Visual Motif Discovery via First-Person Vision

Ryo Yonetani[1(✉)], Kris M. Kitani[2], and Yoichi Sato[1]

[1] The University of Tokyo, Tokyo, Japan
{yonetani,ysato}@iis.u-tokyo.ac.jp
[2] Carnegie Mellon University, Pittsburgh, PA, USA
kkitani@cs.cmu.edu

Abstract. *Visual motifs* are images of visual experiences that are significant and shared across many people, such as an image of an informative sign viewed by many people and that of a familiar social situation such as when interacting with a clerk at a store. The goal of this study is to discover visual motifs from a collection of first-person videos recorded by a wearable camera. To achieve this goal, we develop a commonality clustering method that leverages three important aspects: inter-video similarity, intra-video sparseness, and people's visual attention. The problem is posed as normalized spectral clustering, and is solved efficiently using a weighted covariance matrix. Experimental results suggest the effectiveness of our method over several state-of-the-art methods in terms of both accuracy and efficiency of visual motif discovery.

Keywords: Commonality discovery · First-person video

1 Introduction

We are interested in understanding from a data-driven perspective, what images from a person's visual experience are common among the majority. By developing algorithms for automatically extracting such shared visual experiences, we aim to understand what parts of the physical world are meaningful to people. We denote these shared visual experiences that have significance across many people as *visual motifs*. While visual motifs can include images of physical objects like signs and historic buildings, they can also be images of social situations or observed human activities. Examples of visual motifs are illustrated in Fig. 1.

From a practical perspective, the ability to extract perceptually important images can be useful for such tasks as life-logging, video summarization, scene understanding, and assistive technologies for the blind. Automatically extracting important visual motifs can be helpful for identifying meaningful images for life-logging or summarization. By associating visual motifs to localized regions of the environment, we can inform scene understanding by identifying what parts of the scene are visual 'hot-spots.' The extraction of important visual information in the environment can also be helpful for assistive technologies, by conveying to blind people the information embedded in the visual world [1–3].

B. Leibe et al. (Eds.): ECCV 2016, Part II, LNCS 9906, pp. 187–203, 2016.
DOI: 10.1007/978-3-319-46475-6_12

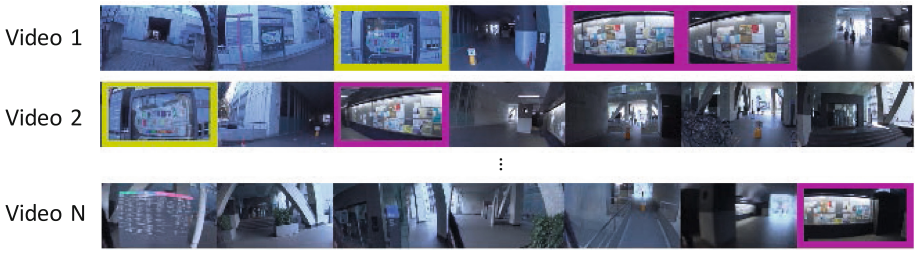

Fig. 1. Examples of visual motif discovery. Two signs annotated with colored rectangles are discovered as visual motifs from first-person videos. (Color figure online)

In this work, we automatically discover visual motifs using wearable cameras (*e.g.*, Google Glass), which we term *visual motif discovery*. Wearable cameras, especially when mounted on people's heads, can capture what people see clearly in the form of first-person point-of-view (POV) videos. This unique viewing perspective of wearable cameras has made it the platform of choice for understanding fine-grained human activities [4–9] and video summarization [10–15].

While it is intuitive that people will share meaningful visual experiences, it is not clear how these visual motifs can be extracted automatically from large first-person video collections. A common approach to discover visual motifs is to use *inter-video similarity*. Typically, a clustering algorithm is used to find cluster centroids corresponding to frequently occurring visual signatures shared across multiple images (*e.g.*, [16,17]). This is particularly problematic for first-person videos that tend to contain many mundane actions such as walking down a bare corridor or looking down at the ground. A straightforward application of clustering produces large clusters of mundane actions. This suggests that we need to discover visual commonalities while weighing them according to their significance. However, this raises the question of how to quantify significance.

To address this fundamental question of significance, we leverage visual cues unique to large collections of first-person videos taken in the same environment. As stated earlier, a large portion of the egocentric visual experience is *frequently* filled with mundane moments. Conversely, important visual motifs are typically distributed *sparsely* through our visual experience. This implies that *intra-video sparseness* is an important characteristic of meaningful visual motifs.

Another important feature of first-person videos is that they capture a person's focus of attention. Here we make the simple observation that when a person needs to acquire important visual information from a scene, she often stops and stays in the same position. Such an action can be observed clearly in the form of ego-motion of first-person videos. This implies that *egocentric attention* measured via camera ego-motion is a salient cue for discovering meaningful motifs.

We integrate the requirements of (1) inter-video similarity, (2) intra-video sparseness, and (3) egocentric attention into a constrained optimization problem to discover visual motifs across a large first-person video collection. In the proposed method, the problem can be formulated as normalized spectral

clustering constrained by an intra-video sparseness prior and cues from the egocentric attention, and is solved efficiently using a weighted covariance matrix. We empirically show that our method can discover meaningful visual motifs while processing a million first-person POV image frames in 90 s.

To the best of our knowledge, this work is the first to introduce the task of discovering visual motifs via first-person vision – significant first-person POV visual experiences shared across many people. The proposed method is tailored to discover visual motifs from a large collection of first-person videos using the constraints of intra-video sparseness and egocentric attention cues. Empirical validation shows that our method outperforms state-of-the-art commonality discovery methods [18,19] on first-person video datasets.

