Master Thesis

Temporal Localization and Spatial Segmentation of Joint Attention in Multiple First Person Videos

(複数人の一人称視点映像における共同注視区間と被注視領域の検出)

Yifei Huang
Advisor: Professor Yoichi Sato
Submission Date: Jan 31st, 2018

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

Advisor
Prof. Yoichi Sato

Abstract

Joint attention often happens during social interactions, in which individuals share focus on the same object. The understanding of joint attention is of great importance for many applications such as group activity analysis and autism diagnosis. This work aims to develop a computer-vision technique for discovering the objects jointly attended by a group of people during social interactions. As a key tool to discover objects of joint attention, we rely on a collection of wearable eye-tracking cameras that provide a first-person view of interaction scenes and also points-of-gaze data of interacting parties. Specifically, the goal is to temporally localize the time interval of joint attention, and to spatially segment the objects of joint attention.

The main challenges of this task lie on three aspects: 1) Although points of gaze data illustrate regions of interest of the camera wearer, the noises in gaze measurement would downgrade the reliability of information provided by points of gaze data, which in turn would affect the correct localization of the object of interest. 2) In natural cluttered scenes, the region around gaze position often includes many unrelated objects, making it hard to identify the attended object. 3) Usually an object is composed by parts with different appearance, which causes ambiguity of segmentation of the attended objects.

To address the above challenges, we propose a new method which alternates temporal localization of joint attention and spatial segmentation of jointly attended objects. The key insight behind the proposed method is that these two sub-tasks are closely coupled and the knowledge of one sub-task facilitates the inference of the other. Technically, we propose a hierarchical conditional random field-based model that observes as input segment proposals extracted from multiple videos, and infers as latent variables which segments are attended in each video and whether joint attention is established between videos. While comparing the visual similarity of segments that are likely being looked at across multiple videos, we also encode the temporal consistency between the appearance of segments that are looked at by individuals and between the binary states of whether joint attention is established. This makes it possible to discover objects of joint attention reliably even when scenes are cluttered, and points of gaze are noisy.

We evaluate our proposed method on a newly collected dataset which contains two-person cases and general cases. Experimental results show that our approach outperforms state-of-the-art methods for co-segmentation and joint attention discovery. Furthermore, we discussed the influence of different scales of noises in gaze measurement. Experimental results demonstrates the robustness of our method even in large

noise of gaze measurements. The failure cases indicate several possible extensions of our method. Although computationally complicated, 3D geometry is a good information that compensates the shortcomings of appearance based method. Using gaze prediction is also a possible extension, since gaze trackers are not convenient enough for larger scale deployment.

Contents

List of Figures

List of Tables

1. Introduction

1.1. Overview

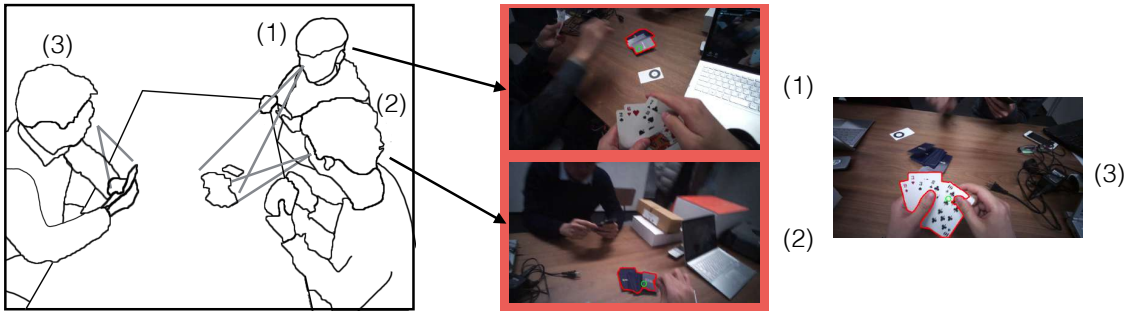

Figure 1.1.: Discovering Objects of Joint Attention. Joint attention between persons (1) and (2) is detected (highlighted in red boundaries) from first-person videos recorded with points of gaze data (green circles in the video frames.)

Joint attention is one of the primitive group behaviors observed during social interactions. In a meeting scene, people sometimes read a document together to share the information. On the street, there is a certain object like a posted notice that attracts attention of multiple pedestrians simultaneously. During group work, people may pay attention the same object at the same time, and the joint attention may happen many times on different task-dependent objects. Understanding when and to what such joint attention is established is crucial for multiple disciplines. For instance, joint attention of children provides an important cue for autism studies [CSBC+97]. Another example would be an automatic robot who would help people get objects of joint attention, or avoid running into those objects since they are important for multiple people [SPS15]. Moreover, locations where a group of people jointly focus can also be used for automatic video summarization [APS+14]. In this work, we aim to develop a computer-vision technique that can automatically discover objects of joint attention from multiple video streams recorded during natural social interactions.

We are particularly interested in using wearable eye-tracking cameras, such as Tobii Glasses[1], as a key tool to discover objects of joint attention. Such eye-tracking cameras can provide *first-person points-of-view videos* that contain what were observed in the camera wearer's field of view, and *points of gaze* data indicating where the

[1] https://www.tobiipro.com/product-listing/tobii-pro-glasses-2/

wearer looked at in the first-person videos (see Figure 1.1). Since joint attention is a group behaviour when different people cast attention onto one same object, the use of multiple cameras equipped with interaction parties is, therefore, promising for recording what they attended jointly during interactions.

Technically, we propose a hierarchical conditional random field-based model that can 1) localize events of joint attention temporally and 2) segment objects of joint attention spatially. We show that by alternating these two procedures, objects of joint attention can be discovered reliably even from cluttered scenes and noisy points-of-gaze data. Experimental results demonstrate that our approach outperforms several state-of-the-art methods for both co-segmentation and joint attention discovery.

1.2. Challenges and Contributions

One fundamental challenge of localizing objects of joint attention using point of gaze data is the inaccuracy of gaze measurement. Since an eye tracker cannot be 100% accurate, it happens frequently that the camera wearer is fixating on an object, while measured gaze position fall outside of the object (E.g., Figure (1.2a)). As a result, although point of gaze data measured by an eye tracker illuminates the parts of the wearer's field of view that receive attention, noisy points-of-gaze data provided by inaccurate eye tracking do not necessarily correspond to where people actually attend to. Because of this noise in gaze measurement, both the temporal localization and the spatial segmentation of objects of joint attention will become difficult.

Another challenge is the description of region being looked at. Naively thinking, given points of gaze data, we can extract features from the region near the points of gaze to describe a scene or an object being looked at, then using spatiotemporal commonality clustering on such feature vectors would be enough to achieve the task of finding objects of joint attention. But 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 inform us the spatial region the camera wearer is looking at, they do not tell which part is the region of objects, nor the spatial size of the region. 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, the size of objects changes drastically in first-person videos because the objects can be seen from different distances. To deal with such size variability, [KYHS16] used a multiscale approach for object-feature extraction. They extract features from multiple areas with different scales around gaze position, generating multiple candidates of objects. However, the direct use of such multiple scales as the region for feature extraction becomes problematic when the scene is cluttered, *i.e.*, when different objects are

(a) Noise in gaze measurement during consecutive frames

(b) cluttered scenes

(c) multiple interpretations of object

Figure 1.2.: Challenges of discovering Objects of Joint Attention.

close to each other. In such case (E.g., Figure (1.2b)), a region with large size would contain other objects as distraction, and a region with small size may only include part of the object being looked at.

Coupled with the former challenge comes a third challenge, which is the multiple interpretations of an object. In natural scenes, objects often have different parts, and each part could also be the object. As shown in Figure (1.2c), if the camera wearer is looking at the bottom part of the teddy bear, it is difficult to tell whether the camera wearer is looking at the red shirt, or looking at the whole bear, from only this single frame of the video. This is an intrinsic problem: "what is an object and what is a part of an object?" In our method, we do not limit the interpretation of an object, or object parts, but we demonstrate that by using our model with inter and intra video temporal consistency, we can obtain a semantically consistent result of an object of joint attention.

To address the aforementioned problems, we present a new approach of discovering objects of joint attention, which alternates temporal localization of joint attention and spatial segmentation of jointly attended objects. The key insight behind the proposed approach is that, given accurate segments of objects being looked at in multiple videos, the visual similarity of the segments provides a strong cue for determining whether or not joint attention is occurring. In turn, given the temporal localization of joint attention, we can know when the visual similarity of the segments should be enforced more strongly than other cues such as proximity to points of gaze. This contributes to better segmentation of jointly attended objects.

We formulate our approach using a conditional random field (CRF) that observes as input segment proposals extracted from multiple videos, and infers which segments are attended in each video and whether joint attention is established as latent variables. Since we use an object segment proposal as a region to extract features, it is possible for us to accurately describe each object even when the scene is cluttered. While comparing the visual similarity of segments that are likely to be a part of objects being looked at across multiple videos, we also evaluate the temporal consistency on which segments are looked at by individuals and if joint attention is established. This makes it possible to discover objects of joint attention reliably even when points of gaze are noisy, and when object can be differently interpreted.

Our main contributions are three-folded and summarized as follows:

- Firstly, we propose a novel model for temporally localizing and spatially segmenting objects being looked at jointly by people. To the best of our knowledge, this is the first work that performs co-segmentation on multiple first person videos with the utilization of point-of-gaze data. Our model addresses the main challenges that arise in the task of temporal localization and spatial segmentation of joint attention: 1) object size variability among objects and views, 2) the noise in gaze measurement, and 3) the different interpretation of object.
- Secondly, We introduce a new dataset of natural social interactions recorded with multiple wearable eye-trackers equipped with interaction parties, which includes annotations of temporal intervals and spatial segments of objects being looked at jointly. To our best knowledge, this is the first dataset that have multiple first person videos, with ground truth gaze data, object segmentation mask and joint attention period labeled.
- Thirdly, we demonstrate that the proposed method achieves state-of-the-art performance on both tasks of temporal localization and spatial segmentation of jointly-attended objects.

1.3. Thesis Outlines

The rest of this thesis is organized as follows. In Chapter 2, we first provide an overview of recent related works on co-segmentation, joint attention estimation and gaze guided computer vision. After that, four closely related methods are described in detail. We then present our method in Chapter 3, our method includes the base model for 2 person cases, and a generalized model for general cases. 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. To analyze the performance of our method in case of the use of other cheaper unreliable gaze trackers, We discuss the performance of our method and other baselines in more severe cases where additional synthesis noise is added into

point-of-gaze data in Chapter 5. Finally, Chapter 6 summarizes this thesis. In Appendix, we show some more graphical results.

2. Related Work

Since one of our goal is to spatially segment the object of joint attention using multiple first person videos, co-segmentation is the closest related topic. [CF13, TJLFF14, BKP+10, DNWZZ13, ZDlTH13, JBP10, ZZC12, TFNR12, QHZN16, WSSS17, SBH+13]. Our goal also includes temporally localize joint attention using point of gaze data, so joint attention [PJS12, SPS15, SPJS13, KYHS16, YCK+17] and gaze guided computer vision [Yar67, XML+15, FRR11, SRSM13, YPS+13, ZTMHL+17, KASB17, LRS+17, SPHSSP17], are also related with our work.

2.1. Co-segmentation

One of the popular computer vision topics which is closely relevant to our work is co-segmentation. Rother *et al.* [RKMB06] originally introduced the idea of co-segmentation, and much work has been done recently [CF13, TJLFF14, DNWZZ13, ZDlTH13, ZJS14]. Ma *et al.* [MLQH17] used an L1-manifold hypergraph joint-cut framework for unsupervised multi-class video co-segmentation. Taniai *et al.* [TSS16] used a hierarchical Markov Random Field to jointly recover co-segmentation and dense per-pixel correspondence between two images. Wang *et al.* [WHS+16] used a spatio-temporal energy minimization formulation for object co-segmentation across multiple videos containing irrelevant frames. Similar to our work, [FXZL14] used general object proposals as candidate regions. They further used a multi-state selection graph model to jointly optimize the segmentation of multiple objects.

However, one basic assumption behind existing co-segmentation methods is that the same object instances should be present under different background contexts for multiple input sources (with some exceptions aimed for dealing with intra-class variability of foreground objects, *E.g.*, [JBP10, RSLP12]). On the other hand, the task of discovering objects of joint attention presupposes that multiple cameras capture exactly the same scene (but possibly from different points of view). This prevents direct applications of existing co-segmentation methods and requires an additional cue to identify objects to be discovered, which is the points-of-gaze data in our work.

In the following subsection, we introduce the most important related study on co-segmentation. This work uses region proposals for video co-segmentation, but without utilizing gaze information.

2.1.1. Object Based Multiple Foreground Video Co-segmentation

Fu *et al.* [FXZL14] presented a video co-segmentation method that uses category-independent object proposals as its basic element, which is similar to our work. They observed that co-segmentation methods based on low-level appearance features may not adequately discriminate between the foreground and background. Also, object-based methods designed for single video segmentation do not take advantage of the joint information between videos.

Consequently, they proposed an Object-based Multiple foreground Video Co-segmentation method (ObMiC). By utilizing an object-based framework they can robustly and meaningfully separate foreground and background regions in images and individual videos, and then by constructing a co-selection graph they can connect each foreground candidate in multiple videos, thus performing the video co-segmentation task. Furthermore, since there are often cases where multiple objects appear in video, they extended the graph model to allow selection of multiple states in each node. This multistate selection graph is additionally able to accommodate the cases of a single foreground and/or a single video, and can be optimized by existing energy minimization techniques.

In the experiments, they evaluated their model in two cases: single foreground video co-segmentation and multiple foreground video co-segmentation. The baselines include co-saliency detection [FCT13], object-based proposals [EH10], object based image co-segmentation [MLLN12], Object-based video segmentation [ZJS13], Multi-class image co-segmentation [JBP12], and Multi-class video co-segmentation [CF13]. By utilizing both inter video and intra video constraints, their method achieves best performance in both cases. However, their method assumes the the existence of a common object proposal among the videos. When common objects exist, but not in all the videos, our method can still extract them, but will also extract an unrelated region in videos where the common object is missing.

