Master Thesis

Discovering Objects of Shared Attention via First-Person Sensing

(一人称視点映像を用いた共同注視物体の検出)

Hiroshi Kera

Advisor: Prof. Yoichi Sato

The University of Tokyo
Department of Information and Communication Engineering
Graduate School of Information Science and Technology

Advisor
Prof. Yoichi Sato

Contents

List of Figures

List of Tables

1. Introduction

1.1. Overview

Figure 1.1.: Examples of the first-person video (left) and that with points-of-gaze (right). (Left) the first-person video shows that the wearer is interested in the book on the desk. (Right) a point of gaze explicitly shows that the camera wearer is looking at the book. Without the point of gaze, one might take the hat in the middle of the view for the object of interested.

Shifts in attention are one of the primal 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 commonly viewed by multiple people. Such objects of shared attention[1] reflect what people attend to from moment to moment and can be used as a cue to understand group activities [FRR11, XML+15]. For instance, a set of objects of shared attention tells what event is going on (*e.g.*, a speaker, handout, projector screen, then it's a meeting or presentation). Shifts of objects of shared attention in cooperative work tell us how the work proceeds; that is, what is built by workers, how its appearance changes as the work goes on, and which tools are needed in each step. In the context of computer-supported cooperative work, the ability to extract objects of shared attention allows us to evaluate how systems mediate collaborative work of people [Ver99]. For instance, we can test

[1] There are several different definitions for shared attention [OBT06, Eme00]. Throughout this thesis, we define shared attention as people's attentions commonly focussing on something at the same moment (see Section 3.1).

a system whether or not it really helps workers to easily understand which objects they are going to manipulate. If the extracted objects of shared attention are the desired one, then it follows that the system is helpful.

In this thesis, we employ computer vision technologies to discover such objects of shared attention. In particular, we utilize wearable cameras and eye trackers mounted on the head of people during interactions. First-person points-of-view videos, or first-person videos, recorded by such cameras can clearly capture what people see (Figure 1.1(left)) and thus can be used for action recognition [FHR12, PR12] and activity summarization [APS+14, CSJ15, YGG12, LG13, XML+15]. 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 (Figure 1.1(right)). This enables localizing important objects spatially and temporally [FLR12, FRR11, SRSM13, XML+15, YPS+13].

Motivated by these advantages of wearable cameras and eye trackers, we propose a method to discover objects of shared attention using multiple wearable cameras and eye-trackers worn by each of interaction parties. With such first-person videos and points-of-gaze data recorded during interactions, our method discovers when a shared attention on objects occurs. The proposed algorithm proceeds as follows. Using .... across multiple videos. Some results from our experiments are illustrated in Figure 1.2. Using points-of-gaze data of each camera wearer, we segment first-person videos into subsequences (called *shots*) by detecting eye movements from one object to another. Then, we perform a commonality clustering to find shots that contain objects with similar appearances across multiple videos (highlighted frames in the figure).

1.2. Challenges and Contributions

Given points-of-gaze data, it is natural to extract visual features from the region around points of gaze to describe objects being viewed. We may then perform a commonality clustering on such feature vectors to discover objects of shared attention across videos. A fundamental problem that arises here is how to appropriately define a region in first-person videos, from which we extract features to describe objects being viewed. Although points of gaze tell us which point the wearer is looking at, they do not tell which part is the region of objects. The region of objects around points of gaze largely depends on object sizes and viewpoints. While some studies use regions of a specific fixed size around points of gaze [FRR11, LYR15, XML+15], 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 see 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

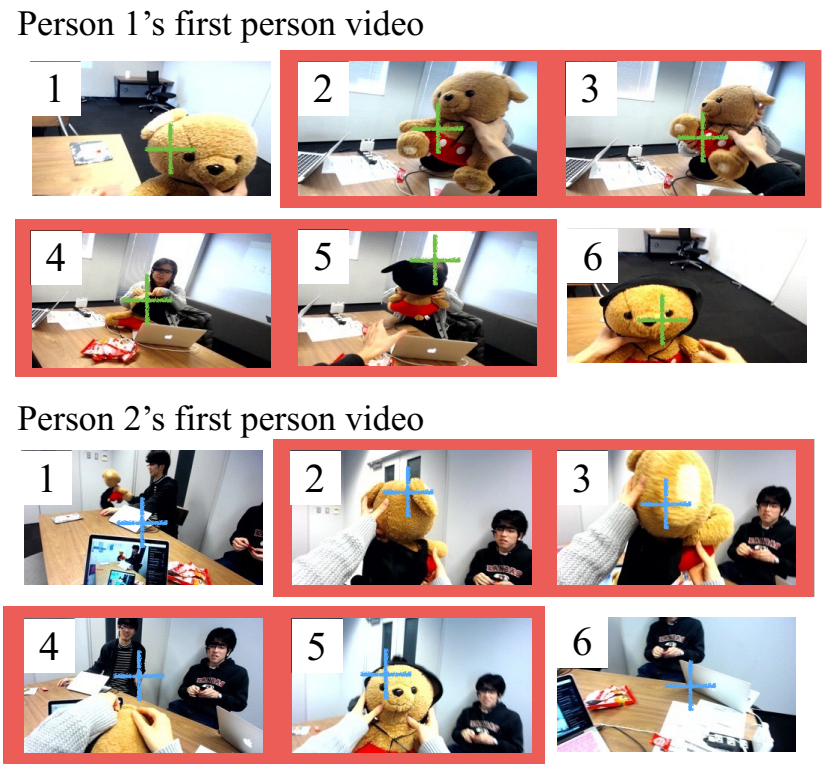

Figure 1.2.: An illustration of objects of shared attention discovery. Objects of shared attention are discovered in multiple first-person videos recorded during interactions (highlighted frames). Points of gaze of camera wearers are annotated by crosses.

only describe a limited part of objects or are affected by plenty of irrelevant background regions. In the former case, it is impossible to match objects across views when each of the observers is looking at the different parts of the same object; in the latter case, backgrounds have a large effect on the result of matching, and thus different backgrounds lead to false negative matching and similar backgrounds lead to false positive matching.

Using improper size of region for extracting object features also becomes problematic for temporal segmentation of videos. We particularly consider a straightforward approach to the segmentation by thresholding similarity of consecutive frames with their regions around points of gaze. Ideally, the similarity sharply drops when the gaze shifted from an object to another one, and thus the videos are segmented into shots each of which covers an attention to one and only one object. Note that it is not unusual that one shifts his/her attention among several important parts of a single object. For instance, when we are reading a paper, we might shift our gaze from a figure to the main text describing the figures. It is desirable that such shifts on a single object are covered by a single shot so that we can extract features of the object by taking into account whole parts of an object. However,

with improper size of region for extracting object features, it becomes difficult to achieve such segmentation reliably. Let us consider an example where a too small region is considered as a region of object for a large object. In that case, a small gaze shift from a part of the object to another part of it can lead to wrong segmentation of the video; that is, the video may be segmented into two shots: One shot before the gaze shift and the other shot after the gaze shift.

To address the aforementioned problems, we introduce a multiscale approach for object-feature extraction. 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 (Figure 3.1(upper row)). These visual features are further used to segment an input video into shots based on several different affinity criteria so that for each attention on objects there is at least one shot that properly covers the attention on a single object. This approach allows 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 is discovered for each scale via unsupervised commonality clustering (Figure 3.1(middle row)). Discovery results are finally integrated across scales to find various sizes of objects of shared attention reliably (Figure 3.1(bottom row)).

The main contributions of this thesis are summarized as follows:

- We introduce a new task of discovering objects of shared attention from first-person videos and points-of-gaze data. To the best of our knowledge, this is the first work that deals with multiple first-person videos recorded with points-of-gaze data. Objects of shared attention tend to reflect contexts of social interactions and thus discovering such objects provide cues that capture the semantics of first-person visions.
- We present a method to discover objects of shared attention using multiscale spatiotemporal tubes as object candidates. Our method addresses the main challenge that arises in the task of discovering objects of shared attention: object-size variability among objects and views.
- We collect a novel dataset containing multiple pairs of first-person videos and points of gaze data to validate the effectiveness of our approach. The dataset contains two- or three-person interactions and various kinds of interactions in several formations of people. To the best of our knowledge, there has been no other dataset that uses multiple points-of-gaze sources in first-person vision tasks.

1.3. Thesis Outlines

This thesis is organized as follows. In Chapter 2, we show an overview of the recent work on first person vision, *i.e.*, computer vision using first-person videos, including

those dealing with single first-person video, multiple first-person videos, and gaze information. We also introduce recent commonality discovery methods, which try to discover *commonalities* from multiple images or videos. In the subsequent sections, we show more details of several recent studies that are most relevant to our work. We then present our method in Chapter 3. In each section, three main building blocks of our approach are described step by step (Figure 3.1): generating multiscale temporal tubes, performing commonality clustering, and aggregating the results of different scales by voting. In Chapter 4, we evaluate our method and show its superiority over other baseline methods. Current limitations are also presented and possible solutions and other modifications are discussed. Finally, Chapter 5 summarizes this thesis. In Appendix, we provide the mathematical background of the commonality clustering, and also show whole graphical results that are omitted in the main part of the thesis.

2. Related Work

2.1. Overview

In this chapter, we review some prior work related to the task of discovering objects of shared 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.* [PJS12, PJS13, PS15] 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 have been used to recognize objects [YPS+13] and actions [FLR12, SRSM13] or to summarize videos by detecting important shots [XML+15]. To the best of our knowledge, our 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 [JBP10, RMBK06, ZJS14], co-localization [TJLFF14], and temporal commonality discovery [CZD12]. Perhaps the most relevant work presented is common-interest person detection from multiple first-person videos [LAZ+15]. 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.

In the following sections, we introduce four important studies that are closely relevant to our work. The first two studies [PS15, LAZ+15] addressed the problem of discovering a common interest in a group of people equipped with wearable cameras. Although neither of them utilizes gaze information, their goal is very similar to ours. The rest two studies [CZD12, TJLFF14] presented methods for commonality discovery; that is, methods to retrieve objects or actions that commonly appear

across images or videos. These two methods will be used as baselines in Chapter 4 to evaluate the effectiveness of our method.

2.2. Social Saliency Discovery

Park *et al.* [PS15] presented a method to predict social saliency, *i.e.*, the likelihood of joint attention[1]. They provided an example of an artificial agent that is trying to go through the crowd of people in a social scene. The agent is expected to plan its trajectory not only to avoid colliding with people but also avoid occluding sights of people. To this end, the agent must understand where is attracting the attention of the social group, *i.e.*, social saliency.

Given a social group and location of each member, they compute *social dipole moment*, which describes the direction of joint attention from the center of the mass of the social group. Social formulation feature is defined in order to describe the distribution of a social group. The authors trained a binary ensemble classifier from a collection of social formation features. With the classifier, a continuous social saliency map of the target scene is generated, which can be regarded as a probabilistic map of the likelihood of joint attention. The authors also presented a method to assort people into their social groups based on their geometric relationship. They first generate candidates of social groups based on the spatial distribution of social members. Then, they solved a minimization problem to select proper set of social groups. The minimization is designed so that the center of different social groups is not too close and also nobody belongs to no more than one group.

In the experiments, they evaluated their method with various social interaction scenes captured by first-person videos. They used the 3D reconstruction of first-person videos to measure joint attention, locations of associated members, and directions they are facing to over time. Their experiments demonstrated that their method is able to discover places in social scenes that attract attentions of people. However, since they do not use gaze information, their method is only able to offer *where* is attracting the attentions of people, but cannot offer *what* is attracting the attentions. In our daily life, it is not unusual that many objects are closely located. People's interest can be shifted from objects to objects with subtle head pose change. For instance, in the Figure 1.1(right), the camera wearer is not looking at the hat in the center of his view, but at the book. This information cannot be obtained without points of gaze. In this way, it is difficult to tell which object is focused on by people without points-of-gaze and just by knowing the social saliency. Furthermore, their work requires 3D models of the social scenes, which are not alway available. In contrast to their work, this thesis presents a method to discover objects of shared attention, where points-of-gaze data illuminate which part in first-person vision the wearers are attended to over time.

[1] Joint attention in their work is similar notion to shared attention in this thesis. See Section 3.1.

2.3. Co-interest Person Detection

Lin *et al.* [LAZ+15] proposed a new problem of discovering co-interest person (CIP) from multiple first-person videos. They defined a co-interest person as that who attracts the attentions of other people. CIP usually plays a central role in the ongoing event of interest, and thus it provides useful information to deal with multiple first-person videos: discovering a person with abnormal behavior for surveillance, detecting a kid with strange behavior for an early finding of development issues, and summarizing discussion for efficient information management and retrieval.

They considered it difficult to identify CIPs with appearance-based matching, so they used motion patterns of people. The main challenge of discovering CIP is that such motion patterns appear in a different way across views. Furthermore, camera motions and person motions are mixed up in first-person videos. In their method, they defined an energy function, and reduce the problem to an energy minimization problem of the Conditional Random Field. The energy function is designed so that the relative position and size of CIP should be consistent, and motion patterns of CIP are correlated across views. Since horizontal flows appear in the inverted directions from the opposite view, opposite horizontal motion directions are merged (*e.g.*, North-West and North-East directions are regarded as the same direction).

In their experiments, each of subjects wore a wearable camera, and perform some actions as a CIP in turn. To demonstrate the effectiveness of motion patterns, all the subjects were in similar clothes. Although their method performs well in discovering CIP, it cannot be extended to general objects in a naive way. Their method generates CIP candidates by using a human detector, but it is much more difficult to construct a general object detector that copes with cluttered scenes including occlusion and harsh lighting conditions. In addition, jointly attended objects are not necessarily moving, while the authors exploit motion patterns to discover a CIP. In our case, we try to discover general objects including objects in static or moving and persons or non-persons. We deal with such general objects by utilizing points of gaze to localize them. Since points of gaze do not provide the whole region of objects, we use multiscale approach and find regions that tightly cover the objects.

