Discovering Objects of Joint Attention via First-Person Sensing

Hiroshi Kera
The University of Tokyo
Tokyo, Japan
kera@iis.u-tokyo.ac.jp

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

Keita Higuchi
The University of Tokyo
Tokyo, Japan
khiguchi@iis.u-tokyo.ac.jp

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

Abstract

The goal of this work is to discover objects of joint attention, i.e., objects being viewed by multiple people using head-mounted cameras and eye trackers. Such objects of joint attention are expected to act as an important cue for understanding social interactions in everyday scenes. To this end, we develop a commonality-clustering method tailored to first-person videos combined with points-of-gaze sources. The proposed method uses multiscale spatiotemporal tubes around points of gaze as a candidate of objects, making it possible to deal with various sizes of objects observed in the first-person videos. We also introduce a new dataset of multiple pairs of first-person videos and points-of-gaze data. Our experimental results show that our approach can outperform several state-of-the-art commonality-clustering methods.

1. Introduction

Shifts in attention are one of the fundamental behaviors during everyday social interactions. For instance, we look at various targets of objects including speakers, handouts, and a projector screen during a meeting in an office. When multiple people cooperatively assemble something big, they continuously pay attention to various objects such as parts to be assembled and tools in their hands. To understand such interactions, we need to find objects jointly viewed by multiple people. Such objects of joint attention reflect what people attend to from moment to moment and can be used as a cue to understand group activities [8, 28]. In the context of computer-supported cooperative work, the ability of extracting objects of joint attention allows us to evaluate how systems mediate collaborative work of people [27].

In this work, we argue that one promising approach to extract objects of joint attention is to use wearable cameras and eye trackers mounted on the head of people during interactions. First-person points-of-view videos recorded by such cameras can clearly capture what people see and thus can be used for action recognition [6, 22] and activity summarization [2, 4, 12, 15, 28]. More importantly, points-of-gaze data measured by an eye tracker often illuminate the parts of the wearer's field of view that receive attention. This enables localizing important objects spatially and temporally [7, 8, 24, 28, 29].

Figure 1. Objects of joint attention are discovered in multiple first-person videos recorded during interactions (highlighted frames). Points of gaze of camera wearers are annotated by crosses.

Motivated by these advantages of wearable cameras and

eye trackers, we introduce a new task of discovering objects of joint attention from multiple first-person videos recorded with additional inputs from eye trackers. We illustrate some results from our experiments in Figure 1. Using points-of-gaze data of each camera wearer, we split first-person videos into a shot sequence by detecting eye movements from one object to another. Then, we find shots that contain objects with similar appearances across multiple videos (highlighted frames in the figure).

One important problem in discovering objects of joint attention is how to define appropriately a region in first-person videos, from which we extract features to describe objects being viewed. While some studies use regions of a specific fixed size around points of gaze [8, 13, 28], comparing directly between fixed-size regions does not always work well due to the variability in the size of objects in first-person videos. In our everyday life, we look from a small tool in our hands to a large poster on the wall. The size of these objects changes even more drastically in first-person videos because the objects can be seen from different distances. As a result, features extracted from fixed-size regions can only describe a limited part of objects or are affected by a large amount of irrelevant background regions. This makes it difficult to temporally segment videos into shots reliably based on objects of focus and to compare the objects among shots.

We introduce a multi-scale approach for object-feature extraction to address this problem. In the proposed method, visual features are extracted around points of gaze with several different areas to take into account the size variability of objects. These visual features are further used to split an input video into shots based on several different affinity criteria. This approach allow us to generate as a candidate of objects, several different scales of spatiotemporal "tubes" around points of gaze, where some of them are expected to match closely actual regions of objects being viewed. A group of tubes with similar features are discovered for each scale via unsupervised commonality clustering. Discovery results are finally integrated across scales to find various sizes of objects of joint attention reliably.

The main contributions of this paper are summarized as follows: (1) we introduce a new task of discovering objects of joint attention from first-person videos; (2) We describe a method we developed to discover objects of joint attention using multiscale spatiotemporal tubes as object candidates; and (3) we present a novel dataset containing multiple pairs of first-person videos and points of gaze data to validate the effectiveness of our approach.

2. Related work

In this section, we review some prior work related to the task of discovering objects of joint attention from first-person videos using points of gaze information. Because wearable cameras and eye trackers have become available at a reasonable price, first-person vision is now one of the emerging topics in computer vision. Similar to our work, Park *et al.* [18, 19, 20] proposed detecting a social focus of attention during group interaction using multiple first-person videos. In their work, the location of social focus was found as an intersection of people's viewing directions computed from 3D camera poses and positions. One important problem is that such intersections may not correspond to a true social focus. For instance, two people's viewing directions can intersect while they are looking at different things behind the intersection. In addition, the use of 3D camera poses and positions often requires a 3D model of the scene that may not always be available.