Our work is different from this work in multiple disciplines. In our case, we use gaze information to explicitly represent the spatial location of the object of interest, and we add intra and inter video constraints to avoid the noise of gaze measurement and to predict object of joint attention. Also, including this work, other works on co-segmentation are all based on the assumption that the same object instances should be present under different background contexts for multiple input sources, but since in our problem setting all first person videos are looking at exactly the same scene, all the previous image/video co-segmentation methods cannot be directly used in our work.

2.2. Joint Attention Estimation

Our work tackles the phenomenon of joint attention, which is of great importance to social cognition [MN07, See11], early language[TF86], and the research of autism[CSBC+97]. The pioneering work is first proposed by Kera *et al.* [KYHS16] for discovering joint attention using eye-tracking cameras and extended in [HCK+17]. The method in this paper is an extension based on [HCK+17], which differs from [KYHS16] in multiple aspects. (1) Unlike our method that takes into account object segments of joint attention, their method estimates when joint attention occurs simply by examining feature similarities of spatio-temporal tubes of different sizes around the points of gaze. As a result, their method tends to be susceptible to error in point-of-gaze measurements and have difficulty in dealing with cluttered scenes: when the points-of-gaze data is wrongly measured, the direct use of spatial location of this wrong gaze position for feature extraction is apparently inappropriate. And when the scene is cluttered, extracting features from a region with large size would contain other objects as distraction, and from a region with small size may only include part of the object being looked at.

Our work shares some technical concepts in terms of measuring visual similarity of regions being looked at, the outputs are completely different. While [KYHS16] only localizes joint attention temporally, we can also segment objects of joint attention spatially. (2) When regarding joint attention of multiple persons, our method is able to treat all the possible sub-groups, while [KYHS16] only considers the joint attention of all people in the group. (3) the new model proposed in this paper furthermore takes logical relations of joint attention into account, which further improves the performance of temporal localization of joint attention.

Other related works include the analysis of "social saliency". Park *et al.* [PJS12, SPS15, SPJS13, SHSP17] introduced the notion of 3D social saliency in order to analyze interactions among people by using their first-person videos. The social saliency is modeled as an intersection of multiple 3D gaze directions. If multiple 3D gaze directions intersect, that indicates there is something near the intersection to which multiple people are attended. However, their method cannot differentiate the case of true joint attention from that of accidentally intersecting gaze directions where people are looking at different things behind the intersection. This is because Park *et al.*'s method is purely 3D geometry based, so intersections of fields-of-view of multiple wearable cameras do not necessarily correspond to objects of joint attention.

We introduce two closely related works in detail in the following subsections. The two works both address the problem of discovering the common interest of multiple people with first person cameras. Although neither of them contains both spatial segmentation and temporal localization, their goal is very similar to ours.

2.2.1. 3D Camera Pose Based Social Saliency Prediction

Park *et al.* [SPS15] presented a method to predict social saliency, *i.e.*, the likelihood of joint attention in 3D space. They provided an example in which an artificial agent that is trying to go through the crowd of people in a social scene. The agent tries 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 the location attracting the attention of the social group, *i.e.*, social saliency.

Given a social group and location of each member, they estimate the direction of joint attention from the center of the mass of the social group. To describe the distribution of a social group, the authors defined social formulation feature, and trained a binary ensemble classifier from a collection of such features. A continuous social saliency map of the target scene is generated with the classifier, which can be seen as the likelihood of joint attention. Given the situation where multiple groups exist simutaneously, 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 and then solved a minimization problem to select proper set of social groups. The minimization is designed so that the centers of different social groups are distant and also, each member in the scene 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 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.

In spite of sharing similar goals, 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(1) and (2), the camera wearers are not looking at the objects in the center of their views, but at the cards. 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 always available and are very costly to compute. In contrast to their work, this thesis presents a method to discover objects of joint attention, where points-of-gaze data illuminate which part in first-person vision the wearers are attended to over time. Also, in our work the spatial segment of the object of joint attention is also estimated, which is more precise than a brief location in 3D space.

2.2.2. Discovering Temporal Interval of Shared Attention

Kera *et al.* [KYHS16] presented a method for temporally localizing objects of shared attention using multiple first person cameras. This is the pioneering work on discovering objects of joint attention, and our work is inspired and extended from this work. Similar to this thesis, shared attention (or joint attention in this thesis) is defined 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) for the establishment of joint attention.

One fundamental challenge then arises in discovering the temporal interval containing objects of joint attention, is how to appropriately define the region of interest, from which to extract object features for commonality comparison. Although the point of gaze data illuminates the location of interest in the video, it does not tell which part is the region of objects. The region of objects around points of gaze largely depends on object sizes and viewpoints. A region too large may include many unrelated background, while a region too small may cause a small gaze shift inside an object be treated as attention shift between objects.

To address these problems, they proposed a multiscale approach for object feature extraction. In particular, 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 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 them to generate 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. Discovery results are finally integrated across scales to find various sizes of objects of shared attention reliably. In their experiments they showed competitive results against many commonality clustering and co-localization methods.

Our work originates from this work, and furthermore extended this work in multiple directions. Technically, we enhanced the method for temporally localization of joint attention. Wider speaking, we achieve another goal which is that our method is also able to spatially segment the objects of joint attention. To be precise, our work is different from this work in three key perspectives: firstly, although in this work the authors used spatiotemporal tubes, the spatial area is only different sizes around gaze position, which may be inappropriate under cluttered scenes. This thesis instead use object segments as region for extracting features, making it possible to find the real object of attention separately from other objects. Secondly, we not only determine the temporal interval of joint attention, but also output the spatial segmentation of joint attention. Thirdly, we extended the definition of joint attention, to enable the detection of joint attention among sub-groups. This work defined joint attention of

multiple people to be established only when *all* of the people in the group jointly attend to a same object, while our work is able to discover all possible states of joint attention, between *part* of people in a group.

2.3. Gaze-guided Computer Vision

The use of gaze information has significantly increased performance of numerous computer vision tasks [FLR12, SRSM13, YPS+13, XML+15, YRLS15], since point-of-gaze data illuminates the region of attention in the video or image. In [SRSM13] higher accuracy of action classification and localization was achieved by using gaze in the form of a weak supervisory signal. In [XML+15] gaze fixation was explored to help video summarization, by allowing generation of personalized summaries. In [YRLS15], eye-gaze patterns are used for multitask clustering to recognize first-person daily actions. By utilizing gaze information, it is possible for us to pinpoint the object of joint attention, enabling us to better spatially segment and temporally localize the objects of joint attention. In contrast to most previous work that used gaze information only from a single person, our work explore the gaze information in a collective way for an interesting topic - temporally localizing moments of joint attention and spatially segmenting objects of joint attention.

The following subsection describes a related work on gaze guided computer vision, which leverages point of gaze information as a cue for first person action recognition and other computer vision tasks like segmentation.

2.3.1. First-person Action Recognition Using Gaze

Fathi *et al.* [FLR12] first adopted gaze information into the task of first-person action recognition. One of the biggest difference between first-person action recognition and general action recognition is the huge ego-motion included in first-person videos. While this issue has been addressed in previous works [FRR11], the relationship between gaze and ego-motion is only first tackled in [FLR12]. As is well known, human attention and gaze are directed in a top-down task-dependent and goal-oriented manner [Yar67], this work utilized this fact and developed a top-down approach that utilizes fixation locations to help better recognize actions.

To describe the relationship between the egocentric action and the gaze location in each frame of an image sequence, they proposed a generative probabilistic model that observes object and appearance features as input, and infers both gaze positions and action labels. In particular, the features they used include object based features, since objects play important roles in action recognition. They also used appearance features to better describe the objects, since same object will have different appearances in different stages of an action. Furthermore, they introduced

future manipulation features, based on the fact in the psychology literature that the gaze is usually ahead of the hands in the hand-eye coordinate system [LH01].

Since there's no first person video dataset containing gaze information together with action object labels at the time, they proposed two datasets, **GTEA Gaze** and **GTEA Gaze+** and used both of them in their experiments. Both datasets are recorded when different subjects are cooking meals, using eye-trackers to capture gaze data and are manually labeled action labels. Compared with **GTEA Gaze**, **GTEA Gaze+** is more organized as they subjects were asked to make certain dishes with fixed recipe. In their experiment, the accuracy of action recognition given gaze (47%) is significantly higher than that without the aid of gaze information (27%). Inversely, gaze prediction results given action labels is also improved. Finally, the joint inference of gaze positions and action labels got best performance in both gaze prediction task and action recognition task.

While this work utilize gaze information, our work utilize gaze information in a collective way. We use synchronized point-of-gaze data from multiple first-person camera wearers for a different task - temporally localize and spatially segment the objects of joint attention.

3. Proposed Method

3.1. Model Architecture

Our model bases a hierarchical CRF that comprises several linear-chain CRFs as a sub-module. For simplicity of explanation, here we exclusively present a simple case for modeling joint attention of two persons. Figure 3.1 (a) depicts the overall architecture of the two-person case. Later we will show that our model can be easily extended to general cases where more than two people exist.

Let $j_t \in \{0, 1\}$ be a latent binary variable indexed by time-frame, where $j_t = 1$ means the two people establish joint attention at frame t and $j_t = 0$ otherwise. For the p-th video recorded by the p-th person (we here consider $p \in \{1, 2\}$ for two-person cases), we denote by $R_t^{(p)} = \{r_{t,1}^{(p)}, r_{t,2}^{(p)}, \dots\}$, a set of region proposals (spatial segments) at frame t. This can be generated by any region proposal method such

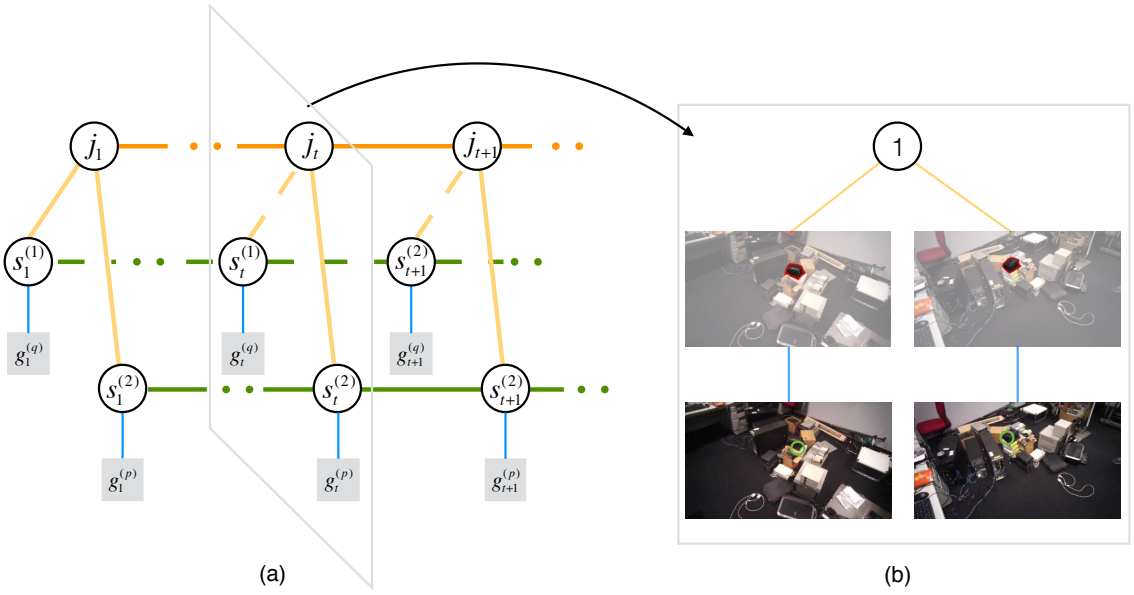

Figure 3.1.: Proposed Hierarchical CRF Model for discovering joint attention of two persons. The model accepts points of gaze $g_t^{(p)}$ as the input (green circles in (b), $p \in \{1, 2\}$) and estimate segments $s_t^{(p)}$ being looked at (red boundaries in (b)) as well as binary state j_t indicating whether the two persons establish joint attention or not.

as selective search [UvdSGS13] that provides spatial segments as object candidates. Then, the object segment looked at by the p-th person is described by $s_t^{(p)} \in R_t^{(p)}$ (*e.g.*, red boundaries in Figure 3.1 (b)). We regard $s_t^{(p)}$ as a latent variable as noisy points of gaze are not necessarily located inside the segment actually being looked at. Finally, we let $\boldsymbol{g}_t^{(p)} \in \mathbb{R}_+^2$ be a 2D point of gaze data at frame t (green circles in Figure 3.1 (b)), which is recorded in synchronization with the p-th video.

Now we construct the proposed model. The p-th sub-module takes points-of-gaze data $G^{(p)} = (\boldsymbol{g}_1^{(p)}, \ldots, \boldsymbol{g}_T^{(p)})$ as observations and segments being looked at $S^{(p)} = (s_1^{(p)} \ldots, s_T^{(p)})$ as latent variables. As a connection across the two sub-modules, two segments $s_t^{(1)}$ and $s_t^{(2)}$ further depend on joint attention variable j_t, which intuitively means that what each person looks at depends on if the two persons look at the same object or not. The objective function is then formulated as follows:

$$\begin{aligned}\Psi(S^{(1)}, S^{(2)}, J \mid G^{(1)}, G^{(2)}) = &\sum_{p\in\{1,2\}} \Psi_{\mathrm{GO}}(S^{(p)} \mid G^{(p)}) \\ &+ \sum_{p\in\{1,2\}} \Psi_{\mathrm{TS}}(S^{(p)}) \\ &+ \Psi_{\mathrm{JA}}(J, S^{(1)}, S^{(2)} \mid G^{(1)}, G^{(2)}) \\ &+ \Psi_{\mathrm{TJ}}(J),\end{aligned} \tag{3.1}$$

where the terms $\Psi_{\mathrm{GO}}, \Psi_{\mathrm{TS}}, \Psi_{\mathrm{JA}}, \Psi_{\mathrm{TJ}}$ are given concretely in the next section.

