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

Yifei Huang, Minjie Cai, Hiroshi Kera, Ryo Yonetani, Keita Higuchi, Yoichi Sato
The University of Tokyo
Tokyo, Japan

Abstract

This work aims to develop a computer-vision technique for understanding objects jointly attended by a group of people during social interactions. As a key tool to discover such objects of joint attention, we rely on a collection of wearable eye-tracking cameras that provide a first-person video of interaction scenes and points-of-gaze data of interacting parties. 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 co-segmentation and joint attention discovery.

1. Introduction

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. The understanding of 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 [4]. Moreover, locations where a group of people jointly focus could also be used for automatic video summarization [1]. 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, 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

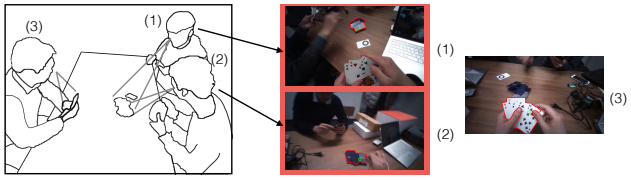

Figure 1. **Discovering Objects of Joint Attention**. Joint attention between persons (1) and (2) is detected (highlighted in red boundaries) from first-person videos recorded with points of gaze data (green circles in the video frames.)

[8][3][13], and *points of gaze* data indicating where the wearer looked at in the first-person videos (see Figure 1). The use of multiple cameras equipped by interaction parties is, therefore, promising for recording what they attended jointly during interactions [26].

One pioneering work along this line of research has been presented recently [11]. By comparing spatiotemporal patches around points of gaze based on their visual similarity, they can localize temporal intervals when joint attention occurred. However, their approach becomes problematic when 1) first-person videos capture cluttered scenes and / or 2) eye tracking is inaccurate, both of which often happen in recording natural social interactions. Under cluttered scenes, spatiotemporal patches around points of gaze may include not only objects being focused on but also surrounding objects or complex backgrounds, making visual features extracted from the patches irrelevant to the objects of focus. Moreover, noisy points-of-gaze data provided by inaccurate eye tracking do not necessarily correspond to where people actually attend. As a result, the straightforward comparison of spatiotemporal patches around points of gaze becomes unreliable.

To address these 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 hierarchical 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. 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 scenes are cluttered, and points of gaze are noisy.

Our main contributions are summarized as follows: firstly, to the best of our knowledge, this work is the first to both temporally localize and spatially segment joint attention. Secondly, we propose a hierarchical CRF that jointly solve the two tasks together. Thirdly, we introduce a new dataset of natural social interactions recorded with multiple wearable eye-trackers equipped by interaction parties, which includes annotations of temporal intervals and spatial segments of objects being looked at jointly. We will make this dataset publicly available.

1.1. Related Work

Co-segmentaion One of the popular computer vision topics closely relevant to our work is co-segmentation, and much work has been done recently [5, 22, 2, 6, 28, 27, 23]. 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.*, [10, 17]). Similar to our work, [9] 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, previous works only focus on foreground objects that are not necessarily the objects of attention. This prevents direct applications of existing co-segmentation methods and requires an additional cue to identify those objects. In this work, we utilize gaze information as an important cue for segmenting objects from multiple videos that are under human attention.

Joint attention estimation Another topic relevant to our work is joint attention estimation which is of great importance to social cognition [12, 18] and the research of autism [4]. Park *et al*. proposed methods [14, 20, 19] to estimate social saliency by modeling human viewpoint as a 3D cone and use the intersections to construct social saliency fields. These methods are however designed to detect intersections of fields-of-view of multiple wearable cameras, which do not necessarily correspond to objects of joint attention. The most relevant work is Kera *et al*. [11] which tried to temporally localize joint attention by comparing commonalities of image appearance around gaze positions from multiple videos. However, their method didn't consider the spatial segmentation of the attended object, which in turn weakened the performance in cluttered scenes. In our method, we temporally localize joint attention of multiple people with spatial segmentation of their attended objects.

2. Proposed Model

Given a collection of pairs of first-person videos and points-of-gaze data recorded in synchronization by multiple people in interactions, we aim to 1) temporally localize when joint attention occurs and 2) spatially segment object instances that people jointly attended to. As stated earlier, we observe that accurate object segmentation guides temporal localization of joint attention by evaluating the visual similarity of segments being looked at, and the prior knowledge about when joint attention occurs will act as a salient cue for segmenting object regions being looked at jointly in each video. These observations motivate us to develop a framework to solve these two tasks alternately. Specifically, we propose a new model based on the conditional random field (CRF) that estimates temporal intervals and spatial segments of jointly-attended objects via alternating optimization.