Points-of-gaze data act as a salient cue to boost various computer vision tasks. Because points of gaze are indicative of important parts in images, they are used to recognize objects [29] and actions [7, 24] or to summarize videos by detecting important shots [28]. To the best of our knowledge, this work is the first to use multiple points-of-gaze sources to discover important objects across multiple videos.

The ability to discover commonalities across multiple images or videos has also been adopted in a variety of computer vision tasks, such as object co-segmentation [10, 23, 30], co-localization [25], and temporal commonality discovery [5]. Perhaps the most relevant work presented is common-interest person detection from multiple first-person videos [14]. Accurate human detection is required to generate candidates of co-interest people. In comparison to this approach, we make use of points-of-gaze information to generate candidates of common objects and do not require any object detectors. This enables co-localizing any categories of objects in a scene.

3. Our method

Our method accepts as input N pairs of first-person videos and points-of-gaze data captured by each of the N people during interactions and outputs time intervals where the same object is viewed in all of the N videos (*i.e.*, an object of joint attention). More formally, we consider a time sequence $\mathcal{T} = [1, 2, \ldots, T]$. Each time $t \in \mathcal{T}$ has N image frames $V_{1,t}, \ldots, V_{N,t}$ and two-dimensional points of gaze $\boldsymbol{g}_{1,t}, \ldots, \boldsymbol{g}_{N,t} \in \mathbb{R}^2$. The goal of this work is to obtain a time interval $\mathcal{J} \subset \mathcal{T}$ where all image frames $\{V_{n,t} \mid t \in \mathcal{J}, n \in [1, 2, \ldots, N]\}$ contain instances of the same object around the corresponding point of gaze $\boldsymbol{g}_{n,t}$.

In Section 3.1, we first explain generating multiscale spatiotemporal tubes from videos to describe objects being viewed. Then, in Section 3.2, we describe how to perform unsupervised commonality clustering on the tubes to discover time intervals where joint attention is likely to occur for each scale. Finally, we introduce a voting scheme to integrate the discovery results across scales in Section 3.3.

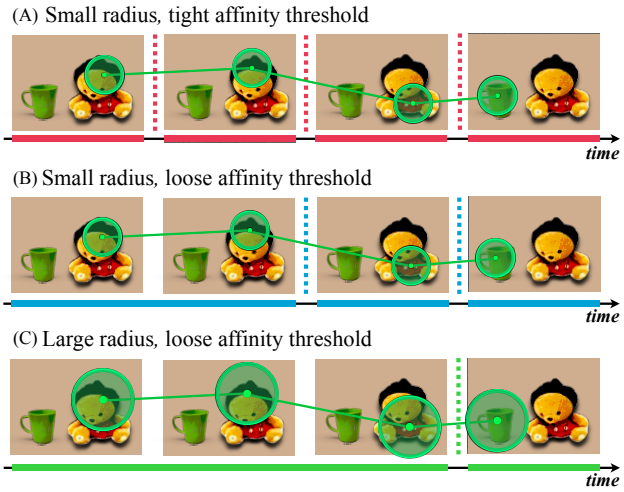

Figure 2. Concept figure of multiscale spatiotemporal tubes. Colored time axes represent time intervals split with several pairs of a radius and an affinity threshold. (A, B) Smaller radius of tubes is more appropriate to extract features from the object on the left side; (C) Larger radius and longer length are needed to cover the object on the right.

3.1. Generating Multiscale Spatiotemporal Tubes

When we see objects, points of gaze are often distributed over important parts of the objects. If we properly split videos into a sequence of shots (sub-sequences of image frames) by detecting eye movements from one object to another, we can then extract visual features from regions around points of gaze to describe objects of focus in each shot. However, the size of regions that match closely to important parts of objects should differ depending on the apparent object sizes in videos. We need to define a proper spatial range around points of gaze for feature extraction so that we can reliably segment videos into shots and compare instances of objects across multiple videos.

We address this problem by generating spatiotemporal tubes along points of gaze at various scales from which we extract features of objects being viewed. As illustrated in Figure 2, we expect that an appropriate combination of spatial and temporal ranges will cover important parts of objects correctly. Let us denote by $\mathcal{F}(V_{n,t})$ a set of local features extracted from image frame $V_{n,t}$. We consider a set of spatial ranges $\mathcal{R} = \{r_1, \ldots, r_{N_r}\}$ that control a radius of spatiotemporal tubes. For each $r \in \mathcal{R}$, a feature vector of what people see in $V_{n,t}$ is then described by $\boldsymbol{s}_{n,t}^{(r)} = \mathcal{H}(\{\boldsymbol{f} \in \mathcal{F}(V_{n,t}) \mid \|\boldsymbol{l}(\boldsymbol{f}) - \boldsymbol{g}_{n,t}\| < r\})$, where $\boldsymbol{l}(\boldsymbol{f}) \in \mathbb{R}^2$ is a spatial location that the feature $\boldsymbol{f}$ is extracted from, and $\mathcal{H}$ is a certain feature-aggregation operator that takes as an input a set of local features, such as a naive histogram and Fisher vector coding [21].