3.1.1. General cases

Our model can be extended to cases where $N \geq 2$ persons are present. Taking $M = N(N-1)/2$ pairs of first-person videos and points-of-gaze data as input, our extended model comprises M linear-chain CRFs as a sub-module. Given $\mathcal{S} = \{S^{(p)} \mid p = 1, \ldots, N\}$, $\mathcal{G} = \{G^{(p)} \mid p = 1, \ldots, N\}$, and $\mathcal{J} = \{J^{(p,q)} \mid p, q = 1, \ldots, N, p \neq q\}$, where $J^{(p,q)}$ denotes the joint attention between p and q-th persons, Eq. (3.1) is then modified as follows:

$$\begin{aligned}\Psi(\mathcal{S}, \mathcal{J} \mid \mathcal{G}) = &\sum_{p\in\{1,\ldots,N\}} \Psi_{\mathrm{GO}}(S^{(p)} \mid G^{(p)}) \\ &+ \sum_{p\in\{1,\ldots,N\}} \Psi_{\mathrm{TS}}(S^{(p)}) \\ &+ \sum_{p,q\in\{1,\ldots,N\}, p\neq q} \Psi_{\mathrm{JA}}(J^{(p,q)}, S^{(p)}, S^{(q)} \mid G^{(p)}, G^{(q)}) \\ &+ \sum_{p,q\in\{1,\ldots,N\}, p\neq q} \Psi_{\mathrm{TJ}}(J^{(p,q)}) \\ &+ \Psi_{\mathrm{LJ}}(\mathcal{J}).\end{aligned} \tag{3.2}$$

In the experiments we apply this extended model to discover joint attention of four persons.

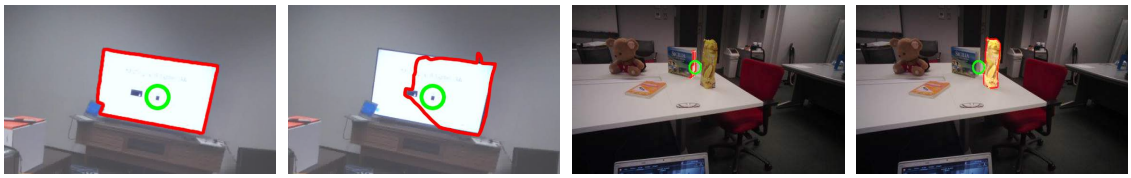

Figure 3.2.: Examples of image segments with different gaze proximity and objectness. Compared with (2), (1) should have a higher score of being the region of interest, since the region's shape is more convex. Compared with (4), (3) should have a higher score of being the region of interest, since the gaze point is closer to the region's center of mass.

3.2. Cues for Discovering Joint Attention

Our technical interests lie in how various cues about inputs (first-person videos and points of gaze data) and outputs (temporal intervals and spatial segments of joint attention) can be incorporated into the proposed model. The previous work [KYHS16] just focuses on the visual similarity of regions being looked at across multiple videos, which becomes problematic under practical cases when videos have cluttered scenes and points of gaze are noisy. In what follows we define the four terms $\Psi_{\mathrm{GO}}, \Psi_{\mathrm{TS}}, \Psi_{\mathrm{JA}}, \Psi_{\mathrm{TJ}}$ to cope with such cases.

3.2.1. Gaze proximity and objectness

Since point-of-gaze data illuminates the region of interest in first person videos, a segment near gaze point should be more likely to be the region of interest, than a segment relatively far away from the gaze point. Also, the object that attracts people's attention should be of regular shape, a random-shaped region is not likely to be the region of interest of the camera wearer, *e.g.*, Figure (3.2). We base this intuition and propose the following term Ψ_{GO}.

Ψ_{GO} describes how likely segment $s_t^{(p)}$ is to be looked at by p-th person given a point of gaze $\boldsymbol{g}_t^{(p)}$ (*gaze proximity*) and how likely the segment is to be an object (*objectness*). We evaluate the gaze proximity by the spatial distance between $s_t^{(p)}$ and $\boldsymbol{g}_t^{(p)}$ while the objectness is measured based on the shape of segments as follows:

$$\Psi_{\mathrm{GO}}(S^{(p)} \mid G^{(p)}) = \sum_{t=1}^{T} \left(\lambda_{\mathrm{GO1}} \frac{\| C(s_t^{(p)}) - \boldsymbol{g}_t^{(p)} \|_2}{|s_t^{(p)}|^{\frac{1}{2}}} + \lambda_{\mathrm{GO2}} \left(1 - \frac{|s_t^{(p)}|}{|H(s_t^{(p)})|} \right) \right), \quad (3.3)$$

where $C(s_t^{(p)})$ is the 2D centroid of segment $s_t^{(p)}$, $H(s_t^{(p)})$ is the convex hull of $s_t^{(p)}$, and $|x|$ is here the area of region x. The second term in the right-hand side intuitively means that a segment with large concavities is less likely to be an object. λ_{GO1} and λ_{GO2} are weight parameters that we will give concretely in Section 3.5.

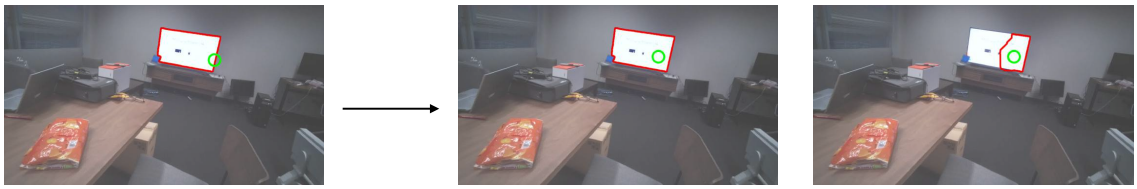

Figure 3.3.: An example of temporal consistency of segments. The segment in "frame t+1 (1)" should be more likely to be the following segment of "frame t", than the segment in "frame t+1 (2)".

3.2.2. Temporal consistency of segments

While the gaze proximity and objectness of segments are evaluated independently for each time frame, segments being looked at should be visually consistent over time as long as the people look at the same object, *e.g.*, Figure (3.3). We, therefore, consider the temporal consistency of segments in Ψ_{TS}. This is measured by the visual similarity of consecutive segments as follows:

$$\Psi_{\mathrm{TS}}(S^{(p)}) = \lambda_{\mathrm{TS}} \sum_{t=1}^{T-1} \left(1 - f_{\mathrm{sim}}(s_t^{(p)}, s_{t+1}^{(p)})\right), \tag{3.4}$$

where λ_{TS} is a weight parameter. The similarity function f_{sim} gives the cosine similarity of appearance-based features extracted from segments, which will be explained in detail in Section 3.5. This cost term helps us to track objects over time even if noisy points of gaze are scattered across various segments in a cluttered scene.

3.2.3. Joint attentionness

Similar to [KYHS16], we introduce the inter-video similarity of segments being looked at. Here we make simple assumptions that 1) when people look at the same object ($j_t = 1$), segments across multiple videos, $s_t^{(1)}$ and $s_t^{(2)}$, should be visually consistent and 2) when people pay attention to objects, their head is kept stable. A positive and a negative example of joint attention is shown in Figure (3.4). These two assumptions are implemented in Ψ_{JA} in the following fashion:

$$\Psi_{\mathrm{JA}}(J, S^{(1)}, S^{(2)} \mid G^{(1)}, G^{(2)}) = \sum_{t=1}^{T} \left(\lambda_{\mathrm{JA1}} Y(j_t, s_t^{(1)}, s_t^{(2)}, \boldsymbol{g}_t^{(1)}, \boldsymbol{g}_t^{(2)}) + \lambda_{\mathrm{JA2}} Z(j_t)\right), \tag{3.5}$$

where $\lambda_{\mathrm{JA1}}, \lambda_{\mathrm{JA2}}$ are weight parameters. The term $Y(j_t, s_t^{(1)}, s_t^{(2)}, \boldsymbol{g}_t^{(1)}, \boldsymbol{g}_t^{(2)})$ measures the visual similarity of the two segments $s_t^{(1)}$ and $s_t^{(2)}$:

$$Y(j_t, s_t^{(1)}, s_t^{(2)}, \boldsymbol{g}_t^{(1)}, \boldsymbol{g}_t^{(2)}) = j_t \left(1 - f_{\mathrm{sim}}(s_t^{(1)}, s_t^{(2)})\right) + (1 - j_t)\alpha(\boldsymbol{g}_t^{(1)}, \boldsymbol{g}_t^{(2)}), \tag{3.6}$$

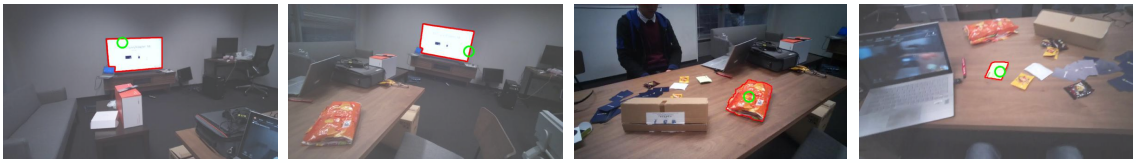

Figure 3.4.: An example of joint attentionness measurement of two persons. In this example, (1) and (2) are likely to have joint attention, while (3) and (4) only have very low probability of having joint attention.

where f_{sim} is given by the cosine similarity between two segments across videos as in Eq. (3.4). The first term in Eq. (3.6) encourages the two segments $s_t^{(1)}, s_t^{(2)}$ to be visually consistent when $j_t = 1$. On the other hand, the second term is needed in order to avoid a trivial solution where j_t becomes always zero. Please note that $\alpha(\boldsymbol{g}_t^{(1)}, \boldsymbol{g}_t^{(2)})$ measures the cosine similarity between regions around points of gaze $\boldsymbol{g}_t^{(1)}$ and $\boldsymbol{g}_t^{(2)}$, instead of $s_t^{(1)}$ and $s_t^{(2)}$. This is because the similarity of the segments $s_t^{(1)}$ and $s_t^{(2)}$ is irrelevant when no joint attention exists, and we expect that the people are more likely to be looking at different objects with different visual appearances. More details on how $\alpha(\boldsymbol{g}_t^{(1)}, \boldsymbol{g}_t^{(2)})$ is computed will be given in Section 3.5.

$Z(j_t)$ in the second term of Eq. (3.5) takes $Z(j_t) = j_t$ if the magnitude of global motion between consecutive frames is over threshold δ_{m} for either of the two videos, and $Z_t(j_t) = 0$ otherwise. This penalizes joint attention that occurs under large head motion and, as a result, allows us to discover joint attention only when the two people keep their head stable.

3.2.4. Temporal consistency of joint attention

Finally, we observe that joint attention typically continues for a certain time, *e.g.*, Figure (3.5). This motivates us to introduce another temporal consistency term Ψ_{TJ} on joint attention variables J as follows:

$$\Psi_{\text{TJ}}(J) = \lambda_{\text{TJ}} \sum_{t=1}^{T-1} |j_t - j_{t+1}|, \tag{3.7}$$

where λ_{TJ} is a weight parameter. Ψ_{TJ} prevents frequent onsets and offsets of joint attention.

3.2.5. Logical consistency of joint attention

In general cases (with 3 or more people), there is logical constraints between states of joint attention of different pairs of people. For example, if person 1 and 2 are sharing joint attention, and meanwhile person 2 and 3 are also sharing joint attention, then it

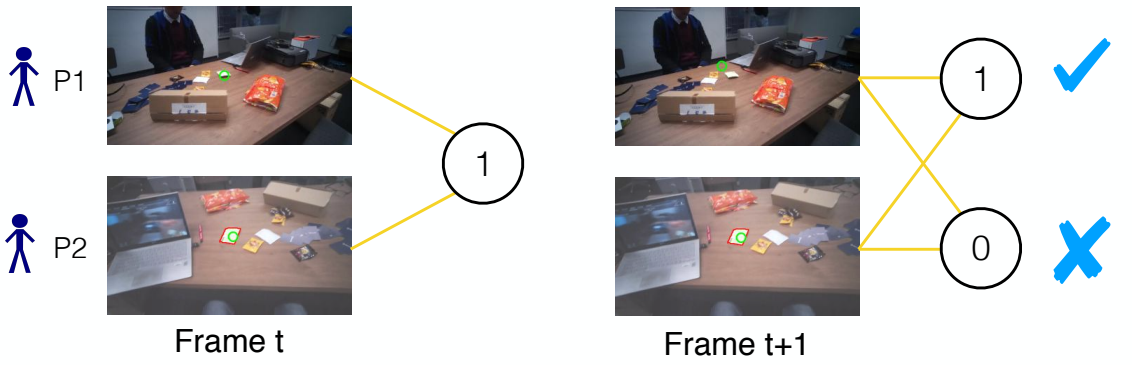

Figure 3.5.: Example of temporal consistency of joint attention. Person P1 and P2 are determined to have joint attention in frame t. In frame t+1, they are more likely to maintain their state of joint attention, than to suddenly change their state of joint atteniton.

is logical to assert that person 1 and 3 should have joint attention as well. Therefore, we introduce the term in Ψ_{LJ} to address the logical consistency of joint attention in general cases:

$$\Psi_{\mathrm{LJ}}(\mathcal{J}) = \lambda_{LJ} \sum_{t=1}^{T} f_{\mathrm{lv}}(\mathcal{J}_t) \tag{3.8}$$

where λ_{LJ} is a weight parameter. The function f_{lv} gives the number of logical violations taking as input the pairwise joint attention states $\mathcal{J}_t = \{j_t^{(p,q)} | p, q \in \{1, \ldots, N\}, p \neq q\}$ for all N persons at time t. The procedure for computing f_{lv} is given in Algorithm 1.

3.3. Parameter learning

Model parameters are jointly learned from training data which weight different components of the model. At the parameter learning stage, annotations of object masks and joint attention periods, and the measured points-of-gaze data are used. We apply the annotated object masks and the measured points-of-gaze to Eq. (3.3) to get gaze proximity and objectness scores. Similarly, we apply the annotated object masks and the joint attention periods to Eq. (3.5) to get joint attentionness scores. Temporal consistency scores of segments and joint attention are obtained by applying object masks and joint attention periods to Eq. (3.4) and Eq. (3.7) respectively. Then, we use a maximum likelihood approach to estimate the optimal model parameters $\{\lambda_{\mathrm{GO1}}, \lambda_{\mathrm{GO2}}, \lambda_{\mathrm{TS}}, \lambda_{\mathrm{JA1}}, \lambda_{\mathrm{JA2}}, \lambda_{\mathrm{TJ}}, \lambda_{\mathrm{LJ}}\}$ with which the model (3.1) achieves highest potential on the training data.

