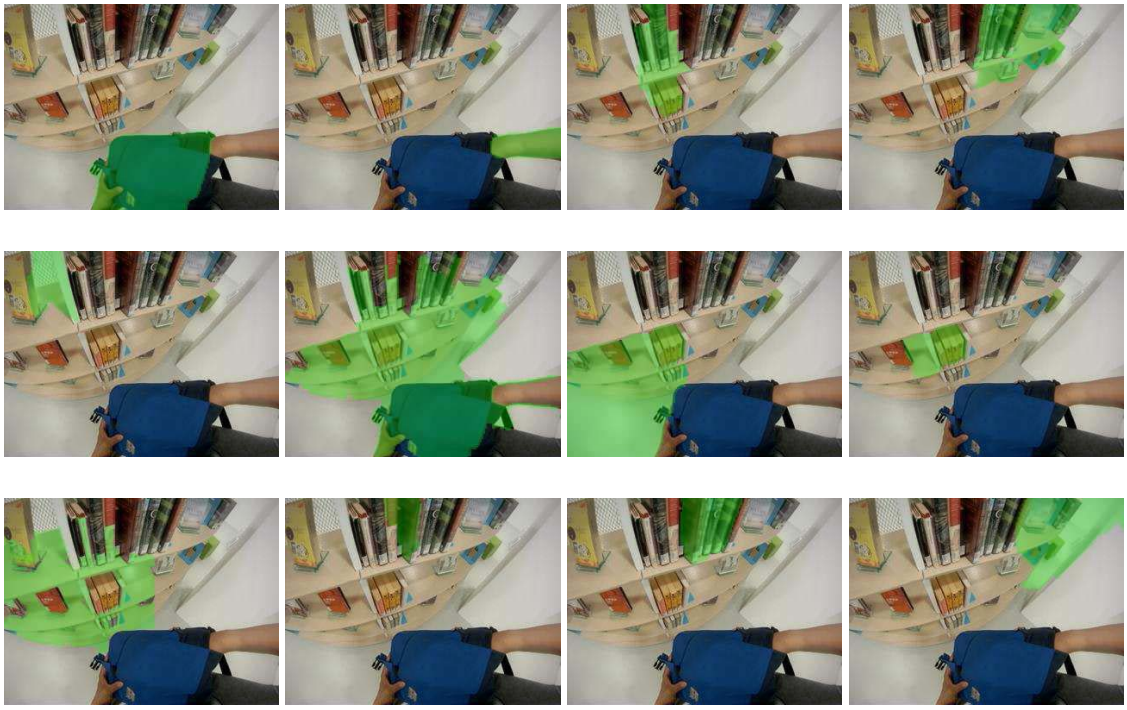


Fig. 5. Several object-like regions produced by CPMC.

We also implement face detection, which is very accurate now, to remove nearly all face regions in these results.

2) Object Clustering: Given remaining region candidates, we construct object clusters via a few middle-level cues and egocentric features, which form a 261-dimension feature vector for every candidate region. 256 dimensions are for Bag-of-Color (BoC), 2 for centroid, 1 for object-like appearance, 1 for shape descriptor and 1 for region size. They are detailed below.

a) Bag-of-color (BoC): To model the color distribution of our interested objects, we exploit Bag-of-Color (BoC) representation [36] because it outperforms a few other signatures in image retrieval. A palette of 256 different colors is learned from 10000 random Flicker images. We then project every pixel in the region onto the color palette according to their nearest color. Post-process, including inverse document frequency (IDF), power law transformation and ℓ_1 normalization [36], is applied to produce the final BoC features.

b) Shape descriptor: To incorporate shape information, we make use of the convex hull of regions and calculate the ratio between the areas of convex hull and the region. This feature is useful to indicate how regular a region is.

c) Gestalt score: As stated in [33], scores output by the learned ranking function from CPMC reflect the Gestalt cues, which indicate the likelihood of a region being an object. It helps discriminate between objects and others.

d) Region size: We also take region size into account. For numerical stability, absolute region size is replaced by the ratio between region and image sizes.

e) Location: To model the location of each object candidate, we compute centroids of regions and normalize them with respect to the image.

After obtaining these features for all object candidates, we use an unsupervised clustering algorithm in the $\mathbb{R}^{261}$ feature space to identify repeating objects across all frames. Since the number of objects contained in the video is unknown, nonparametric Bayesian clustering by Dirichlet Process Mixture Model (DPMM) [37], is adopted. DPMM is a hierarchical Bayesian model and its stick-breaking construction [38] is given below.

1, Draw $v_i \sim Beta(1, \alpha), \quad i = \{1, 2, \ldots, \infty\}$
2, Draw $\eta_i \sim G_0, \quad i = \{1, 2, \ldots, \infty\}$
3, For the n-th data point:
   (a) Draw $z_n \sim Mult(\pi(\mathbf{v}))$.
   (b) Draw $x_n \sim p(x_n|\eta_{z_n})$.