2.4. Temporal Commonality Discovery

Chu *et al.* [CZD12] introduced a new problem of discovering the temporal commonality between a pair of videos. For instance, there should be scenes of kisses in romance movies, and their method discovers such scenes as a temporal commonality across videos. More formally, their method, named TCD, tries to find pairs $(b_i, e_i), (i \in \{1, 2\})$ of the start point and end point of the most similar subsequence pairs between two videos.

The main challenge is its computational complexity. Testing all possible combinations requires a huge computational cost. The authors proposed the branch and

bound algorithm to efficiently find the optimal solution. The authors first defined a rectangle set $\mathcal{R} = [B_1, E_1, B_2, E_2]$, where B_i is the range of the start point to be considered in the i-th video, and E_i is the counterpart at the end point. They then start from a rectangle set that includes all possible rectangles. In the iterative fashion, a rectangle set is split into two new rectangle sets and evaluated the range of similarity they can take. Rectangle sets with higher similarity are evaluated in preference to others. The search terminates when there is only one rectangle in the rectangle set to be considered.

TCD suffers from the scalability problem. When it deals with three or more streams, its computational cost sharply increases. More precisely, when it deals with N videos where each of video has $n_i, (i = 1, 2, \ldots, N)$ frames, then its computational complexity in the worst case is $O\left(\prod_{i=1}^{N} n_i^2\right)$ (though their branch-and-bound-based approach is rather efficient in practice). That does not suit for our task where commonality in a group of people are dealt with. On the other hand, our approach is more efficient than their one. Instead of considering all videos at a time, we first consider pairwise commonality between each of video pairs and integrate them later. Furthermore, we segment videos into shots based on the attention on objects, which greatly reduces the number of feature vectors to be considered compared to directly considering whole frames.

2.5. Co-localization

Tang *et al.* [TJLFF14] proposed a method for the co-localization problem, the goal of which is simultaneously localizing an object common across multiple images. The distinct point of their method is that it allows input images to contain *noisy* images, *i.e.*, images that do not contain the same object with (most of) the others.

In their method, they first generate bounding boxes in each input image using an off-the-shelf object proposal method. Each bounding box is a candidate that potentially contains a common object across images. The authors then introduced various objective functions based on saliency-based priors, similarity across images and boxes, and discriminability within each image and box. Finally, they solved a minimization problem for the co-localization that is formulated by aggregating all the objective functions above.

Note that their method for co-localization is not tailored for videos but for images. Although a video can be seen as a sequence of images (or *frames*), their method cannot be simply applied to our task of discovering objects of shared attention from multiple videos. The first reason is that their method will not keep temporal consistency; that is, common objects across frames should appear for sensible length of time interval across multiple videos. Furthermore, without points-of-gaze, there can be common objects across videos but not interested by people. Another problem is that their co-localization method assumes that the backgrounds of objects are

different, but this is not the case of our problem. It is highly possible that people looking at the same or different objects in similar backgrounds. In Chapter 4, we will apply this co-localization method for our task and demonstrate that it only shows a limited performance even when points-of-gaze data are used.

3. Proposed Method

3.1. Definition of Shared Attention

Throughout the thesis, we define shared attention as events that multiple people are looking at the identical object within a certain time interval. Here *objects* include boxes, tables, walls, persons, projected screens, and so forth. Interactions among the people are not required (while these are expected to exist). According to [OBT06, Eme00], shared attention requires that people attending to objects are mutually aware of the attentions of other people to the objects. However, we here do not force our definition of shared attention to satisfy this requirement, because it is difficult to know whether people are actually aware of attentions of others. In our dataset introduced in Section 4.1, we collected shared attentions during interactions among people, where the mutual attention requirement is hopefully satisfied, but the existence of interactions was not taken into account when the ground truth labels were manually annotated.

Shared attention is another term that is closely related to shared attention. While Emery [Eme00] distinguished join attention from shared attention as a special case of shared attention, the author also mentioned that these two terms had been used interchangeably in the literature. In the field of first-person vision, Arev *et al.* [APS+14] and Park *et al.* [PS15] used shared attention instead of shared attention. Similar to our definition, there used shared attention to refer to people's simultaneous attention during interactions, but they did not provide any specific definition.

3.2. Problem Setting

Suppose that N persons are involved in interactions. During the interactions, they commonly attend to objects such as a book at passing one from another, projected screen on the wall at a meeting, and snacks on a table at break time. If the time interval where people are looking at such objects is given, we are able to identify objects of shared attention by referring to points of gaze during the interval. Therefore, our goal is now to discover time intervals where objects are commonly attended by people, from N pairs of first-person videos and points-of-gaze data recorded by head-mounted cameras and eye-tracker worn by each of the people. In other words,

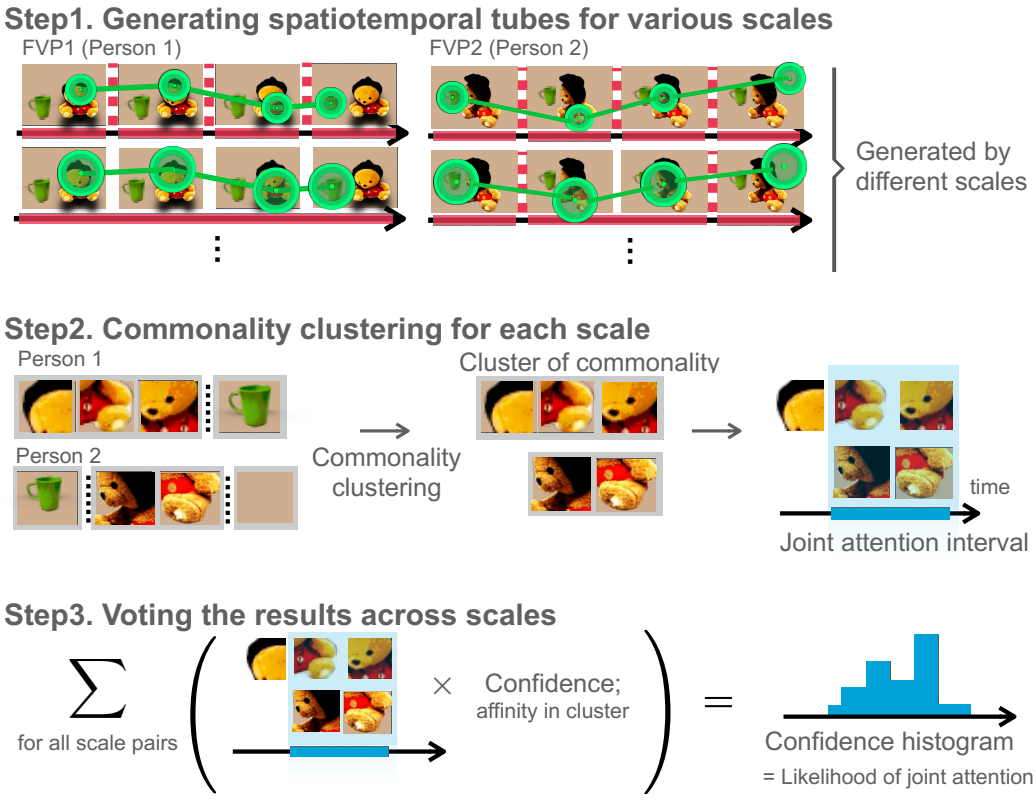

Figure 3.1.: Pipeline of the proposed method

our method accepts as input such N pairs of first-person videos and points-of-gaze data and outputs time intervals where the same object is viewed in all of the N videos (*i.e.*, an object of shared attention). More formally, given N pairs of first-person videos and points-of-gaze data $\{(V_k, \boldsymbol{g}_k)\}_{k=1,\ldots N}$, where V_k and $\boldsymbol{g}_k$ are lists of frames and two-dimensional points at each time $t \in \mathcal{T} = [1, 2, \ldots, T]$, our goal 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 the subsequent sections, we describe each of the key steps of our method as illustrated in Figure 3.1. In Section 3.3, we first explain generating multiscale spatiotemporal tubes from videos to describe objects being viewed. Then, in Section 3.4, we describe how to perform unsupervised commonality clustering on the tubes to discover time intervals where shared attention is likely to occur for each scale. Finally, we introduce a voting scheme to integrate the discovery results across scales in Section 3.5.

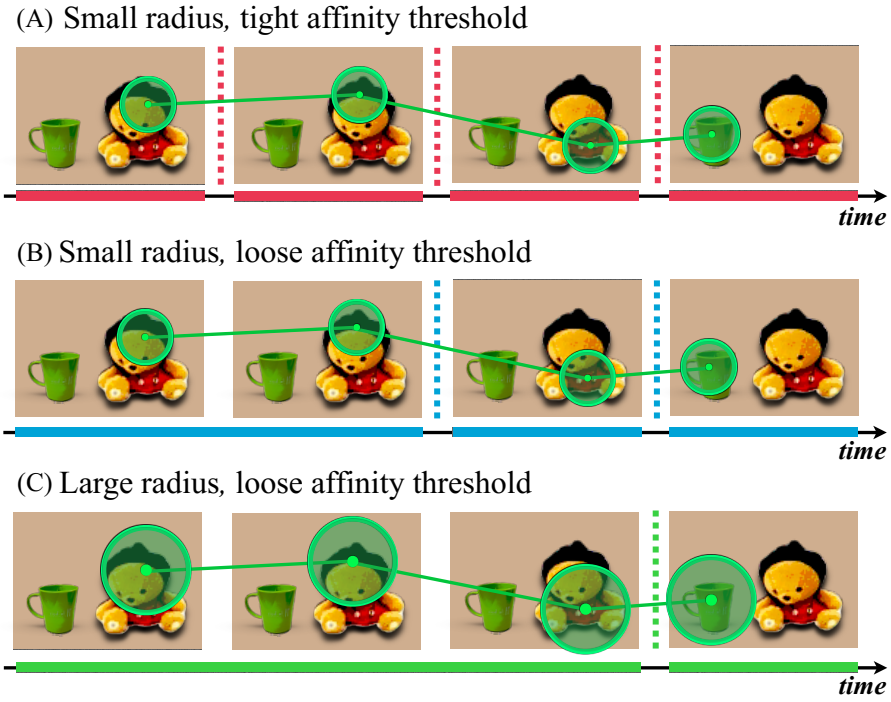

Figure 3.2.: Concept figure of multiscale spatiotemporal tubes. Colored time axes represent time intervals split with several pairs of a radius and an affinity threshold. (A, B) Smaller radius of tubes is more appropriate to extract features from the object on the left side; (C) Larger radius and longer length are needed to cover the object on the right.

3.3. Generating multiscale Spatiotemporal Tubes

When we see objects, points of gaze are often distributed over important parts of the objects. If we properly segment 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 3.2, 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 features extracted from the region around a point of gaze $\boldsymbol{g}_{n,t}$ in the 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 features around points of gaze, such as a naive histogram and deep features of deep neural networks [LSD15, SZ14, SLD16].

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 the 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:

$$\boldsymbol{s}_{n,k}^{(p_n)} = \mathcal{H}(\{\boldsymbol{f} \in \mathcal{F}(V_{n,t}) \mid \|t \in j_{n,k}^{(p_n)}, \boldsymbol{l}(\boldsymbol{f}) - \boldsymbol{g}_{n,t}\| < r_n\}).$$

3.4. Commonality Clustering on Tubes

To discover objects of shared 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 3.5 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.

$$A = \begin{pmatrix} O & C \\ C^\top & O \end{pmatrix}, \tag{3.1}$$

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 3.6. Similar to normalized spectral clustering [NJW01], 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 [CSJ15], co-clusters can be obtained via spectral clustering with the Laplacian matrix $L = D - A$. Refer to Appendix A for the details. 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 shared attention are observed sparsely during interactions, the maximal-biclique-based approach proposed in [CSJ15] 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 shared 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:

$$\mathcal{J}^{(p_1,p_2)} = (\cup_{k \in K_1} j_{1,k}^{(p_1)}) \cap (\cup_{k \in K_2} j_{2,k}^{(p_2)}). \tag{3.2}$$

Note that co-clusters discovered using the affinity A in Eq. (3.1) always contain tubes from both of the two videos. If no intersections are found in Eq. (3.2) at a certain combination of scales (p_1, p_2), the result from that scale setting is just ignored in the subsequent voting scheme.

3.5. 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 shared 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 shared 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:

$$\begin{aligned} c_t &= \sum_{p_1,p_2 \in \mathcal{R} \times \Theta} c^{(p_1,p_2)} \delta(t, \mathcal{J}^{(p_1,p_2)}), \\ \delta(t, \mathcal{J}^{(p_1,p_2)}) &= \begin{cases} 1 & t \in \mathcal{J}^{(p_1,p_2)}, \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{3.3}$$

The time interval including objects of shared 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 3.4 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.

3.6. Implementations

In this section, we describe details of the implementation of our method. We first briefly describe important implementations in Sections 3.6.1 and 3.6.2, and then provide other minor implementations in Section 3.6.3.

3.6.1. Features

To describe appearances of objects, we used HSV color histogram, a deep feature of Fully Convolutional Network (FCN; [LSD15, SLD16]), and time interval.

HSV color histogram We used HSV color histograms as a low-level appearance-based feature robust for spatial fluctuations and rotations. In our task, objects can be looked at in different views, and thus spatial information is not always helpful. For instance, when an object is passed from a person to another person, the object will be viewed upside-down from the receiver. 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 of a frame that is the nearest to the mean of the frames in each shot.

