Ego-Surfing First Person Videos

Ryo Yonetani
The University of Tokyo
Tokyo, Japan
yonetani@iis.u-tokyo.ac.jp

Kris M. Kitani
Carnegie Mellon University
Pittsburgh, PA, USA
kkitani@cs.cmu.edu

Yoichi Sato
The University of Tokyo
Tokyo, Japan
ysato@iis.u-tokyo.ac.jp

Abstract

We envision a future time when wearable cameras (e.g., small cameras in glasses or pinned on a shirt collar) are worn by the masses and record first-person point-of-view (POV) videos of everyday life. While these cameras can enable new assistive technologies and novel research challenges, they also raise serious privacy concerns. For example, first-person videos passively recorded by wearable cameras will necessarily include anyone who comes into the view of a camera – with or without consent. Motivated by these benefits and risks, we develop a self-search technique tailored to first-person POV videos. The key observation of our work is that the egocentric head motions of a target person (i.e., the self) are observed both in the POV video of the target and observer. The motion correlation between the target person's video and the observer's video can then be used to uniquely identify instances of the self. We incorporate this feature into our proposed approach that computes the motion correlation over supervoxel hierarchies to localize target instances in observer videos. Our proposed approach significantly improves self-search performance over several well-known face detectors and recognizers. Furthermore, we show how our approach can enable several practical applications such as privacy filtering, automated video collection and social group discovery.

1. Introduction

New technologies for image acquisition, such as wearable eye glass cameras or lapel cameras, can enable new assistive technologies and novel research challenges but may also come with latent social consequence. Around the world hundreds of millions of camera-equipped mobile phones can be used to capture special moments in life. By the same token, a mobile phone image containing a GPS position embedded EXIF tag can also be used to determine where and when you captured that image and could be used as a means of violating one's privacy. Novel wearable camera technologies (*e.g.*, the Google Glass or the Narrative lapel camera) also offer a new paradigm for keeping a visual record of everyday life in the form of *first-person point-of-view (POV) videos*, and can be used to aid human productivity, such as automatic activity summarization [4, 14, 16] and assistive systems [15, 25, 27]. But as in the case of mobile phone cameras, wearable cameras also come with hidden social implications and the inevitable risk of unintended use. For example, wearable cameras which passively capture everyday life will necessarily include videos of people – with or without consent. Without the proper mechanisms and technologies to preserve privacy, wearable cameras run the risk of inadvertently capturing sensitive information.

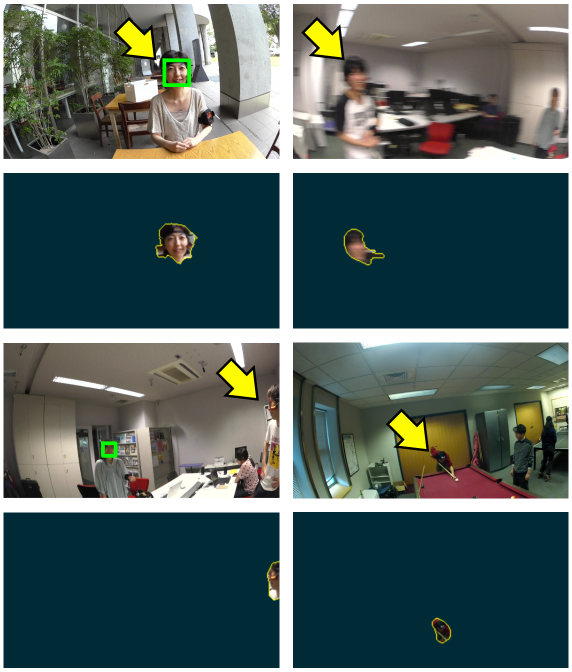

Figure 1. Robust self-search results. Self instances (**yellow arrows**) are detected (**unmasked regions**) despite heavy occlusion (**bottom-left**), motion blur (**top-right**) and extreme pose (**bottom-right**), where face recognition fails (**green rectangles**).

Keeping in mind both the benefits and risks of using wearable cameras, we argue that one important technology

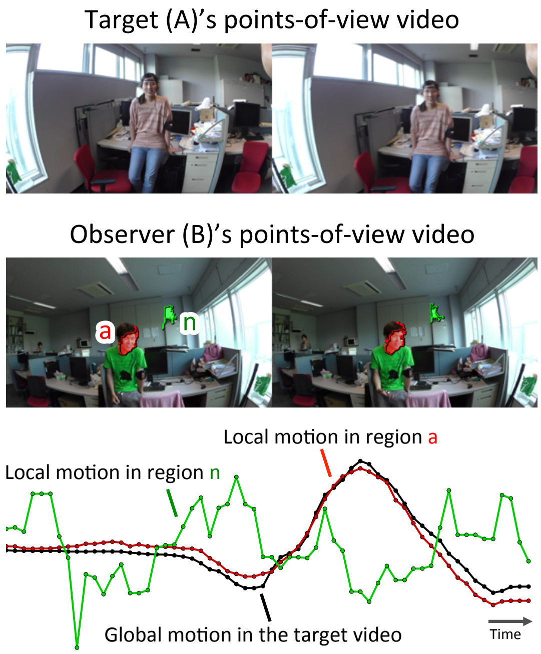

Figure 2. Global motion in the target (A)'s points-of-view video and local motions in the observer (B)'s video. The local motion in the target region **a** is highly correlated with the global motion compared to that in the non-target region **n**.

to develop is the ability to automatically search large repositories of first-person POV videos for the videos of a single user. Much like ego-surfing enables us to perform an Internet search with our own name, we believe that self-search in first-person videos can empower users to monitor and manage their own personal data. To this end, we develop a video-based self-search technique tailored to first-person videos. Since the appearance of people in the first-person videos often comes under heavy occlusions, motion blur and extreme pose changes (see Fig. 1), we require a robust approach beyond what can be accomplished by face recognition alone [13, 35, 36].