Related Work. The method of discovering commonalities in multiple images is adopted in many computer vision tasks such as common object discovery (co-segmentation or co-localization) [16–18,20–25], co-summarization [26], co-person detection [27], temporal commonality discovery [19], and popularity estimation [28]. They often generate candidates of commonalities (*e.g.*, superpixels, bounding-box proposals, video shots), in which the significance of each candidate is evaluated based on objectness or saliency. Significance measurements are also essential in automatic video summarization. Each video shot is evaluated for its significance using the presence of important persons and objects [10] or interestingness [12]. In contrast to previous work, we take advantage of using first-person videos by leveraging egocentric attention cues as a more natural feature for measuring the subjective significance of a visual experience. Although we limit our study to the use of a single wearable camera, the additional use of an eye tracker could help us to further understand visual attention [15].

In the field of first-person vision, recent studies proposed the use of multiple wearable cameras recording at the same time to estimate the joint focus of attention [29–31]. Accurate poses and positions of wearable cameras enabled by geometric information of the environment are used to find intersections of people's attention directions. In contrast, we focus on discovering shared visual experiences across many individuals without the use of temporal synchronization or assumption of interactive scenarios.

2 Discovering Visual Motifs from First-Person Videos

Suppose that we are given a collection of first-person videos recorded by many people in a certain environment (*e.g.*, a university campus). The goal of this work is to discover visual motifs specific to the environment: significant visual experiences shared across multiple people such as an image of an informative sign viewed by many people or that of a familiar social situation such as when interacting with a clerk at the university bookstore.

To discover visual motifs, we propose a method based on an unsupervised commonality clustering framework. We accept a collection of videos (a sequence of image frames) as input and output a cluster of images corresponding to

visual motifs. In Sect. 2.1, we describe how image frames observed across multiple videos are analyzed using the clustering framework to discover visual motifs while taking into account inter-video similarity and intra-video sparseness. Then in Sect. 2.2, we outline a method for detecting pauses in visual attention as an egocentric attention cue, and describe how that information can be used to inform the proposed method of significant image frames. We further present a technique for increasing the computational efficiency of our method through the use of weighted covariance matrix for clustering in Sect. 2.3. Finally, we describe an incremental framework for discovering multiple visual motifs from a large video collection in Sect. 2.4.

2.1 Discovering Common Scenes

We first describe a general commonality clustering framework (*e.g.*, [18,20,23–26]) for discovering common scenes from multiple videos. This framework integrates the concepts of inter-video similarity and intra-video sparseness.

Let $\boldsymbol{f}_t^{(i)} \in \mathbb{R}^V$ be a V-dimensional feature vector describing a scene of the t-th image frame in the i-th video. We denote a sequence of scene features extracted from the i-th video as $F^{(i)} = [\boldsymbol{f}_1^{(i)}, \ldots, \boldsymbol{f}_{T^{(i)}}^{(i)}]^\top \in \mathbb{R}^{T^{(i)} \times V}$, where $T^{(i)}$ is the number of image frames. The moments when a common scene is observed in the i-th video are described by an indicator vector $\boldsymbol{x}^{(i)} = [x_1^{(i)}, \ldots, x_{T^{(i)}}^{(i)}]^\top \in \{0,1\}^{T^{(i)}}$ where $x_t^{(i)}$ takes 1 if the t-th image frame includes the common scene.

Our goal is to estimate the indicator vector $\boldsymbol{x}^{(i)}$ for N given videos. To this end, we define the inter-video similarity by an affinity matrix, $W_{ij} \in \mathbb{R}^{T^{(i)} \times T^{(j)}}$, where the (t,t')-th entry of W_{ij} is given by an affinity function $\sigma(\boldsymbol{f}_t^{(i)}, \boldsymbol{f}_{t'}^{(j)}) \in \mathbb{R}$ (*e.g.*, a dot product or a radial basis function). We also introduce a degree matrix $D_i = \mathrm{diag}(d_1^{(i)} \ldots, d_{T^{(i)}}^{(i)})$ where $d_t^{(i)} = \sum_{t'=0}^{T^{(i)}} \max(\sigma(\boldsymbol{f}_t^{(i)}, \boldsymbol{f}_{t'}^{(i)}), 0)$. This degree matrix describes the inverse of intra-video sparseness; $d_t^{(i)}$ will increase when the t-th image frame of the i-th video is similar to the other frames in the same i-th video. The inter-video similarity W_{ij} and the inverted intra-video sparseness D_i are further combined across all combinations of N videos:

$$W = \begin{bmatrix} W_{11} & \cdots & W_{1N} \\ \vdots & \ddots & \vdots \\ W_{N1} & \cdots & W_{NN} \end{bmatrix} \in \mathbb{R}^{T_{\mathrm{all}} \times T_{\mathrm{all}}}, \quad D = \begin{bmatrix} D_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & D_N \end{bmatrix} \in \mathbb{R}^{T_{\mathrm{all}} \times T_{\mathrm{all}}}, \tag{1}$$

where $T_{\mathrm{all}} = \sum_i T^{(i)}$ is the total number of image frames. Likewise, we stack $\boldsymbol{x}^{(i)}$ to summarize the indicators across multiple videos:

$$\boldsymbol{x} = [(\boldsymbol{x}^{(1)})^\top, \ldots, (\boldsymbol{x}^{(N)})^\top]^\top \in \{0,1\}^{T_{\mathrm{all}}}. \tag{2}$$

By maximizing the sum of inter-video similarities $\boldsymbol{x}^\top W \boldsymbol{x}$ with respect to $\boldsymbol{x}$, we can find a scene frequently observed across multiple videos. At the same time, the scenes sparsely distributed in each video can be found by minimizing

the inverse of intra-video sparseness $\boldsymbol{x}^\top D\boldsymbol{x}$ on $\boldsymbol{x}$. These two requirements can be satisfied simultaneously by solving the following maximization problem.

$$\boldsymbol{x} = \arg\max_{\boldsymbol{x}} \frac{\boldsymbol{x}^\top W\boldsymbol{x}}{\boldsymbol{x}^\top D\boldsymbol{x}} \quad \text{s.t. } \boldsymbol{x} \in \{0,1\}^{T_{\text{all}}}. \tag{3}$$