Note that sampling follows i.i.d., where $\mathbf{v} = [v_1, v_2, \ldots, v_\infty]$ is the set of infinite Beta-distributed random variables for constructing stick proportions $\pi(\mathbf{v}) = [\pi_1(\mathbf{v}), \pi_2(\mathbf{v}), \ldots, \pi_\infty(\mathbf{v})]$. Specifically, $\pi_i(\mathbf{v}) = v_i \prod_{j=1}^{i-1}(1-v_j)$. η_i are hyper-parameters of the i-th mixture component. G_0 is a base measure and α is the concentration parameter. In our case, the feature of each object candidate is x_n and one mixture component $p(x_n|\eta_{z_n})$ takes the form of a multivariate Gaussian distribution. G_0 is the Gaussian-Wishart distribution, which is the conjugate prior for the multivariate Gaussian component. Since the samples from the Dirichlet Process (DP) [39] are discrete, data generated from DPMM can be partitioned naturally according to the distinct values of the sampled parameters. The number of mixture components is thus random and grows with new observed data, making DPMM a flexible and powerful model.

To estimate the posterior distribution in DPMM, truncated stick-breaking representation and mean-field variational inference [40] are applied. For the DPMM setting, grid search

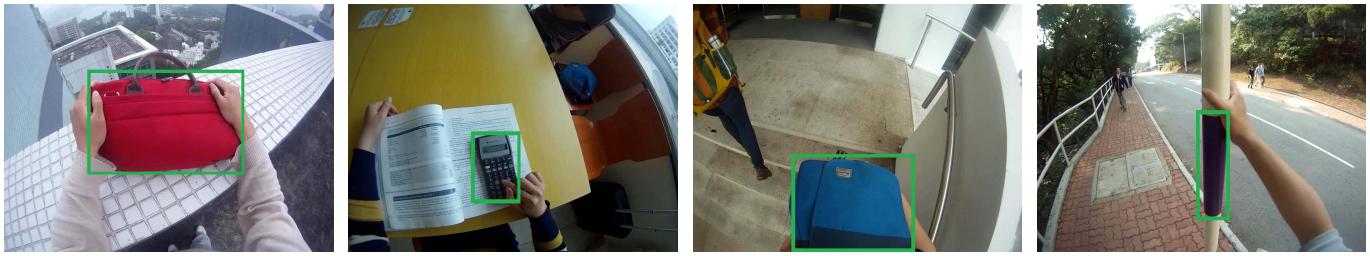


Fig. 6. Four resulting wearer-interacted objects. It includes personal and non-personal objects.

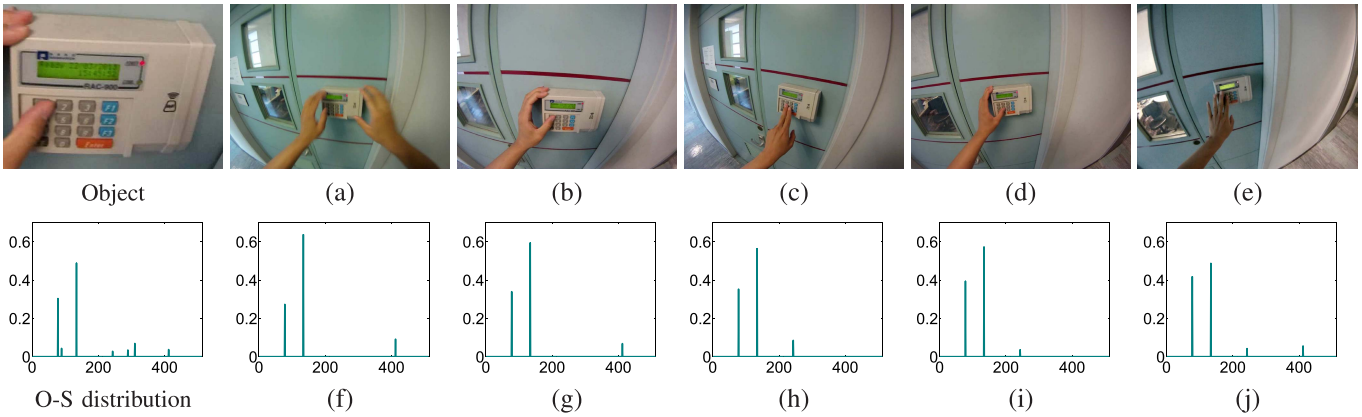


Fig. 7. Non-personal Object. (a)-(e) are five representative frames containing an electronic lock. Their corresponding sparse codes (likelihood) in 512 latent scenes are shown in (f)-(j). "O-S distribution" is referred to as $P(S|O)$.

is performed in the interval of 1-100 and with step size 5 to obtain the best concentration parameter α. The truncation level in our case is empirically set to 500. Moreover, to avoid sensitivity of initialization, we repeat variational inference with hyper-parameters learned from last run. For all video clips, DPMM automatically finds about 100 to 150 clusters. The results from this step are call *wearer-interacted objects*. A few examples are shown in Fig. 6.

3) Analysis: Why can this two-step framework discover wearer-interacted object effectively? First, interaction model outputs high quality frame-level candidate regions. Without it, common regions, such as sky, road, and tree, will be mistakenly clustered. Note a large number of wrong candidate regions could lead to poor clustering performance.

Second, object clustering can find commonness of one object when it appears in disconnected frames, complying to the common knowledge that general personal objects tend to present at similar viewing directions and distances with respect to the wearable camera. This property makes the mid-level features in different frames comparable.