A time interval where spatiotemporal tubes are defined is given by temporally segmenting videos into shots based on a frame-wise feature $\boldsymbol{s}_{n,t}^{(r)}$ with multiple thresholds. Specifically, we compute affinities between consecutive frames $\boldsymbol{s}_{n,t-1}^{(r)}, \boldsymbol{s}_{n,t}^{(r)}$ and find shot boundaries where the affinities are below one of a set of affinity thresholds $\theta \in \Theta$. These multiple thresholds allow us to segment videos into shots based on objects of focus while considering a variety of similarities among multiple objects in a scene.

As a result, we obtain a sequence of spatiotemporal tubes for each video given a certain combination of spatial range and affinity threshold parameters. We describe the time interval of k-th tube by $j_{n,k}^{(p_n)} \subset \mathcal{T}$, where $p_n = (r_n, \theta_n) \in \mathcal{R} \times \Theta$ is a specific combination of parameters used for extracting features from the n-th video. Finally, visual features of objects being viewed in the k-th shot are extracted by aggregating features in the tube: $\boldsymbol{s}_{n,k}^{(p_n)} = \mathcal{H}(\{\boldsymbol{f} \in \mathcal{F}(V_{n,t}) \mid \|t \in j_{n,k}^{(p_n)}, \boldsymbol{l}(\boldsymbol{f}) - \boldsymbol{g}_{n,t}\| < r_n\})$.

3.2. Commonality Clustering on Tubes

To discover objects of joint attention, we perform unsupervised commonality clustering on feature vectors $\boldsymbol{s}_{n,k}^{(p_n)}$ extracted from spatiotemporal tubes. In what follows, we particularly focus on the two-person case (*i.e.*, $N = 2$) for the sake of simplicity. We will discuss in Section 3.3 how our method can be extended to more than two-person cases.

For each combination of scale parameters p_1, p_2, we aim to find a "co-cluster" of spatiotemporal tubes that have similar features. To this end, we first define an affinity matrix between tubes across a pair of videos.

$$A = \begin{pmatrix} O & C \\ C^\top & O \end{pmatrix}, \tag{1}$$

where the (i, j)-th entry of the matrix C is given by the affinity between $\boldsymbol{s}_{1,i}^{(p_1)}$ and $\boldsymbol{s}_{2,j}^{(p_2)}$. A concrete affinity function will be given in Section 3.4. Similar to normalized spectral clustering [17], we also introduce a degree matrix D: a diagonal matrix where the i-th diagonal element is given by the sum of the entries in the i-th row of A. Then, as described in [4], co-clusters can be obtained via normalized spectral clustering with the Laplacian matrix $L = D - A$. In practice, we perform the two-class clustering and select one co-cluster whose members have higher affinities. Note that in a particular situation where objects of joint attention are observed sparsely during interactions, the maximal-biclique-based approach proposed in [4] can also be applied.

Given the co-cluster of tubes for scale parameter combination p_1, p_2, the time interval where an object of joint attention is likely to be observed, $\mathcal{J}^{(p_1,p_2)} \subset \mathcal{T}$, is determined as follows. Let us denote by K_n a set of tube indices in n-th video belonging to the discovered co-cluster.

Recall that the k-th tube of n-th video is defined in interval $j_{n,k}^{(p_n)} \subset \mathcal{T}$. The interval $\mathcal{J}^{(p_1,p_2)}$ is then obtained by finding all the intersections of intervals between a pair of videos:

$$\mathcal{J}^{(p_1,p_2)} = (\cup_{k\in K_1} j_{1,k}^{(p_1)}) \cap (\cup_{k\in K_2} j_{2,k}^{(p_2)}). \quad (2)$$

Note that co-clusters discovered using the affinity A in Eq. (1) always contain tubes from both of the two videos. If no intersections are found in Eq. (2) at a certain combination of scales (p_1, p_2), the result from that scale setting is just ignored in the subsequent voting scheme.

3.3. Voting across Multiple Scales

Finally, we integrate discovered time intervals $\mathcal{J}^{(p_1,p_2)}$ across all the scale combinations $\mathcal{R} \times \Theta$ to discover objects of joint attention with the variability in their size. To this end, for each scale setting, we weigh how likely the discovered co-cluster of spatiotemporal tubes includes objects of joint attention. More specifically, we design a confidence score $c^{(p_1,p_2)}$ computed by the sum of affinities among spatiotemporal tubes corresponding to $j_{n,k}^{(p_n)} \subset \mathcal{J}^{(p_1,p_2)}$. This score increases when tubes in the co-cluster are more similar.

The confidence scores are then summed up per frame $t \in \mathcal{T}$ to construct a confidence histogram. This histogram is aimed at describing in which time intervals we observe more confident co-clusters:

$$c_t = \sum_{p_1,p_2\in\mathcal{R}\times\Theta} c^{(p_1,p_2)} \delta(t, \mathcal{J}^{(p_1,p_2)}), \quad (3)$$

$$\delta(t, \mathcal{J}^{(p_1,p_2)}) = \begin{cases} 1 & t \in \mathcal{J}^{(p_1,p_2)} \\ 0 & \text{otherwise.} \end{cases} \quad (4)$$

The time interval including objects of joint attention $\mathcal{J}$ is derived by binarizing $c_1 \ldots, c_T$ with a certain threshold.