FCN deep feature To introduce high-level information, we used deep features extracted a layer of FCN. FCN used here was trained for object classification, and thus it extracts abstract information that takes into account spatial details and ignores unnecessary spatial noise for classification. We expect HSV color histograms and FCN deep feature to complement each other. We utilized the output of the pool4 layer of the FCN as an FCN deep feature. A 512-dimensional vector was constructed by taking the spatial average of the 3D tensor of the output of pool4 layer, and then flattening. We used an FCN model with three-stream and eight-pixel prediction stride net, which were pre-trained on PASCAL VOC dataset [EGW$^+$] and distributed by the authors [LSD15, SLD16]. We adopted FCN rather than widely used VGG model [SZ14] for its two advantages: (1) it takes into account spatial detail of input images and (2) it fully consists of convolutional layers so that any size of images can be input.

Time interval We took 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 $\frac{1}{\sqrt{|j_{n,k}^{(p_n)}|}}$ 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 $\boldsymbol{s}_{n,k}^{(p_n)}$.

3.6.2. Affinity matrix

Suppose the i-th frame $\boldsymbol{s}_i$ (or shot) are described by k features $\left\{\boldsymbol{s}_i^{(1)}, \boldsymbol{s}_i^{(2)}, \ldots, \boldsymbol{s}_i^{(k)}\right\}$. We defined the distance $\tilde{w}_{ij}$ between $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ as follows:

$$\tilde{w}_{ij} = \sum_{l=1}^{k} a_l \frac{\|\boldsymbol{s}_i^{(l)} - \boldsymbol{s}_j^{(l)}\|}{\eta_l},$$

where η_l is the mean of all distance values with respect to l-th feature, and a_i is the tunable hyper-parameter for providing different weights for each feature . The η_l was introduce to rescale distance values so that all distance values vary similar range in each feature. We then defined the affinity between $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ by $\exp\left(-\rho\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|\right)$, where $\|\cdot\|$ is the Euclidean distance and ρ is set to the median of all distance values.

3.6.3. Other details

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. All visual features are extracted from rectangular regions instead of circular region. This is because FCN is trained with rectangular images and circular boundaries are not assumed. To maintain the input image size in moderate range, we generated an empty image in 300x300 resolution, and bind to it a rectangular region around points of gaze. We avoided enlarging the rectangular region to keep the object size in realistic size. Affinity thresholds were obtained by computing 10th, 30th, and 50th percentiles of all the affinities for each video. Output confidence histograms are scaled so that each of them ranges from zero to one.

4. Experiments

To evaluate the effectiveness of our approach, we built a novel dataset containing multiple pairs of first-person videos and points-of-gaze data. To the best of our knowledge, this dataset is the first to use multiple points-of-gaze sources in first-person vision tasks. The experiments demonstrate that our approach can outperform several state-of-the-art commonality clustering methods on the task of discovering objects of shared 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. Refer to Figures 4.2, 4.3, 4.4, and those in Appendix B for the overview for the interactions in dataset.

During each recording, subjects were asked to establish shared 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 commonly 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 shared 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 commonly. 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 [KPB14] 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 3.6 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 shared 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. (3.3) 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 multiscale 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.* [CZD12] 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 FCN deep features around points of gaze as well as a time interval feature vector for each frame. As for the FCN feature, we used the output of pool4 layer. 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.* [TJLFF14] 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

Figures 4.2 and 4.3 show some of the results of our approach in the SbS sequences and FtF sequences, respectively. Refer to Appendix **??** for the rest of the results. In

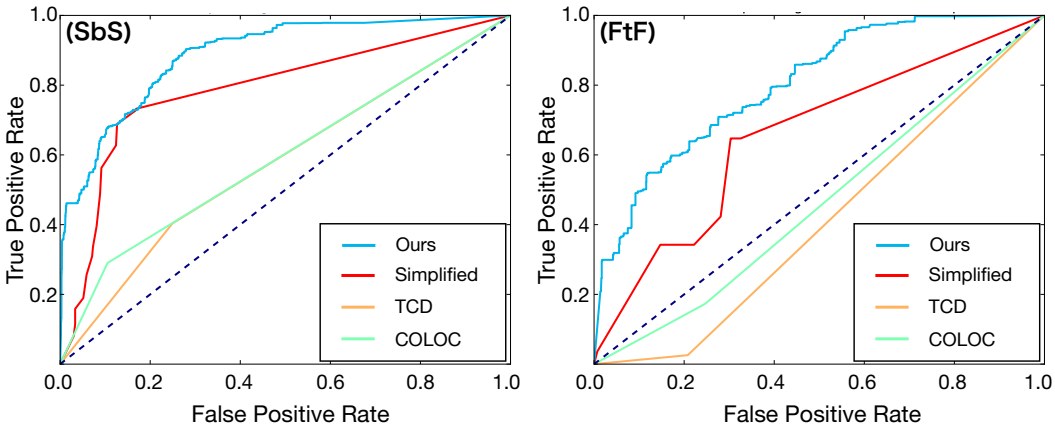

Figure 4.1.: ROC curves of the proposed and baseline methods

each example, subjects were involved in the following interaction: Figure 4.2(A) a subject pointed to a camera on a table and get the other subject sitting next to him to look at it. Figure 4.2(B) a subject looked at a projector screen and spoke to the other subject to see it; Figure 4.2(C) two subjects saw a teddy bear from different points of view; Figure 4.3(A) two subjects saw a bottle from different points of view; Figure 4.3(B) two subjects sitting face to face exchanged a book; and Figure 4.3(C) a subject asked the other subject in front to put a block into a cylindrical box.

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 bottle in Figure 4.3(A) to a large projector screen in Figure 4.2(B). We were also able to deal with cases when the size of object instances was drastically different, as shown in Figures 4.2(A)(C) and Figure 4.3(C). By using points of gaze to limit the location of features to be extracted and compared, we can discover objects of shared attention even when background scenes are greatly similar across videos, such as in example Figures 4.2(A)(C). This unique property of our approach is unlike many standard object co-localization and co-segmentation methods [JBP10, RMBK06, TJLFF14, ZJS14] that assume background scenes are different across images.

We also present quantitative evaluations based on ROC curves and AUC scores in Figure 4.1 and Table 4.1. On average, our method using multiscale 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.

Method	SbS	FtF	Avg.
(1) COLOC ($r = 50, \theta = 10$) [TL14]	0.592	0.463	0.528
(2) COLOC ($r = 25, \theta = 10$) [TL14]	0.574	0.622	0.598
(3) TCD ($r = 50$) [CZD12]	0.577	0.409	0.493
(4) TCD ($r = 25$) [CZD12]	0.480	0.441	0.461
(5) Simplified ($r = 15, \theta = 10$)	0.789	0.657	0.723
(6) Simplified ($r = 25, \theta = 10$)	0.779	0.701	0.740
Ours	**0.889**	**0.803**	**0.846**

Table 4.1.: 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.

4.4. More than Two-Person Cases

Figure 4.4 shows how our method can work on cases where three subjects are present in a scene. In example (A), one subject manipulated a box and asked the other subjects to look at the box. In (B), a teddy bear was passed from one subject to another followed by a third subject paying attention to the interaction. For both cases, our method successfully discovered the objects of shared attention, while the size 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 score on the three-person sequences was 0.89.

4.5. Failure Cases and Possible Extensions

Figures 4.2 and 4.3 include some failure cases. Discovering objects that were barely observed in first-person videos was difficult (*e.g.*, the book in hands in example Figure 4.3(E)). Moreover, false-positive responses were observed when subjects kept looking at textureless regions like in Figures 4.2(C). Some other failure cases were present in Figure 4.5. In example (A), our method failed to detect a shared attention on a bag that appeared differently across videos due to harsh lighting conditions. Note that since we included FCN deep feature to describe objects, this false negative is rather eased. In Section 4.6, we will show that without FCN feature, our method gives much fewer votes for this time interval.

In example (B), there are an amount of false positive votes for the time interval where two persons are looking at each other during chatting. They look similar to each other and thus our method could not distinguish between them.

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 [PJS12], 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 [LAZ+15].

Another interesting extension is to use segmentation around fixation points [MAF09] or object proposal [CZLT14] 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. Although our method also not so efficient with multiple videos,

4.6. Feature Comparison

Features	SbS	FtF	Tri	Avg.
hsv	**0.910**	0.803	0.832	0.848
fc7	0.713	0.604	0.729	0.682
pool5	0.642	0.597	0.732	0.657
pool4	0.783	0.595	0.771	0.717
pool3	0.783	0.595	0.769	0.716
time	0.823	0.776	0.696	0.765
hsv+time	0.900	**0.814**	0.846	0.853
fc7+time	0.816	0.724	0.729	0.756
pool5+time	0.780	0.722	0.718	0.740
pool4+time	0.889	0.751	0.764	0.801
pool3+time	0.829	0.680	0.782	0.774
hsv+fc7	0.903	0.799	0.851	0.851
hsv+pool5	0.871	0.810	0.847	0.842
hsv+pool4	0.899	0.771	0.818	0.829
hsv+pool3	0.829	0.762	0.843	0.811
hsv+fc7+time	0.909	0.800	0.853	0.854
hsv+pool5+time	0.899	0.805	0.873	0.859
hsv+pool4+time	0.889	0.803	**0.890**	**0.861**
hsv+pool3+time	0.899	0.796	0.883	0.859

Table 4.2.: Feature comparison among HSV color histograms, time feature, deep features of FCN extracted from different layers, and their combinations.

In this section, we compare several features and their combinations to describe the appearance of objects. Table 4.2 shows the AUC scores of the proposed approach

with different feature sets. We tested all possible combinations among HSV color histograms, time feature, and the FCN deep features extracted from fc7, pool5, pool4, or pool3 layers. Refer to Section 3.6.1 for the details of the features.

The following tendencies are observed:

- HSV color histograms perform pretty nice even without combined with other features.
- FCN deep features do not work well by themselves, but they improve the AUC scores when they are combined with HSV and time features. Especially, the AUC scores in the three-person sequences (Tri) largely increased.
- Among FCN layers, features extracted from pool4 works the best in most cases.

It is surprising that we can achieve high AUC scores with simple HSV color histograms, even though it has only 48-dimensions. A possible reason for this is that in our problem settings, the appearances of objects are largely different across views. In addition, the points of gaze can fluctuate with head motion or measurement error. HSV color histograms are robust for these appearance variability and gaze fluctuations.

The second observation implies that FCN deep features can not provide discriminative features for objects that are not included in a training dataset. However, the increase of AUC score when they are combined with others suggests that FCN features contain complementary information from HSV color histograms.

As for the third observation, we consider that middle layer (pool4) of FCN performed better because it balances spacial details and semantics. The lower layer (pool3) does not extract abstract information and not reflect semantics, while the higher layer (fc7) loses spatial details that are not useful for classifying objects into object classes in the training dataset.

It is obvious that HSV feature cannot distinguish two different objects that share similar color distributions. Figure 4.6(A) provides an example for this problem. In the end of the video (highlighted in blue), a subject shifted his attention from a checker board, cylindrical box, and cardboard box, while the other subjects were looking at a block in his hand. In this case, HSV+time voted much more than HSV+time+FCN did. The reason is that the objects viewed by the first subject, especially the cylindrical box and cardboard box, look similar in color to the block in the second subject's hand. In our dataset, there are not so many cases that subjects pay attention at the same time to different objects that are similar in color. However, it is not unusual that such situations occur in our everyday life. Testing in such cases is left to the future work to improve our method.

The histogram in the upper row of Figure 4.6(B) is the same one as that in Figure 4.5(A). As mentioned in Section 4.5, a bag looks in different color across views due to the lighting condition, and thus there are false negative votes for the time interval of the first highlighted column. As can be seen from Figure 4.6(B), without

FCN feature there are much fewer votes for the time interval. This result suggests that FCN feature can provide information complementary to that of HSV color histograms.

In summary, simple HSV color histograms work well in the current dataset, while it is necessary to be tested other cases where people simultaneously pay attentions to different objects similar in color. FCN deep features contain information complementary to HSV color histograms, and those extracted from middle layer are useful since they balance the spatial details and abstract semantics.

4.7. Limitations

In this section, we discuss several limitations of our method yet to be solved. Currently, our method can discover objects of shared attention under moderate conditions. However, there are some cases where interactions are recorded in rather harsh but not unusual conditions as follows:

- An object has largely different appearances across views due to lighting conditions or its design.
- There are many objects that have the same (or similar) appearances.

In the former case, it is difficult to match objects across views by their appearances. For example, a white bag being viewed by a person may appear much darker when being viewed against the light by another person on the opposite side. Another example is that a book read by a person may appear in completely different way from another person on the opposite side because of the white pages and its colorful covers. As for the second limitations, our method will discover false positive objects of shared attention when people pay their attention at once to objects that share the same appearance. These two limitations are both resulted from that objects are described only by their visual features. To remedy these problems, an interesting extension is to incorporating other types of features that do not rely on object appearances such as geometric relationship among people and motion patterns of objects. When a geometric relationship between head-mounted cameras is available by preliminarily scanning a scene like [PJS12], 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 [LAZ+15].

Another limitation is the scalability of our method. Currently, our approach tries all combinations of the scales and persons. With N_p persons, N_r radii, and N_θ affinity thresholds, our method find a cluster of commonality for ${}_{N_p}C_2 N_r N_\theta$ pairs. In practical, we only need to try the small number of scales, so the main problem is the size of group of people. While our experiments demonstrated that our multiscale approach is effective for the object-size variability problem, it only works with the moderate size of a social group. The number of pairs to consider increases as the size

of group grows. If it is necessary for some applications to discover objects that are being looked at by many people, then our method still has a room for improvement.