D. Personal Object Score

Given $\mathbf{h}[x_t = S_j]$ obtained as sparse code $\boldsymbol{\beta}_{t,j}$ in Sec. III-B and $\boldsymbol{\Omega}_i$, we calculate the object-scene distribution $P(x|y = O_i) = \{P(x = S_1|y = O_i), \ldots, P(x = S_j|y = O_i)\}$ using Eq. (4) and normalize it to a unit vector ($\sum_j P(x = S_j|y = O_i) = 1$). Before that, we truncate values smaller than 0.1 in $\mathbf{h}[x_t = S_j]$ to 0, to only cope with dominant LLC codes.

Since personal objects generally accompany us from scene to scene, their occurrence frequencies should be higher than those without much interaction. According to information theory, sparsity of the occurrence distribution is measured by entropy

$$H[P(x|y = O_i)] = -\sum_j P(x = S_j|y = O_i) \ln[P(x = S_j|y = O_i)]. \quad (9)$$

A denser distribution leads to a smaller entropy value. For example, the digital lock in Fig. 7, which is not a personal object, finds entropy 0.5143 because it appears only in a specific one or two door views, while the value of bag in Fig. 8 is 3.7497, as it is used in much more scenes.

1) Relative Entropy: Entropy in Eq. (9) treats all scenes equally important. But in many cases objects appear in common (important) scenes should be assigned higher weights compared to those appear in rare scenes. For example, a personal laptop – very important object to the camera wearer – is possibly used in apartment, office and meeting rooms.

We count the frequencies of different scenes to measure their importance. Scene distribution is expressed as

$$P(x = S_j) = \frac{1}{Z} \sum_t \mathbf{h}[x_t = S_j], \quad (10)$$

where Z is a normalization factor to make $\sum_j P(x = S_j) = 1$. Given the wearer-scene distribution, we introduce the relative

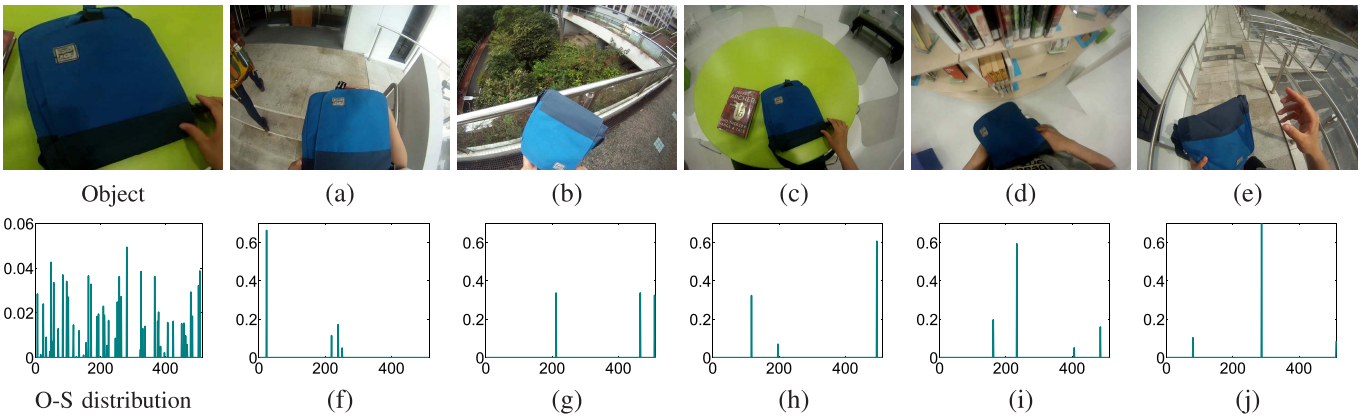


Fig. 8. Personal Objects. (a)-(e) are five representative frames containing a bag. Their corresponding sparse codes (likelihood) in 512 latent scenes are shown in (f)-(j). "O-S distribution" refers to $P(S|O)$.

Fig. 9. Personal object scores by Eq. (12).

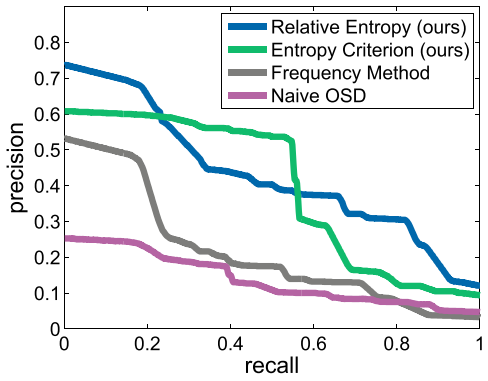


Fig. 10. Precision-recall curves on our dataset.

entropy in a form of Kullback-Leibler divergence

$$KL(P(x = S_j|y = O_i)||P(x = S_j)) = \sum_j P(x = S_j|y = O_i) \ln[\frac{P(x = S_j|y = O_i)}{P(x = S_j)}]. \quad (11)$$

In comparison to entropy in Eq. (9), the relative one considers scene importance. $KL(P(x = S_j|y = O_i)||P(x = S_j))$ ranges in $[0, +\infty]$. To line up with scoring measure in range $[0, 1]$, we apply

$$\text{score}(O_i) = \exp(-\eta KL[P(x = S_j|y = O_i)||P(x = S_j)]), \quad (12)$$

where η is set to 3, and personal object receive large values in this measure in general. Personal objects are more likely to receive high scores.