Equation (3) can be solved via normalized spectral clustering [32,33] with two-clusters or normalized cuts [34]. We first compute two eigenvectors $Y = [\boldsymbol{y}_1, \boldsymbol{y}_2] \in \mathbb{R}^{T_{\text{all}} \times 2}$ of the matrix $L = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$ for the two largest eigenvalues[1]. Each row of the eigenvectors Y is then divided into common-scene and non-common-scene clusters via k-means clustering, where the centroid of the common-scene cluster is more distant from the origin. Cluster assignments are finally used for $\boldsymbol{x}$ such that $x_t^{(i)} = 1$ if and only if the corresponding elements of Y belong to the common-scene cluster. Importantly, the eigenvalue problem on L can be solved efficiently using various sparse eigensolvers (*e.g.*, the Lanczos method) since L is typically sparse and only two eigenvectors are required [34].

2.2 Learning to Detect Egocentric Attention Cues

The framework described above discovers common scenes that are not always significant to people, such as a hallway to reach a visual motif. In order to identify significant parts of visual experiences in videos, we focus on a specific but yet commonly occurring moment when people pause to acquire important visual information from a scene (*e.g.*, looking at maps or purchasing something from vending machines). We can detect such pausing actions taken by camera wearers by observing ego-motion of first-person videos. Detection results can then be used as an *egocentric attention* cue to constrain the clustering process.

Formally, the egocentric attention cue is given for each frame by a score $p_t^{(i)} \in [0,1]$ that increases if a camera wearer is more likely to take a pausing action at the t-th frame of the i-th video. Similar to D in Eq. (1), this egocentric attention cue is extended to handle multiple videos: $P = \text{diag}(P_1, \ldots, P_N)$ where $P_i = \text{diag}(p_1^{(i)}, \ldots, p_{T^{(i)}}^{(i)})$. We then constrain the clustering process by solving the eigenvalue problem on the following matrix L':

$$L' = (D^{-\frac{1}{2}} P^{\frac{1}{2}}) W (D^{-\frac{1}{2}} P^{\frac{1}{2}}) = AWA. \tag{4}$$

The indicator $\boldsymbol{x}$ obtained from L' can maximize not only the inter-video similarity and intra-video sparseness, but also the sum of egocentric attention cues.

Pausing actions are detected as follows. We observe that people's heads can remain stable for a long period when people stay in the same locations and move quickly for a short period when actively scanning visual information. As illustrated in Fig. 2, these trends are observed clearly in the ego-motion of first-person videos. As shown in the left of the figure, we compute motion vectors on a 10×5 grid following [7]. By smoothing these motion vectors over time using a set

[1] Similar to [33], we compute the eigenvectors for the largest eigenvalues of L instead of the smallest eigenvalues of $I - L$.

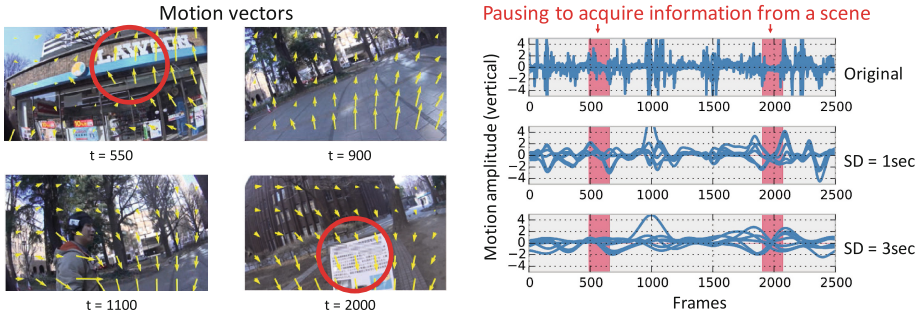

Fig. 2. Left: motion vectors computed on a 10×5 grid for several image frames where visual motifs are annotated with red circles. Right: amplitudes of motion vectors along the vertical direction smoothed over time with a set of Gaussian filters. (Color figure online)

of Gaussian filters with several standard deviations, we can see smaller motion amplitudes for larger deviations when people are pausing to acquire information from a scene (the right of the figure). The proposed method learns a pausing-action detector from these motion vectors. In a learning step, we apply the set of Gaussian filters independently to the horizontal and vertical elements as well as the magnitude of the motion vectors. A set of smoothed vectors and the original vectors for each frame are then aggregated to serve as a feature vector, and are learned using a binary classifier. Note that learning of the detector needs to be carried out only once and is not necessary for each environment or person.

In a testing step, the decision score of the detector obtained for each image frame is fed to the following postprocessing pipeline to generate egocentric attention cue P. A Gaussian filter is first applied to a sequence of decision scores to ensure its temporal smoothness. We then adopt a power normalization (*i.e.*, $\mathrm{sgn}(p')\sqrt{\mathrm{abs}(p')}$ for a smoothed score p') to encourage small decision peaks and a sigmoid function to suppress extremely strong ones. Finally, we scale the sequence into $[0, 1]$ to use it as the diagonal entries of P.

2.3 Efficient Clustering Using Weighted Covariance Matrix

The clustering process in Sect. 2.1 should be conducted efficiently since a large collection of first-person videos is often required for visual motif discovery. Reliable motifs that are not just attractive to a limited number of people can be obtained from a video collection containing as many recordings as possible. In addition, each video can have a large number of frames when a camera wearer keeps recording everyday life (*e.g.*, [35]). As a result, the number of total frames T_{all} will inevitably become huge. While the eigenvalue problem on L' in Eq. (4) can be solved in linear time to T_{all} by using sparse eigensolvers, there is a critical bottleneck in computing an affinity matrix W; the time complexity of computing W is $\mathcal{O}(T_{\mathrm{all}}^2 V)$. This computation is required for most cases when one tries to cluster commonalities based on pairwise affinities [18,20,23–26].