This voting scheme can be extended to cases where more than two people are present, as follows. We first conduct the commonality clustering presented in Section 3.2 for all the pairs of videos. Then, the confidence histogram is built by aggregating confidence scores over multiple scales as well as multiple video pairs. Intuitively, the more people see the same object in a certain frame t, the higher the score is given to c_t. We show in Section 4.4 how this voting scheme works on three-person cases.

3.4. Implementations

Here, we briefly describe some important implementations of our method. More details are described in the appendix. To describe the appearances of objects, we use dense RootSIFT descriptors [1] encoded by the improved Fisher vector [21] and HSV color histograms. Furthermore, we take into account time intervals to avoid matching tubes observed at a completely different time. A time interval of a shot $j_{n,k}^{(p_n)}$ is represented by a T-dimensional feature vector, whose t-th element takes $\frac{1}{|j_{n,k}^{(p_n)}|}$ if $t \in j_{n,k}^{(p_n)}$ (where $|j_{n,k}^{(p_n)}|$ is the number of image frames in $j_{n,k}^{(p_n)}$) and otherwise zero. All these features are aggregated to form feature vectors $\boldsymbol{s}_{n,k}^{(p_n)}$. Note that we used only the color histograms for per-frame features $\boldsymbol{s}_{n,t}^{(p_n)}$ in video-shot segmentation because it performed better.

Another important implementation is the affinity function used in the video-shot segmentation and commonality clustering. We define the affinity between two features $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ by $\exp(-\rho\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|)$, where $\|\cdot\|$ is the Euclidean distance and ρ is set to the median of all distance values.

4. Experiments

To evaluate the effectiveness of our approach, we built a new dataset containing multiple pairs of first-person videos and points-of-gaze data. The experiments demonstrate that our approach can outperform several state-of-the-art commonality clustering methods on the task of discovering objects of joint attention in various interaction scenes.

4.1. Data Collection

Our new dataset consists of 29 sequences of two- and three-person interaction scenes recorded in three different environments. Each subject was equipped with a head-mounted camera and an eye tracker to record first-person videos and points-of-gaze data collectively. To the best of our knowledge, this dataset is the first to use multiple points-of-gaze sources in first-person vision tasks.

During each recording, subjects were asked to establish joint attention on various objects such as books, projector screens, and faces, like they do in their everyday interaction. Specific types of interactions included object exchanges, pointing by hands followed by shifts in attention, and jointly looking at a person who came into a room. In two-person sequences, subjects took one of two formations: side-by-side (SbS) and face-to-face (FtF). In the SbS sequences, two subjects sat next to each other where objects of joint attention were located in front of the subjects. As for the FtF sequences, subjects were facing each other across from the objects to be looked at jointly. In the three-person sequences, subjects were positioned in a triangle at difference distances. In the dataset, we have 14 SbS, seven FtF, and eight triangle sequences.

We used the Pupil Lab eye trackers [11] to record HD-resolution first-person videos with points-of-gaze data at 30 fps. All videos and gaze data were synchronized manually. While the length of each sequence varied from 40 to 120 seconds, we downsampled all the videos and points-of-gaze data to have 500 frames per sequence. This makes

the length of time-interval feature vectors presented in Section 3.4 equal for all the sequences. Each video was downsized to 320x180 before feature extraction to reduce computational cost. Eye trackers were calibrated before each recording session. Missing gaze data due to eye blinks or tracking failures were filled with linear interpolation.

Each sequence was manually annotated with ground truth labels of time intervals where all subjects looked at the same object. More specifically, we annotated a binary label to the frames based on whether objects of joint attention were located within a 15-pixel radius around points of gaze at the 320x180 resolution.

4.2. Evaluation Scheme and Baselines

We calculate the area under ROC curves (AUC scores) on confidence histograms and binary ground truth labels to evaluate how accurately our outputs in Eq. (4) can capture correct time intervals. First, we present a comparison of our method with some baseline methods on two-person sequences (*i.e.*, SbS and FtF). We implemented the following three methods for the baselines.

Simplified version of our method. To provide evidence for the effectiveness of using a multi-scale approach, we implemented the simplified version of our method that used only a single combination of a spatial radius and an affinity threshold. In the experiments, we *manually* selected one parameter combination for each formation that produced the highest AUC score.

Temporal commonality discovery. Chu *et al.* [5] introduced the temporal commonality discovery (TCD) method to extract a pair of common temporal patterns from two input videos via branch and bound. We performed the TCD to find a pair of time intervals with similar object-feature patterns from a pair of videos. We extracted HSV color histograms and RootSIFT Fisher vectors around points of gaze as well as a time interval feature vector for each frame. Similar to the aforementioned simplified version, we manually selected one radius to extract features that produced the highest AUC score for each formation.