In order the account for the high variability in self appearance in first-person POV videos, we propose to use motion as our primary feature. The key insight of our proposed work is that the first-person video of the target user can act as a unique identifier to enable a target-specific search over a repository of the first-person videos. We give a concrete example in Fig. 2. Consider the case where the target (self) individual A is conversing with another person (which we call observer B). When A shakes his head, it induces large global motion (the camera moves from left to right; black line in Fig. 2) in the video. Now, from the perspective of observer B, we expect to see the same shake pattern but in the form of a local motion pattern in the B's POV video (we see the target individual A shaking his head at region **a** in Fig. 2). This correlation between the global motion of target A's video and the local motion of observer B's video indicates that these two videos are indeed related. Furthermore, this correlation is expected to increase only in the target regions. This illustrates the general insight that the ego-motion of the target is a unique signature which can be used to localize the self in observer videos.

Based on this insight, we develop a novel motion-correlation-based approach to search and localize target individuals from a collection of first-person POV videos. The algorithm takes as input the target's POV video, and retrieves as output all the instances of that target individual from observer POV videos. Our algorithm proceeds as follows. First, a supervoxel hierarchy [32, 33] is generated over all videos as target candidates. Second, each supervoxel is evaluated to compute their 'targetness' based on the correlations between local motion patterns inside a supervoxel and the global motion pattern of target video. Third, supervised classifier refines the 'targetness' scores by taking into account generic targetness and potential under- or over-segmentation of supervoxels. The supervoxel labeling task is posed as a binary-class Bayesian inference problem, where the likelihood and prior are respectively modeled by the motion correlation and supervoxel classifier.

Experimental results show that our self-search approach significantly outperforms baseline face recognizers. Furthermore, to show the potential impact that self-search can have on assistive technologies, we provide three proof-of-concept application scenarios: (1) privacy filtering, (2) automated video collection and (3) social group discovery.

Related Work

The idea of searching for a specific person in images or video has been addressed in several areas of computer vision. On such area is the topic of *person re-identification* in the context of visual surveillance. An extensive survey of the field can be found in [11, 29, 31]. One common approach to re-identification is to utilize visual signatures of a specific person, such as color and texture, to find specific individuals in images. Since many approaches presuppose a surveillance scenario, the features and approaches often depend on the assumption of a static camera (single POV, constant background, *etc*.). With the exception of work using active cameras [23], re-identification approaches are not designed to deal with extreme camera motion.

Recent work in the area of *egocentric vision* focused on human interactions have utilized person identification as a feature of their approaches. Many studies have relied on off-the-shelf face detectors and recognizers [2, 3, 8, 14, 19, 20, 21, 34]. In many of these scenarios, the use of face detection is justified since people are engaged in conversation and the first-person POV camera is relatively stable.

Poleg *et al*. [20] recently proposed a person identification method based on the correlation of head motions in first-person videos. Their method relies on people detection to track the head of each target candidate, making it challenging to reliably perform the identification when a person is mobile and significant egomotion is induced. By contrast, our approach does not require explicit detection but directly examines the correlations at the supervoxel level.

Another alternative method for identifying specific individuals is the use of geometric 2D [12, 17] or 3D information [18, 19]. An accurate map can be used to compute the precise location of individuals with wearable cameras and the location can be used to estimate visibility in other cameras. While these approaches work for local areas (a room or building) it is not currently feasible to learn maps on a global scale (an entire country).

To the best of our knowledge, this is the first to address the topic of self-search in first-person videos with significant egomotion, where face recognition and geometric localization are not applicable (*i.e.*, people with a high variability of appearances, captured by the cameras with significant motion, without any restriction on recorded places).

2. Correlation-based Self-search

Given a collection of first-person POV videos, our goal is to search for a target person (*i.e.*, the self) across many videos. Specifically, we wish to localize where the target appears in a collection of first-person POV videos. Here we consider two types of videos: videos recorded (1) by the target (*target video*) and (2) by observers (*observer videos*).

To perform a search, the target video is used as a unique identifier to compare against all observer videos. As mentioned above, an appearance-based search, such as face recognition, will fail when the face undergoes heavy occlusion, motion blur or when the face is not frontal parallel to an observer's camera. Instead we use the head motion pattern of the target which is more robust to these types of distortion. Global motion induced by ego-motion in the target video can be matched against local motions observed at candidate regions in observer videos. Instead of generating the candidates via detection [20], we utilize a more general supervoxel representation [32, 33]. Matched regions with high correlation are suggested as target instances.

2.1. Hierarchical Approach to Evaluate Targetness

The main technical challenge of this work arises when localizing target instances from the correlation evaluated at each supervoxel of observer videos. Namely, an image region corresponding to a target person is not necessarily found as one supervoxel, but likely to be under- or over-segmented. For example, supervoxels under-segmenting a target region merge with a part of backgrounds. To solve this problem, our approach generates a hierarchy of supervoxels to seek preferable segmentation of targets. While evaluating the correlation for each supervoxel, we also learn a discriminative model to consider generic targetness of supervoxels and avoid under- or over-segmentation of target regions potentially involved among the hierarchy. We frame an overall procedure as a binary-class Bayesian inference problem. That is, we aim to refine the likelihood derived from correlation-based targetness by the prior targetness learned in the discriminative model.

Denote a pixel as $x \in \mathbb{R}^{W\times H\times T}$ where W, H, T is the width, height and temporal length of an observer video. Each observer video is processed with a hierarchical supervoxel segmentation algorithm, such that we have a set of supervoxels $\mathcal{V} = \{v^{(1)}, \dots, v^{(N)}\}$ for each video. The supervoxel $v^{(n)}$ is a connected component (a spatiotemporal region) in the video. Since we retain the entire hierarchical segmentation tree, each pixel x belongs to a set of supervoxels $\mathcal{V}$, $\mathcal{V}_x = \{v^{(i)} \mid x \in v^{(i)}, v^{(i)} \in \mathcal{V}\}$.