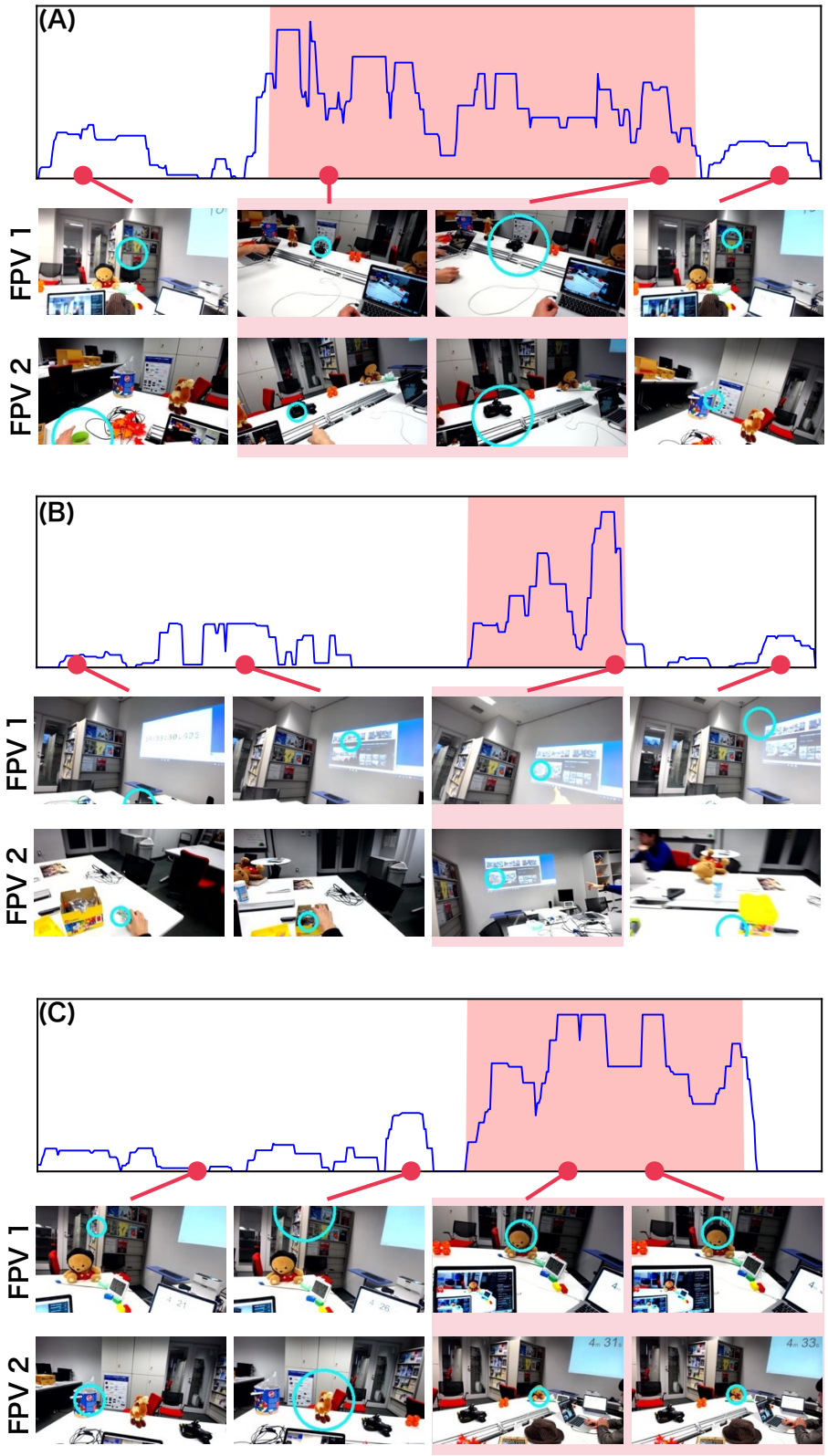

Figure 4.2.: Confidence histograms and image frames in the SbS sequences. Time intervals and image frames where objects of shared attention were observed are highlighted in pink. Blue circles denote regions attended by subjects. We selected the radius from the scale pair that gives the highest confidence score at each time point.

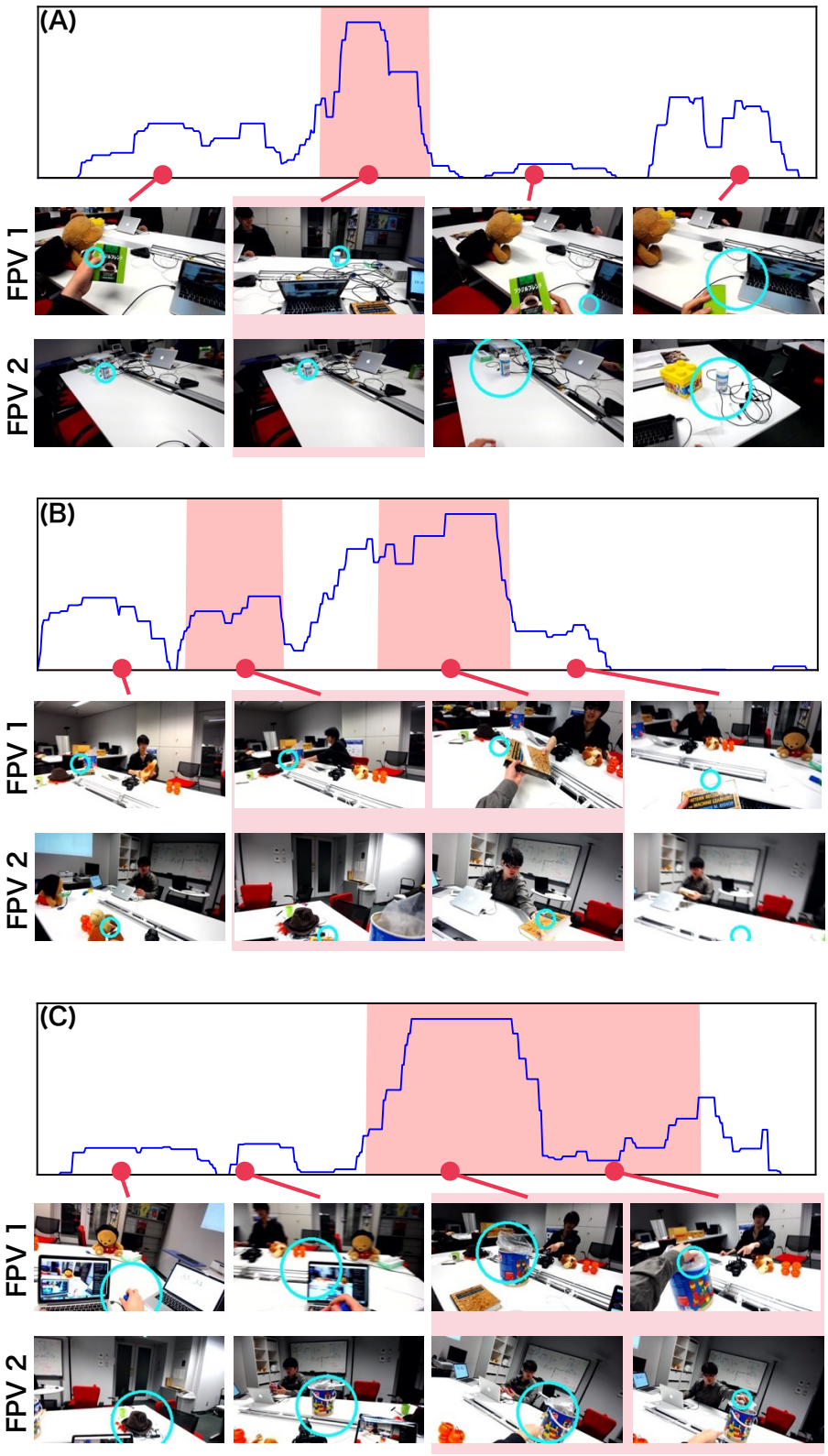

Figure 4.3.: Confidence histograms and image frames in the FtF sequences. Time intervals and image frames where objects of shared attention were observed are highlighted in pink. Blue circles denote regions attended by subjects. We selected the radius from the scale pair that gives the highest confidence score at each time 34point.

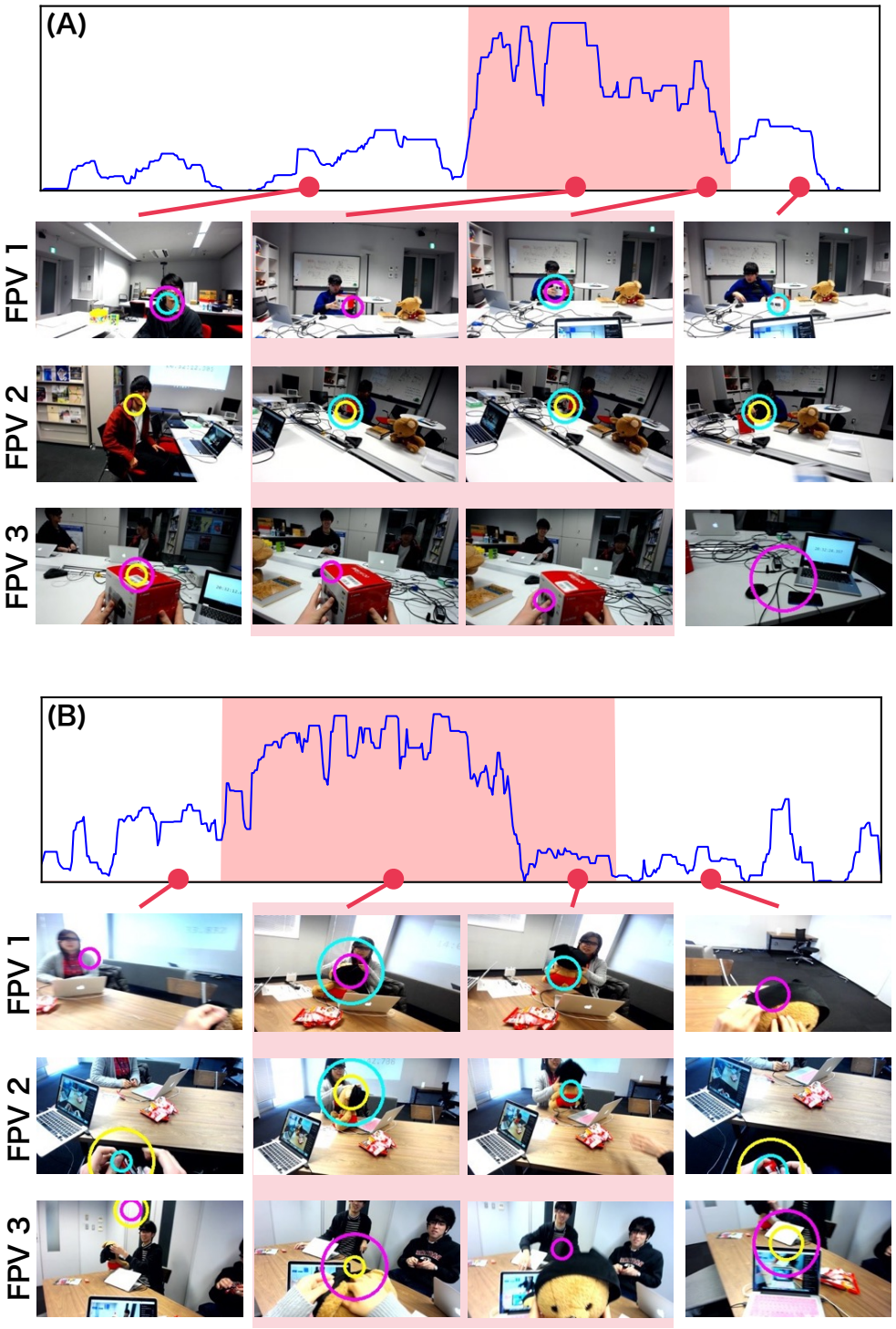

Figure 4.4.: Confidence histograms and image frames. Time intervals and image frames where objects of shared attention were observed are highlighted in pink. Circles denote regions attended by subjects. We selected the radius from the scale pair that gives the highest confidence score at the time point. Blue, purple, and yellow circles correspond to video pairs of video 1 and 2, video 1 and 3, and video 2 and 3, respectively.