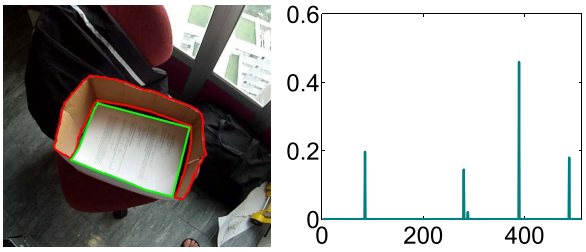


Fig. 11. A falsely detected wearer-interacted object. (a) Representative frame. The regions with the red and green boundaries are detected as skin and wearer-interacted object respectively. (b) Object-scene distribution with personal object score 0.13.

a) Discussion: We discuss the relationship between Eqs. (9) and (11). We rewrite Eq. (11) as

$$\text{score}(O_i) = \frac{\exp(-\eta \sum_j P(x = S_j|y = O_i) \ln[P(x = S_j|y = O_i)])}{\exp(-\eta \sum_j P(x = S_j|y = O_i) \ln[P(x = S_j)])} = \frac{\exp(-\eta H[P(x|y = O_i)])}{\exp(-\eta \sum_j P(x = S_j|y = O_i) \ln[P(x = S_j)])}. \quad (13)$$

The numerator of Eq. (13) is exactly Eq. (9) with an exponent operation. Compared to Eq. (9), Eq. (11) takes the scene importance prior into consideration regarding the additional term $\exp(-\eta \sum_j P(x = S_j|y = O_i) \ln[P(x = S_j)])$. Obviously, if an object appears in important scene S_v and $P(x = S_v)$ is large, the score of Eq. (12) becomes large.

A few examples are demonstrated in Fig. 9. We compare the results using relative and absolute entropy measures in the next Section.

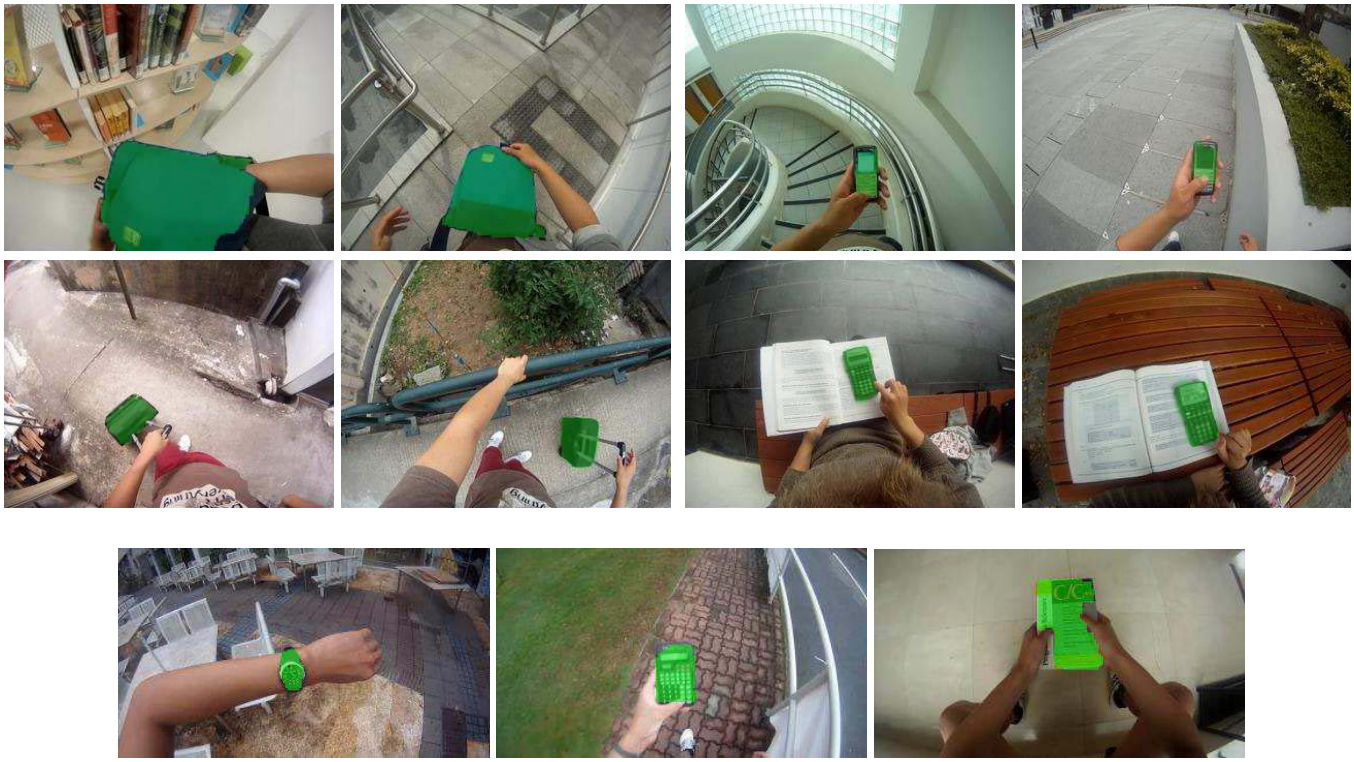


Fig. 12. Representative frames with personal objects detected (masked in green).