We define a binary assignment variable a_x for each pixel in the observer video. Likewise, we define a binary assignment variable $a_{v^{(i)}}$ for every supervoxel $v^{(i)}$ in the observer video. The posterior probability of a pixel assignment given a target video V_{G} in terms of the supervoxel assignment variables covering that pixel is defined as:

$$\begin{aligned} P(a_x|V_{\mathrm{G}}) &\propto \prod_{v^{(i)}\in\mathcal{V}_x} P(a_{v^{(i)}} \mid V_{\mathrm{G}}) && (1)\\ &\propto \prod_{v^{(i)}\in\mathcal{V}_x} P(V_{\mathrm{G}} \mid a_{v^{(i)}})P(a_{v^{(i)}}). && (2)\end{aligned}$$

The goal of our approach is to find the pixel-wise binary labels that maximize this posterior. The individual supervoxel posteriors can be further decomposed using Bayes rule to obtain Eq. (2). We will estimate the likelihood $P(V_{\mathrm{G}} \mid a_{v^{(i)}})$ using a modified cross correlation score to deal with head motions in first-person videos (Sec. 2.2). The prior term $P(a_{v^{(i)}})$ learns various statistics of supervoxels indicating a trait for generic targetness and incorrect segmentation (Sec. 2.3).

2.2. Correlation-based Likelihood Computation

We introduce a novel correlation measure between the global motion patterns in target videos and local motion patterns in observer videos to estimate the likelihood $P(V_{\mathrm{G}} \mid a_{v^{(i)}})$ in Eq. (2). Let us begin with specific procedures to obtain those motion patterns. To calculate the global motion patterns in target videos, we follow [31] and first estimate homographies between consecutive frames. These homographies can generate global motion vectors at each pixel for every frame (see Appendix A for details). We then average these vectors for each frame to describe the global motion pattern by a sequence of two-dimensional (*i.e.*, horizontal and vertical) vectors according to [20].

We also calculate global motion vectors in observer videos to estimate 'residual' local motion vectors. We first compute dense optical flows such as [7], and then subtract the global motion vectors from these flows to obtain the local motion vectors. Finally, the local motion pattern in each supervoxel is computed by averaging the local motion vectors for each frame, but this time over regions where the supervoxel defines. Since vertical motions of target and observer videos are inverted (*e.g.*, head motions by nodding down appear as upper global motions in target videos), we inverted vertical elements of local motions.

Formally, let us denote the global motion at frame t as $\boldsymbol{g}_t \in \mathbb{R}^2$. In addition, we describe the local motion in a supervoxel at frame t by $\boldsymbol{l}_t \in \mathbb{R}^2$ (the ID of supervoxels is omitted without loss of generality). We consider the global and local motion patterns within the interval where the supervoxel defines (*i.e.*, global motion patterns are assumed to be cropped adaptively). We respectively denote them as $G_{b:e} = (\boldsymbol{g}_b, \dots, \boldsymbol{g}_e)$ and $L_{b:e} = (\boldsymbol{l}_b, \dots, \boldsymbol{l}_e)$, where b and e are the beginning and ending frames of supervoxel. In addition, both $G_{b:e}$ and $L_{b:e}$ are assumed to have zero mean by subtracting their mean value beforehand.

The key observation for our correlation measure is that, global motions in target videos are usually consistent with target head motions in first-person videos. Considering this observation, we compute the correlation between global and local motions on the subspace spanned by global motions. Since this subspace is uniquely characterized by the global motions, projecting local motions onto that effectively eliminates many irrelevant local motions. Indeed, first-person videos recording everyday life involve many motions other than target head motions. For example, hand gestures of observers will also induce local motions in their POV videos.

Particularly, we regard $\boldsymbol{g}_t$ as a single sample and perform a principal component analysis on $G_{b:e}$. The eigenvector corresponding to the larger eigenvalue is the subspace indicating the dominant orientation of global motions. We denote it as $\boldsymbol{s} \in \mathbb{R}^2$. Then, the cross correlation on the subspace is computed as follows:

$$C(G_{b:e}, L_{b:e}) = \frac{\sum_{t=b}^{e} \boldsymbol{g}_t^{\mathrm{T}} \boldsymbol{s} \cdot (\boldsymbol{l}_t)^{\mathrm{T}} \boldsymbol{s}}{\sqrt{\sum_{t=b}^{e} (\boldsymbol{g}_t^{\mathrm{T}} \boldsymbol{s})^2} \cdot \sqrt{\sum_{t=b}^{e} ((\boldsymbol{l}_t)^{\mathrm{T}} \boldsymbol{s})^2}}. \quad (3)$$

In practice, we need to evaluate videos (*i.e.*, computing the correlations) in a streaming manner to save memory resources as often done in supervoxel segmentation [33]. To that end, we first split videos uniformly into the sequence of short-length intervals. As a result, $L_{b:e}$ is also split into $L_{b_1:e_1}, \dots, L_{b_Q:e_Q}$ where $b_1 = b, e_q + 1 = b_{q+1}, e_Q = e$. Then, we evaluate Eq. (3) for each interval and weighted average them instead of $C(G_{b:e}, L_{b:e})$:

$$C'(G_{b:e}, L_{b:e}) = \sum_q \frac{e_q - b_q + 1}{e - b + 1} C(G_{b_q:e_q}, L_{b_q:e_q}). \quad (4)$$

Finally, we scale $C'(G_{b:e}, L_{b:e})$ into the range of $[0, 1]$ to deal it with the likelihood in Eq. (2).