Co-localization. We also adopted a co-localization method (COLOC) proposed by Tang *et al.* [26] as another baseline. Originally, the COLOC generates object proposals for each image and finds a group of proposals that are similar. Instead of object proposals, we used spatiotemporal tubes for each video. The tubes were constructed and evaluated in the same way as in the simplified method.

4.3. Results

Figure 3 shows some of the results of our approach. In each example, subjects were involved in the following interaction: (A) a subject showed a book to the other subject sitting next to him so that they could read it together; (B) two subjects sitting side-by-side looked at a green mug held by another person; (C) a subject looked at a projector screen and spoke to the other subject to see it; (D) two subjects saw a teddy bear from different points of view; (E) two subjects sitting face to face exchanged a book; and (F) a subject asked the other subject in front to put a block into a cylindrical box.

Method	SbS	FtF	Avg.
(1) COLOC ($r = 15, \theta = 50$) [26]	0.57	0.50	0.53
(2) COLOC ($r = 15, \theta = 10$) [26]	0.52	0.71	0.61
(3) TCD ($r = 50$) [5]	0.50	0.48	0.49
(4) TCD ($r = 15$) [5]	0.48	0.48	0.48
(5) Simplified ($r = 50, \theta = 10$)	0.75	0.47	0.61
(6) Simplified ($r = 15, \theta = 30$)	0.63	**0.82**	0.73
Ours	**0.87**	0.79	**0.83**

Table 1. AUC scores of the proposed and baseline methods. Combinations of spatial radius r and affinity threshold θ were manually selected to provide the highest AUC score in SbS sequences ((1), (3), (5)) and FtF ones ((2), (4), (6)) in baselines.

We found that higher confidence scores were given to correct time intervals in many cases. Our method worked robustly on various sizes of objects from a small mug in (B) to a large projector screen in (C). We were also able to deal with cases when the size of object instances were drastically different, as shown in (D)(E)(F). By using points of gaze to limit the location of features to be extracted and compared, we can discover objects of joint attention even when background scenes are greatly similar across videos, such as in example (D). This unique property of our approach is unlike many standard object co-localization and co-segmentation methods [10, 23, 25, 30] that assume background scenes are different across images.

We also present quantitative evaluations based on ROC curves and AUC scores in Figure 4 and Table 1. On average, our method using multi-scale spatiotemporal tubes performed the best. Among the baseline methods, the combination of scale parameters (r and θ) that provided the highest AUC scores were different between SbS and FtF sequences. This indicates the necessity of considering multiple scales to cope with various sizes of objects in videos.

4.4. More than Two-Person Cases

Figure 5 shows how our method can work on cases where three subjects are present in a scene. In example (A), a teddy bear was passed from one subject to another followed by a third subject paying attention to the interaction. In (B), one subject manipulated a box and asked the other subjects to look at the box. For both cases, our method successfully discovered the objects of joint attention, while the size