Algorithm 1: $ComputeLogicalViolation(\mathcal{J}_t)$

$\mathcal{X} \leftarrow \{\{p,q\} | j_t^{(p,q)} = 1, j_t^{(p,q)} \in \mathcal{J}_t\}$;
$n_0 \leftarrow |\mathcal{X}|$;
repeat
  $\mathcal{S} \leftarrow \emptyset$;
  foreach *pair of* X_i, X_j **in** $\mathcal{X}$ **do**
    $P \leftarrow X_i \cap X_j$;
    $Q \leftarrow X_i \cup X_j$;
    if $P \neq \emptyset$ **and** $Q \setminus P \notin \mathcal{X}$ **then**
      $\mathcal{S} \leftarrow \mathcal{S} \cup Q \setminus P$;
    end
  end
  $\mathcal{X} \leftarrow \mathcal{X} \cup \mathcal{S}$;
until $\mathcal{S} = \emptyset$;
return $|\mathcal{X}| - n_0$;

3.4. Model inference

Minimizing Eq. (3.1) with respect to $S^{(1)}, S^{(2)}, J$ gives us both of the temporal localization and the spatial segmentation of objects being looked at jointly. Here we describe the details of model optimization for the two-person case for simplicity of description. The two-person case can be extended easily to general cases of more than two persons by summing the energy functions of all possible pairs and adding the term $\Psi_{\mathrm{LJ}}(\mathcal{J})$. Since exhaustive search on the space of all possible combinations of object segments $S^{(1)}, S^{(2)}$ and joint attention states J is computationally intractable, we take an alternative inference algorithm to optimize the model. We divide the whole optimization procedure into three parts, each of which can be optimized separately using Viterbi algorithm [SM+12]:

Initialization At the beginning, we use gaze proximity, objectness, and temporal consistency of the object segments of attention to initialize $S^{(1)}$ and $S^{(2)}$ independently:

$$S^{(1)*}, S^{(2)*} = \arg\min_{S^{(1)}, S^{(2)}} \sum_{p \in \{1,2\}} \Psi_{\mathrm{GO}}(S^{(p)} \mid G^{(p)}) + \sum_{p \in \{1,2\}} \Psi_{\mathrm{TS}}(S^{(p)}) \tag{3.9}$$

Note that this part is also used as *Baseline 2* in our paper.

Temporal localization Fixing object segments obtained from the initialization part or the spatial segmentation part, we temporally localize joint attention by utilizing joint attentionness (visual similarity between object segments of two videos), and temporal consistency of joint attention:

$$J^* = \arg\min_{J} \Psi_{\mathrm{JA}}(J \mid S^{(1)}, S^{(2)}, G^{(1)}, G^{(2)}) + \Psi_{\mathrm{TJ}}(J) \tag{3.10}$$

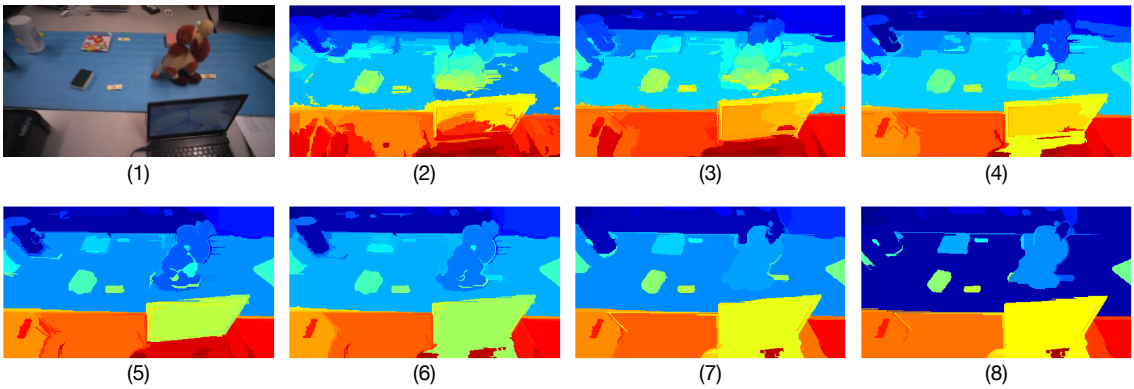

Figure 3.6.: Example output of Selective Search [UvdSGS13]. (1) is the original image, and (2) - (8) are the visualized different hierarchy of the output of Selective Search. This method enables us to generate object segment proposals of different scales.

Spatial segmentation Fixing joint attention states obtained from the temporal localization part, we optimize object segments using information as in the initialization part, and also the information from the other video if joint attention happens.

$$\begin{aligned} S^{(1)*}, S^{(2)*} = & \arg\min_{S^{(1)}, S^{(2)}} \sum_{p \in \{1,2\}} \Psi_{\mathrm{GO}}(S^{(p)} \mid G^{(p)}) \\ & + \sum_{p \in \{1,2\}} \Psi_{\mathrm{TS}}(S^{(p)}) + \Psi_{\mathrm{JA}}(S^{(1)}, S^{(2)} \mid J) \end{aligned} \tag{3.11}$$

As summarized in Algorithm 2, the initialization part is executed only once at the beginning. After that, we alternatively run the temporal localization part and spatial segmentation part until the change rate of J is below a certain threshold ξ (set as 0.02).

Algorithm 2: Alternative inference algorithm

Result: Optimized $S^{(1)}, S^{(2)}$ and J
Initialize segmentation $S^{(1)}$ and $S^{(2)}$ using Eq. (3.9) ;
while *Change rate* $\geq \xi$ **do**
    Optimize J by fixing $S^{(1)}, S^{(2)}$ using Eq. (3.10);
    Optimize $S^{(1)}, S^{(2)}$ by fixing J using Eq. (3.11);
    Estimate *Change rate* of J;
end

3.5. Implementation Details

We generate segment sets $R_t^{(p)}$ by Selective Search [UvdSGS13] per frame, making use of region masks and the bounding boxes. An example of the output region proposal is shown in Figure (3.6). For the region features extracted for comparing visual similarity, we first compute 1000-dimensional deep descriptors by feeding a pre-trained deep neural network (we used pre-trained network model from [SZ14]) using the cropped, warped box with the background of the region masked out (with the mean image), like [HAGM14]. We also compute HSV color histogram by discretizing each color channel into 16 bins, normalizing them independently and then concatenating them into 48-dimensional feature vectors. We then concatenate those features to form a final 1048 dimensional feature vector for comparing region visual similarity.

To compute global motion of videos we use the Lucas-Kanade method, and set the threshold to $\delta_m < 1.5$. For computing $\alpha(\boldsymbol{g}_t^{(1)}, \boldsymbol{g}_t^{(2)})$, we first compare the visual similarity of circular regions around points of gaze at multiple scales (15, 25, 50 pixel-radius) similar to [KYHS16] and then select the maximum similarity.

4. Experiments

To evaluate the performance of the proposed approach on both tasks of temporal localization and spatial segmentation of jointly attended objects, we collected a new dataset that recorded realistic social interaction scenes with multiple wearable eye-tracking cameras. We divide our dataset into two parts: two persons cases and general cases, to test our proposed algorithms on such two situations.

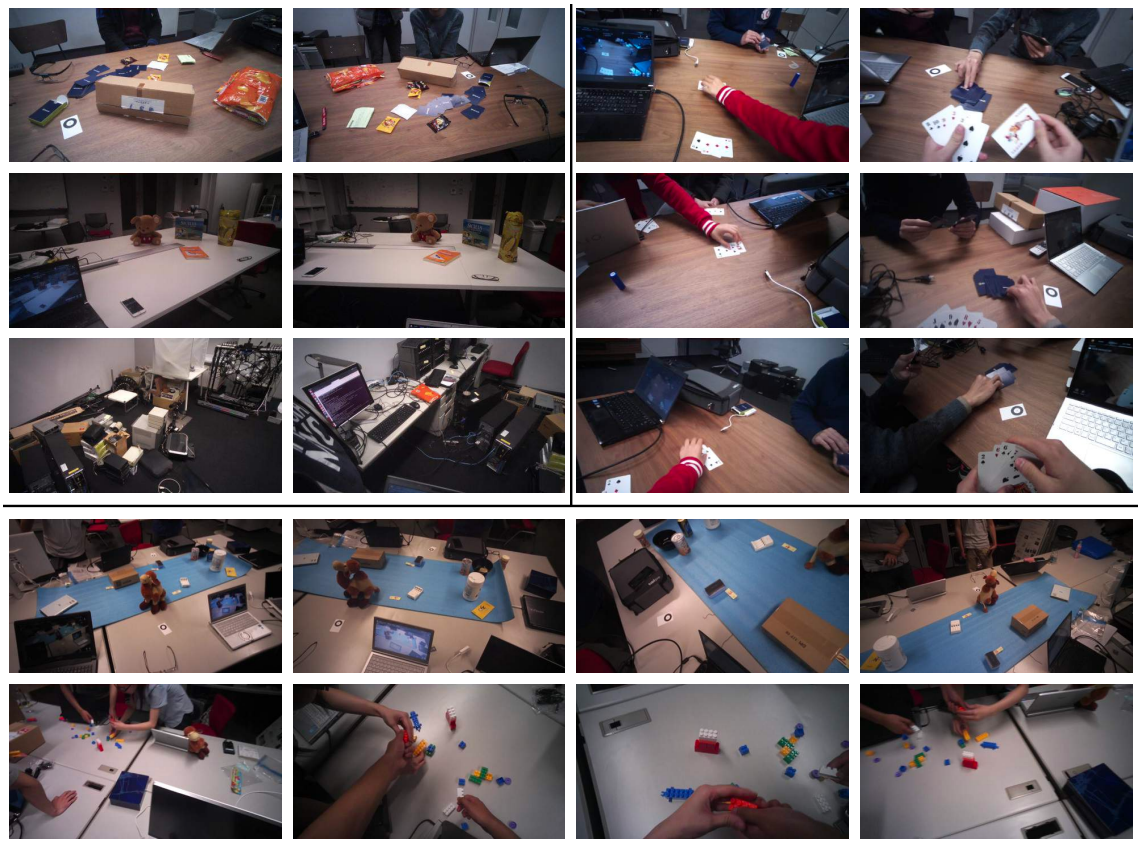

Figure 4.1.: Example images in our dataset. Our dataset include two-person cases (top-left), and general cases: three-person cases on the top-right, and four-person cases at bottom.

4.1. Two persons cases

Following [KYHS16], we first address the cases where two persons in interactions establish joint attention under several different formations and will particularize more general cases in the following section.

4.1.1. Experimental Setting

For each of the formations **side-by-side (SbS)** and **face-to-face (FtF)** originally presented in [KYHS16], we further divide it into two different scenarios where people shift attention between objects with large head motion or small head motion. For the first scenario (SbS), objects are placed close to each other, which requires only a slight shift in attention with little head motion. For the other scenario, objects are placed on two tables distant to each other, which induces large head rotations (over 90 degrees) to shift attention between the objects. As a result, we evaluate the methods for four different recording conditions in total: **SbS-large**, **SbS-small**, **FtF-large**, **FtF-small**.

Our dataset is collected in four different indoor environments. For each environment, we use a diverse set of objects for joint attention. During each recording, subjects were asked to establish joint attention on different objects placed at different locations, just as what they do in everyday interactions. In the two-person cases, 24 pairs of first-person videos and points-of-gaze data were recorded in total. The dataset is now publicly available online[1]. Each participant was equipped with a Tobii Pro Glass 2 that was calibrated and manually synchronized for each recording. During recording, one subject first describes or interacts with an object, and then the other subject turns attention to it. This is repeated several times to form a whole sequence. Videos were recorded at 25-fps with the resolution of 1920×1080. Ground-truth labels for temporal localization were annotated by manual inspections from all participants of each recording. Then we used GrabCut [RKB04] to generate binary masks of objects being looked at jointly for a total of 1250 sampled frames, as ground-truth labels for the segmentation task.

We first address the task of spatially segmenting jointly attended objects. The intersection-over-union (IoU) ratio is used as an evaluation metric. We adopt the following three baselines:

ObMiC [FXZL14]. This method is one of the most relevant method to our work as it used region proposals and considered temporal consistency for co-segmenting objects across multiple videos. We introduce this baseline to see how points-of-gaze information guides the segmentation of objects being looked at jointly since this baseline method is not utilizing the information from point-of-gaze data.

[1] https://github.com/cai-mj/UTJA_dataset

Figure 4.2.: Segmenting Objects of Joint Attention: Examples. Red boundaries indicate jointly-attended object segments and green circles describe points of gaze. The first two rows describe the ground truth segments for the two input videos. The remaining rows show segmentation results in the second video, performed by our method and the three other baselines.

Baseline1. In order to see if points-of-gaze information alone works well for segmenting objects of joint attention, this simplified version of the proposed model employs the only Ψ_{GO}, the first term of Eq. (3.1), while abnegating temporal consistency between segments and the effect of joint attention.

Baseline2. In this baseline, we aim to see how the cue of temporal consistency helps stable segmentation under cluttered scenes and noisy points of gaze. Specifically, we use Ψ_{GO} and Ψ_{TS}, which means that we optimize multiple linear-chain CRF sub-modules independently without considering the cues about joint attention.

Quantitative results are shown in Table 4.1. The proposed model clearly outperforms ObMiC [FXZL14] that did not use gaze information. The proposed model also performs consistently better than the two baselines, indicating the necessity of temporal consistency cue Ψ_{TS} and joint attention cues $\Psi_{\mathrm{JS}}, \Psi_{\mathrm{TJ}}$. By comparing the four recording conditions, it can be seen that FtF formations are generally more challenging than SbS ones. This is typically due to a large difference of viewpoints between the two persons in the FtF formation, causing object appearance inconsistent across videos. In addition, the segmentation performance often degrades under large head motion due to unstable eye tracking and motion blur.

Method	FtF-large	FtF-small	SbS-large	SbS-small	Avg.
ObMiC [FXZL14]	0.287	0.212	0.065	0.336	0.225
Baseline1	0.552	0.599	0.681	0.691	0.631
Baseline2	0.611	0.629	0.723	0.726	0.672
Ours	**0.633**	**0.660**	**0.730**	**0.735**	**0.690**

Table 4.1.: Quantitative Comparisons on Segmentation Task of Two-person Cases: Intersection-over-union (IoU) ratio for four different recording conditions of two persons.