To address this problem, we use a compact weighted covariance matrix instead of the large L'. The only requirement to use this technique is to define an affinity function by a dot-product, *i.e.*, $\sigma(\boldsymbol{f}_t^{(i)}, \boldsymbol{f}_{t'}^{(j)}) \triangleq (\boldsymbol{f}_t^{(i)})^\top \boldsymbol{f}_{t'}^{(j)}$. Let us introduce a data matrix stacking all feature vectors: $F = [(F^{(1)})^\top, \ldots, (F^{(N)})^\top]^\top \in \mathbb{R}^{T_{\text{all}} \times V}$. We define W by $W = FF^\top$. Then, L' is rewritten as follows:

$$L' = AWA = (AF)(AF)^\top \in \mathbb{R}^{T_{\text{all}} \times T_{\text{all}}}. \tag{5}$$

Now we introduce a covariance matrix of F weighted by A:

$$C = (AF)^\top (AF) \in \mathbb{R}^{V \times V}. \tag{6}$$

Crucially, the eigenvectors of L' needed for spectral clustering can be obtained from those of the weighted covariance matrix C [36,37]. Given $z_i \in \mathbb{R}^V$ as an eigenvector of C for the i-th largest eigenvalue, the corresponding eigenvector $y_i \in \mathbb{R}^{T_{\text{all}}}$ of L' can be reconstructed by $y_i \propto AFz_i$. The time complexity to compute C is $\mathcal{O}(V^2 T_{\text{all}})$, which is much smaller than that of L' when $V \ll T_{\text{all}}$. When F is designed such that each feature dimension is less correlated, C is sparse and the eigenvalue problem on C can be solved efficiently. One limitation in using the weighted covariance matrix is that visual motifs should be linearly separable from other scenes in a feature space because we do not introduce any nonlinearity in the affinity function σ. We therefore use a high-level feature tailored to linear classifiers such as the Fisher vector [38] in F.

2.4 Discovering Multiple Visual Motifs

We have so far described how to discover a single visual motif from videos. Our method can be further extended to an incremental framework that allows videos to have multiple motifs. Specifically, we iteratively discover the most probable motif while updating C based on the discovery result.

Suppose that a visual motif is discovered in the form of $\boldsymbol{x}_k \in \{0,1\}^{T_{\text{all}}}$ at the current k-th step. Here we denote the i-th row of matrix AF and vector $\boldsymbol{x}$ as $AF[i]$ and $\boldsymbol{x}[i]$, respectively. Then, the degree of how the k-th motif biases the original C is explained by $C_k = \sum_{i \in \{j | \boldsymbol{x}_k[j]=1\}} (AF[i])^\top (AF[i])$. Other motifs can therefore be discovered in the subsequent $k+1$-th step by updating $C \leftarrow C - C_k$. We also deflate A of selected frames to be zero (*i.e.*, $A \leftarrow A \cdot \text{diag}(\boldsymbol{1} - \boldsymbol{x}_k)$) so that they will not be selected again.

One important problem for discovering multiple motifs is the termination of iterative discovery, *i.e.*, how to estimate the number of motifs in a video collection. Some studies on spectral clustering proposed to observe eigenvalues to determine the number of clusters [32,39]. Intuitively, eigenvalues are large as long as the number of clusters is below that of actual groups. In our method, the eigenvalues will become small when no more common scenes remain in videos. We also observe that egocentric attention cue P can indicate the number of motifs. If the attention cue of selected frames is small, these frames are not likely to be a visual motif but a scene incidentally observed across multiple videos.

Algorithm 1. Discovering multiple visual motifs

Require: Feature F, degree matrix D, egocentric attention cue P, threshold $e_{\min}$.
Ensure: Set of indicator vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$.

```
Compute A = D^{-1/2} P^{1/2}.
Compute C = (AF)^T (AF).
Set k = 0
repeat
    Find two eigenvectors of C for the two largest eigenvalues, Z = [z_1, z_2].
    Compute Y = AFZ.
    Conduct two-clusters k-means clustering on Y to obtain x_k.
    Compute e = λ (x_k^T P x_k) / (x_k^T x_k), where λ is the largest eigenvalue in Step 5.
    Compute C_k = Σ_{i ∈ {j | x_k[j]=1}} (AF[i])^T (AF[i])
    Update C ← C − C_k.
    Update A ← A · diag(1 − x_k).
    Update k ← k + 1.
until e < e_min
```

To incorporate these two criteria, we define a *confidence score* for the k-th motif by $e = \lambda \frac{\boldsymbol{x}_k^\top P \boldsymbol{x}_k}{\boldsymbol{x}_k^\top \boldsymbol{x}_k}$, where λ is the largest eigenvalue obtained in the eigenvalue problem of the k-th step. We discover visual motifs iteratively as long as e is above a pre-defined threshold $e_{\min}$. A complete algorithm to discover multiple visual motifs is described in Algorithm 1. After running the algorithm, we finally refine the results (a sequence of indicator vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$) by omitting some indicator vectors that only select the image frames from a single video.

3 Experiments

To evaluate the effectiveness of the proposed method, we constructed a dataset composed of multiple first-person videos during a navigation task in several different environments. We also tested our method on a dataset recorded by people during social interactions [40]. The experimental results show that the proposed method successfully improves upon both accuracy and efficiency of visual motif discovery compared to several state-of-the-art methods [18,19].

3.1 Navigation Dataset

Because many prior studies on first-person vision focused on how their method could be generalized to a variety of environments (*e.g.*, GeorgiaTech Egocentric Activities [4], JPL First-Person Interaction [6], CMU Social Saliency [29]), there are few datasets of first-person videos that have been recorded many times in one environment. One prospective dataset that we will test in Sect. 3.5 is First-Person Social Interactions [40]. However, the number of visual motifs is not sufficient for quantitatively evaluating the accuracy of visual motif discovery.