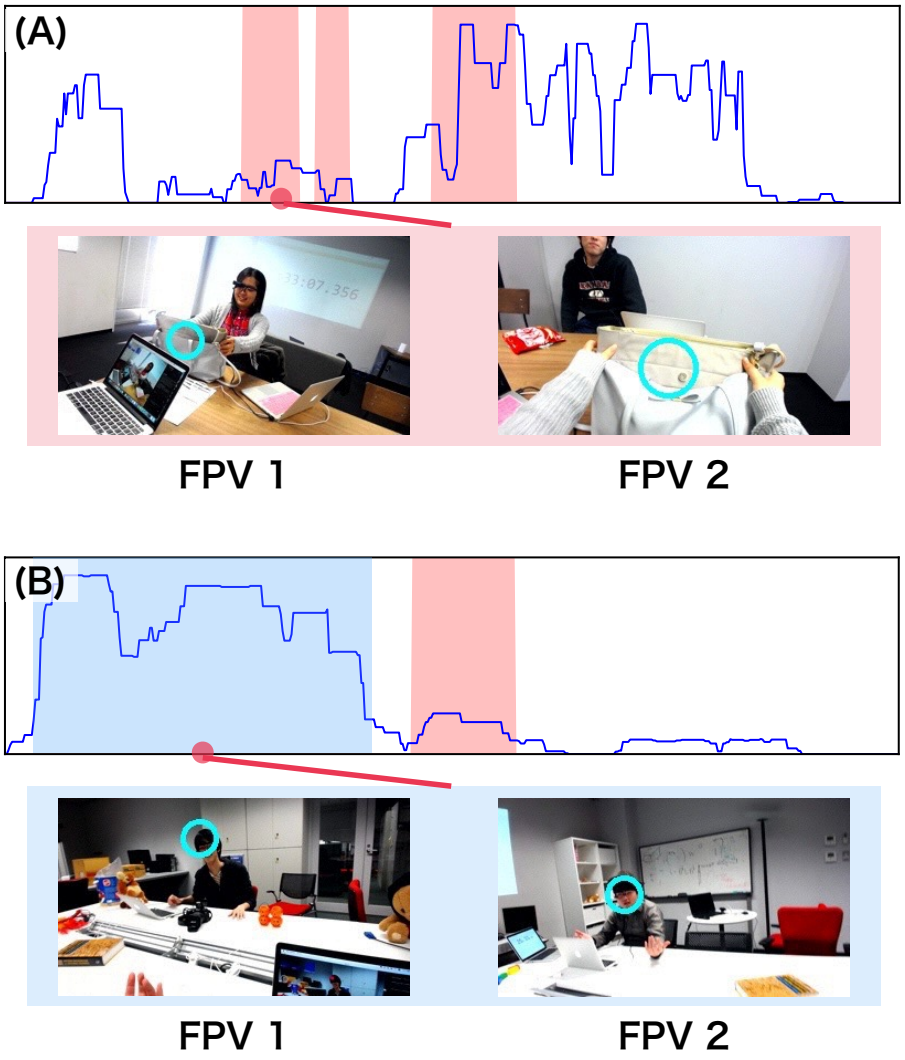

Figure 4.5.: Confidence histograms and image frames for failure cases. (A) False-negative detection and (B) False-positive detection highlighted in blue.

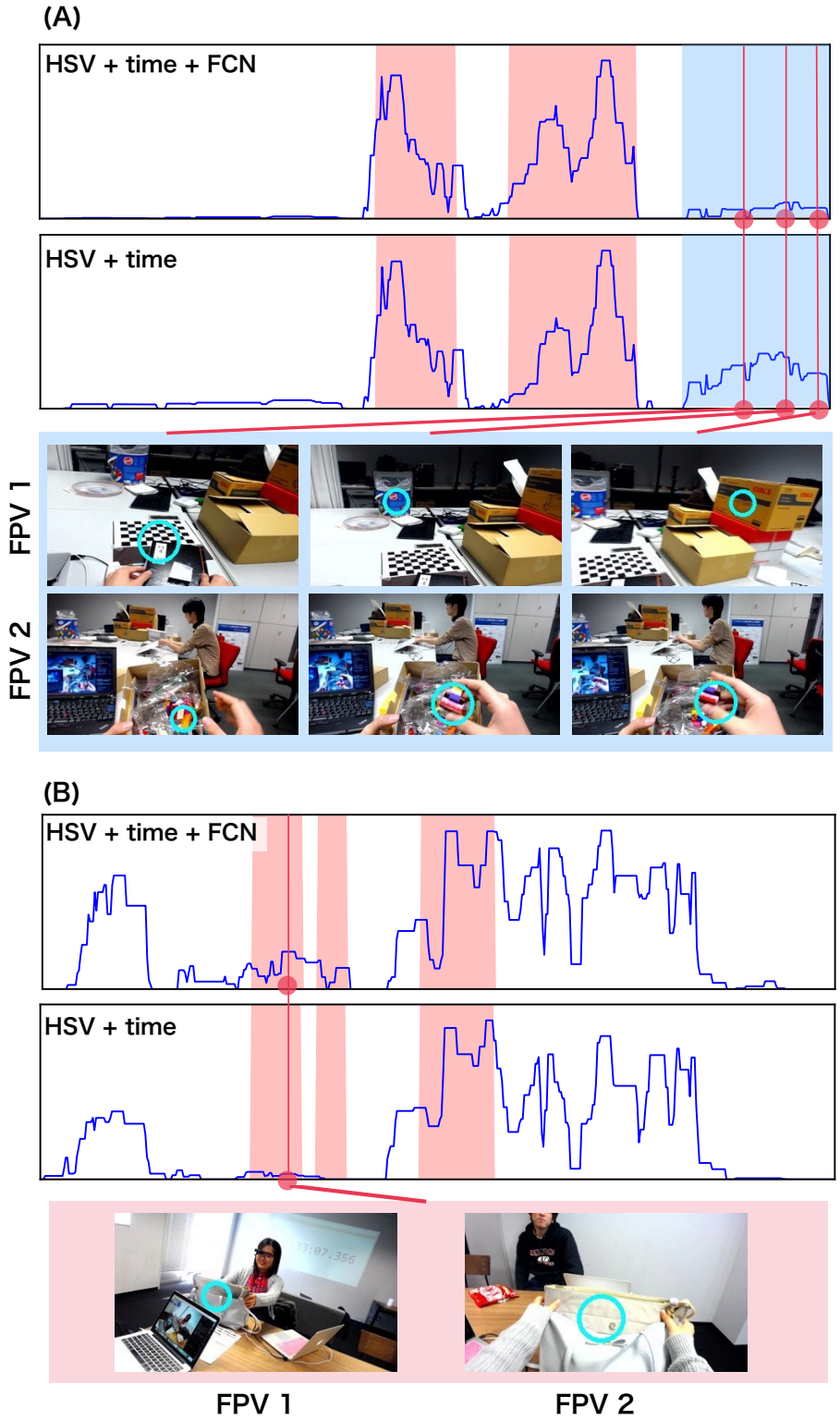

Figure 4.6.: Vote comparison between feature set with or without FCN feature. (A) In the end of the sequence highlighted in blue, two subjects are looking at different objects that share similar color distributions. (B) FCN deep feature eases the false negative votes that are caused by color variability in appearance due to the lighting condition.

5. Conclusion and Future work

In this thesis, we introduced a novel task of discovering objects of shared attention in multiple first-person videos. Since objects of shared attention reflect the contexts of the social interactions in our daily life, discovering such objects should be helpful for the further understanding of first-person visions. The main challenge to be solved for this task is how to deal with the object-size variability across objects and views. The key idea of our approach presented in this thesis is to segments videos into multiscale spatiotemporal tubes. Our experimental results demonstrated the effectiveness of our multiscale approach over several state-of-the-art commonality discovery methods. For this evaluation, we collected a novel dataset that provides multiple first-person videos and points-of-gazed data, which record various interactions by two or three persons in several formations. We also obtained some insights for which features to use: simple HSV color histograms work well; while FCN deep features did not work well by themselves, it provides some information complementary to HSV color histograms, resulting in the increase in AUC scores when combined with HSV color histograms; deep features extracted from the middle layer of FCN works the best compared to those extracted from other layers.

As discussed in Section 4.7, our method has several limitations yet to be solved. Some limitations are caused by that our method only uses the appearance-based feature to describe objects. With such features, it is difficult to match objects which largely differ in their appearances across views due to lighting conditions or object designs. Another limitation is the scalability of our method. Currently, our method works with interactions in a moderate size of group. In this thesis, we demonstrated that multiscale approach is the key for this task. How to prune unnecessary person pairs and scale pairs before conducting costly computation is one of the next key steps for efficient discovery.

Based on the insights we obtained from the results of the experiments, we list up several future directions of this work.

Incorporating non-appearance-based features As already suggested in Section 4.7, incorporating non-appearance-based features will be helpful for discovering objects of shared attention in more difficult cases. With geometric relationships (*e.g.*, where he/she is, which direction he/she is facing to) among people in a group, we can find objects of shared attention in a more accurate way. We can avoid matching different objects that share similar appearance on the ground that two persons are facing to different places. We can also avoid wasteful computation with such information

when people are obviously looking at the different places. Motions patterns also are helpful in some cases. The object size can drastically change when objects move at approaching, receding or being exchanged. Gaze may not be exactly on the objects when they move fast. In such cases, correlation of the motion patterns across views provides important cues for matching the objects.

Generating candidates of objects of shared attention via object proposals As some results of our experiments show, there are false positive matching between untextured parts of objects (*e.g.*, a surface of a table and that of a wall). To avoid such matching, it is promising to use object proposal methods [CZLT14] for generating candidates of objects to be matched across views. We can weigh object proposals by their size, *objectness* (a sort of saliency), and distance from points-of-gaze. While it is uncertain that we can obtain nice object proposals under cluttered background as it always the case in first-person videos, this is a still interesting extension.

Modifying the task: Discovering objects attended by a subgroup of people In this thesis, we aimed to discover objects of shared attention, which was defined as those being looked at by *all* of the members of the group. An interesting modification of this task is to discover objects that are looked at by part of the group. In cooperative work, for instance, a group may split to subgroups to complete different tasks in parallel. Members of the subgroups might be split, merged or exchanged as the work goes on. In such cases, discovering objects commonly attended to by subgroups will provide important cues for discovering subgroups, understanding group dynamics, and how the overall tasks are completed. In this way, this modified task might be more challenging but can lead to another fruitful result.

Acknowledgments

本研究を進めるにあたり, 多くの方からご指導をいただきました．特に指導教官の佐藤洋一先生，並びに米谷竜先生，樋口啓太先生に最大の感謝を贈ります．修士の２年間にわたり継続的で丁寧な指導を受け多くのことを学びました．佐藤先生には研究全体の方向や問題の本質に関わる重要な助言を多くいただきました．また明快なプレゼンテーションのための発表構成やデザインに関しても多くコメントをいただきました．分野のトップ会議に参加するという貴重な経験もさせていただきました．米谷竜先生，樋口啓太先生には実際に研究を進めるにあたり様々な面でサポートをしていただきました．米谷先生は特に本稿も含めた様々な原稿の執筆において，文章の一文一文にわたり詳細なコメントと丁寧な修正をしていただきました．どのように論理を展開するか，文と文の間の論理の飛躍をどのように埋めていくかなどを学びました．樋口先生には特に実装面についてお世話になりました．また研究室で気さくに声をかけていただいて，楽しい研究生活を支えていただきました．その他の研究室のメンバーにも感謝を述べます．谷合竜典さんには楽しい SNS ライフを支えていただきました．同期の杉田さんと中野くんにも，ともに２年間を過ごしてくれたことに感謝します．データセットを集めるにあたり協力していただいた諸先輩方にも感謝します.

研究を進めている途中には様々なアップダウンがありました．ここまで辿り着くことができたのはひとえに，佐藤先生，米谷先生，樋口先生の親身なご指導，研究室の面々と学内外の友人や両親の誠実な支えのおかげです．ここに挙げきることのできなかったその他の方々にもあわせて感謝を述べます．ありがとうございました.

2017 年 02 月 02 日

A. Mathematical Background of Commonality Clustering

We here describe the mathematical background of the normalized spectral clustering for a bipartite graph that is used in Section 3.4 as a commonality clustering method. First, we introduce spectral clustering, which is formulated as an eigenvalue problem of graph laplacian matrix (Appendix A.1). We then present a variant of spectral clustering, known as normalized spectral clustering (Appendix A.2). At the commonality discovery, what we want is similar instances across two sets (in our case, two videos), which can be reduced to a bipartite graph. In this case, normalized spectral clustering can be efficiently conducted via singular value decomposition (SVD) (Appendix A.3).

A.1. Spectral Clustering

Spectral clustering [NJW01, Lux07] is a widely used clustering method, which utilizes spectral of an affinity matrix of data. We here focus on two-class clustering since our method uses two-class commonality clustering. Given affinity matrix $W \in \mathbb{R}^{n \times n}$ of n data points, two important matrices, degree matrix D and graph laplacian L are defined as follows:

$$D \coloneqq \begin{pmatrix} \sum_{j=0}^{n} w_{1j} & & \\ & \ddots & \\ & & \sum_{j=0}^{n} w_{nj} \end{pmatrix},$$
$$L \coloneqq D - W,$$

where w_{ij} is the (i, j)-th entry of W, and D is a diagonal matrix whose i-th diagonal entry is the sum of entries of the i-th row of W. Note that all entries of W are assumed to be positive.

The algorithm of the spectral clustering for two-class case is as follows:

Spectral Clustering (2-class case)

1. With graph laplacian L computed from the affinity matrix, solve the following minimization problem:

$$\boldsymbol{z}^* = \arg\min_{\boldsymbol{z} \in \mathbb{R}^n} \boldsymbol{z}^\top L \boldsymbol{z}, \quad \text{s.t. } \boldsymbol{z}^\top \boldsymbol{z} = 1, \boldsymbol{z}^\top \mathbf{1}_n = 0, \tag{A.1}$$

where $\mathbf{1}_n$ is an n-dimensional all-one vector.

2. Apply k-means clustering [Bis06] for $\boldsymbol{z}^*$ (in this case, $k = 2$).

The i-th entry of $\boldsymbol{z}^*$ is tied with the i-th data point. If the i-th and j-th entires of $\boldsymbol{z}^*$ belong to the same cluster after the second step, then that indicates the corresponding data points belongs to the same cluster.

To understand what Eq. (A.1) means, let us see the following five properties of the graph laplacian L:

Properties of the Graph Laplacian

1. The objective function of Eq. (A.1) is the weighted sum of square distance among entires of $\boldsymbol{z}$.

$$\boldsymbol{z}^\top L \boldsymbol{z} = \frac{1}{2} \sum_{i,j} w_{ij}(z_i - z_j)^2 = \sum_{i<j} w_{ij}(z_i - z_j)^2 \geq 0. \tag{A.2}$$

   This can be easily derived from the definition of L.