Method	FtF-large (%)		FtF-small (%)		SbS-large (%)		SbS-small (%)		Avg. (%)
	P	R	P	R	P	R	P	R	F1 score
Kera *et al.*	74.5	89.7	69.7	**93.8**	72.9	**96.5**	67.1	83.4	79.0
Ours	**91.9**	**92.8**	**84.7**	86.5	**94.3**	92.6	**79.7**	**98.7**	**89.3**

Table 4.2.: Quantitative Comparisons on Temporal Localization Task: Precision (P) and recall (R) scores for each condition as well as the F1 score averaged over all the conditions.

Figure 4.2 shows some qualitative results of our experiment on two person cases. Without using gaze information, ObMiC [FXZL14] may not be able to correctly localize the attended object, resulting a poor performance on IoU. As shown in the examples (a) and (b), the proposed model is able to find the correct object segment even when the noisy point-of-gaze is outside the object of attention by taking into account temporal consistency. Baseline methods under-segment or over-segment objects in the examples (c), (d), and (e), (f), while our method can perform a stable segmentation thanks to the cues of joint attention. Additionally as shown in example (d), our method is able to segment a consistent interpretation of the object of joint attention: the whole teddy bear, but not jumping between the interpretations of "the head of the bear" and "the whole bear".

More importantly, we observe in the experiments that the per-frame score of objective function monotonically decreases at each step of iteration (see Figure 4.3), which validates our claim that accurate segmentation guides accurate temporal localization, and vice versa.

4.1.2. Temporal Localization Task

Next, we address the task of temporal localization of joint attention. Here we compare our approach against [KYHS16] which is the only relevant work for the same task to the best of our knowledge. As shown in Table 4.2, the baseline method [KYHS16] is prone to obtain higher recall/lower precision scores, indicating

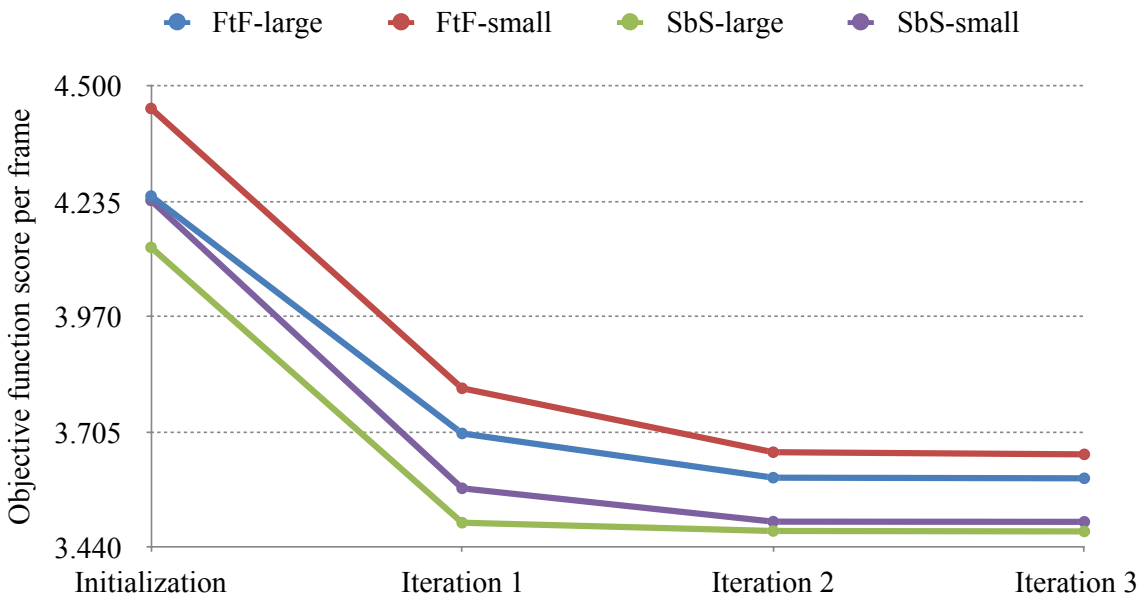

Figure 4.3.: Per-frame objective function score at each iteration. Note that not all video pairs enter Iteration 3, and the scores of those which terminate at Iteration 2 are treated as static in Iteration 3.

that irrelevant temporal intervals tend to be judged as joint attention periods. On the other hand, our approach can obtain more balanced precision and recall scores and a much higher F1 score (by more than 10%).

By comparing four different recording conditions, the performance on FtF cases are worse than that on SbS cases due to object appearance disparity. However, our method performs better under large head motion since head motion works as a constraint (as shown in Eq. (3.5)) in predicting joint attention.

4.2. General cases

We also collected three and four-persons interaction data to evaluate the extended version of our model presented in Section 3.1.

4.2.1. Experiment Settings

Without loss of generality, we use three and four person cases to represent general cases where a group of people collaboratively doing tasks. In three or four person cases, people take round formulation naturally (triangle formulation for three people, and square formulation for four). In three persons case, three participants were asked to sit in triangle formation around a table, as shown in Figure (1.1), to play a card game. In four persons cases, participants took square formation, and were asked to

perform tasks such as passing-receiving, assembling building blocks. The formation makes it hard to separate people as FtF or SbS, so we evaluate performances people-orientedly. Here we collected 11 groups of first person videos with synchronized points-of-gaze data, in which there are 8 four-person groups and 3 three-person group. We manually synchronize the videos between different persons and label the ground truth of all the data. Other settings are congruent with that of two person cases.

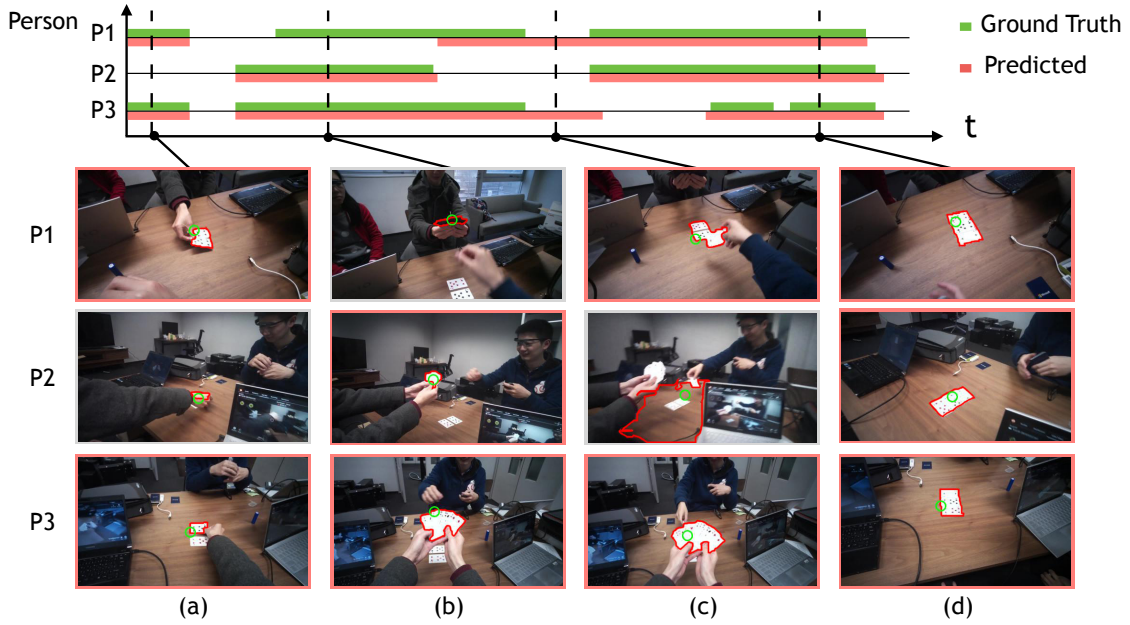

Figure 4.4.: Joint Attention Discovery for Three Persons Case. The top half shows the ground truth and predicted results of temporal localization. The bottom half depicts some segmentation results in pink boundaries and points of gaze in green circles. Images highlighted in pink borders are judged as joint attention periods by the proposed model.

4.2.2. Three persons cases

As three person cases are only the sub-problems of four person cases, without loss of generality, we only show the qualitative results on both tasks for the sake of simplicity.

Figure (4.4) depicts qualitative results. We confirm that joint attention is discovered correctly when (a) persons P1 and P3 jointly pay attention to the same card in P3's hand and (d) P1, P2, P3 all look at the same card on the table. On the other hand, false negative (P1 and P2, P3 are jointly looking at the same cards but P1 is not considered as one participant of joint attention) and false positive (P1 and P3 are looking at different set of cards but they are determined to be establishing joint attention in this time period) results are found in (b) and (c), respectively. These

Method	Person1	Person2	Person3	person4	Avg.
ObMiC	0.016	0.087	0.039	0.181	0.081
Baseline1	0.524	0.446	0.635	0.438	0.511
Baseline2	0.656	0.572	0.696	0.606	0.632
Huang *et al.*	0.707	0.592	**0.740**	0.601	0.660
Ours	**0.713**	**0.622**	0.721	**0.613**	**0.667**

Table 4.3.: Quantitative Comparisons on Segmentation Task: Intersection-over-union (IoU) for four different persons in four persons cases.

failure cases imply some potential limitations of our approach, which we will discuss in the next chapter.

4.2.3. Four person cases

We mainly use four person cases to analyze the performance of our extended model, since three person cases are only a sub-problem of four person cases. In the temporal localization task, we again use IoU as evaluation metric. Other than the two baselines discussed in section 4.1.1, we add another baseline (Huang *et al.*[HCK+17]), which is our model without the extension of logical relationship. In this baseline, we would like to see how the accuracy of joint attention state effect the segmentation IoU ratio. We show qualitative results of both tasks in Figure (4.5). We can conclude from the figure that our method successfully detected the case of no joint attention in (a), the case of joint attention of two two-person sub-group in (b), the joint attention of a three-person sub-group in (c), and the case where all of the four persons establish joint attention in (e). (d) represents a failure case where P1, P2 and P3 are all casting attention to the teddy bear, while P3 is not determined to be establishing joint attention together with P1 and P2, because of the viewpoint difference enlarged the appearance difference, and our method is purely appearance based.

Quantitative results are shown in Table (4.3). The proposed model outperforms all baselines on average IoU ratios. From the comparison of Huang *et al.*and our method with the extension of logical relationship, we can conclude two points. Firstly, the use of logical relationship benefits the temporal localization of joint attention. Secondly, (shown in Figure 4.6) with more accurate temporal localization of joint attention state, the spatial segmentation result can also be improved.

Quantitative results of the temporal localization task is shown in Figure (4.6). We use three baselines here. Other than [KYHS16] and [HCK+17], we consider the segment only method (segment-based), which is our model only using only Ψ_{GO}, the first term of Eq.(3.3). Here we consider several definitions of joint attention: pair is the smallest condition where we consider the joint attention of each pair of persons. Triplet indicates we consider joint attention from a triplet's perspective (e.g. there

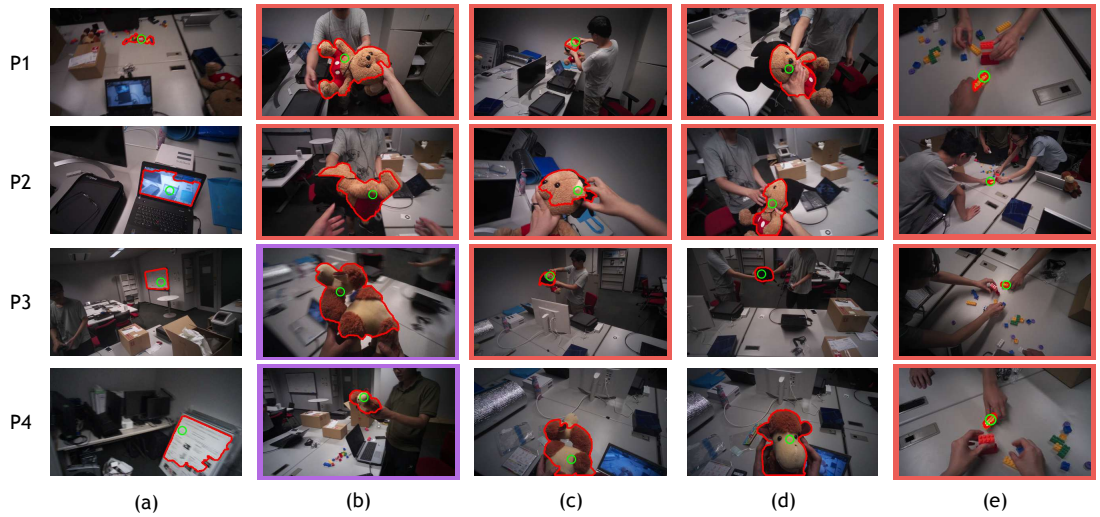

Figure 4.5.: Examples of temporal localization and spatial segmentation results of 4 persons cases dataset. In each image, green circle indicates gaze position. Object segment is highlighted using red boundaries inside each image. Images borders highlited using same color are judged as joint attention periods by our proposed model.

are three different triplets in a group of four): whether three people jointly attend to one object simultaneously. Quadruplet is the case where we consider the group as a whole, whether all of the four people attend to the object or not. We observed that our method achieves best performance (in F1 scores) in all cases. From the comparison of [KYHS16] and the segment-based method, we can see that the use of image segment as regions for feature extraction is a better choice for describing the object of interest, since as we have discussed, less background and less regions of other objects would be included. From the comparison of our model with [HCK+17], we can conclude that the use of logical consistency of joint attention is helpful for the temporal localization task.