We therefore introduce a new **Navigation** dataset that contains multiple recordings of 21 visual motifs. First-person videos were taken in six different

environments, where three to four subjects were assigned to each of the environments. Subjects joined in on a navigation task as follows. They visited several pre-defined places with attractive physical objects such as map signs, vending machines, and the entrance of a store. They were asked to look at what was described on these objects (*e.g.*, reading a map sign). They were able to take arbitrary positions, poses, and head motions when looking at the objects. This made the appearance of motifs sufficiently variable for each recording. We only instructed the subjects to look at objects from a reasonable distance to restrain them from acquiring information in an extremely unusual way, such as reading signs from an extremely long (or short) distance. They were also allowed to visit other places that were not pre-defined. In total, 44 first-person videos were recorded at 30 fps. The time of each recording was 90 to 180 s and the total number of image frames for each environment was on average 27784.0. To complete the feature extraction steps in a reasonable time, each video was resized to a resolution of 320×180.

Ground truth labels of visual motifs were given as follows. We first annotated the time intervals when image frames captured pre-defined objects roughly at the center. Then, we refined the intervals so that the acceleration of head motion was locally minimum and maximum at the beginning and end of the intervals, respectively. The average time when subjects were judged as looking at the objects was 5.8 s. These annotations were also used for learning a detector for egocentric attention cues. Note that we confirmed by manual inspection that our dataset did not contain other visual motifs (*i.e.*, images of other physical objects seen by the majority of subjects) that were not in the pre-defined set.

3.2 Implementations

One important implementation of our method is the design of scene features. As we stated in Sect. 2.3, the features should have the potential of linearly separating visual motifs from other unimportant scenes to use a weighted covariance matrix. In the experiments, the following two types of features were used:

SIFT + Fisher vector (FV). RootSIFT descriptors [41] were sparsely sampled from each image frame. They were then fed to the principal component analysis (PCA) with 64 components and the Fisher vector [38] with the 128-component Gaussian mixture model (GMM) followed by power and L2 normalizations. As the features were rather high dimensionally (16384 dimensions), we adopted the sparse random projection [42] to project the features onto a 1024-dimensional feature space. We trained the PCA and GMM components for each environment independently.

CNN feature (CNN). A convolutional neural network (CNN) trained with the MIT Places database [43] was used as a feature descriptor. To investigate how the pre-trained CNN can be used to extract high-level features that could cope with the variability of motif appearances, we utilized the *fc6* layer outputs of the pre-trained network as a 4096-dimensional feature.

Note that both features were scaled for each environment so that each feature dimension had zero-mean and unit-variance. Based on these features, we implemented two variants of the proposed method: **Ours (FV)** and **Ours (CNN)**. For both methods, we set $e_{\text{th}} = 0.5$ which empirically worked well.

To enable an egocentric attention cue, we trained a linear support vector machine. Specifically, we split our dataset into two subsets based on environment IDs (videos of three environments recorded by three subjects and those of the other three environments by four subjects) and trained a pausing-action detector with one subset to test the other. Note that subjects and environments did not overlap between training and testing subsets. The standard deviations used for a set of Gaussian filters were 1 and 3 s in feature extraction, and 1 second in postprocessing. The impact of using the egocentric attention cue was validated with the following two degraded versions of our method: (1) **SC**, which uses a covariance matrix $C = (D^{-\frac{1}{2}}F)^{\top}(D^{-\frac{1}{2}}F)$ and a confidence score $e = \lambda$ to remove the effect of egocentric attention cue P in Algorithm 1, and is equivalent to standard normalized spectral clustering; and (2) **EgoCue**, which directly uses $p_t^{(i)}$ as a confidence score for each frame, which can be regarded as a simple supervised learning method to detect visual motifs.

Baselines. Two state-of-the-art methods on commonality discovery served as baselines. One is the temporal commonality discovery method (**TCD**) [19][2]. Given a pair of videos, **TCD** discovers a pair of temporal intervals that include similar feature patterns. In the experiments, **TCD** was applied to all combinations of videos in a collection. Each image frame was then given a confidence score of visual motifs based on how many times the frame was discovered as a visual motif. We also took into account the egocentric attention cue of discovered frames, $p_t^{(i)}$, as follows. If the t-th frame of the i-th video was discovered by the combinations of K other videos, the frame obtained a confidence score of $Kp_t^{(i)}$.

The other baseline is the object co-localization method (**COLOC**) [18][3]. This method discovers a single common object observed across multiple images by selecting one of the object proposals generated per image. Each proposal has a prior score given by objectness, and objects are discovered with a confidence score. Instead of the object proposals per image, we used image frames of a given video as a proposal. The prior score of each frame proposal was then given by $p_t^{(i)}$ instead of the objectness used in [18]. Importantly, our implementation of **COLOC** discovered, as a visual motif, only a single image frame for each video. However, visual motifs were observed for consecutive frames of a certain length. We therefore found the consecutive frames around the discovered frame via temporal dilation, where the dilation size was learned from training subsets.

[2] http://humansensing.cs.cmu.edu/wschu/project_tcd.html.
[3] http://ai.stanford.edu/~kdtang/.

3.3 Detecting Visual Motifs

We first compared how well the methods could detect *any* visual motifs. To evaluate detection performance, we extended the confidence score e defined originally *per visual motif* in **Ours (FV)**, **Ours (CNN)**, **SC**, and **COLOC** to that defined *per frame* such as that given in **EgoCue** and **TCD**. Specifically, video frames discovered as a certain visual motif were given the confidence score of that motif; otherwise, they were given 0. These per-frame confidence scores were used to calculate precision-recall curves and average precision scores.