2.3. Learning the Prior of Targetness

The prior $P(a_{v(i)})$ in Eq. (2) is introduced to consider generic targetness of supervoxels while avoiding potential incorrect segmentation among supervoxel hierarchies, such as under- or over-segmentation of target regions. This prior is learned from many pairs of observer videos and corresponding masks annotating target regions. We first extract positive and negative feature samples respectively from target and non-target regions, and then train a binary classifier. This classifier produces a posterior probability of supervoxels belonging to the target class, *i.e.*, $P(a_{v(i)})$. It also penalizes supervoxels that have feature values different from those of positive samples in the training data.

To capture a trait of generic targetness and incorrect segmentation, we extract the following features from supervoxels and underlying local motions.

Spatiotemporal sizes of supervoxels. The spatial area of supervoxels averaged over frames as well as their temporal length are used for features. These features indicate extremely short and small supervoxels that over-segment a target.

Sparseness and frequency of local motions. Local motions are expected to appear in very limited locations when the supervoxels under-segment targets. In other words, the sparseness of (per-pixel) local motions can be used as a trait for under-segmentation. We therefore calculate the standard deviation of local motions for each frame as a measure of sparseness. The obtained deviations are then averaged over time to serve as a feature. Furthermore, since people tend to move frequently as they interact with others, we also use the standard deviation of local motion patterns, $\sqrt{\sum_{t=b}^{e} ((\boldsymbol{l}_t)^{\mathrm{T}} \boldsymbol{s})^2}$, as a cue for generic targetness.

While manual annotations of target people are required here, this prior is independent of specific people and backgrounds. Therefore, learning of the prior needs to be carried out only once, and is not necessary for each target video.

3. Experiments

We first built a new first-person video dataset to evaluate how our approach performs on the self-search in details. We also evaluated its effectiveness on the CMU-group first-person video dataset used in [4, 18, 19]. Implementation details are described in Appendix B, and we would like to particularly note that our algorithm was able to run on the videos of size 320x180 (our dataset) and 320x240 (CMU dataset) while baseline face detectors and recognizers required full resolution videos (1920x1080 and 1280x960).

Data annotations and evaluations. We manually annotated image regions corresponding to a person's head at every 0.5 second. This is because local motions corresponding to ego-motion should be observed in a head region. These annotations also served as a supervised label for learning the prior $P(a_{v^{(i)}})$. Any classifier will work for the prior but we used the linear discriminant analysis because it performed the best. Since we evaluated two completely different datasets, we used one for training the prior to test the other. It ensured that any person and scene in test subsets did not appear in training ones. When evaluating performance, we referred to a set of target and observer videos as *session*. That is, the session is different if we swap targets for observers. For each session, we calculated the area under the receiver-operator characteristic curve (AUC) based on the pixel-wise comparisons between annotations and targetness scores, $P(a_x = 1 \mid V_{\mathrm{G}})$.

3.1. Evaluations on Our Dataset

Our dataset comprises 30 sessions. The number of participants wearing the cameras were two or three. These participants stayed at the same positions but often changed their poses. Interactions were recorded at 60fps, 30sec, under 4 indoor (I1 - I4) and 4 outdoor scenes (O1 - O4)[1].

Fig. 3 visualizes some of our target search results. In these results, we first searched the woman wearing a pink shirt from different POVs in I3. Then, we searched the man wearing a green shirt across scenes I1, O2 and O4. Finally, we tried to specify different targets in the same frame in O4. Our method successfully detected targets even if their appearances were drastically changed due to extreme changes of poses and unstable illuminations. To visualize target instances (in the 4th column of the figure), we applied GrabCut [22] on pixel-wise targetness to obtain regions with the maximum targetness.

Tab. 1 describes the AUC scores averaged over each scene. Several off-the-shelf face detectors and recognizers served as baselines. We used the mixtures-of-tree based facial landmark detector [37] (**ZR** in the table) and the Haar-cascade face detector [30] (**VJ**). **VJ** combined frontal and profile face models to permit head pose variations. We also employed face recognizers based on local binary pattern histograms [1] (**VJ+LBPH**) and Fisher face [6] (**VJ+FisherFace**). These two recognizers learned target faces from different sessions and ran on the detection results of **VJ**. Our approach obviously outperformed these baseline methods. While face recognizers eliminated incorrect detection of faces, they also failed to recognize target faces due to the high variability of facial appearances.

In addition, we implemented the following three variants of our method to see the contribution of each component.

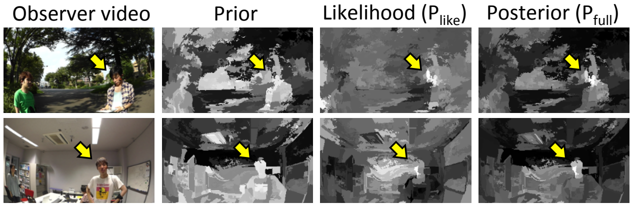

Figure 4. Comparisons between a prior, likelihood $\mathbf{P}_{\mathrm{like}}$ and posterior $\mathbf{P}_{\mathrm{full}}$. Target instances are annotated with yellow arrows.

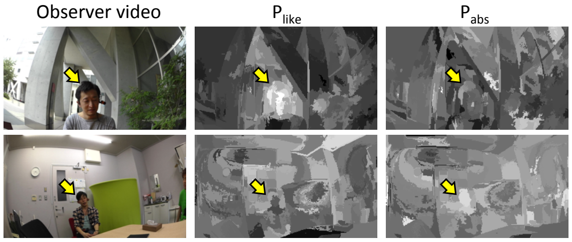

Figure 5. Comparisons between subspace cross correlations $\mathbf{P}_{\mathrm{like}}$ and amplitude-based cross correlations $\mathbf{P}_{\mathrm{abs}}$. Target instances are annotated with yellow arrows.

$\mathbf{P}_{\mathrm{full}}$. Evaluating $P(a_x \mid V_{\mathrm{G}})$ based on Eq. (2) and Eq. (4).

$\mathbf{P}_{\mathrm{like}}$. Using $P(V_{\mathrm{G}} \mid a_{v^{(i)}})$ instead of $P(V_{\mathrm{G}} \mid a_{v^{(i)}})P(a_{v^{(i)}})$ in Eq. (2) to see how the correlation likelihood and learned prior work.