2.1. Model Architecture

Our model bases a hierarchical CRF that comprises several linear-chain CRFs as a sub-module. Figure 2 (a) depicts the overall architecture. We exclusively consider a simple case where we have the only two sub-module CRFs for modeling joint attention of two persons. We will extend our model later for 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 as selective search [24] 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 2 (b)). We regard $s_t^{(p)}$ as a latent variable as noisy points of gaze are not nec-

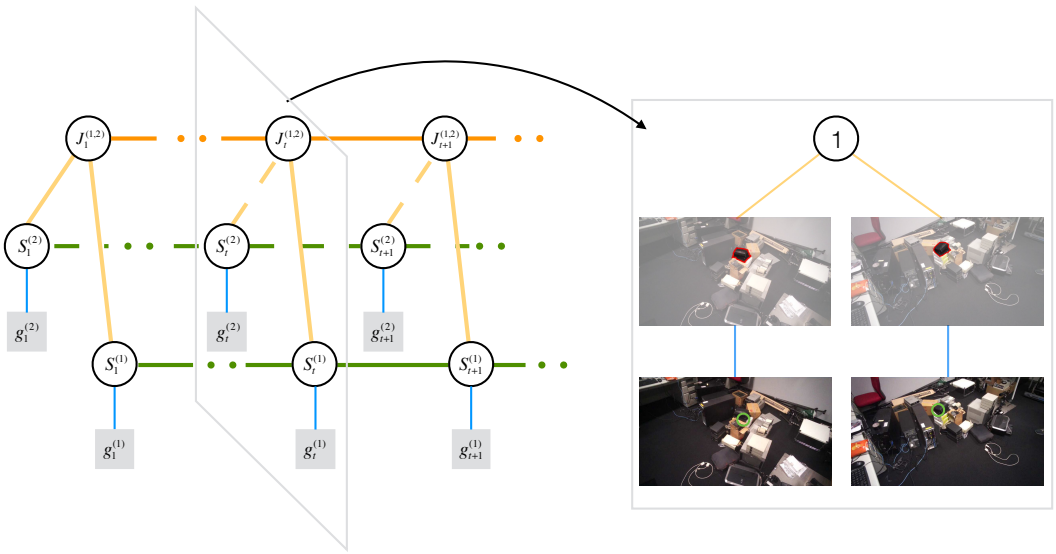

Figure 2. **Proposed Hierarchical CRF Model** for discovering joint attention of two persons. The model accepts points of gaze $g_t^{(p)}$ as the input (green circles in (b), $p \in \{1, 2\}$) and estimate segments $s_t^{(p)}$ being looked at (red boundaries in (b)) as well as binary state j_t indicating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

essarily 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 2 (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 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:

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

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

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. (1) is then modified as follows:

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

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

2.2. 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 [11] 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.

Gaze proximity and objectness 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:

$$\Psi_{\mathrm{GO}}(S^{(p)} \mid G^{(p)}) = \sum_{t=1}^{T} \left(\lambda_{\mathrm{GO1}} \frac{\|C(s_t^{(p)}) - \boldsymbol{g}_t^{(p)}\|_2}{|s_t^{(p)}|^{\frac{1}{2}}} + \lambda_{\mathrm{GO2}} \left(1 - \frac{|s_t^{(p)}|}{|H(s_t^{(p)})|}\right) \right), \tag{3}$$

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 2.4.

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. We, therefore, consider the temporal consistency of segments in Ψ_{TS}. This is measured by the visual similarity of consecutive segments as follows:

$$\Psi_{\mathrm{TS}}(S^{(p)}) = \lambda_{\mathrm{TS}} \sum_{t=1}^{T-1} \left(1 - f_{\mathrm{sim}}\left(s_t^{(p)}, s_{t+1}^{(p)}\right)\right), \tag{4}$$

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 2.4. 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.