Since it was difficult to run the two baselines on **Navigation** in a reasonable time, we also constructed a smaller dataset, **Navigation-1**, which was cropped from **Navigation** to include a single visual motif per video. For each video, we cropped a shot including the time interval of visual motifs with a margin of 10 s (*i.e.*, 300 frames) before and after the interval. As a result, 21 collections of videos were generated. On **Navigation-1**, we detected a single motif with each method. We then tested the proposed method as well as its degraded versions on **Navigation**, where multiple motifs were discovered for evaluation.

Table 1 lists the average precision scores for all methods. Note that for **SC**, **TCD**, and **COLOC**, we describe the results using the **FV** feature, which were better than those using the **CNN** feature. The left of Fig. 3 also depicts precision-recall curves. Overall, the proposed method clearly outperformed the two baselines on **Navigation-1**. We also confirmed that this was achieved given the combination of common scene discovery and egocentric attention cue because the performance of **SC** and **EgoCue** was quite limited. The **FV** feature worked comparably to **CNN** on **Navigation-1** and the best on **Navigation**. The number of visual motifs pre-defined in **Navigation** and discovered with **Ours (FV)** are compared in Table 2. For most cases, our method could estimate the number of visual motifs accurately.

We also evaluated the computation times of each method, as shown on the right of Fig. 3. We generated videos of various numbers of frames simply by

Table 1. Average precision scores. **Navigation-1**: 21 video collections each of which includes single motif. **Navigation**: six video collections all including multiple motifs.

	Ours (FV)	Ours (CNN)	SC	EgoCue	TCD [19]	COLOC [18]
Navigation-1	0.77	**0.79**	0.64	0.67	0.63	0.63
Navigation	**0.77**	0.70	0.60	0.60	-	-

Table 2. Number of visual motifs pre-defined per environment on **Navigation** (top) and that discovered with **Ours (FV)** (bottom).

	Env 1	Env 2	Env 3	Env 4	Env 5	Env 6
# motifs	4	4	4	3	3	3
# discovered	4	4	5	3	3	4

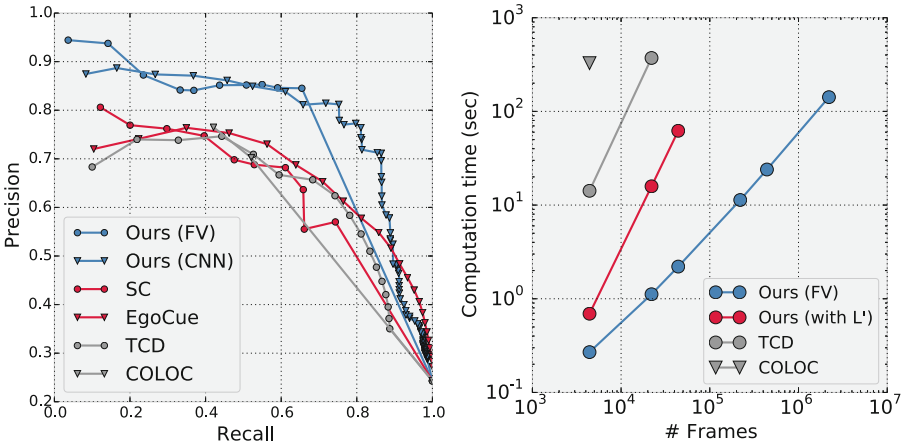

Fig. 3. Left: Precision-recall curves of methods on **Navigation-1**. Right: Computation times. For both **TCD** and **COLOC** we used codes available on authors' websites. **Ours (FV)** and **Ours (with L)** were implemented in Python. All methods were tested on MacPro with 2.7-GHz 12-Core Intel Xeon E5.

concatenating our videos multiple times. To show the impact of using a weighted covariance matrix, we also tested a variant of the proposed method that relied on L' in Eq. (5) instead of C in Eq. (6), which we referred to as **Ours (with L')** in the figure. Since the time complexity to compute the weighted covariance is linearly proportional to the number of image frames, the proposed method is an order-of-magnitude faster than the others. Note that high framerate videos are necessary only when computing ego-motion. Once egocentric attention cues are given, the clustering process can work under much lower framerates. If all videos are downsampled to 1 fps, our method can find visual motifs from 10 h of recording in 1 s.

3.4 Distinguishing Multiple Motifs

Next, we show how our method can distinguish multiple visual motifs. In Fig. 4, we describe confusion matrices and average accuracies of **Ours (FV)** and **SC** for each environment. To obtain the confusion matrices, we assigned pre-defined (*i.e.*, ground-truth) motifs to discovered (predicted) ones via linear assignment. By using an egocentric attention cue, we successfully classified visual motifs for many environments. As shown in Fig. 5, visual motifs were matched regardless of the points-of-view or parts of pre-defined objects that could be observed. Examples 3 and 6 also suggest that our method works well even when people are interacting with others and making frequent head and hand motions.

Figure 6 presents other examples of visual motifs. The use of high-level features such as **FV** allows us to match motifs even when a few changes were made in their appearance. Most failure cases were due to undiscovered instances (*i.e.*, incorrectly classified as unimportant background scenes). We found that these

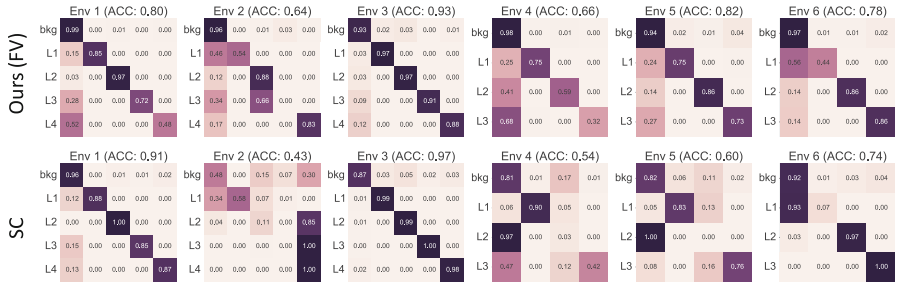

Fig. 4. Confusion matrices and average accuracies for multiple visual motif discovery on the six environments (Env 1, . . . , Env 6) in **Navigation**. Annotated labels **L1** to **L4** indicate ID of visual motifs while **bkg** denotes other background scenes.