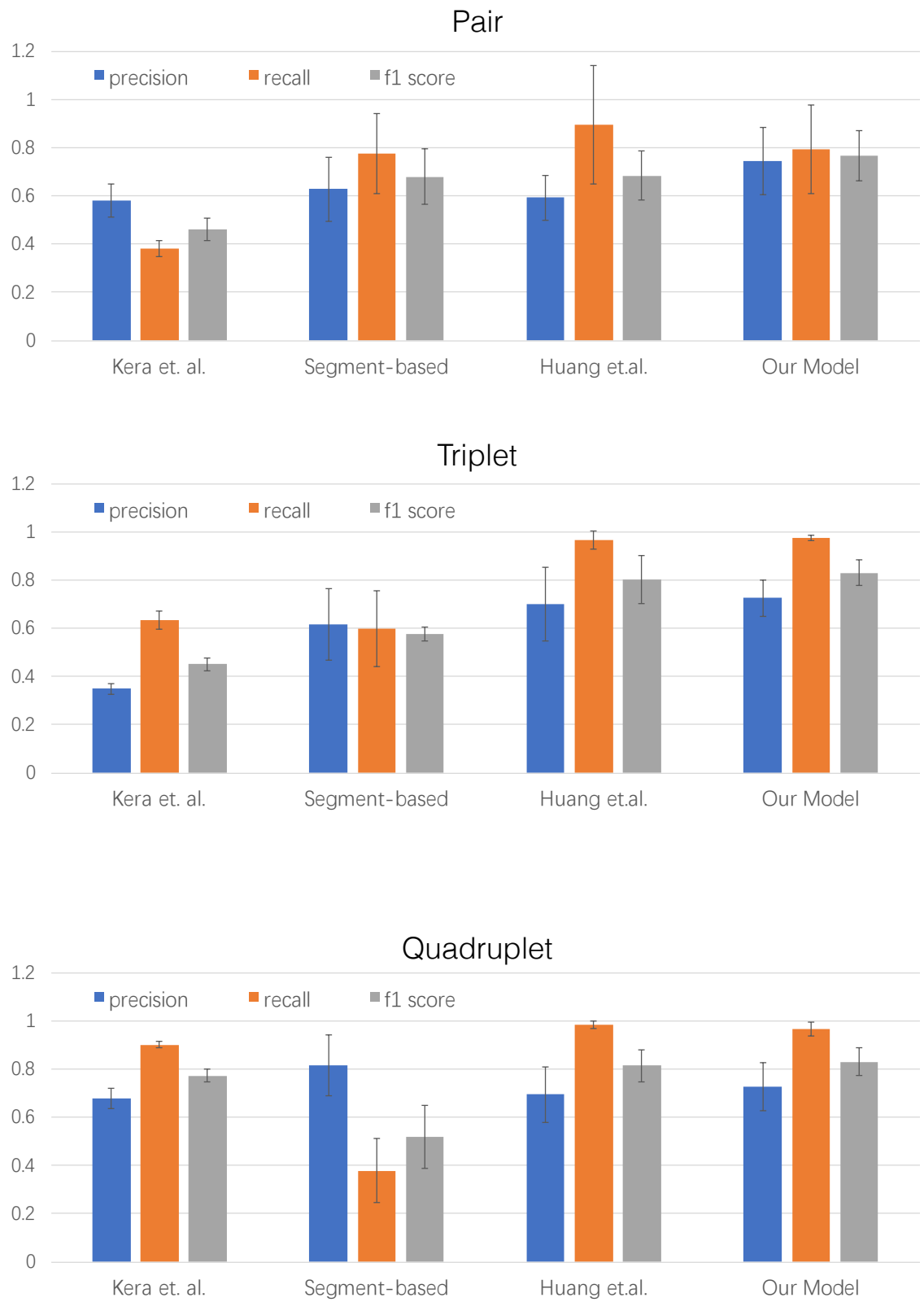

Figure 4.6.: Quantitative Comparisons on Temporal Localization Task: Precision-recall scores for different methods and different definitions of joint attention.

5. Discussion

5.1. Impact of noise in gaze measurements

Our dataset is collected using a relatively expensive wearable eye-tracking camera - Tobii Glass 2, which has more stable gaze measurement than other cheaper wearable eye-tracking devices. Such device is not suitable for large scale deployment. Although appropriate for larger scale deployment, a cheaper wearable eye-tracking camera may include larger noise in gaze measurement.

To have a better knowledge about the impact of noise in gaze measurements on the task performance, which may come from a cheaper eye-tracking device, calibration error, or a more complex interaction scene, we add additional synthetic Gaussian noises of different scales to the points-of-gaze data recorded by Tobii Glass 2 in our dataset. We then compare the performance of temporal localization task and spatial segmentation task between our method and several baselines: **Huang** *et al.* is our model without considering logic relationships, and **Segment based** is our simplified model that only uses gaze information, the first term of Eq. (3.3). We use the four person dataset in this comparison experiment.

We add Gaussian noises with zero means and different standard deviations ($\sigma = 2.5, 5, 7.5, 10, 12.5, 15$ pixels) to simulate different scales of gaze noises. Note that the reduced image resolution in process is 480×270 pixels, and the average distance between the measured gaze position and the nearest object border is 10.28 pixels. As noise scale increases, our method's performance degrade but still gets the best performance.

We also add Gaussian noises with different means ($\mu = [2.5, 2.5], [5, 5], ..., [15, 15]$) to simulate different scales of gaze bias, possibly caused by calibration error. Quantitative results are shown in Figure (5.1). With small gaze bias so that the mean of the noised gaze positions is within most of the attended objects (*i.e.*, $\mu \leq [10, 10]$), our method can still get a reliable performance. With larger gaze bias, our method can still outperform the two baselines.

According to Figure (5.1), our method consistently outperform the two baselines temporal localization task. We observe our method have a high recall score when noise scale or bias is larger than 10 pixels. Since gaze position get consistently out of object boundary in such circumstances, our method tend to look at the background and consider all the frames as joint attention. This indicates a weakness of out method, which we will discuss in the following section.

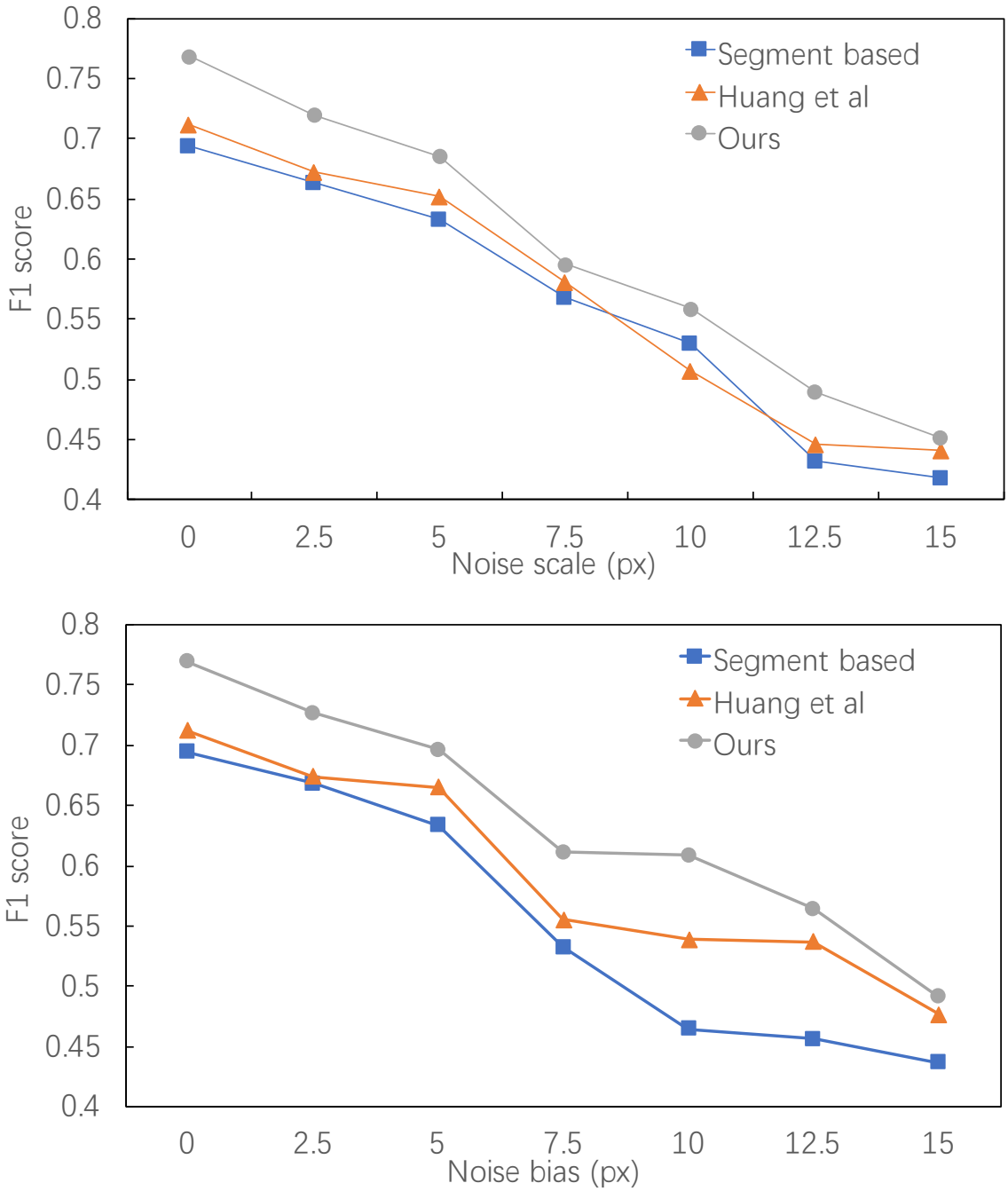

Figure 5.1.: Impact of gaze noise: F1 scores of temporal localization task with different types of noise added to gaze data. Noise scale is indicated by standard deviation of zero-mean Gaussian noises in pixels (px) with image resolution of 480×270. Noise bias is indicated by Gaussian mean in pixels with standard deviation $= 1$.

5.2. Limitation of appearance-based methods

While the proposed approach outperforms existing co-segmentation [FXZL14] and joint-attention discovery [KYHS16] methods, there are some limitations on our appearance-based approach. First, currently we can't segment objects with quite dissimilar appearances from different viewpoints. This limitation causes the failure in Figure 4.4 (b) and degrades the performance in the FtF conditions in Table 4.1. In addition, different objects with similar appearance, like the cards in Figure 4.4 (c), cannot be distinguished by our approach. Finally, our assumption about stable head pose during joint attention will not always hold for more challenging scenarios where people can move (*e.g.*, walking) during interactions. One possible solution to address these limitations is by making use of 3D geometric relationship of the people, though it requires costly computation for stable 3D reconstructions. We leave this for our future work.

6. Conclusion and Future work

In this thesis, we propose a new method for temporally localizing and spatially segmenting objects of joint attention in multiple first person videos recorded with gaze data. Since objects of joint 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. Since objects of joint attention also reflect group behaviour, the deeper research on discovering objects of joint attention will help the anthropological or psychological analysis of human group behaviours. The main challenges to be solved for this task is how to deal with the object size variability across objects, the noise in gaze measurement, and how to deal with different interpretations of one object.

The key idea of our approach presented in this thesis is to use object segment proposals together with intra and inter video consistency. The use of object segment proposals can address the challenge of object size variability, and the utilization of inter and intra video consistency tackles the noise in gaze measurement, and can produce a semantically consistent interpretation of objects. For the general cases with more than two persons, we furthermore add the logical consistency of joint attention. The two coupled tasks are solved together in a unified framework, which alternates temporal localization of joint attention and spatial segmentation of jointly attended objects. A new dataset is collected for evaluating the performance of different methods. Our experimental results include the experiment on two person cases and more general cases. The experiment results demonstrate that our approach is able to achieve state-of-the-art performance in both temporal localization task and spatial segmentation task. Since our dataset is collected using a relatively expensive equipment which is not always available for larger scale applications, we simulated the situations when a cheaper equipment is used to capture point-of-gaze data, by adding synthetic noise into the point-of-gaze data. Our extra experiment results demonstrate that our method works better than the baselines even when point-of-gaze data is noisy.

We also analyzed the limitations of our method and will tackle those in our future work. 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, also it is difficult to distinguish the situation where different objects share too similar visual features 6.1. Another limitation is the assumption of stable head movement, which is not always true in real social interactions. Currently, our

Figure 6.1.: Example failure cases produced by our method. (1) and (2) are different pieces of paper, but due to their visual similarity, they are predicted as object of joint attention by our method, even though they are not one object. (3) and (4) are the same set of cards. However, since their appearance appear too differently due to the two subject's view point difference, these cards are not predicted as object of joint attention by our method.

method only works well when the object is static so that the camera wearer would have stable head motion during fixation. However, if in the case where two persons are both looking at a moving car, our method will fail. Additionally, although the use of gaze data provided a significant cue for the object of attention, the gaze data itself is not always available.

Based on the analysis of limitations and 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 discussed in Section 5.2, incorporating non-appearance-based features will be helpful for temporal localizing objects of joint attention in more difficult cases, and in turn enhance the performance of spatial segmentation task. With geometric relationships (*E.g.*, where he/she is, which direction he/she is facing to) among people in a group, we can avoid matching different objects that share similar appearance that two persons are facing to different places. We can also avoid wasteful computation with such information when people are obviously looking at different directions.

Enabling more general motion situations In this thesis, we only consider the case where both people are looking at a static object as joint attention. However, as is discussed in Section 5.2, a more realistic setting should not limit the camera wearer's head motion. An obvious extension of this work is to enable more general motion situations. This requires the joint use of appearance feature and motion patterns. When two persons match motion patterns and share similar appearance features, it is possible that joint attention is established upon a moving object.

Using gaze estimation Although relatively accurate in gaze measurement, the Tobii Glasses 2 we used in this thesis is relatively expensive for daily use. An interesting extension of this work would then be detecting objects of joint attention without using measured gaze data, but using estimated gaze data. The use of

estimated gaze data would certainly affect the performance of the task, but would make this research more applicable.

Using object re-identification The problem of re-identification [ZYH16] is an emerging topic in recent years. Although most of the work only deals with person or pedestrians, the idea of re-identification is very suitable for our task of discovering objects of joint attention. During our experiment, we observed that during one specific task, there are certain objects that will attract joint attention multiple times. By adopting the idea of object re-identification, we may be able to find the most relevant object of joint attention, which may be of use for future studies on psychology and group behaviour.

Acknowledgments

Foremost, I would like to express my sincere gratitude to my advisor Prof. Yoichi Sato for the continuous support of my Master's study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor during my study.