$\mathbf{P}_{\mathrm{abs}}$. Using $\|\boldsymbol{g}_t\|$, $\|\boldsymbol{l}_t\|$ instead of $\boldsymbol{g}_t^{\mathrm{T}}\boldsymbol{s}$, $(\boldsymbol{l}_t)^{\mathrm{T}}\boldsymbol{s}$ in $\mathrm{P}_{\mathrm{like}}$ to see how the subspace correlation works.

Correlation likelihood and prior. $\mathbf{P}_{\mathrm{full}}$ demonstrated better performance than $\mathbf{P}_{\mathrm{like}}$ on most of the scenes. Fig. 4 depicts some comparisons between prior $\prod_{v^{(i)} \in \mathcal{V}_x} P(a_{v^{(i)}})$, likelihood $\mathbf{P}_{\mathrm{like}}$ and posterior targetness $\mathbf{P}_{\mathrm{full}}$. The prior was able to roughly eliminate non-target regions (dark regions in the second column) and sometimes highlighted target heads (yellow arrows). On the other hand, the likelihood in the third column localized targets but often generated false positives in non-target regions (bottom of the figure). Consequently, the two terms were complementarily, resulting in better final results (fourth column). Among the features designed in Sec. 2.3, we found that the spatial size of supervoxels and the sparseness of local motions were particularly discriminative.

On the subspace cross correlation. In many cases, the subspace cross correlation $\mathbf{P}_{\mathrm{like}}$ worked effectively compared to the amplitude-based correlation $\mathbf{P}_{\mathrm{abs}}$. As depicted in the above of Fig. 5, many false-positives are eliminated in $\mathbf{P}_{\mathrm{like}}$. However, $\mathbf{P}_{\mathrm{like}}$ was less robust when targets often inclined their head (bottom of the figure), violating the assumption that the cameras were mounted horizontally.

[1] We captured videos by Panasonic HX-A500. The dataset is available at http://www.hci.iis.u-tokyo.ac.jp/datasets/.

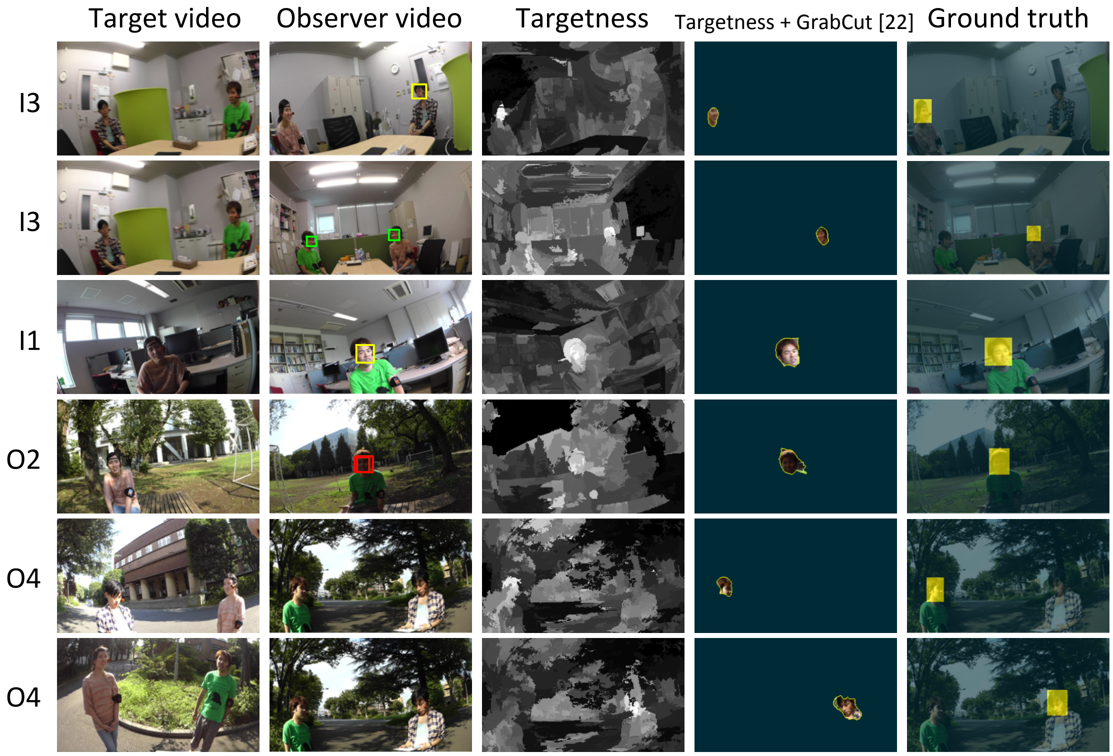

Figure 3. Target search results on our dataset. [I3] searched the woman wearing a pink shirt from different POVs in the same scene. [I1, O2, O4] searched the man wearing a green shirt across multiple scenes. [O4] specified different targets in the same frame. Results of **VJ** [30], **VJ+LBPH** [1] and **VJ+FisherFace** [6] are respectively depicted as green, red and yellow rectangles in the second column.

Table 1. AUC scores averaged over scenes. **ZR**: Mixtures-of-tree based facial landmark detector [37]. **VJ**: Haar-cascade face detector [30]. **VJ+LBPH**: face recognition using local binary pattern histograms [1]. **VJ+FisherFace**: face recognition using the fisher face [6].