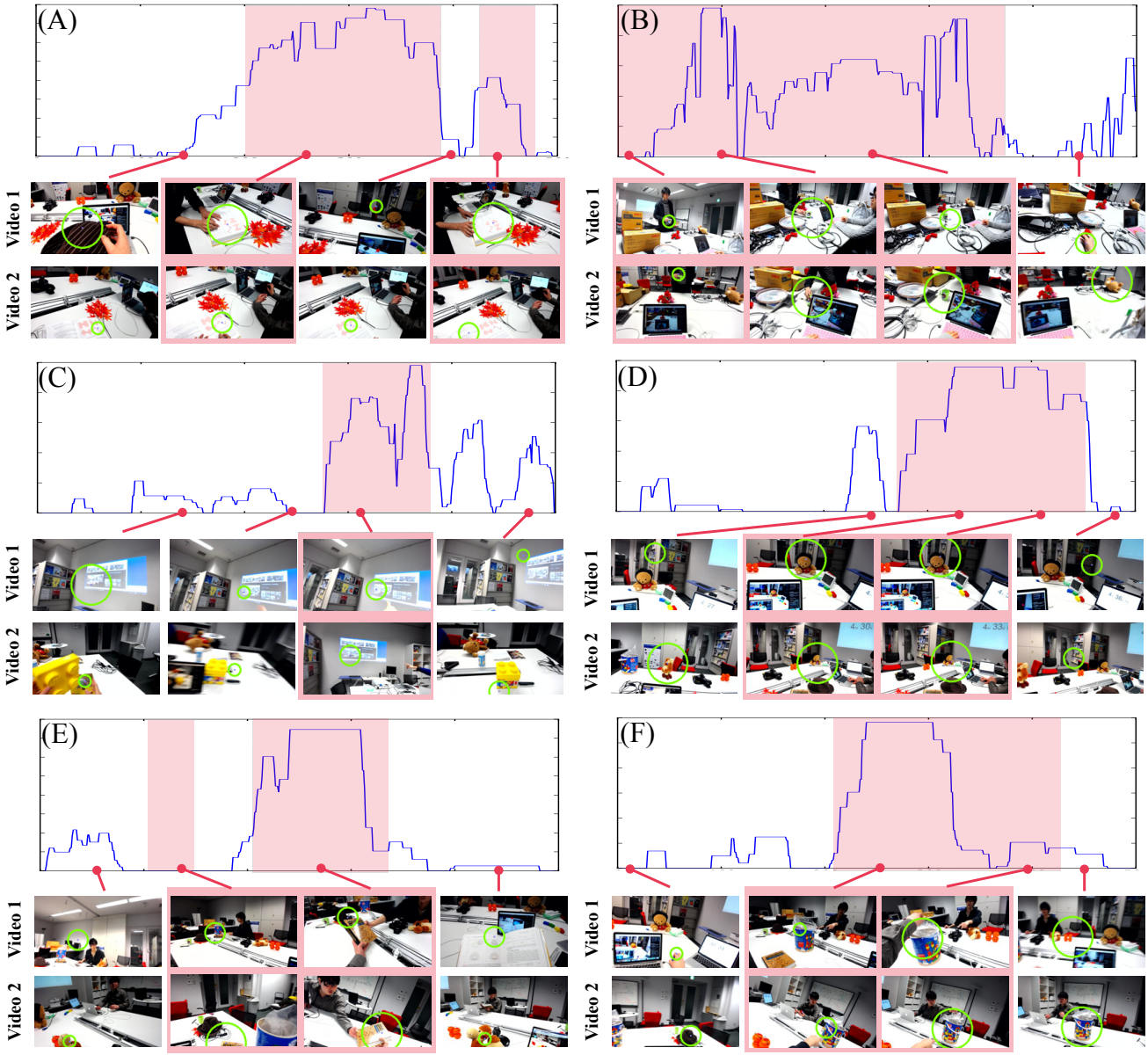

Figure 3. Confidence histograms and image frames. Time intervals and image frames where objects of joint attention were observed are highlighted in pink. Green circles denote regions attended by subjects. We selected the radius from the scale pair that gives the highest confidence score at each time point.

of object instances varied significantly among videos (*e.g.*, larger instances in the point of view of the person holding an object and smaller instances in the other people's points of view). The AUC scores on the three-person sequences were on average 0.74.

4.5. Failure Cases and Possible Extensions

Figure 3 includes some failure cases. Discovering objects that were barely observed in first-person videos was difficult (*e.g.*, the book in hands in example (E)). Moreover, false-positive responses were observed when subjects kept looking at textureless regions like in (C). Some other failure cases were present in Figure 6. In example (A), two subjects looked at objects that had a similar appearance but that were located in different places. Our method is not yet able to distinguish such pairs of objects because it relies on only the visual appearances of a scene. Another case of false-negative detection is depicted in (B); objects appeared