IV. Experiments

We collect 14 videos clips recorded by 7 subjects (4 males and 3 females) using a head-mounted GoPro Hero 2 camera with resolution 1280×720 in 30fps. All subjects are college students from different disciplines. In order to collect realistic and individual-specific data, we did not give them instructions, and encouraged them to behave as usual. The activities involve discussing, walking, reading, and dining. The scenes subjects captured are mainly in campus including library, dormitory, classroom, restaurant and campus path.

Our captured first-person videos are 11+ hours long in total. For data compression, we extract the 10^{th}, 20^{th} and 30^{th} frames in each second; all frames are resized to 800×450, sufficient for object and scene representation.

To annotate data, the question about what objects are used is answered by all subjects. We also asked the 7 subjects to annotate their videos by drawing a tight bounding box surrounding each personal object and adjusting it for every 10 frames. Finally, 25 objects are annotated including mobile phone, wallet, bag, cup, book, laptop, ID card, keyboard, to name a few.

A. Evaluation and Results

We evaluate personal object scores. We output detected wearer-interacted objects whose scores are higher than a threshold τ. For each ground truth bounding box, if an output personal object region largely overlaps with it, the object in that frame is regarded as correctly detected. The overlapping criterion with regard to the ground truth bounding box R_1 and detected region R_2 is defined as $(R_1 \cap R_2)/(R_1 \cup R_2) > 0.6$.

TABLE I
AUC of Different Methods

method	AUC
Frequency Method	14.36
Naive OSD	20.56
Proposed Entropy Scores by Eq. (9)	38.57
Proposed Relative Entropy Score by Eq. (12)	42.14

We vary threshold τ and count how many ground truth bounding boxes are detected to produce the precision-recall curve.

1) Baseline Methods: We design two baseline methods to compare with. The first one ignores the object-scene distribution, and instead counts the appearing times of objects as their scores. We denote it as the *frequency method*. In the second baseline method, we employ a naive technique to obtain the object-scene distribution. Instead of using $\mathbf{h}[x_t = S_j]$, we set $\mathbf{1}[x_t = S_j]$ to estimate scene property of a frame. That is, we select 512 frames to represent the scene where all frames are clustered into 500 groups using GMM. Each frame is assigned to a scene depending on the closest scene cluster using the nearest neighbor. For wearer-interacted object discovery, we employ K-means clustering instead of DPMM with cluster number 100. We call this method *naive OSD*.

The first baseline method is to verify the effectiveness of object-scene distribution. The second baseline tests steps in sparse latent scene extraction and DPMM object discovery. We report our performance using entropy scores in Eq. (9) and relative entropy scores in Eq. (12).

2) Results: We plot precision-recall curves of the two baseline methods and ours on the datasets in Fig. 10,