	I1 (2 persons)	I2 (2 persons)	I3 (3 persons)	I4 (3 persons)	O1 (2 persons)	O2 (2 persons)	O3 (3 persons)	O4 (3 persons)	All
$\mathbf{P}_{\text{full}}$	**0.88**	**0.89**	0.75	**0.72**	**0.91**	**0.88**	**0.90**	**0.70**	**0.79**
ZR [37]	0.57	0.60	0.60	0.64	0.63	0.66	0.69	0.62	0.62
VJ [30]	0.66	0.73	0.66	0.62	0.75	0.77	0.73	0.62	0.67
VJ+LBPH [1]	0.50	0.50	0.51	0.54	0.57	0.55	0.55	0.50	0.52
VJ+FisherFace [6]	0.50	0.50	0.51	0.54	0.50	0.52	0.50	0.51	0.51
$\mathbf{P}_{\text{like}}$	0.85	0.87	0.64	0.61	0.83	0.85	0.81	0.65	0.71
$\mathbf{P}_{\text{abs}}$	0.68	0.75	**0.77**	0.63	0.73	0.68	0.78	0.62	0.70

3.2. Evaluations on the CMU Dataset

We also evaluated the effectiveness of our approach on the CMU-group first-person video dataset. 11 participants formed groups to (1) play pool, (2) play table tennis, (3) sit on couches to chat and (4) talk to each other at a table. They changed their poses and positions, and often disappeared from observer videos, standing for a more difficult scenario than our dataset. For 9 videos available for analyses, we used 3861st-5300th frames (30sec, at 48fps), where two people were involved in playing pool, three for table tennis, and the remaining four for the chat at the couches. 18 sessions served for evaluation in total.

Tab. 2 describes AUC scores and Fig. 6 visualizes example results. Again, our approach significantly improved the performance over face detectors and recognizers. We also found that limited performance in $\mathbf{P}_{\text{like}}$ and $\mathbf{P}_{\text{abs}}$ was due to the frequent disappearance of targets from observer videos. It suggests one limitation of correlation-based approaches; we must observe targets for a certain long time in observer videos to stably compute correlations as pointed out in [20].

4. Applications

The self-search over a repository of first-person videos is an important pre-processing in many studies and has a large potential impact on a variety of practical applications, including but not limited to privacy protection [20, 26], au-

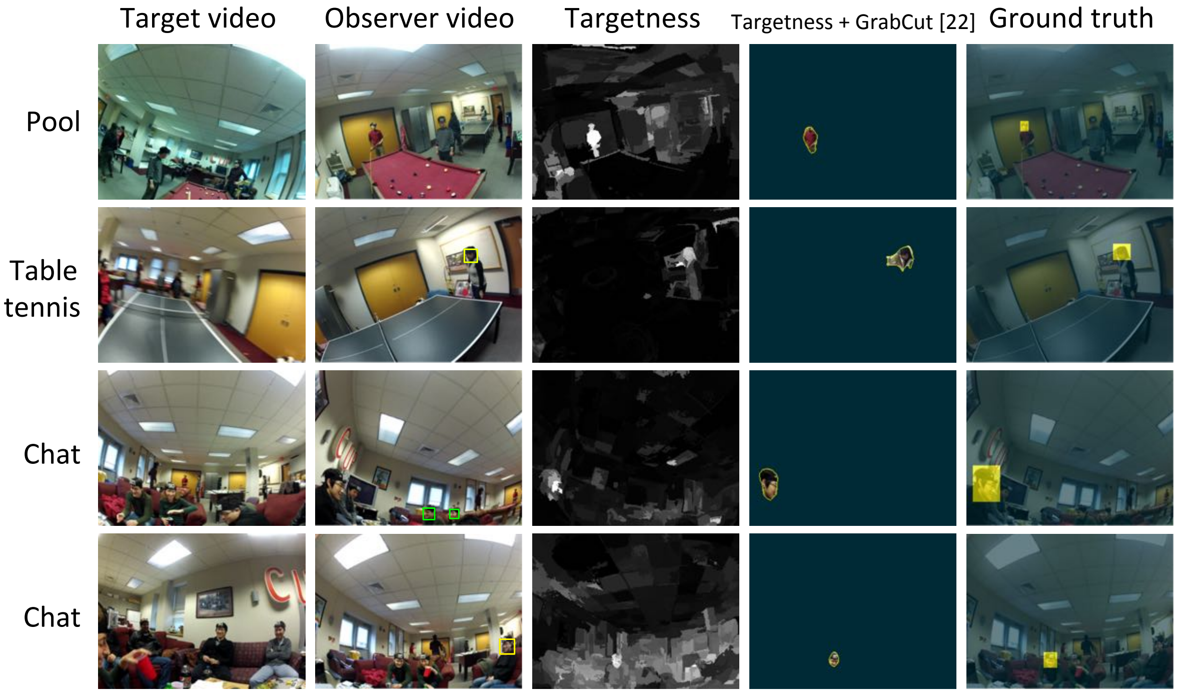

Figure 6. Target search results on the CMU dataset. Results of **VJ** [30], **VJ+LBPH** [1] and **VJ+FisherFace** [6] are respectively depicted as green, red and yellow rectangles in the second column.

Table 2. AUC scores averaged over scenes on the CMU dataset. **ZR**: Mixtures-of-tree based facial landmark detector [37]. **VJ**: Haar-cascade face detector [30]. **VJ+LBPH**: face recognition using local binary pattern histograms [1]. **VJ+FisherFace**: face recognition using the fisher face [6]. No face recognition results were provided for Pool data because each participant was observed in the only one session.

	Pool (2 persons)	Table tennis (3 persons)	Chat (4 persons)
$\mathbf{P}_{\text{full}}$	**0.83**	**0.80**	**0.79**
ZR [37]	0.49	0.50	0.54
VJ [30]	0.52	0.56	0.58
VJ+LBPH [1]	-	0.52	0.53
VJ+FisherFace [6]	-	0.51	0.53
$\mathbf{P}_{\text{like}}$	0.54	0.47	0.61
$\mathbf{P}_{\text{abs}}$	0.50	0.55	0.64

tomated video summarization [4, 14, 34], and social interaction analyses [3, 8, 18, 19, 21]. While the focus of this paper is to establish a new correlation-based search, we would like to suggest several proof-of-concept application scenarios.