2. L is symmetric and positive-semidefinite. The symmetrisity is obvious from the definition of L. The positive-semidefiniteness is the result of Eq. (A.2): For arbitrary $\boldsymbol{z} \in \mathbb{R}^n$, $\boldsymbol{z}^\top L \boldsymbol{z} \geq 0$ holds.

3. The last equality of Eq. (A.2) obviously holds when $\boldsymbol{z}$ is n-dimensional all-one vector $\mathbf{1}_n$.

4. L has n non-negative real-valued eigenvalues $0 \leq \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$, since L is symmetric and positive-semidefinite.

5. Due to the properties 3 and 4, L has $\mathbf{1}_n$ as an eigenvector of $\lambda_1 = 0$.

With the property 1, Eq. (A.1) can be interpreted as follows.

- Minimizing $\boldsymbol{z}^\top L \boldsymbol{z}$ forces $z_i - z_j$ to be zero (or small) for large w_{ij}, which implies the i-th and j-th entries of $\boldsymbol{z}$ take the same (or similar) value when the i-th and j-th data points show high affinity. As a consequence, these points with high affinity will be clustered into the same one after the k-means clustering step.
- The constraint $\boldsymbol{z}^\top \mathbf{1}_n = 0$ forces $\boldsymbol{z}$ has both positive and negative entries. As the result, data points corresponding to positive entries of $\boldsymbol{z}$ form a cluster, and those corresponding to negative entries form another cluster.

- Minimizing $\boldsymbol{z}^\top L\boldsymbol{z}$ also forces w_{ij} to be small when the i-th and j-th points belong to different clusters, since z_i and z_j have different sign and thus $(z_i - z_j)^2 > 0$, which inevitably increases the objective quantity. In other words, it assigns the i-th and j-th point into different clusters when their affinity is low.

The $\boldsymbol{z}^*$ that minimizes the objective function of Eq. (A.1) can be obtained by solving eigenvalue problem.

$$L\boldsymbol{z} = \lambda\boldsymbol{z}. \tag{A.3}$$

Since $\boldsymbol{z}^\top L\boldsymbol{z} = \lambda$, $\boldsymbol{z}^*$ is the eigenverctor corresponding to the second smallest eigenvalue λ_2.

So far, we have introduced spectral clustering for two-cluster case. For general k clusters, one only need to take eigenvectors corresponding to the second to $(k+1)$-th smallest eigenvalues. The whole procedures are summarized as follows:

Spectral Clustering (k-class case; eigenvalue problem formulation)

1. With graph laplacian L computed from the affinity matrix, solve the following eigenvalue problem for $l = 2, \ldots (k+1)$:

   $$L\boldsymbol{z}_l = \lambda_l\boldsymbol{z}_l,$$

   where $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$.

2. Construct a matrix $Z = (\boldsymbol{z}_2 \ \ldots \ \boldsymbol{z}_{k+1})$, whose i-th column corresponds to $\boldsymbol{z}_{i+1}$, and apply k-means clustering for Z by regarding that each row represents the corresponding data point.

A.2. Normalized Spectral Clustering

Normalized spectral clustering [NJW01, Dhi01, Lux07] is a variant of the spectral clustering, which takes into account different cluster size among clusters. Instead of the graph laplacian, normalized spectral clustering uses a normalized graph laplacian:

$$L_{\text{norm}} \coloneqq D^{-\frac{1}{2}}LD^{-\frac{1}{2}} = I - D^{-\frac{1}{2}}WD^{-\frac{1}{2}}, \tag{A.4}$$

where D is a degree matrix, W is an affinity matrix and L is a graph laplacian. Normalized spectral clustering can be done by simply replacing L with L_{norm} in the

spectral clustering algorithm[1]. The properties 2 to 4 of the L hold for L_{sym}. As for the property 1, the following relation holds:

$$\boldsymbol{z}^\top L_{\mathrm{norm}}\boldsymbol{z} = \frac{1}{2}\sum_{i,j} w_{ij}\left(\frac{z_i}{\sqrt{d_i}} - \frac{z_j}{\sqrt{d_j}}\right)^2,$$

where d_i is the i-th diagonal entry of D. In the normalized graph laplacian case, minimizing the objective quantity $\boldsymbol{z}^\top L_{\mathrm{norm}}\boldsymbol{z}$ can be viewed as maximizing $\boldsymbol{z}^\top D^{-\frac{1}{2}} W D^{-\frac{1}{2}}\boldsymbol{z}$ because

$$\boldsymbol{z}^\top L_{\mathrm{norm}}\boldsymbol{z} = \boldsymbol{z}^\top L_{\mathrm{norm}}\boldsymbol{z} = \boldsymbol{z}^\top\left(I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}\right)\boldsymbol{z} = 1 - \boldsymbol{z}^\top D^{-\frac{1}{2}} W D^{-\frac{1}{2}}\boldsymbol{z},$$

where the second term is always non-negative since all entries of $D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$ are positive. As a consequence, the eigenvalue problem Eq. (A.3) becomes,

$$\begin{aligned} & L_{\mathrm{sym}}\boldsymbol{z} = \lambda\boldsymbol{z}, \\ & \iff (1 - D^{-\frac{1}{2}} W D^{-\frac{1}{2}})\boldsymbol{z} = \lambda\boldsymbol{z}, \\ & \iff D^{-\frac{1}{2}} W D^{-\frac{1}{2}}\boldsymbol{z} = \sigma\boldsymbol{z}, \end{aligned} \tag{A.5}$$

where $\sigma \coloneqq 1 - \lambda$. Instead of selecting the eigenvector of the second smallest eigenvalue, that of the second *largest* eigenvalue should be adopted. For k-class clustering, the eigenvectors of the second to $(k+1)$-th largest eigenvalues is used.

Let us look into the new objective $\boldsymbol{z}^\top D^{-\frac{1}{2}} W D^{-\frac{1}{2}}\boldsymbol{z}$. This can be expand to the following form.

$$\begin{aligned} \boldsymbol{z}^\top D^{-\frac{1}{2}} W D^{-\frac{1}{2}}\boldsymbol{z} &= \frac{1}{2}\sum_{i,j}\left(\frac{w_{ij}}{\sqrt{d_i}\sqrt{d_j}}\right) z_i z_j, \\ &= \frac{1}{2}\sum_{i,j}\sqrt{\frac{w_{ij}}{d_i}}\sqrt{\frac{w_{ji}}{d_j}} z_i z_j. \end{aligned}$$

w_{ij}/d_i is the ratio of w_{ij} to the sum of i-th row of W. Therefore, normalized spectral clustering gives higher importance on pairs with larger *normalized* affinity, *i.e.,* pairs whose affinity covers large fraction in the rows.

A.3. Normalized Spectral Clustering for Bipartite Graph

Suppose there are two groups of data points, and we want to perform clustering to discover similar data across groups. What we want to consider is not the affinity

[1] According to [NJW01, Lux07], it is useful to normalized each row of the matrix that constructed by eigenvectors before k-means clustering when there is a large variance in the diagonal entries of D. We omitted here this step since this is less likely to happen in our case.

between data points within a group, but those between groups. Such situation can be seen as clustering on a bipartite graph, where each of the group corresponds to different vertex set, and edges exists only between the vertex sets. In this case, the affinity matrix of data points can be modeled as follows:

$$W = \begin{pmatrix} O & C \\ C^\top & O \end{pmatrix},$$

where O is zero matrix, and $C \in \mathbb{R}^{n_1 \times n_2}$ is a group-between affinity matrix, whose (i, j)-th entries is the affinity of the i-th data of one group and j-th data of the other group, and n_k is the number of data in the k-th data group. The left hand side of Eq. (A.5) becomes

$$\begin{aligned} D^{-\frac{1}{2}} W D^{-\frac{1}{2}} \boldsymbol{z} &= \begin{pmatrix} D_1^{-\frac{1}{2}} & O \\ O & D_2^{-\frac{1}{2}} \end{pmatrix} \begin{pmatrix} O & C \\ C^\top & O \end{pmatrix} \begin{pmatrix} D_1^{-\frac{1}{2}} & O \\ O & D_2^{-\frac{1}{2}} \end{pmatrix} \boldsymbol{z}, \\ &= \begin{pmatrix} O & D_1^{-\frac{1}{2}} C D_2^{-\frac{1}{2}} \\ D_2^{-\frac{1}{2}} C^\top D_1^{-\frac{1}{2}} & O \end{pmatrix} \boldsymbol{z}, \end{aligned}$$

where D_i are the degree matrices of i-th data group. By defining the first n_1 entries of $\boldsymbol{z}$ as $\boldsymbol{z}_1$ and the rest n_2 entries as $\boldsymbol{z}_2$ (*i.e.*, $\boldsymbol{z} := \left(\boldsymbol{z}_1^\top \ \boldsymbol{z}_2^\top\right)^\top$), we obtain from Eq. (A.5),

$$\begin{aligned} D_1^{-\frac{1}{2}} C D_2^{-\frac{1}{2}} \boldsymbol{z}_2 &= \sigma \boldsymbol{z}_1, \\ D_2^{-\frac{1}{2}} C^\top D_1^{-\frac{1}{2}} \boldsymbol{z}_1 &= \sigma \boldsymbol{z}_2. \end{aligned}$$

These equations are what defines SVD of the normalized group-between affinity matrix $\tilde{C} := D_1^{-1/2} C D_2^{-1/2}$. The $\boldsymbol{z}_1, \boldsymbol{z}_2$ are the left and right singular vectors respectively, and σ is the corresponding singular value. For k-class clustering, the singular vector pairs of the second to $(\log_2\lceil k \rceil + 1)$- th singular value are used ($\lceil \cdot \rceil$ is the ceiling function). Not k but only $\log_2\lceil k \rceil$ singular vector pairs are required because each pair contains bi-modal information.

The k-class normalized spectral clustering for bipartite graph is summarized as follows.

Normalized Spectral Clustering for Bipartite Graphs

1. With the normalized group-between affinity matrix $\tilde{C} := D_1^{-1/2} C D_2^{-1/2}$ computed from the graph-between affinity matrix, perform SVD and obtain the left and right singular vector pairs of the second to $(\log_2\lceil k \rceil + 1)$-th largest singular value.

2. Construct a matrix $Z = \left(Z_1^\top \; Z_2^\top\right)^\top$ by vertically stacking Z_1 and Z_2, where Z_1 (or Z_2) is the the left (or right) singular matrix, whose columns are the left (or right) singular vectors obtained at the previous step. Apply k-means clustering for Z by regarding each row represents the corresponding data point[2].

Refer to [Dhi01] for more details.

[2] According to [NJW01, Lux07], it is useful to normalized each row of Z when there is a large variance in the diagonal entries of D. In such case, instead of the row normalization, we can also use $Z = \left((D_1^{-1/2}Z_1)^\top, (D_2^{-1/2}Z_2)^\top\right)$ as [Dhi01]. We omitted here these steps since this is less likely to happen in our case.

B. Other Results

Here, we provide the graphical results that are omitted in Chapter 4.

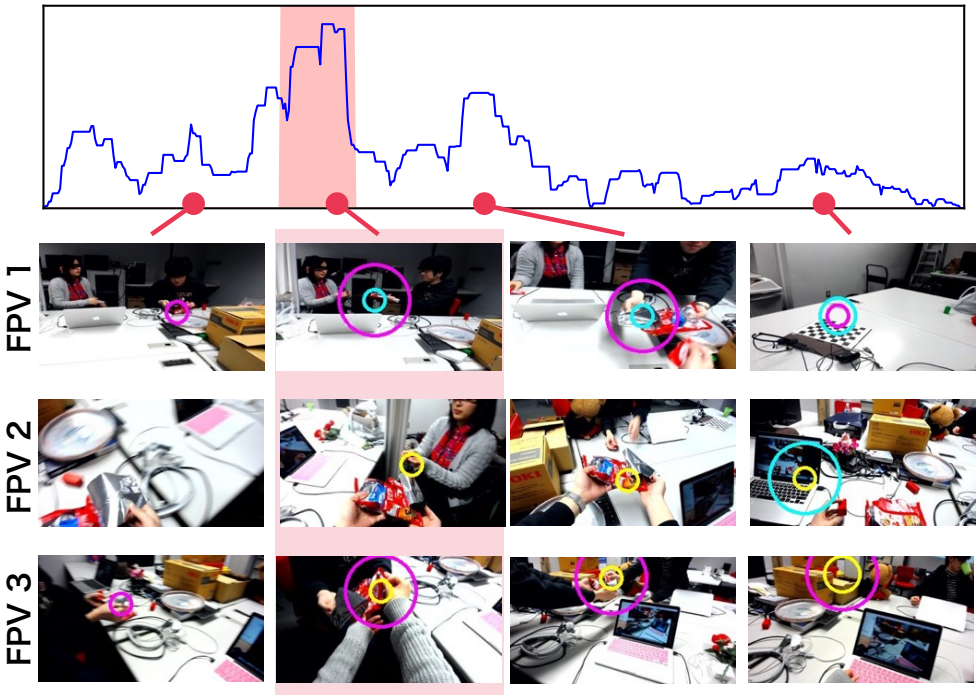

Figure B.1.: Other results omitted in the main text.