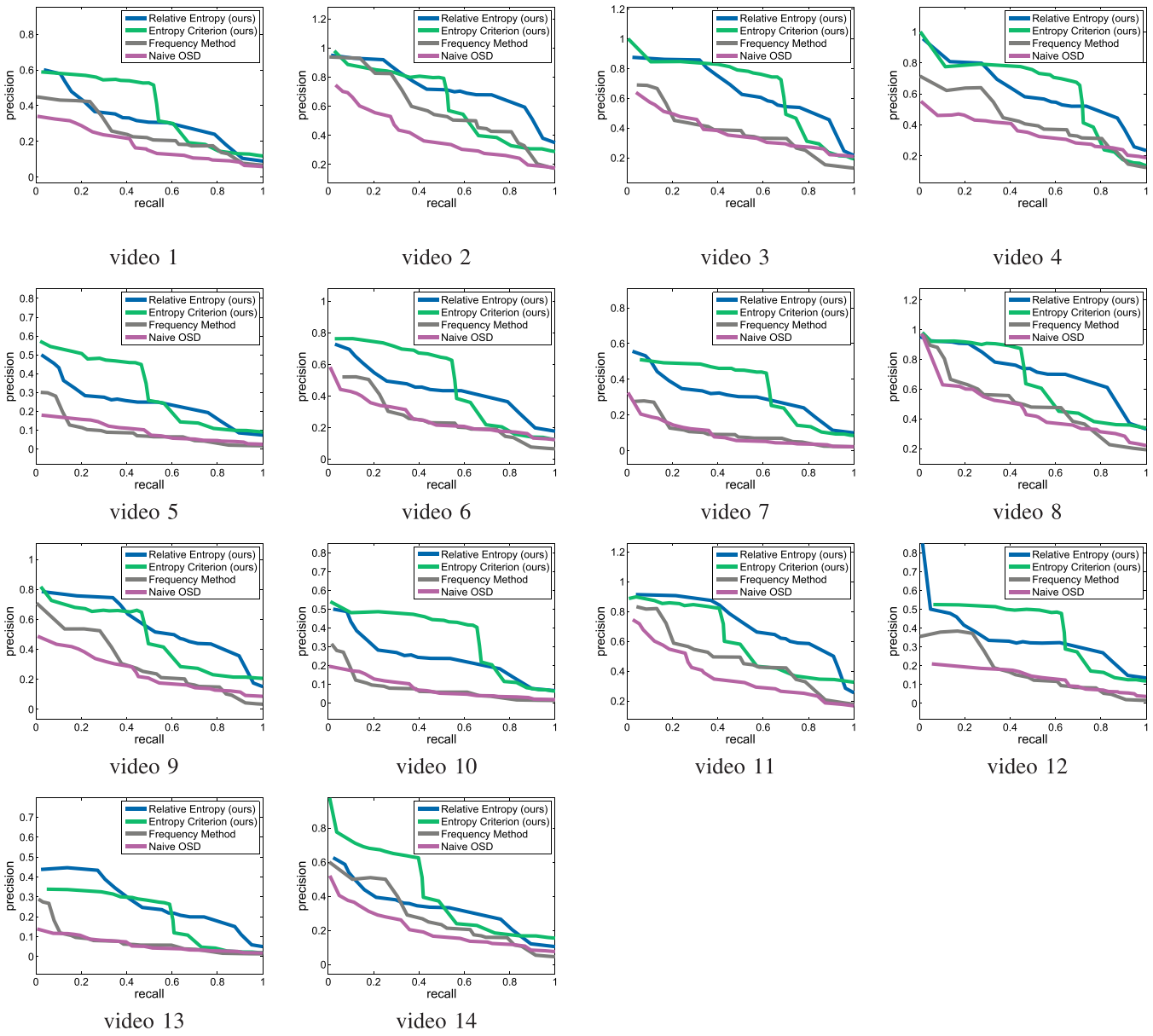


Fig. 13. Precision-recall curves for 14 individual videos.

TABLE II
AUC ON 14 INDIVIDUAL VIDEOS. "ES" REFERS TO THE ENTROPY SCORES OF EQ. (9). RES REFERS TO THE RELATIVE ENTROPY SCORES OF EQ. (12)

video	No. 1	No. 2	No. 3	No. 4	No. 5	No. 6	No. 7	No. 8	No. 9	No. 10	No. 11	No. 12	No. 13	No. 14
Frequency Method	0.259	0.596	0.364	0.442	0.090	0.243	0.098	0.493	0.326	0.078	0.464	0.183	0.074	0.293
Naive OSD	0.184	0.380	0.360	0.357	0.094	0.266	0.096	0.477	0.252	0.082	0.365	0.130	0.063	0.209
ES (ours)	0.380	0.620	0.680	0.641	0.315	0.483	0.332	0.660	0.474	0.355	0.609	0.363	0.198	0.435
RES (ours)	0.322	0.740	0.656	0.624	0.251	0.443	0.310	0.740	0.560	0.247	0.704	0.360	0.280	0.335

which manifest that our method outperforms the alternatives. The precision-recall curves of each individual videos are shown in Fig. 13.

A few results are shown in Fig. 12 and video samples are included in Youtube.[1] We implement integration for precision-recall curves to calculate the area under precision-recall curves (AUC). AUC values are listed in Table I. AUCs of each individual videos are tabulated in Table II.

Our method outperforms the "frequency method" significantly. It is because many wearer-interacted objects, such as public facilities, are not personal ones, although they are frequently used. For example, the electronic lock shown in Fig. 7 is not removable from the door and thus can present only in one scene. If there is no scene distribution consideration,

[1]http://www.youtube.com/watch?v=WRdd6Eh3DGI&feature=youtu.be

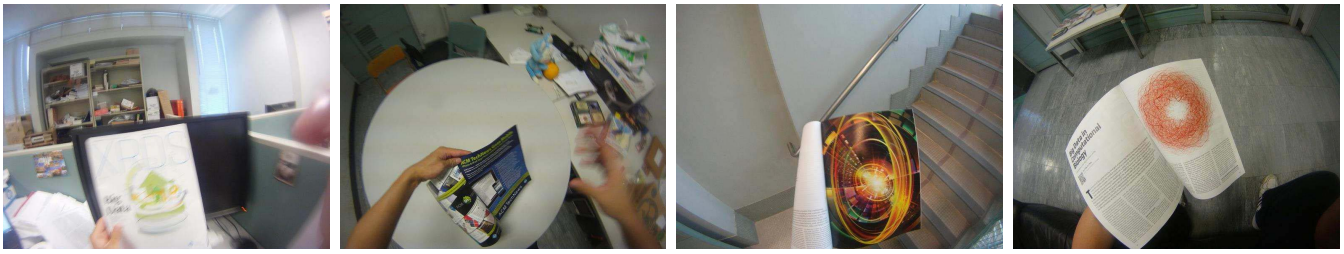


Fig. 14. A mis-detection example: a deforming magazine with changing content.

this object will be regarded as a personal one. Adopting the absolute entropy cannot tackle this problem. To test the importance of latent scene extraction and DPMM object discovery, we exclude one of them and observe the performance drops. Excluding respectively latent scene extraction and DPMM object discovery result in AUCs 28.4 and 31.7.

For the second baseline method, K-mean cannot determine an appropriated cluster number, making a cluster possibly be broken into several small ones. In scene extraction, naive assignment cannot model scene transition that is universal in first-person videos.

We also experimented with the video object discovery [41] method on our data. The AUC is only 2.16 because it does not distinguishing between personal and other objects.

3) Computational Cost: Our method can process one frame per 2 seconds on a single thread. The latent scene extraction and wearer-interacted object discovery take about 30% and 60% computation time respectively. Part of our system can be parallelized, such as sparse coding in latent scene extraction and object discovery. Our parallel computing implementation can process about 7 frames per second on a workstation with 24 CPU cores.