Privacy filtering based on the self localization. As mentioned in Sec. 1, wearable cameras have a potential risk of recording videos of people without consent. To preserve their privacy from unintended recording, one solution is to localize and blur self instances. Specifically, one can localize oneself in other POV videos to blur the regions where targetness is high (above of Fig. 7). Alternatively, blurring other than target regions prevents us from violating others' privacy (bottom of Fig. 7).

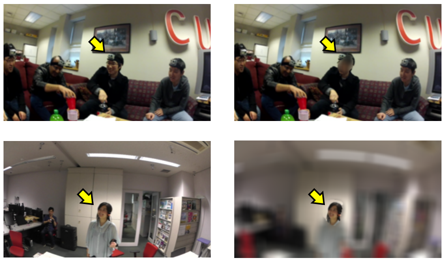

Figure 7. Privacy filtering based on the self localization. Target instances are annotated with yellow arrows.

Video collector with correlation signatures. While the prior considers generic targetness, the correlation in the likelihood serves as powerful signatures of specific targets. One practical usage of such correlation signatures is to collect video clips including targets from a large pool of first-person videos. We developed an automated video collector, which accepted target videos as a query. Example results on the CMU dataset are shown in Fig. 8. This way, we can collect the videos of ourselves from various POVs. Correlation signatures were defined as follows: for each video clip, we first computed median scores of correlation targetness

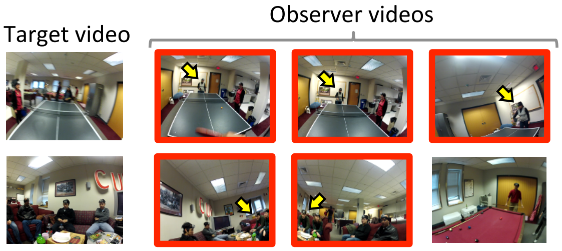

Figure 8. Automated video collector with correlation signatures. We used target videos as a query, and retrieved 3-best video clips. Target instances are annotated with yellow arrows. Correct retrievals are highlighted by red frames.

Table 3. Accuracies of social group discovery and the estimated numbers of the groups (the numbers in parentheses. The correct group number is 3) on the CMU dataset.

Interval (sec)	3	15	30
Proposed	**0.66 (3.27)**	**0.76 (3.00)**	**0.78 (3.00)**
GIST [28]	0.58 (1.72)	0.70 (2.20)	0.69 (2.80)

$P(V_{\mathrm{G}} \mid a_{v(i)} = 1)$ for each pixel within a certain temporal interval (*e.g.*, 15sec), and then calculated their maximum value within a frame. These signatures enabled us to highlight target instances appearing at one place for a long time.

Social group discovery based on correlation signatures. Another usage of correlation signatures is an unsupervised discovery of social groups from a collection of first-person videos. We pose this problem as a clustering problem where affinities between people are determined from the correlation signatures. To estimate the number of clusters (*i.e.*, groups) as well as a cluster assignment for each video, we adopted the affinity propagation algorithm [10]. Tab. 3 shows the accuracies (*i.e.*, ratios of true-positive and true-negative over all the assignments) and estimated numbers of groups. We split the CMU dataset into short clips of several different length for the evaluation. Since many objects in a background are different among groups (see Fig. 6), we implemented an unsupervised scene clustering as a baseline. Specifically, we used the GIST scene descriptor [28][2] encoded into a bag-of-words representation for features of the affinity propagation. Our approach improves the performance of group discovery regardless of video lengths.

5. Conclusions

This paper introduced a novel correlation-based approach to the problem of self-search for first-person videos. Experimental results proved that our search was able to localize self instances robustly even if well-known face recognizers were unavailable. One limitation of our approach is that it is not well suited for crowd scenes where many people may be moving in the same way (*e.g.*, too much correlation across multiple individuals) and individuals are only visible for short periods of time (*e.g.*, not enough signal). Localizing people in these types of videos will require a richer set of features. We leave this for future work.

Extending the self-search to a huge repository of first-person videos leads to many novel applications, including but not limited to what are suggested in Sec. 4. For example, searching people across first-person videos recorded at a variety of places around the world will illuminate their social activities in life, which have never been pursued by any visual surveillance. This application also raises new computer vision problems, such as supervoxel segmentation on large-scale video streams, camera-pose registration in a wide-spread area based on self-search and localization, and visual recognition from first-person multiple POV videos.

A. Motion estimation

To estimate global and local motions in Sec. 2.2, we followed the dense trajectory estimation presented in [31], except for local motion elimination by people detection. First, we associated consecutive frames with a homography matrix. Dense optical flows such as [7] thresholded by good-features-to-track criterion [24] as well as SURF features [5] produced the points. We adopted the optical flow estimation based on polynomial expansion [7]. To run the method in a reasonable time, we resized all the video frames into 160x90 (our dataset) and 160x120 (CMU dataset).

B. Implementation details

Supervoxel hierarchies $\mathcal{V}_x$ were computed by the streaming version of graph-based hierarchical oversegmentation [9, 33] implemented in LIBSVX[3]. To use default parameters provided with the codes, we resized frames into 320x180 (our dataset) and 320x240 (CMU dataset). We empirically used the supervoxels in 1, 5, 9, 13-th layers of 13 layers in total. The interval length for calculating the correlation $C'(G_{b:e}, L_{b:e})$ was set to 0.5sec. We chose to use a sigmoid function to scale the correlation into the form of likelihood, *i.e.*, $P(V_{\mathrm{G}} \mid a_{v(i)} = 1) = 1/(1 + \exp(-C'(G_{b:e}, L_{b:e})))$. We also tested several other options, such as different interval lengths (1sec and 2sec) and the linear scaling of $C'(G_{b:e}, L_{b:e})$, but confirmed that they did not affect overall performance.

[2] We used the code available at http://lear.inrialpes.fr/software.