Joint attentionness Similar to [11], 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. These two assumptions are implemented in Ψ_{JA} in the following fashion:

$$\Psi_{\mathrm{JA}}(J, S^{(1)}, S^{(2)} \mid G^{(1)}, G^{(2)}) = \sum_{t=1}^{T} \left(\lambda_{\mathrm{JA1}} Y(j_t, s_t^{(1)}, s_t^{(2)}, \boldsymbol{g}_t^{(1)}, \boldsymbol{g}_t^{(2)}) + \lambda_{\mathrm{JA2}} Z(j_t)\right), \tag{5}$$

where $\lambda_{\mathrm{JA1}}, \lambda_{\mathrm{JA2}}$ are two 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 $s_t^{(2)}$:

$$Y(j_t, s_t^{(1)}, s_t^{(2)}, \boldsymbol{g}_t^{(1)}, \boldsymbol{g}_t^{(2)}) = j_t(1 - f_{\mathrm{sim}}(s_t^{(1)}, s_t^{(2)})) + (1 - j_t)\alpha(\boldsymbol{g}_t^{(1)}, \boldsymbol{g}_t^{(2)}) \tag{6}$$

where f_{sim} is given by the cosine similarity between two segments across videos as in Eq. (4). The first term in Eq. (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 2.4.

$Z(j_t)$ in the second term of Eq. (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.

Temporal consistency of joint attention Finally, we observe that joint attention typically continues for a certain time. This motivates us to introduce another temporal consistency term Ψ_{TJ} on joint attention variables J as follows:

$$\Psi_{\mathrm{TJ}}(J) = \lambda_{\mathrm{TJ}} \sum_{t=1}^{T-1} |j_t - j_{t+1}|, \tag{7}$$

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

2.3. Model inference

Here we describe the model inference for the two-person case for simplicity of description. Minimizing Eq. (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. 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 [21]:

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:

$$S^{(1)*}, S^{(2)*} = \underset{S^{(1)}, S^{(2)}}{\arg\min} \sum_{p \in \{1,2\}} \Psi_{\mathrm{GO}}(S^{(p)} \mid G^{(p)}) + \sum_{p \in \{1,2\}} \Psi_{\mathrm{TS}}(S^{(p)}) \tag{8}$$

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:

$$J^* = \arg\min_{J} \Psi_{\mathrm{JA}}(J \mid S^{(1)}, S^{(2)}, G^{(1)}, G^{(2)}) + \Psi_{\mathrm{TJ}}(J) \tag{9}$$

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.

$$\begin{aligned} S^{(1)*}, S^{(2)*} = &\arg\min_{S^{(1)}, S^{(2)}} \sum_{p\in\{1,2\}} \Psi_{\mathrm{GO}}(S^{(p)} \mid G^{(p)}) \\ &+ \sum_{p\in\{1,2\}} \Psi_{\mathrm{TS}}(S^{(p)}) + \Psi_{\mathrm{JA}}(S^{(1)}, S^{(2)} \mid J) \end{aligned} \tag{10}$$

As summarized in Algorithm 1, 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.

Algorithm 1: Alternative inference algorithm

Result: Optimized $S^{(1)}, S^{(2)}$ and J
Initialize segmentation $S^{(1)}$ and $S^{(2)}$ using Eq. (8) ;
while *Change rate* $\geq \xi$ **do**
  Optimize J by fixing $S^{(1)}, S^{(2)}$ using Eq. (9);
  Optimize $S^{(1)}, S^{(2)}$ by fixing J using Eq. (10);
  Estimate *Change rate* of J;
end

2.4. Implementation Details

We generate region proposals $R_t^{(p)}$ by Selective Search [24] per frame, using region masks instead of bounding boxes. We used "single strategy" as stated in [24] for speed. For efficient inference, we perform pre-processing to filter out most region proposals that are probably irrelevant to the objects of attention. We compute a score for each region proposal based on Eq. (3), and keep only 16 ones with highest scores as the final region candidates. As for the region features extracted for comparing visual similarity, we first compute 144-dimensional Local Intensity Order Pattern (LIOP) [25] descriptors in a 5×5 grid and then pool them spatially by the Fisher Vector [15] with a 64-component Gaussian Mixture Model (GMM) learned from randomly sampled descriptors. We then concatenate HSV color histogram discretized into 16 bins for each color channel independently (*i.e.*, 48-dimensional features). To compute global motion of videos we use the Lucas-Kanade method, and set the threshold to $\delta_m < 1.5$. The weight parameters are set to $\{\lambda_{\mathrm{GO1}}, \lambda_{\mathrm{GO2}}, \lambda_{\mathrm{TS}}, \lambda_{\mathrm{JA1}}, \lambda_{\mathrm{JA2}}, \lambda_{\mathrm{TJ}}\} \leftarrow \{1, 1.5, 2, 2, 10, 0.25\}$ empirically. The change rate threshold ξ for the optimization is set to 0.02. For $\alpha(\boldsymbol{g}_t^{(1)}, \boldsymbol{g}_t^{(2)})$, we computed the similarity of circular regions around points of gaze at multiple scales (15, 25, 50 pixel-radius) similar to [11] and gave its maximum similarity.