B. More Discussions

When we set threshold τ to 0, all detected wearer-interacted objects are regarded as personal objects. We found the precision is already as high as 73.9%, indicating our wearer-interacted objects cover many personal ones and form a good candidate pool with a reasonable size. We nevertheless observe a few false detection examples. In Fig. 11, the paper is close to the yellow region and is looked at by the subject for more than one second. These incorrectly detected regions have low personal scores because it appears in one or a small number of scenes, which leads to a sparse object-scene distribution (see Fig. 11).

V. Conclusion and Limitation

We have exploited a new topic of discovering personal and frequently used objects. Object-scene distribution has been proposed based on the observation that personal objects generally accompany us in many scenes. We provided solutions to extract scenes and objects from first-person videos in an unsupervised manner. It is notable that our latent scene learning scheme may be applied to other computer vision systems.

Our method unsurprisingly has a number of limitations. If a personal object is seldom used, its discovery from the egocentric videos could be difficult. For example, keeping the mobile phone always in the pocket or simply using it for a very short period of time, will fail our method. Some objects that change their appearances greatly in different scenes may also lead to mis-detection. One failure example is the deforming magazine with turned pages shown in Fig. 14. This limitation could be partly alleviated by including more egocentric information in future work.

References

[1] C.-H. Ling, Y.-M. Liang, C.-W. Lin, Y.-S. Chen, and H.-Y. M. Liao, "Human object inpainting using manifold learning-based posture sequence estimation," *IEEE Trans. Image Process.*, vol. 20, no. 11, pp. 3124–3135, Nov. 2011.

[2] F. Lecumberry, A. Pardo, and G. Sapiro, "Simultaneous object classification and segmentation with high-order multiple shape models," *IEEE Trans. Image Process.*, vol. 19, no. 3, pp. 625–635, Mar. 2010.

[3] C.-H. Chuang and W.-N. Lie, "A downstream algorithm based on extended gradient vector flow field for object segmentation," *IEEE Trans. Image Process.*, vol. 13, no. 10, pp. 1379–1392, Oct. 2004.

[4] E. Maggio and A. Cavallaro, "Learning scene context for multiple object tracking," *IEEE Trans. Image Process.*, vol. 18, no. 8, pp. 1873–1884, Aug. 2009.

[5] C. Huang and P. Salama, "Error concealment for shape in MPEG-4 object-based video coding," *IEEE Trans. Image Process.*, vol. 14, no. 4, pp. 389–396, Apr. 2005.

[6] C. Jung and C. Kim, "A unified spectral-domain approach for saliency detection and its application to automatic object segmentation," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 1272–1283, Mar. 2012.

[7] D. M. Weber and D. P. Casasent, "Quadratic Gabor filters for object detection," *IEEE Trans. Image Process.*, vol. 10, no. 2, pp. 218–230, Feb. 2001.

[8] R. Strickland and H. I. Hahn, "Wavelet transform methods for object detection and recovery," *IEEE Trans. Image Process.*, vol. 6, no. 5, pp. 724–735, May 1997.

[9] J. Luo and D. Crandall, "Color object detection using spatial-color joint probability functions," *IEEE Trans. Image Process.*, vol. 15, no. 6, pp. 1443–1453, Jun. 2006.

[10] V. Vilaplana, F. Marques, and P. Salembier, "Binary partition trees for object detection," *IEEE Trans. Image Process.*, vol. 17, no. 11, pp. 2201–2216, Nov. 2008.

[11] C. Sun and K.-M. Lam, "Multiple-kernel, multiple-instance similarity features for efficient visual object detection," *IEEE Trans. Image Process.*, vol. 22, no. 8, pp. 3050–3061, Aug. 2013.

[12] Z. Guo and Z. J. Wang, "Cross-domain object recognition via input-output kernel analysis," *IEEE Trans. Image Process.*, vol. 22, no. 8, pp. 3108–3119, Aug. 2013.

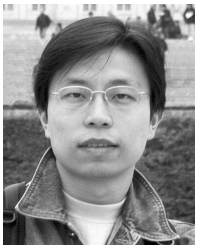
[13] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[14] L. Li, W. Huang, I. Y.-H. Gu, and Q. Tian, "Statistical modeling of complex backgrounds for foreground object detection," *IEEE Trans. Image Process.*, vol. 13, no. 11, pp. 1459–1472, Nov. 2004.