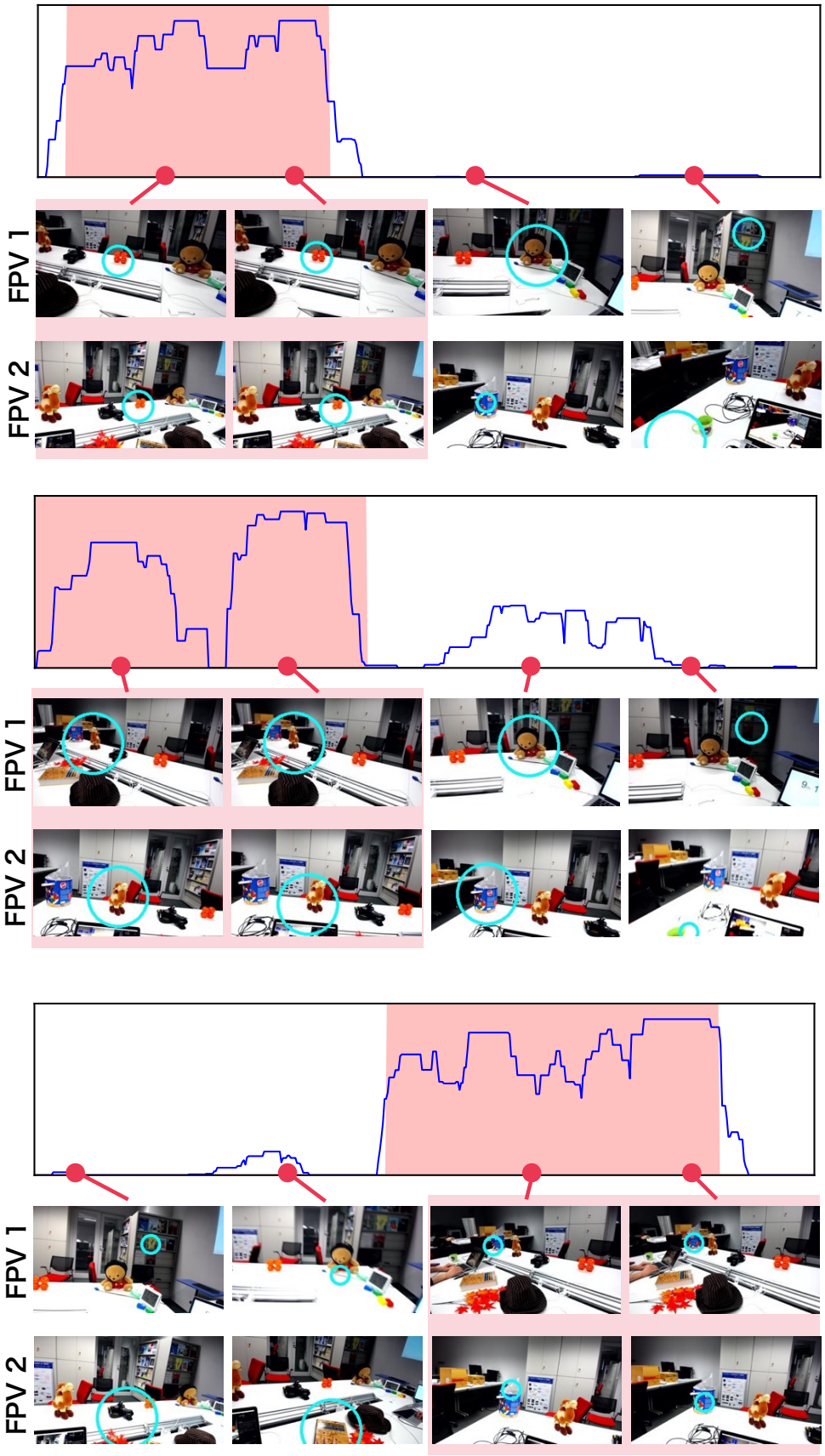

Figure B.2.: Other results omitted in the main text.

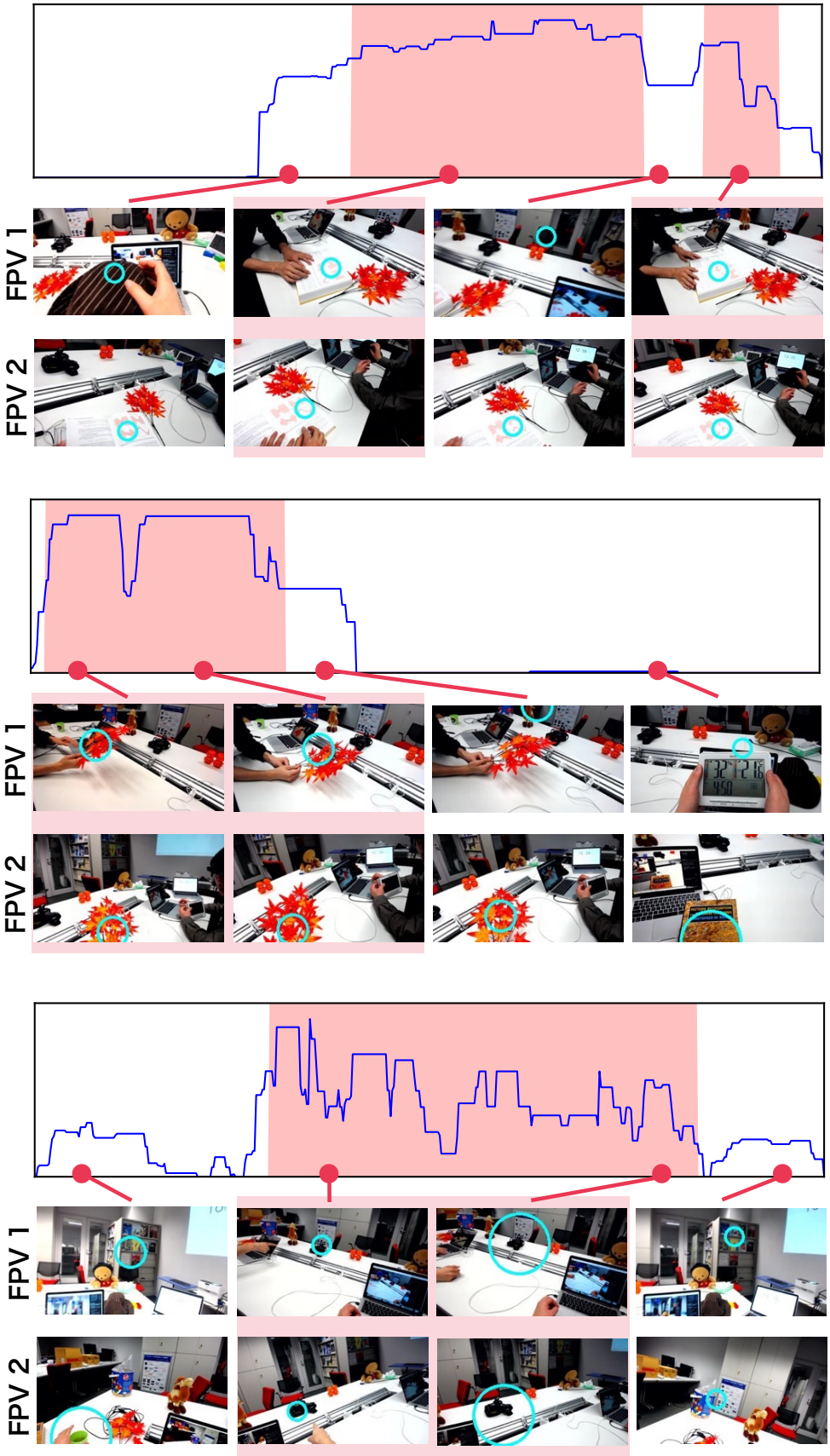

Figure B.3.: Other results omitted in the main text.

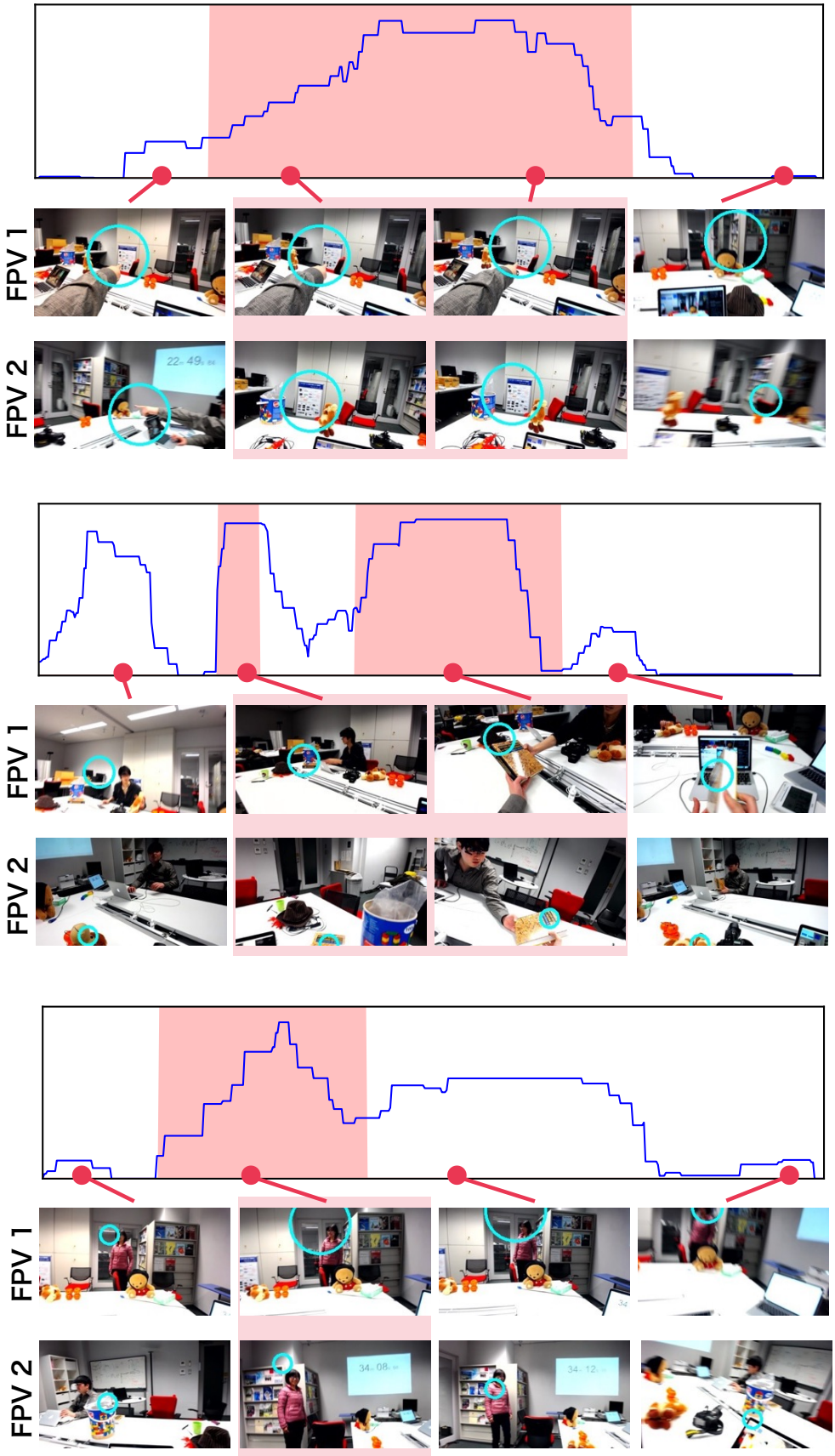

Figure B.4.: Other results omitted in the main text.

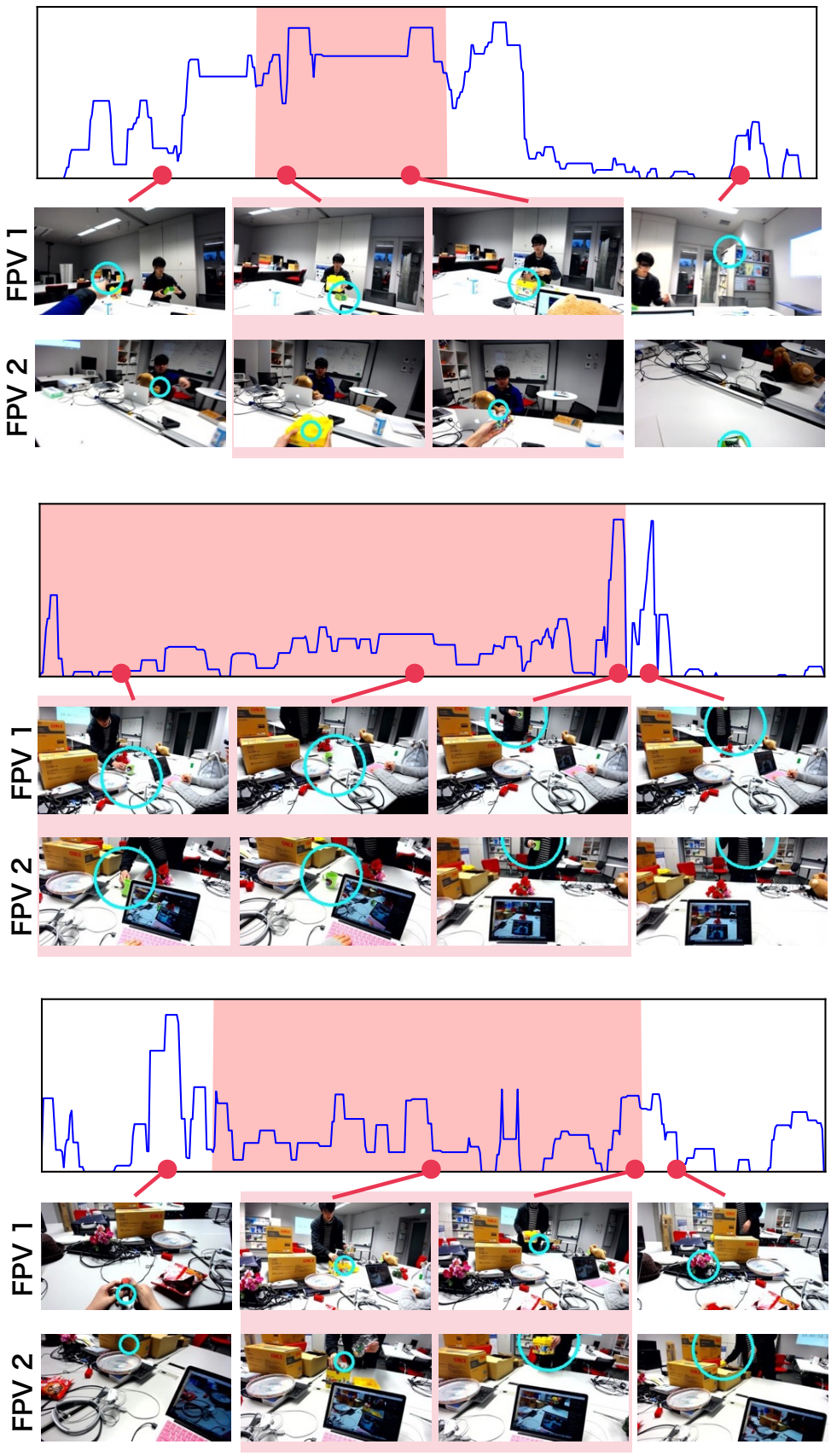

Figure B.5.: Other results omitted in the main text.