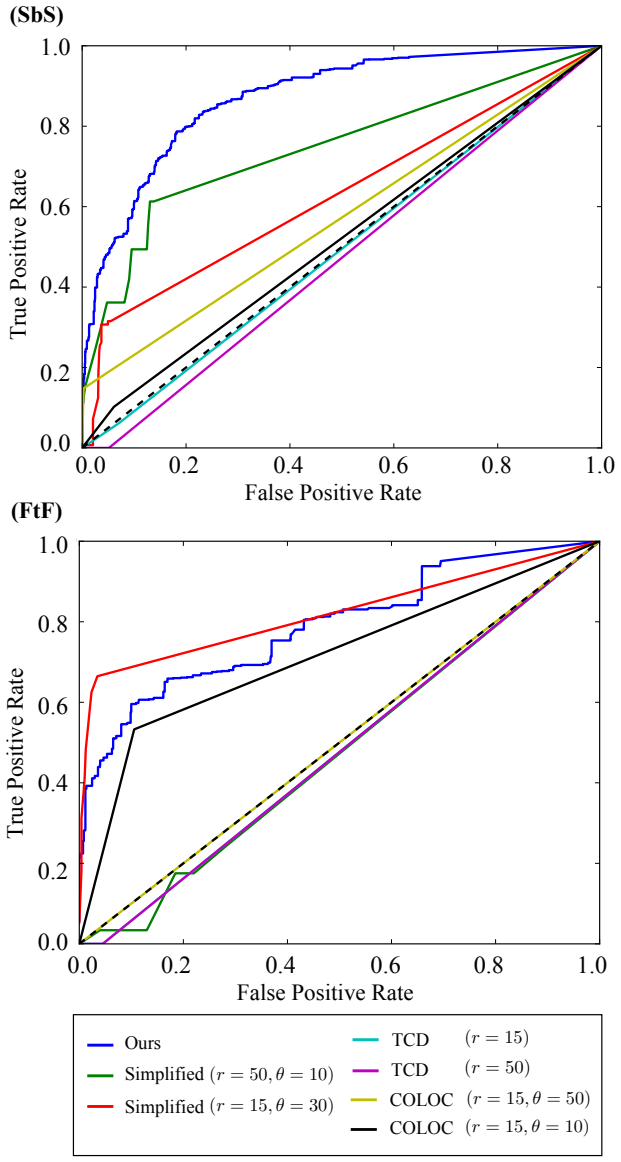

Figure 4. ROC curves of the proposed and baseline methods

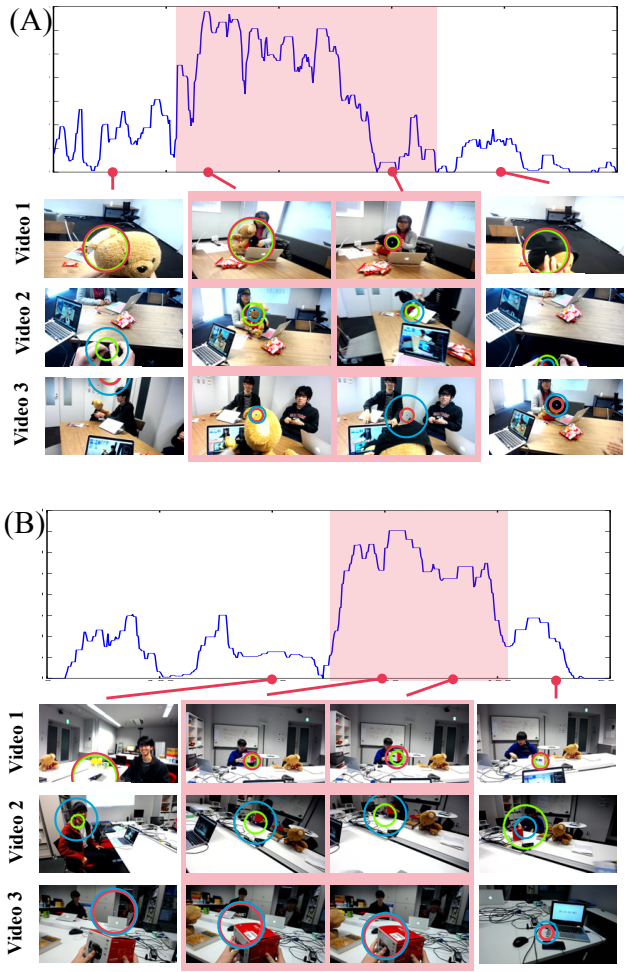

Figure 5. Confidence histograms and image frames. Time intervals and image frames where objects of joint attention were observed are highlighted in pink. Circles denote regions attended by subjects. We selected the radius from the scale pair that gives the highest confidence score at the time point. Green, red, and blue circles correspond to video pairs of video 1 and 2, video 1 and 3, and video 2 and 3, respectively.

differently across videos due to lighting conditions.

Introducing more sophisticated features such as R-CNN features [9] to handle higher level information such as object classes may ease the issue in (B). Incorporating other types of features that do not rely on object appearances is also an interesting extension. When a geometric relationship between head-mounted cameras is possible by preliminarily scanning a scene like [18], we will be able to distinguish objects placed at a different location. If we particularly focus on objects in motion (*e.g.*, objects carried by hands), motion patterns can also be a salient cue [14].

Another interesting extension is to use segmentation around fixation points [16] or object proposal [3] instead of spatiotemporal tubes. The former extracts objects around points of gaze by segmentation, while the latter provides bounding boxes for object-like regions, which both allow us to avoid the size variability issue while considering cluttered backgrounds. However, these approaches may not be directly applied to our problem because they are not always good at dealing with non-salient or non-textured objects.

5. Conclusions

In this work, we introduced a novel task of discovering objects of joint attention in multiple first-person videos. Our experimental results demonstrated the effectiveness of our multiscale approach over several state-of-the-art common-

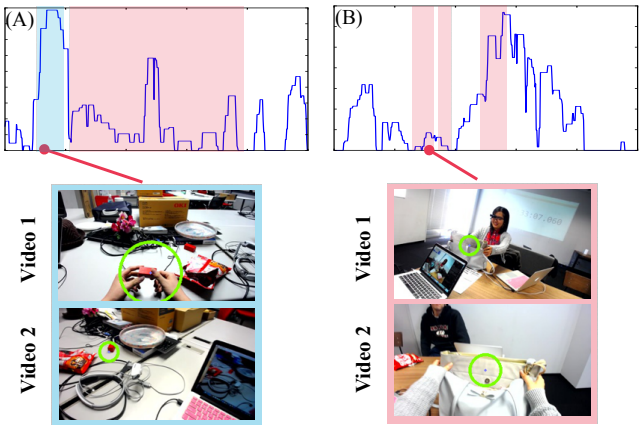

Figure 6. Confidence histograms and image frames for failure cases. (A) False-positive detection highlighted in blue and (B) False-negative detection.

ality discovery methods. Our future work will be to improve feature description and object-candidate generation. Another important direction is to develop an efficient algorithm to discover objects of joint attention in real time.