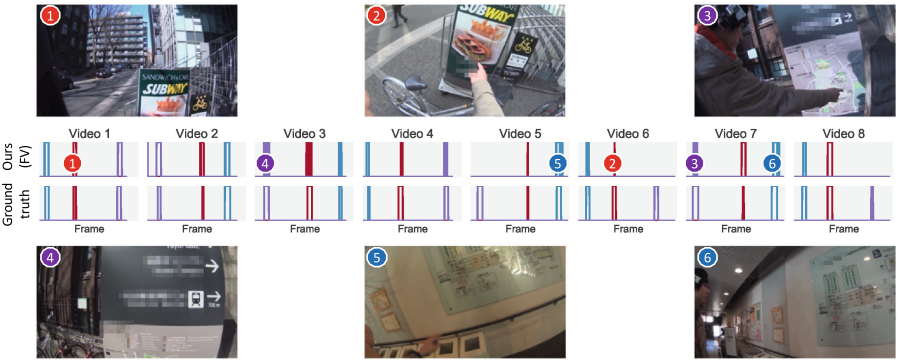

Fig. 5. Example of multiple visual motif discovery (on Env 4). Three motifs were discovered in timeline (colored plots in the middle) for each of eight videos in a video collection. Some discovered image frames are depicted at top and bottom. (Color figure online)

failures occurred when pre-defined objects were observed at different locations (the menu board in the fourth column of Fig. 6) or when they were not salient (the navigation sign in the fifth column).

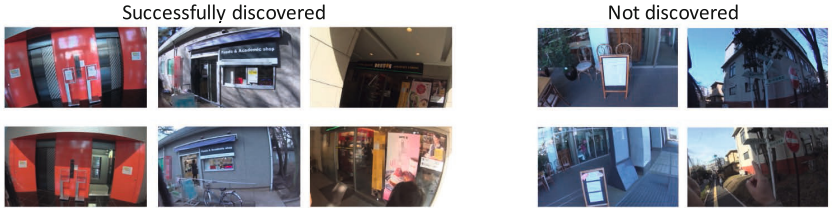

Fig. 6. Examples of successfully-discovered and non-discovered motifs. Videos in each row were recorded on different days and times.

3.5 Examples on First-Person Social Interactions Dataset

While we mainly focused on visual motifs for a navigation task in the experiments, our method can be used to discover different types of visual motifs given a video of other tasks. In particular, Fig. 7 shows the results of visual motif discovery on the First-Person Social Interactions dataset [40], in which a group of people participated in social interactions in an amusement park. We chose three collections of videos from the dataset in which a group of people interacted with each other at several places.

Fig. 7. Some visual motifs found in First-Person Social Interactions dataset [40].

At a cafeteria, our method discovered a situation in which camera wearers were (1) waiting in line, (2) interacting with a clerk, and (3) preparing a dish on a table. Our method also found at an entrance, a situation of (4) waiting in line and (5) interacting with others. Interestingly, the method was able to find (6) a photographer jointly looked at by multiple camera wearers, which was similar to co-person detection [27]. All these situations correspond to the shared visual experiences across multiple camera wearers, while they are quite different from those found in the navigation tasks mentioned in previous sections.

4 Conclusions

We introduced a new task of visual motif discovery from a large collection of first-person videos and developed an efficient method tailored to the task. The proposed method can discover visual motifs more accurately and an order-of-magnitude faster than other state-of-the-art methods.

There are several possible extensions leading to interesting directions for future work. While we focused on a specific class of visual motifs observed when people paused to acquire information, there are other significant moments shared across people such as when carrying important belongings by hand, meeting with friends, *etc.* First-person videos can be used to recognize many types of actions, *e.g.*, not only pausing but using hands [4,5,8,9] and conversing with others [40], which are all informative for recognizing a variety of visual motifs. In addition,

by combining geometric information enabled by visual simultaneous localization and mapping or GPS, our method will be able to distinguish visual motifs that are visually the same but observed at different locations (*e.g.*, the same signs placed at different entrance gates). Another interesting direction for future work is to extend our visual motif discovery method to work in an online manner. This allows us to handle extremely long recordings and makes it possible to extract a variety of visual motifs that we observe in everyday life.

Acknowledgments. This research was supported by CREST, JST and Kayamori Foundation of Informational Science Advancement.

References

1. Leung, T.S., Medioni, G.: Visual navigation aid for the blind in dynamic environments. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 153–158 (2014)
2. Tang, T.J.J., Li, W.H.: An assistive eyewear prototype that interactively converts 3D object locations into spatial audio. In: Proceedings of ACM International Symposium on Wearable Computers (ISWC), pp. 119–126 (2014)
3. Templeman, R., Korayem, M., Crandall, D., Kapadia, A.: Placeavoider: steering first-person cameras away from sensitive spaces. In: Proceedings of Annual Network and Distributed System Security Symposium (NDSS) (2014)
4. Fathi, A., Li, Y., Rehg, J.M.: Learning to recognize daily actions using gaze. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7572, pp. 314–327. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33718-5_23
5. Pirsiavash, H., Ramanan, D.: Detecting activities of daily living in first-person camera views. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2847–2854 (2012)
6. Ryoo, M.S., Matthies, L.: First-person activity recognition: what are they doing to me? In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2730–2737 (2013)
7. Poleg, Y., Arora, C., Peleg, S.: Temporal segmentation of egocentric videos. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2537–2544 (2014)
8. Saran, A., Teney, D., Kitani, K.M.: Hand parsing for fine-grained recognition of human grasps in monocular images. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1–7 (2015)
9. Cai, M., Kitani, K.M., Sato, Y.: A scalable approach for understanding the visual structures of hand grasps. In: Proceedings of IEEE International Conference on Robotics and Automation (ICRA), pp. 1360–1366 (2015)
10. Lee, Y.J., Ghosh, J., Grauman, K.: Discovering important people and objects for egocentric video summarization. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1346–1353 (2012)
11. Lu, Z., Grauman, K.: Story-driven summarization for egocentric video. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2714–2721 (2013)