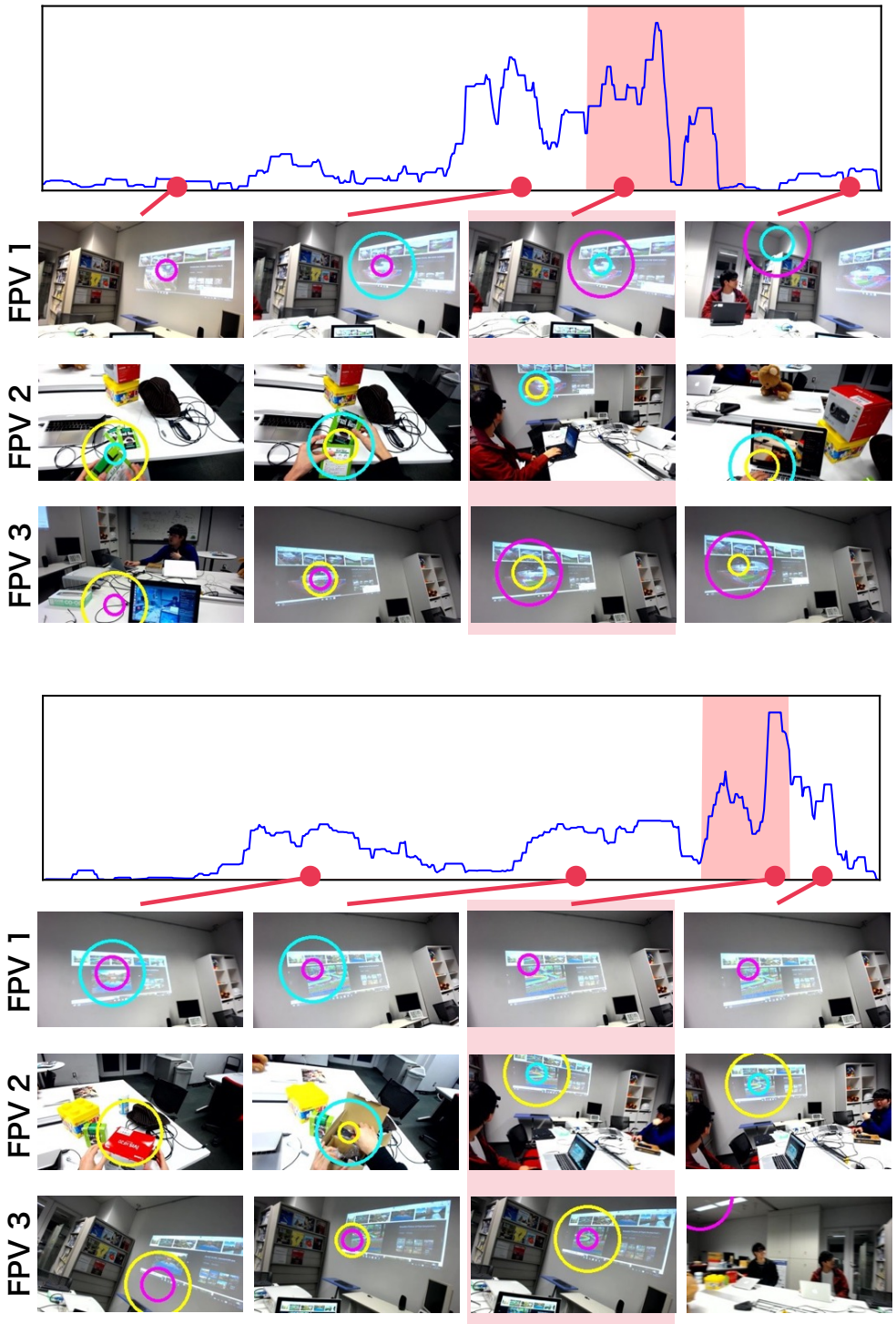

Figure B.6.: Other results omitted in the main text.

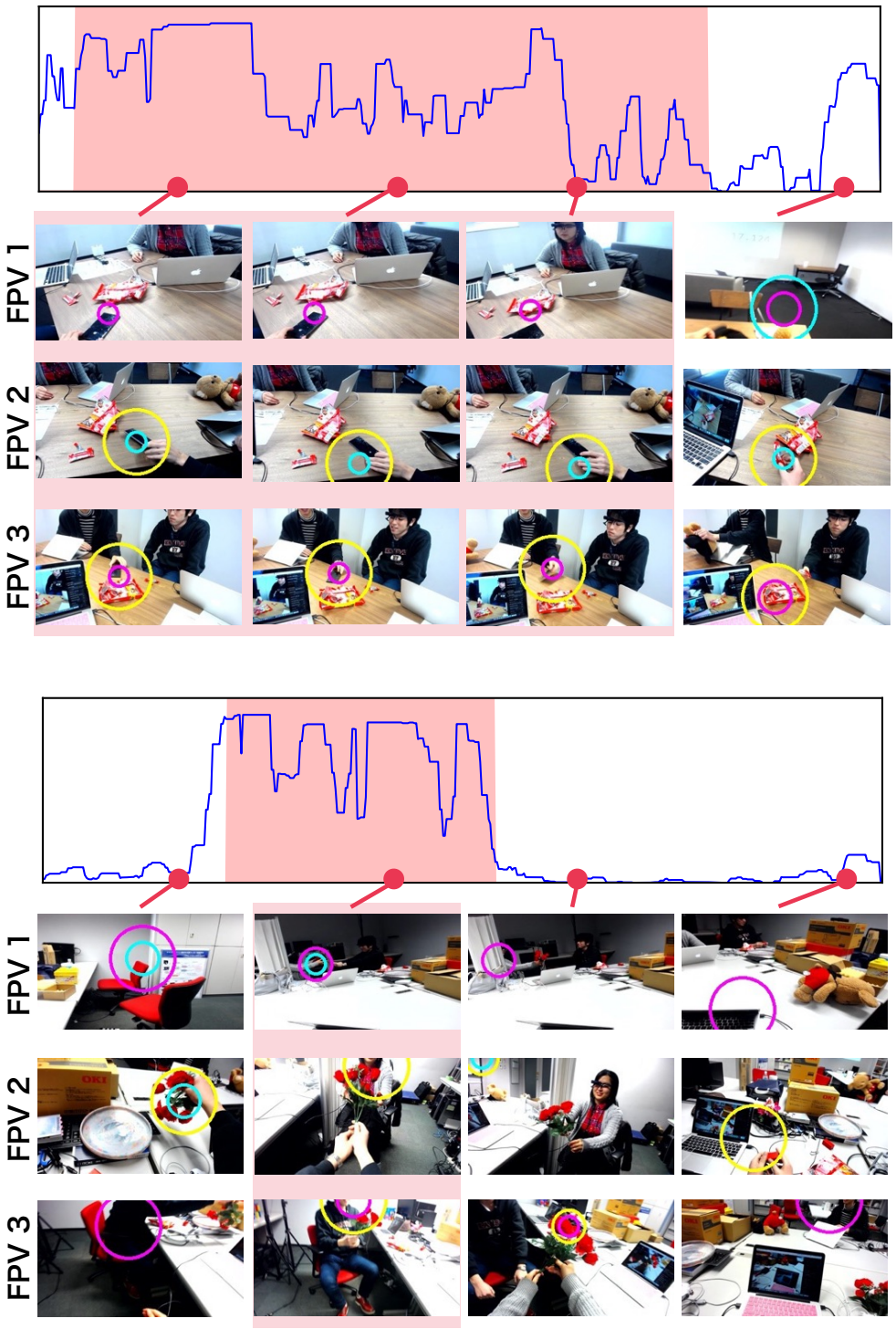

Figure B.7.: Other results omitted in the main text.

Bibliography

[APS+14] Ido Arev, Hyun S. Park, Yaser Sheikh, Jessica Hodgins, and Ariel Shamir. Automatic editing of footage from multiple social cameras. *ACM Transactions on Graphics*, 33(4): 81:1–81:11, 2014.

[Bis06] Christopher M. Bishop. Pattern recognition. *Machine Learning*, 1–28, 2006.

[CSJ15] Wen-Sheng Chu, Yale Song, and Alejandro 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.

[CZD12] Wen-Sheng Chu, Feng Zhou, and Fernando De la Torre Frade. Unsupervised temporal commonality discovery. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 373–387, 2012.

[CZLT14] Ming-Ming Cheng, Ziming Zhang, Wen-Yan Lin, and Philip 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.

[Dhi01] Inderjit S Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 269–274, 2001.

[EGW+] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[Eme00] Nathan J. Emery. The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience & Biobehavioral Reviews*, 24(6): 581–604, 2000.

[FHR12] Alireza Fathi, Jessica K. Hodgins, and James 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.

[FLR12] Alireza Fathi, Yin Li, and James M. Rehg. Learning to recognize daily actions using gaze. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 314–327, 2012.

[FRR11] Alireza Fathi, Xiaofeng Ren, and James M. Rehg. Learning to recognize objects in egocentric activities. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3281–3288, 2011.

[JBP10] Armand Joulin, Francis Bach, and Jean Ponce. Discriminative clustering for image co-segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1943–1950, 2010.

[KPB14] Moritz Kassner, William Patera, and Andreas 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.

[LAZ+15] Yuewei Lin, Kareem Abdelfatah, Youjie Zhou, Xiaochuan Fan, Hongkai Yu, Hui Qian, and Song Wang. Co-interest person detection from multiple wearable camera videos. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 4426–4434, 2015.

[LG13] Zheng Lu and Kristen Grauman. Story-driven summarization for egocentric video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2714–2721, 2013.

[LSD15] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015.

[Lux07] Ulrike Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4): 395–416, 2007.

[LYR15] Yin Li, Zhefan Ye, and James M. Rehg. Delving into egocentric actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 287–295, 2015.

[MAF09] Ajay Mishra, Yiannis Aloimonos, and Cheong L. Fah. Active segmentation with fixation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 468–475, 2009.

[NJW01] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, pages 849–856, 2001.

[OBT06] Sanae Okamoto-Barth and Masaki Tomonaga. *Development of Joint Attention in Infant Chimpanzees*, pages 155–171. 2006.

[PJS12] Hyun S. Park, Eakta Jain, and Yaser Sheikh. 3d social saliency from head-mounted cameras. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, pages 1–9, 2012.

[PJS13] Hyun S. Park, Eakta Jain, and Yaser Sheikh. Predicting primary gaze behavior using social saliency fields. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3503–3510, 2013.

[PR12] Hamed Pirsiavash and Deva 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.

[PS15] Hyun S. Park and Jianbo Shi. Social saliency prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4777–4785, 2015.

[RMBK06] Carsten Rother, Tom Minka, Andrew Blake, and Vladimir 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.

[SLD16] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99): 1–1, 2016.

[SRSM13] Nataliya Shapovalova, Michalis Raptis, Leonid Sigal, and Greg 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.

[SZ14] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[TJLFF14] Kevin Tang, Armand Joulin, Li-Jia Li, and Li 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.

[TL14] Titus J. J. Tang and Wai 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.

[Ver99] Roel 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, 1999.

[XML+15] Jia Xu, Lopamudra Mukherjee, Yin Li, Jamieson Warner, James M. Rehg, and Vikas 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.

[YGG12] Lee J. Yong, Joydeep Ghosh, and Kristen 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.

[YPS+13] Kiwon Yun, Yifan Peng, Dimitris Samaras, Gregory J. Zelinsky, and Tamara 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.

[ZJS14] Dong Zhang, Omar Javed, and Mubarak Shah. Video object co-segmentation by regulated maximum weight cliques. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 551–566, 2014.

List of Publications

1. Hiroshi Kera, Ryo Yonetani, Keita Higuchi, and Yoichi Sato, "Discovering Objects of Joint Attention via First-Person Sensing," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops on Egocentric Vision, pp.7-15, Jul. 2016
2. Hiroshi Kera, Ryo Yonetani, Keita Higuchi, and Yoichi Sato, "Discovering Objects of Shared Attention via First-Person Sensing," In Extended Abstract of Meeting on Image Recognition and Understanding (MIRU), Aug. 2016