A. Implementation Details

Here are some details on our implementations. When constructing HSV color histograms, we discretized each color channel into 16 bins and normalized them independently. They were then aggregated and normalized again to form 48-dimensional histogram vectors. For features of spatiotemporal tubes, we used the histogram vector at the median frame in each shot. To obtain RootSIFT Fisher vectors, we first applied PCA to SIFT descriptors to have 64 dimensions. The number of GMM components trained for Fisher vectors was also 64. We adopted the L2 and power normalizations on the Fisher vectors by following [21].

In video-shot segmentation, we preliminarily applied a median filter with a kernel size of 15 to a sequence of affinities to cope with outliers. After the shot segmentation, we removed some shots whose length was shorter than 15 frames. A set of spatial radius parameters was set to $\mathcal{R} = \{15, 25, 50\}$ in pixels. Affinity thresholds were obtained by computing 10th, 30th, and 50th percentiles of all the affinities for each video.

Acknowledgements

This research was supported by CREST, JST.

References

[1] R. Arandjelovic and A. Zisserman. Three things everyone should know to improve object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2911–2918, 2012. 4324

[2] 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. 4321

[3] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. H. S. Torr. BING: Binarized normed gradients for objectness estimation at 300fps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3286–3293, 2014. 4327

[4] W.-S. Chu, Y. Song, and A. Jaimes. Video co-summarization: Video summarization by visual co-occurrence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3584–3592, 2015. 4321, 4323

[5] W.-S. Chu, F. Zhou, and F. De la Torre Frade. Unsupervised temporal commonality discovery. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 373–387, 2012. 4322, 4325

[6] A. Fathi, J. K. Hodgins, and J. M. Rehg. Social interactions: A first-person perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1226–1233, 2012. 4321

[7] A. Fathi, Y. Li, and J. M. Rehg. Learning to recognize daily actions using gaze. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 314–327. Springer-Verlag, 2012. 4321, 4322

[8] A. Fathi, X. Ren, and J. Rehg. Learning to recognize objects in egocentric activities. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3281–3288, June 2011. 4321, 4322

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 580–587, 2014. 4327

[10] A. Joulin, F. Bach, and J. Ponce. Discriminative clustering for image co-segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1943–1950, 2010. 4322, 4325

[11] M. Kassner, W. Patera, and A. Bulling. Pupil: An open source platform for pervasive eye tracking and mobile gaze-based interaction. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, pages 1151–1160, 2014. 4324

[12] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1346 – 1353, 2012. 4321

[13] Y. Li, Z. Ye, and J. M. Rehg. Delving into egocentric actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 287 – 295, 2015. 4322

[14] Y. Lin, K. Abdelfatah, Y. Zhou, X. Fan, H. Yu, H. Qian, and S. Wang. Co-interest person detection from multiple wearable camera videos. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 4426 – 4434, 2015. 4322, 4327

[15] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2714–2721, 2013. 4321

[16] A. Mishra, Y. Aloimonos, and C. L. Fah. Active segmentation with fixation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 468–475. IEEE, 2009. 4327

[17] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, pages 849–856, 2001. 4323

[18] H. S. Park, E. Jain, and Y. Sheikh. 3d social saliency from head-mounted cameras. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, pages 1–9, 2012. 4322, 4327

[19] H. S. Park, E. Jain, and Y. Sheikh. Predicting primary gaze behavior using social saliency fields. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3503 – 3510, 2013. 4322

[20] H. S. Park and J. Shi. Social saliency prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4777–4785, 2015. 4322

[21] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 143–156, 2010. 4323, 4324, 4328

[22] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2847–2854, 2012. 4321

[23] C. Rother, T. Minka, A. Blake, and V. Kolmogorov. Cosegmentation of image pairs by histogram matching - incorporating a global constraint into mrfs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 993–1000, 2006. 4322, 4325

[24] N. Shapovalova, M. Raptis, L. Sigal, and G. Mori. Action is in the eye of the beholder: Eye-gaze driven model for spatio-temporal action localization. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, pages 2409–2417, 2013. 4321, 4322

[25] K. Tang, A. Joulin, L.-J. Li, and L. Fei-Fei. Co-localization in real-world images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1464–1471, 2014. 4322, 4325

[26] T. J. J. Tang and W. H. Li. An assistive eyewear prototype that interactively converts 3d object locations into spatial audio. In *Proceedings of the ACM International Symposium on Wearable Computers (ISWC)*, pages 119–126, 2014. 4325

[27] R. Vertegaal. The gaze groupware system: mediating joint attention in multiparty communication and collaboration. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 294–301. ACM, 1999. 4321

[28] J. Xu, L. Mukherjee, Y. Li, J. Warner, J. M. Rehg, and V. Singh. Gaze-enabled egocentric video summarization via constrained submodular maximization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2235–2244, 2015. 4321, 4322

[29] K. Yun, Y. Peng, D. Samaras, G. J. Zelinsky, and T. L. Berg. Studying relationships between human gaze, description, and computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 739–746, 2013. 4321, 4322

[30] D. Zhang, O. Javed, and M. Shah. Video object cosegmentation by regulated maximum weight cliques. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 551–566, 2014. 4322, 4325