Besides my advisor, I would like to thank the rest of my paper's co-authors: Minjie Cai, Ryo Yonetani, Hiroshi Kera and Keita Higuchi, for their encouragement, insightful comments and help.

My sincere thanks also goes to Dr. Bo Zheng, for offering me the summer internship opportunities in his group and leading me working on diverse exciting projects.

I thank all of my fellow labmates in Sato Lab, for the stimulating discussions, and for all the fun we have had in the last two years. My special thanks goes to Binhua Zuo, Dailin Li, Zhenqiang Li, Ya Wang and Jiehui Wang, for their collaboration of helping me prepare the dataset used in the paper.

Last but not the least, I would like to thank my family: my parents Shaoshan Huang and Zhihong Xie, for giving birth to me at the first place and supporting me spiritually throughout my life.

A. Additional Results

Our dataset is collected in three different indoor environments. For each environment, we use a diverse set of objects for joint attention. Figure (A.1) and Figure (A.2) show some additional qualitative results on different environments and different object sets. In all figures, green lines represent the ground truth temporal interval of joint attention. The broken parts of the green lines indicate there's no joint attention during such time. Red lines demonstrate the predicted temporal interval of joint attention status in a similar fashion. In each experiment we manually visualize some samples of spatial segmentation results produced by our method. The visualized samples include both successful cases and failure cases.

We also show some figures of three and four person cases. Figure (A.3) shows an example qualitative result on a three person case. In (a) of this figure, the three subjects are actually jointly attending the smart-phone placed on top of the laptop. However, Although our method successfully detected the temporal interval of joint attention, but the spatial segment is incorrect. As we have already discussed in chapter 1, the multiple interpretation of object is a big challenge here. We may think the laptop and the smart-phone as a same combined object, or we may think they are two separated objects. Our experimental result is consistent with our claim that we output a semantically consistent spatial segment of the object. Figure (A.5) and (A.4) show two examples of four person cases. In four person cases, since there may be multiple sub-groups jointly attending at different objects, we distinguish them by using slightly different colors. In (a) of Figure (A.5), P1 and P2 are not determined to have joint attention since P2 has large head motion, and is actually taking a glance at the can P1 is looking at.

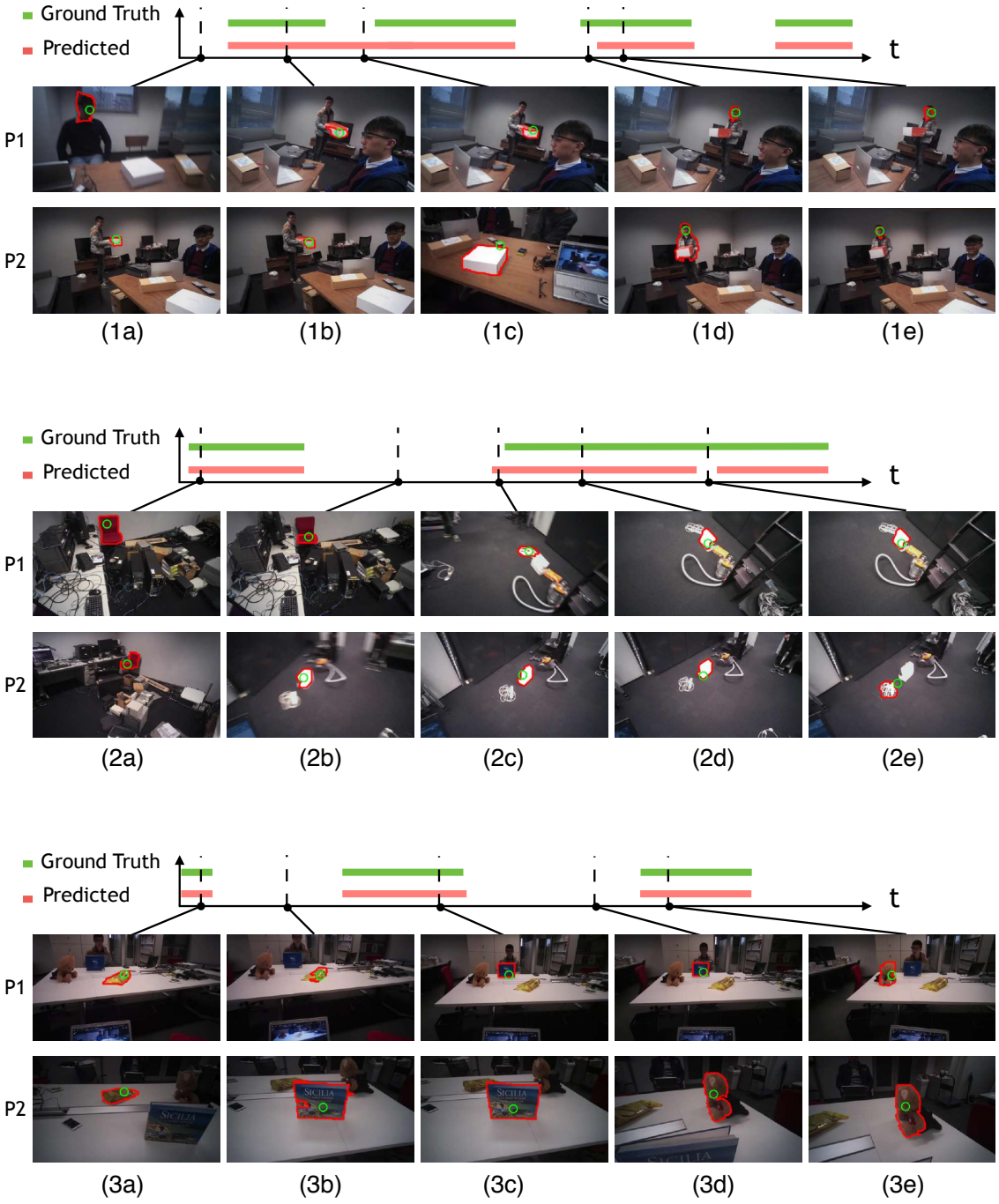

Figure A.1.: Qualitative results in different indoor environments.

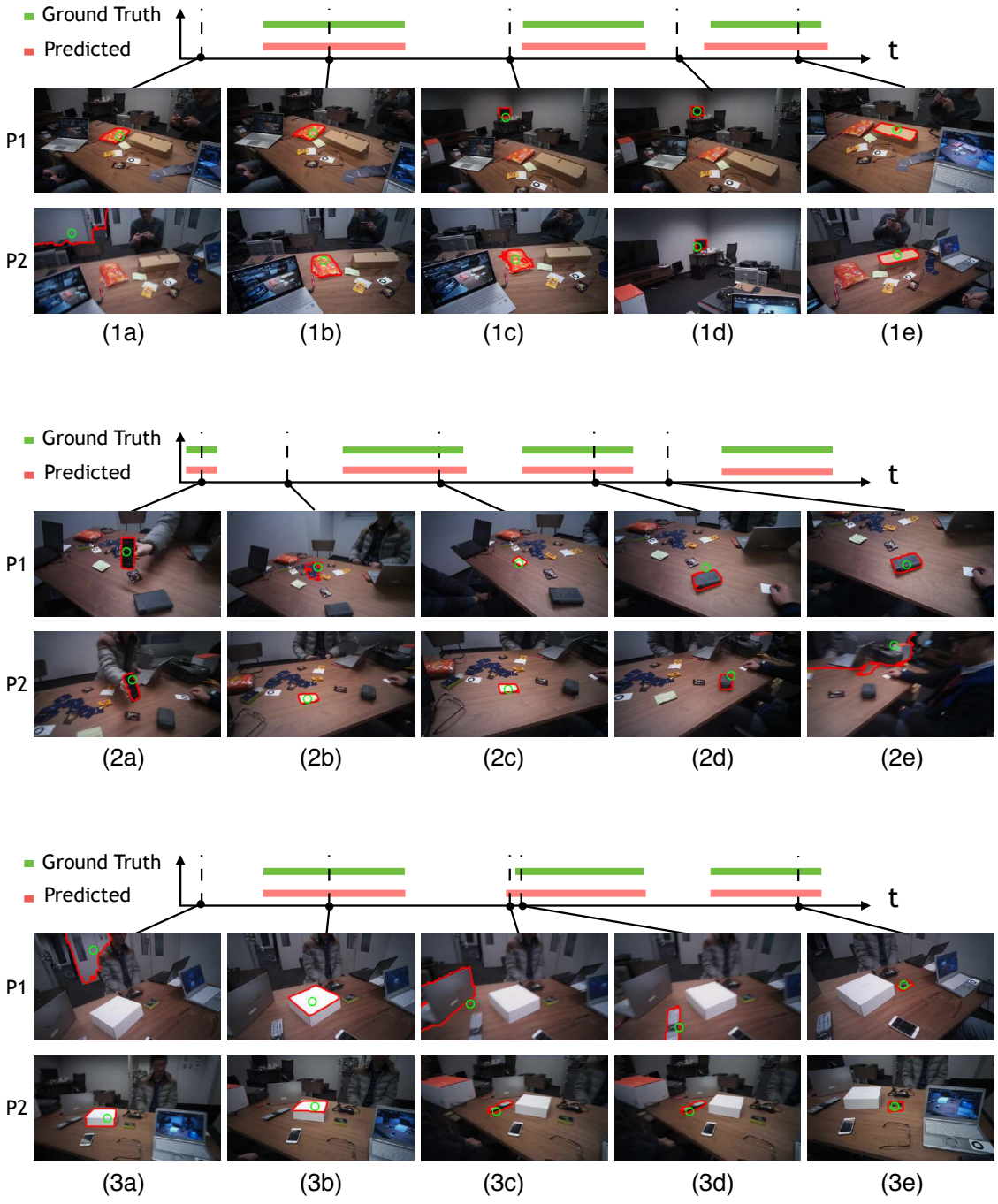

Figure A.2.: Qualitative results for different object sets in the same indoor environment.

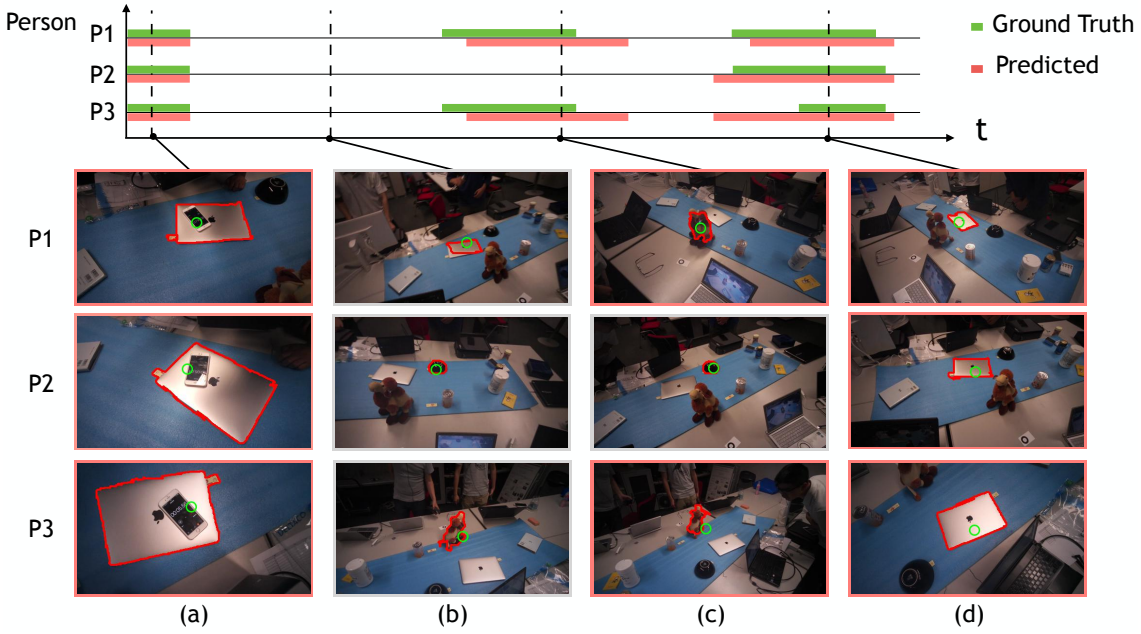

Figure A.3.: Additional qualitative results for three person cases.

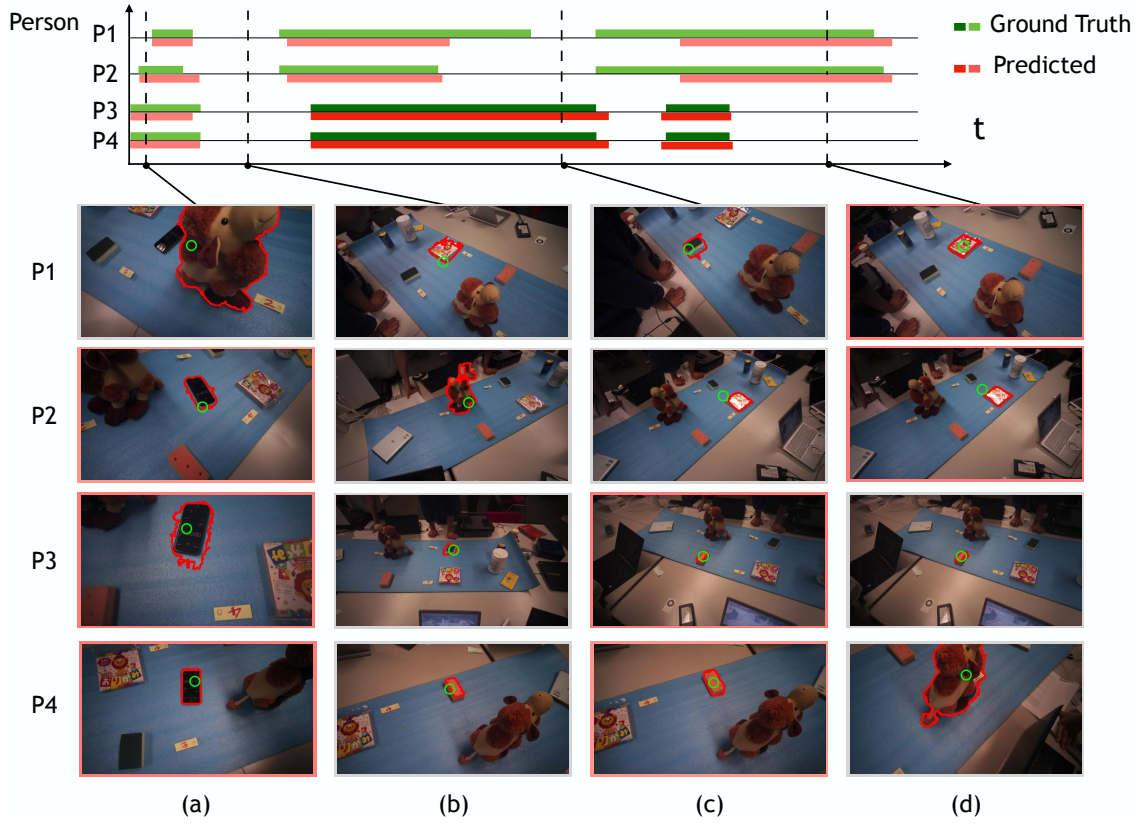

Figure A.4.: Additional qualitative results for four person cases.