[15] H. Kong, J.-Y. Audibert, and J. Ponce, "Detecting abandoned objects with a moving camera," *IEEE Trans. Image Process.*, vol. 19, no. 8, pp. 2201–2210, Aug. 2010.
[16] J. Yang *et al.*, "Discovering primary objects in videos by saliency fusion and iterative appearance estimation," *IEEE Trans. Circuits Syst. Video Technol.*, accepted.
[17] G. Zhao, J. Yuan, and G. Hua, "Topical video object discovery from key frames by modeling word co-occurrence prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 1602–1609.
[18] L. Wang, G. Hua, R. Sukthankar, J. Xue, and N. Zheng, "Video object discovery and co-segmentation with extremely weak supervision," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 640–655.
[19] A. Faktor and M. Irani, "'Clustering by composition'—Unsupervised discovery of image categories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 6, pp. 1092–1106, Jun. 2014.
[20] J. Yuan, G. Zhao, Y. Fu, Z. Li, A. K. Katsaggelos, and Y. Wu, "Discovering thematic objects in image collections and videos," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 2207–2219, Apr. 2012.
[21] B. Schiele, N. Oliver, T. Jebara, and A. Pentland, "An interactive computer vision system DyPERS: Dynamic personal enhanced reality system," in *Proc. Int. Conf. Comput. Vis. Syst.*, 1999, pp. 51–65.
[22] T. Starner, J. Weaver, and A. Pentland, "Real-time American sign language recognition using desk and wearable computer based video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 12, pp. 1371–1375, Dec. 1998.
[23] K. Aizawa, K. Ishijima, and M. Shiina, "Summarizing wearable video," in *Proc. ICIP*, Oct. 2001, pp. 398–401.
[24] H. Goto and M. Tanaka, "Text-tracking wearable camera system for the blind," in *Proc. 10th Int. Conf. Document Anal. Recognit.*, Jul. 2009, pp. 141–145.
[25] L. Han, Z. Li, H. Zhang, and D. Chen, "Wearable observation supporting system for face identification based on wearable camera," in *Proc. IEEE Int. Conf. Comput. Sci. Inf. Technol. (ICCSIT)*, vol. 7. Jul. 2010, pp. 91–95.
[26] X. Ren and C. Gu, "Figure-ground segmentation improves handled object recognition in egocentric video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 3137–3144.
[27] K. Zhan, F. Ramos, and S. Faux, "Activity recognition from a wearable camera," in *Proc. Int. Conf. Control Autom. Robot. Vis. (ICARCV)*, Dec. 2012, pp. 365–370.
[28] Z. Lu and K. Grauman, "Story-driven summarization for egocentric video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 2714–2721.
[29] Y. J. Lee, J. Ghosh, and K. Grauman, "Discovering important people and objects for egocentric video summarization," in *Proc. CVPR*, Jun. 2012, pp. 1346–1353.
[30] J. Wang, J. Yang, K. Yu, F. Lv, T. S. Huang, and Y. Gong, "Locality-constrained Linear Coding for image classification," in *Proc. CVPR*, Jun. 2010, pp. 3360–3367.
[31] A. Bosch, A. Zisserman, and X. Muñoz, "Representing shape with a spatial pyramid kernel," in *Proc. CIVR*, 2007, pp. 401–408.
[32] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res.*, vol. 11, pp. 19–60, Jan. 2010.
[33] J. Carreira and C. Sminchisescu, "Constrained parametric min-cuts for automatic object segmentation," in *Proc. CVPR*, Jun. 2010, pp. 3241–3248.
[34] M. J. Jones and J. M. Rehg, "Statistical color models with application to skin detection," *Int. J. Comput. Vis.*, vol. 46, no. 1, pp. 81–96, Jan. 2002.
[35] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1409–1422, Jul. 2012.
[36] C. Wengert, M. Douze, and H. Jégou, "Bag-of-colors for improved image search," in *Proc. ACM Multimedia*, 2011, pp. 1437–1440.
[37] C. E. Antoniak, "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems," *Ann. Statist.*, vol. 2, no. 6, pp. 1152–1174, Nov. 1974.
[38] J. Sethuraman, "A constructive definition of Dirichlet priors," *Statist. Sinica*, vol. 4, pp. 639–650, Mar. 1994.
[39] T. S. Ferguson, "A Bayesian analysis of some nonparametric problems," *Ann. Statist.*, vol. 1, no. 2, pp. 209–230, Mar. 1973.
[40] D. M. Blei and M. I. Jordan, "Variational inference for Dirichlet process mixtures," *Bayesian Anal.*, vol. 1, no. 1, pp. 121–143, 2006.
[41] M. Grundmann, V. Kwatra, M. Han, and I. Essa, "Efficient hierarchical graph-based video segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 2141–2148.

Cewu Lu (M'13) received the B.S. degree from the Chongqing University of Posts and Telecommunications, in 2006, the M.S. degree from the Graduate University of Chinese Academy of Sciences, in 2009, and the Ph.D. degree in computer science and engineering from The Chinese University of Hong Kong, in 2013. He is currently a Research Professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University. He won the Fourth place in ILSVRC 2014 among 38 teams. He received the best paper award of NPAR 2012 and served as an Associate Editor of journal *Gate to Computer Vision and Pattern Recognition* and reviewers of several major computer vision and graphics conferences and journals, such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and *ACM Transactions on Graphics*. His research interests include activity recognition, object detection, and image/video processing.

Renjie Liao received the B.S. degree in automation science and electrical engineering from Beihang University, China, in 2011. He is currently pursuing the M.Phil. degree with the Department of Computer Science and Engineering, The Chinese University of Hong Kong. His research interests include computer vision and machine learning.



Jiaya Jia (SM'09) received the Ph.D. degree in computer science from The Hong Kong University of Science and Technology, in 2004. He is currently a Professor with the Department of Computer Science and Engineering, The Chinese University of Hong Kong (CUHK). He heads the research group focusing on computational photography, machine learning, practical optimization, and low- and high-level computer vision. He received the Young Researcher Award 2008 and Research Excellence Award 2009 from CUHK. He currently serves as an Associate Editor of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and served as an Area Chair of ICCV 2011 and ICCV 2013. He was on the technical paper program committees of SIGGRAPH Asia, ICCP, and 3DV for several times, and the Co-Chair of the Workshop on Interactive Computer Vision, in conjunction with ICCV 2007.