[3] http://www.cse.buffalo.edu/ jcorso/r/supervoxels/

Acknowledgment

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

References

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 28(12):2037–2041, 2006.

[2] S. Alletto, G. Serra, S. Calderara, and R. Cucchiara. Head pose estimation in first-person camera views. In *Proc. International Conference on Pattern Recognition (ICPR)*, pages 1–6, 2014.

[3] S. Alletto, G. Serra, S. Calderara, F. Solera, and R. Cucchiara. From ego to nos-vision: Detecting social relationships in first-person views. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 594–599, 2014.

[4] I. Arev, H. S. Park, Y. Sheikh, J. Hodgins, and A. Shamir. Automatic editing of footage from multiple social cameras. *ACM Transactions on Graphics*, 33(4):81:1–81:11, 2014.

[5] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features. *Computer Vision and Image Understanding (CVIU)*, 110(3):346–359, 2008.

[6] P. N. Belhumeur, J. a. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 19(7):711–720, 1997.

[7] G. Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Proc. Scandinavian Conference on Image Analysis (SCIA)*, pages 363–370, 2003.

[8] A. Fathi, J. K. Hodgins, and J. M. Rehg. Social Interactions: A First-person Perspective. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1226–1233, 2012.

[9] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision (IJCV)*, 59(2):167–181, 2004.

[10] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–976, 2007.

[11] S. Gong, M. Cristani, C. Loy, and T. Hospedales. The re-identification challenge. In *Person Re-Identification*, Advances in Computer Vision and Pattern Recognition, pages 1–20. Springer London, 2014.

[12] J. Hesch and S. Roumeliotis. Consistency analysis and improvement for single-camera localization. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 15 – 22, 2012.

[13] E. Hjelmås and B. K. Low. Face detection: A survey. *Computer Vision and Image Understanding (CVIU)*, 83(3):236 – 274, 2001.

[14] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1346 – 1353, 2012.

[15] T.-S. Leung and G. Medioni. Visual navigation aid for the blind in dynamic environments. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 153 – 158, 2014.

[16] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2714–2721, 2013.

[17] A. Murillo, D. Gutierrez-Gomez, A. Rituerto, L. Puig, and J. Guerrero. Wearable omnidirectional vision system for personal localization and guidance. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 8 – 14, 2012.

[18] H. S. Park, E. Jain, and Y. Sheikh. 3D Social Saliency from Head-mounted Cameras. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1–9, 2012.

[19] H. S. Park, E. Jain, and Y. Sheikh. Predicting Primary Gaze Behavior using Social Saliency Fields. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 3503 – 3510, 2013.

[20] Y. Poleg, C. Arora, and S. Peleg. Head Motion Signatures from Egocentric Videos. In *Proc. Asian Conference on Computer Vision (ACCV)*, pages 1–15, 2014.

[21] J. Rehg, G. Abowd, A. Rozga, M. Romero, M. Clements, S. Sclaroff, I. Essa, O. Ousley, Y. Li, C. Kim, H. Rao, J. Kim, L. Presti, J. Zhang, D. Lantsman, J. Bidwell, and Z. Ye. Decoding children's social behavior. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3414 – 3421, 2013.

[22] C. Rother, V. Kolmogorov, and A. Blake. "grabcut": Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics*, 23(3):309–314, 2004.

[23] P. Salvagnini, L. Bazzani, M. Cristani, and V. Murino. Person re-identification with a ptz camera: An introductory study. In *Proc. IEEE International Conference on Image Processing (ICIP)*, pages 3552 – 3556, 2013.

[24] J. Shi and C. Tomasi. Good features to track. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 593 – 600, 1994.

[25] T. J. J. Tang and W. H. Li. An assistive eyewear prototype that interactively converts 3d object locations into spatial audio. In *Proc. ACM International Symposium on Wearable Computers (ISWC)*, pages 119–126, 2014.

[26] R. Templeman, M. Korayem, D. Crandall, and A. Kapadia. Placeavoider: Steering first-person cameras away from sensitive spaces. In *Proc. Annual Network and Distributed System Security Symposium (NDSS)*, 2014.

[27] Y. Tian, Y. Liu, and J. Tan. Wearable navigation system for the blind people in dynamic environments. In *Proc. Cyber Technology in Automation, Control and Intelligent Systems (CYBER)*, pages 153 – 158, 2013.

[28] A. Torralba, K. Murphy, W. Freeman, and M. Rubin. Context-based vision system for place and object recognition. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 273 – 280, 2003.

[29] R. Vezzani, D. Baltieri, and R. Cucchiara. People reidentification in surveillance and forensics: A survey. *ACM Compututing Surveys*, 46(2):29:1–29:37, 2013.

[30] P. Viola and M. Jones. Robust real-time object detection. *International Journal of Computer Vision (IJCV)*, 57(2):137–154, 2004.

[31] H. Wang and C. Schmid. Action Recognition with Improved Trajectories. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 3551 – 3558, 2013.

[32] C. Xu and J. J. Corso. Evaluation of super-voxel methods for early video processing. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1202–1209, 2012.

[33] C. Xu, C. Xiong, and J. Corso. Streaming Hierarchical Video Segmentation. In *Proc. European Conference on Computer Vision (ECCV)*, pages 1–14, 2012.

[34] Z. Ye, Y. Li, A. Fathi, Y. Han, A. Rozga, G. D. Abowd, and J. M. Rehg. Detecting eye contact using wearable eye-tracking glasses. In *Proc. ACM Conference on Ubiquitous Computing (UbiComp)*, pages 699–704, 2012.

[35] C. Zhang and Z. Zhang. A survey of recent advances in face detection. Technical Report MSR-TR-2010-66, 2010.

[36] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, 35(4):399–458, 2003.

[37] X. Zhu and D. Ramanan. Face detection, pose estimation and landmark localization in the wild. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2879 – 2886, 2012.