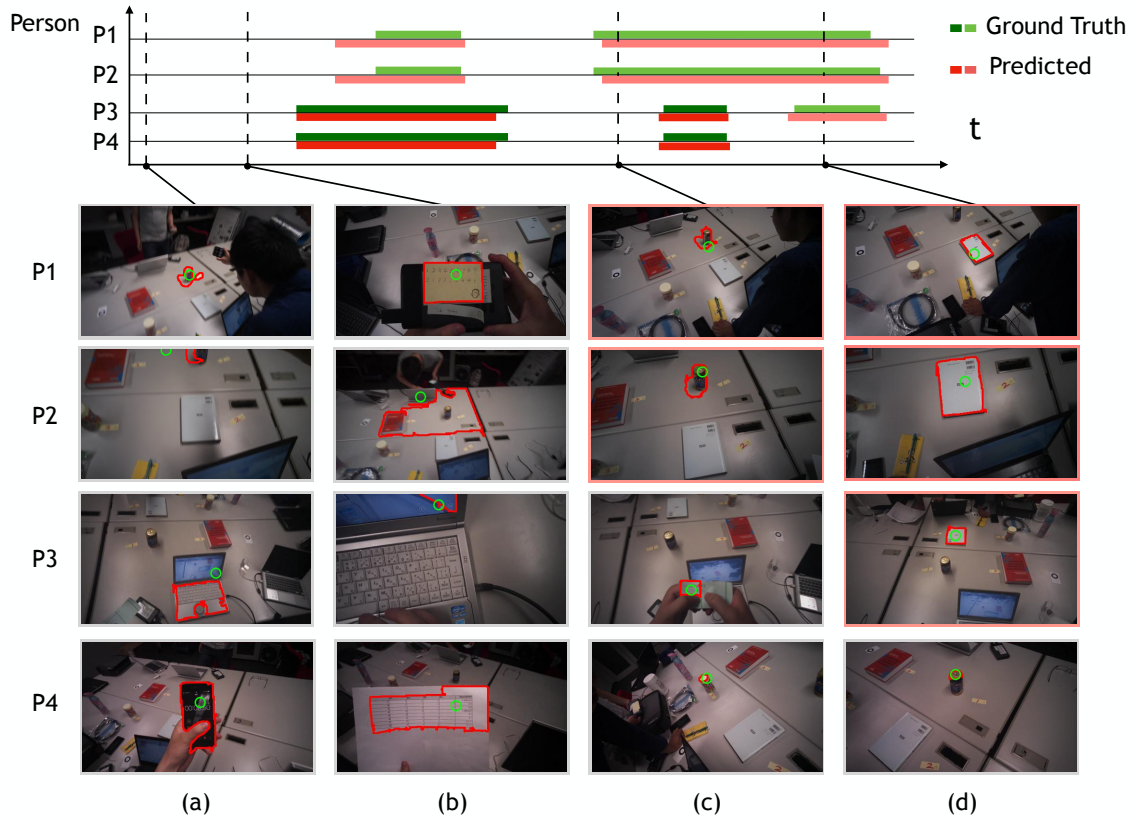

Figure A.5.: Additional qualitative results for four person cases. In case of different sub-group of joint attention appears, the temporal state of joint attention is denoted using 2 different colors as shown in the legends.

References

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

[BKP+10] Dhruv Batra, Adarsh Kowdle, Devi Parikh, Jiebo Luo, and Tsuhan Chen. icoseg: Interactive co-segmentation with intelligent scribble guidance. In *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, pages 3169–3176. IEEE, 2010.

[CF13] Wei-Chen Chiu and Mario Fritz. Multi-class video co-segmentation with a generative multi-video model. In *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, pages 321–328, 2013.

[CSBC+97] Tony Charman, John Swettenham, Simon Baron-Cohen, Antony Cox, Gillian Baird, and Auriol Drew. Infants with autism: an investigation of epathy, pretend play, joint attention, and imitation. *Developmental psychology*, 33(5):781, 1997.

[DNWZZ13] Jifeng Dai, Ying Nian Wu, Jie Zhou, and Song-Chun Zhu. Cosegmentation and cosketch by unsupervised learning. In *Proc. of IEEE Int. Conf. Computer Vision*, pages 1305–1312, 2013.

[EH10] Ian Endres and Derek Hoiem. Category independent object proposals. *Proc. of European Conf. Computer Vision*, pages 575–588, 2010.

[FCT13] Huazhu Fu, Xiaochun Cao, and Zhuowen Tu. Cluster-based co-saliency detection. *IEEE Transactions on Image Processing*, 22(10):3766–3778, 2013.

[FLR12] Alireza Fathi, Yin Li, and James M Rehg. Learning to recognize daily actions using gaze. In *Proc. of European Conf. Computer Vision*, pages 314–327. Springer, 2012.

[FRR11] Alireza Fathi, Xiaofeng Ren, and James M Rehg. Learning to recognize objects in egocentric activities. In *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, pages 3281–3288. IEEE, 2011.

[FXZL14] Huazhu Fu, Dong Xu, Bao Zhang, and Stephen Lin. Object-based multiple foreground video co-segmentation. In *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, pages 3166–3173, 2014.

[HAGM14] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In *Proc. of European Conf. Computer Vision*, pages 297–312. Springer, 2014.

[HCK+17] Yifei Huang, Minjie Cai, Hiroshi Kera, Ryo Yonetani, Keita Higuchi, and Yoichi Sato. Temporal localization and spatial segmentation of joint attention in multiple first-person videos. In *ICCV Workshop on Egocentric Perception, Interaction and Computing*, 2017.

[JBP10] Armand Joulin, Francis Bach, and Jean Ponce. Discriminative clustering for image co-segmentation. In *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, pages 1943–1950, 2010.

[JBP12] Armand Joulin, Francis Bach, and Jean Ponce. Multi-class cosegmentation. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 542–549, 2012.

[KASB17] Nour Karessli, Zeynep Akata, Bernt Schiele, and Andreas Bulling. Gaze embeddings for zero-shot image classification. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[KYHS16] Hiroshi Kera, Ryo Yonetani, Keita Higuchi, and Yoichi Sato. Discovering objects of joint attention via first-person sensing. In *Proc. of IEEE Workshop on Egocentric Vision*, pages 7–15, 2016.

[LH01] Michael F Land and Mary Hayhoe. In what ways do eye movements contribute to everyday activities? *Vision research*, 41(25):3559–3565, 2001.

[LRS+17] George Leifman, Dmitry Rudoy, Tristan Swedish, Eduardo Bayro-Corrochano, and Ramesh Raskar. Learning gaze transitions from depth to improve video saliency estimation. In *Proc. of IEEE Int. Conf. Computer Vision (ICCV)*, Oct 2017.

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

[MLLN12] Fanman Meng, Hongliang Li, Guanghui Liu, and King Ngi Ngan. Object co-segmentation based on shortest path algorithm and saliency model. *IEEE transactions on multimedia*, 14(5):1429–1441, 2012.

[MLQH17] Jizhou Ma, Shuai Li, Hong Qin, and Aimin Hao. Unsupervised multi-class co-segmentation via joint-cut over $l\_\{1\}$-manifold hyper-graph of discriminative image regions. *IEEE Transactions on Image Processing*, 26(3):1216–1230, 2017.

[MN07] Peter Mundy and Lisa Newell. Attention, joint attention, and social cognition. *Current directions in psychological science*, 16(5):269–274, 2007.

[PJS12] Hyun Soo Park, Eakta Jain, and Yaser Sheikh. 3d social saliency from head-mounted cameras. In *Proc. of Int. Conf. Neural Information Processing Systems (NIPS)*, pages 422–430, 2012.

[QHZN16] Rong Quan, Junwei Han, Dingwen Zhang, and Feiping Nie. Object co-segmentation via graph optimized-flexible manifold ranking. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 687–695, 2016.

[RKB04] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Trans. on Graphics*, 23(3):309–314, 2004.

[RKMB06] Carsten Rother, Vladimir Kolmogorov, Tom Minka, and Andrew Blake. Cosegmentation of image pairs by histogram matching - incorporating a global constraint into mrfs. In *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, pages 993–1000, 2006.

[RSLP12] J. C. Rubio, J. Serrat, A. Lopez, and N. Paragios. Unsupervsed co-segmentation through region matching. In *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, pages 749–756, 2012.

[SBH+13] Qi Song, Junjie Bai, Dongfeng Han, Sudershan Bhatia, Wenqing Sun, William Rockey, John E Bayouth, John M Buatti, and Xiaodong Wu. Optimal co-segmentation of tumor in pet-ct images with context information. *IEEE trans. medical imaging*, 32(9):1685–1697, 2013.

[See11] Axel Seemann. *Joint Attention: New Developments in Psychology, Philosophy of Mind, and Social Neuroscience*. MIT Press, 2011.

[SHSP17] Shan Su, Jung Pyo Hong, Jianbo Shi, and Hyun Soo Park. Predicting behaviors of basketball players from first person videos. In *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, volume 2, page 3, 2017.

[SM+12] Charles Sutton, Andrew McCallum, et al. An introduction to conditional random fields. *Foundations and Trends in Machine Learning*, 4(4):267–373, 2012.

[SPHSSP17] Shan Su, Jung Pyo Hong, Jianbo Shi, and Hyun Soo Park. Predicting behaviors of basketball players from first person videos. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[SPJS13] Hyun Soo Park, Eakta Jain, and Yaser Sheikh. Predicting primary gaze behavior using social saliency fields. In *Proc. of IEEE Int. Conf. Computer Vision*, pages 3503–3510, 2013.

[SPS15] Hyun Soo Park and Jianbo Shi. Social saliency prediction. In *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, pages 4777–4785, 2015.

[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 *Advances in Neural Information Processing Systems*, 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.

[TF86] Michael Tomasello and Michael Jeffrey Farrar. Joint attention and early language. *Child development*, pages 1454–1463, 1986.

[TFNR12] David Tsai, Matthew Flagg, Atsushi Nakazawa, and James M Rehg. Motion coherent tracking using multi-label mrf optimization. *International Journal of Computer Vision*, 100(2):190–202, 2012.

[TJLFF14] Kevin Tang, Armand Joulin, Li-Jia Li, and Li Fei-Fei. Co-localization in real-world images. In *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, June 2014.

[TSS16] Tatsunori Taniai, Sudipta N Sinha, and Yoichi Sato. Joint recovery of dense correspondence and cosegmentation in two images. In *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, pages 4246–4255, 2016.

[UvdSGS13] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013.

[WHS+16] Le Wang, Gang Hua, Rahul Sukthankar, Jianru Xue, Zhenxing Niu, and Nanning Zheng. Video object discovery and co-segmentation with extremely weak supervision. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2016.

[WSSS17] Wenguan Wang, Jianbing Shen, Hanqiu Sun, and Ling Shao. Vicos2: Video co-saliency guided co-segmentation. *IEEE Trans. on Circuits and Systems for Video Technology*, 2017.

[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*, pages 2235–2244, 2015.

[Yar67] Alfred L Yarbus. *Eye Movements during Perception of Complex Objects*. Springer, 1967.

[YCK+17] Youngjae Yu, Jongwook Choi, Yeonhwa Kim, Kyung Yoo, Sang-Hun Lee, and Gunhee Kim. Supervising neural attention models for video captioning by human gaze data. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[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*, pages 739–746, 2013.

[YRLS15] Yan Yan, Elisa Ricci, Gaowen Liu, and Nicu Sebe. Egocentric daily activity recognition via multitask clustering. *IEEE Trans. on Image Processing*, 24(10):2984–2995, 2015.

[ZDlTH13] Feng Zhou, Fernando De la Torre, and Jessica K Hodgins. Hierarchical aligned cluster analysis for temporal clustering of human motion. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35(3):582–596, 2013.

[ZJS13] Dong Zhang, Omar Javed, and Mubarak Shah. Video object segmentation through spatially accurate and temporally dense extraction of primary object regions. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 628–635, 2013.

[ZJS14] Dong Zhang, Omar Javed, and Mubarak Shah. Video object co-segmentation by regulated maximum weight cliques. In *Proc. of European Conf. Computer Vision*, pages 551–566. Springer, 2014.

[ZTMHL+17] Mengmi Zhang, Keng Teck Ma, Joo Hwee Lim, Qi Zhao, and Jiashi Feng. Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[ZYH16] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.

[ZZC12] Bao Zhang, Handong Zhao, and Xiaochun Cao. Video object segmentation with shortest path. In *Proc. of ACM Int. Conf. Multimedia*, pages 801–804, 2012.

List of Publications

1. Yifei Huang, Minjie Cai, Hiroshi Kera, Ryo Yonetani, Keita Higuchi, and Yoichi Sato, "Temporal Localization and Spatial Segmentation of Joint Attention in Multiple First-Person Videos, " In Proceedings of the IEEE International Conference on Computer Vision Workshop on Egocentric Perception, Interaction and Computing, pp.1-9, Oct. 2017
2. Yifei Huang, Minjie Cai, Hiroshi Kera, Ryo Yonetani, Keita Higuchi, and Yoichi Sato, "Temporal Localization and Spatial Segmentation of Joint Attention in Multiple First-Person Videos," In Extended Abstract of Meeting on Image Recognition and Understanding (MIRU), Aug. 2017