12. Gygli, M., Grabner, H., Riemenschneider, H., Gool, L.: Creating summaries from user videos. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8695, pp. 505–520. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10584-0_33
13. Arev, I., Park, H.S., Sheikh, Y., Hodgins, J., Shamir, A.: Automatic editing of footage from multiple social cameras. ACM Trans. Graph. **33**(4), 81:1–81:11 (2014)
14. Xiong, B., Kim, G., Sigal, L.: Storyline representation of egocentric videos with an applications to story-based search. In: Proceedings of IEEE International Conference on Computer Vision (ICCV), pp. 4525–4533 (2015)
15. Xu, J., Mukherjee, L., Li, Y., Warner, J., Rehg, J.M., Singh, V.: Gaze-enabled egocentric video summarization via constrained submodular maximization. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2235–2244 (2015)
16. Joulin, A., Bach, F., Ponce, J.: Discriminative clustering for image co-segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1943–1950 (2010)
17. Zhou, F., De la Torre Frade, F., Hodgins, J.K.: Hierarchical aligned cluster analysis for temporal clustering of human motion. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) **35**(3), 582–596 (2013)
18. Tang, K., Joulin, A., Li, L.J., Fei-Fei, L.: Co-localization in real-world images. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1464–1471 (2014)
19. Chu, W.-S., Zhou, F., Torre, F.: Unsupervised temporal commonality discovery. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7575, pp. 373–387. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33765-9_27
20. Joulin, A., Tang, K., Fei-Fei, L.: Efficient image and video co-localization with Frank-Wolfe algorithm. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8694, pp. 253–268. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10599-4_17
21. Rother, C., Minka, T., Blake, A., Kolmogorov, V.: Cosegmentation of image pairs by histogram matching - incorporating a global constraint into mrfs. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 993–1000 (2006)
22. Vicente, S., Rother, C., Kolmogorov, V.: Object cosegmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2217–2224 (2011)
23. Zhang, D., Javed, O., Shah, M.: Video object co-segmentation by regulated maximum weight cliques. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8695, pp. 551–566. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10584-0_36
24. Fu, H., Xu, D., Zhang, B., Lin, S.: Object-based multiple foreground video co-segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3166–3173 (2014)
25. Wang, L., Hua, G., Sukthankar, R., Xue, J., Zheng, N.: Video object discovery and co-segmentation with extremely weak supervision. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 640–655. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10593-2_42
26. Chu, W.S., Song, Y., Jaimes, A.: Video co-summarization: video summarization by visual co-occurrence. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3584–3592 (2015)

27. Lin, Y., Abdelfatah, K., Zhou, Y., Fan, X., Yu, H., Qian, H., Wang, S.: Co-interest person detection from multiple wearable camera videos. In: Proceedings of International Conference on Computer Vision (ICCV), pp. 4426–4434 (2015)
28. Ortis, A., Farinella, G.M., D'amico, V., Addesso, L., Torrisi, G., Battiato, S.: Recfusion: automatic video curation driven by visual content popularity. In: Proceedings of ACM International Conference on Multimedia (MM), pp. 1179–1182 (2015)
29. Park, H.S., Jain, E., Sheikh, Y.: 3D social saliency from head-mounted cameras. In: Proceedings of Advances in Neural Information Processing Systems (NIPS), pp. 1–9 (2012)
30. Park, H.S., Jain, E., Sheikh, Y.: Predicting primary gaze behavior using social saliency fields. In: Proceedings of IEEE International Conference on Computer Vision (ICCV), pp. 3503–3510 (2013)
31. Park, H.S., Shi, J.: Social saliency prediction. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4777–4785 (2015)
32. Luxburg, U.: A tutorial on spectral clustering. Stat. Comput. **17**(4), 395–416 (2007)
33. Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: analysis and an algorithm. In: Proceedings of Advances in Neural Information Processing Systems (NIPS), pp. 849–856 (2001)
34. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **22**(8), 888–905 (2000)
35. Singh, K.K., Fatahalian, K., Efros, A.A.: Krishnacam: using a longitudinal, single-person, egocentric dataset for scene understanding tasks. In: Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–9 (2016)
36. Ghodsi, A.: Dimensionality Reduction a Short Tutorial. Department of Statistics and Actuarial Science, University of Waterloo, Ontario (2006)
37. Murakami, H., Kumar, B.: Efficient calculation of primary images from a set of images. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **PAMI–4**(5), 511–515 (1982)
38. Perronnin, F., Sánchez, J., Mensink, T.: Improving the Fisher kernel for large-scale image classification. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6314, pp. 143–156. Springer, Heidelberg (2010). doi:10.1007/978-3-642-15561-1_11
39. Zelnik-manor, L., Perona, P.: Self-tuning spectral clustering. In: Proceedings of Advances in Neural Information Processing Systems (NIPS), pp. 1601–1608 (2004)
40. Fathi, A., Hodgins, J.K., Rehg, J.M.: Social interactions: a first-person perspective. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1226–1233 (2012)
41. Arandjelovic, R., Zisserman, A.: Three things everyone should know to improve object retrieval. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2911–2918 (2012)
42. Li, P., Hastie, T.J., Church, K.W.: Very sparse random projections. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 287–296 (2006)
43. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. In: Proceedings of Advances in Neural Information Processing Systems (NIPS), pp. 487–495 (2014)