3. 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.

3.1. Experimental Setting

Following [11], we mainly address the cases where two persons in interactions establish joint attention under several different formations. For each of the formations **side-by-side (SbS)** and **face-to-face (FtF)** originally presented in [11], we further divide it into two different scenarios where people pay attention to objects with large head motion or small one. For one 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. For the other scenario, objects are placed close to each other, which requires only a slight shift in attention with little head motion. As a result, we evaluate the methods for four different recording conditions in total: **SbS-large**, **SbS-small**, **FtF-large**, **FtF-small**.

24 pairs of first-person videos and points-of-gaze data were recorded in total. Each participant was equipped with a Tobii Pro Glass 2 that was calibrated and manually synchronized for each recording. Videos were recorded at 25-fps with the resolution of 1920×1080. Ground-truth labels for temporal localization were annotated by manual inspections. Then we used GrabCut [16] 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. Model parameters were chosen via grid search on a separate set of data (four scenarios from one pair of subjects), and the method was evaluated with the rest of the data.

3.2. Jointly-Attended Object Segmentation Task

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

Method	FtF-large	FtF-small	SbS-large	SbS-small	Avg.
ObMiC [9]	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 1. **Quantitative Comparisons on Segmentation Task**: Intersection-over-union (IoU) for four different recording conditions of two persons.

ObMiC [9]. This method is one of the most relevant methods 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. It should be noted that [9] used a different method [7] to generate object proposals, and since they are highly coupled, we could not substitute [7] with Selective Search used in our method. However, as stated in [24] (section 5.2.2), the segmentation performance of Selective Search is even slightly worse than [7]. Therefore our method do not obtain extra advantage against [9] by using different object proposals.

Baseline1. In order to see how 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. (1).

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 1. The proposed model clearly outperforms ObMiC [9] that did not use gaze information. The proposed model also performs consistently better than Baseline1 and Baseline2, indicating the necessity of temporal consistency cue Ψ_{TS} and joint attention cues $\Psi_{\mathrm{JA}}, \Psi_{\mathrm{TJ}}$. 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.

Figure 4 shows some qualitative results. 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), while our method can perform a stable segmentation thanks to the cues of joint attention.

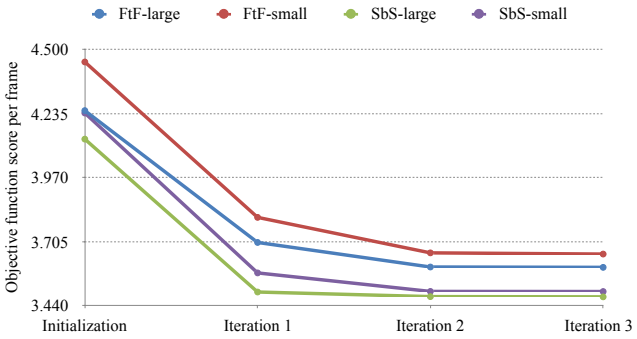

Figure 3. Per-frame objective function score at each iteration. Note that not all video pairs enter Iteration 3, and the scores of those which terminate at Iteration 2 are treated as static in Iteration 3.

More importantly, we observe in the experiments that the per-frame score of objective function monotonically decreases at each step of iteration (see Figure 3), which validates our claim that accurate segmentation guides accurate temporal localization, and vice versa.

3.3. Temporal Localization Task

Next, we address the task of temporal localization of joint attention. Here we compare our approach against [11] which is the only relevant work for the same task to the best of our knowledge. As shown in Table 2, the baseline method [11] is prone to obtain higher recall/lower precision scores, indicating 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. Most importantly, we observe in the experiments that the cost function of the proposed method in Eq. (1) monotonically decreases at each step during the alternative minimization. This result shows that better segmentation guides better temporal localization, and vice versa.

We also collected three-persons interaction data to evaluate the extended version of our model presented in Section 2.1. Specifically, three participants were asked to sit in triangle formation around a table, as shown in Figure 1, to play a card game. Figure 5 depicts qualitative results. We

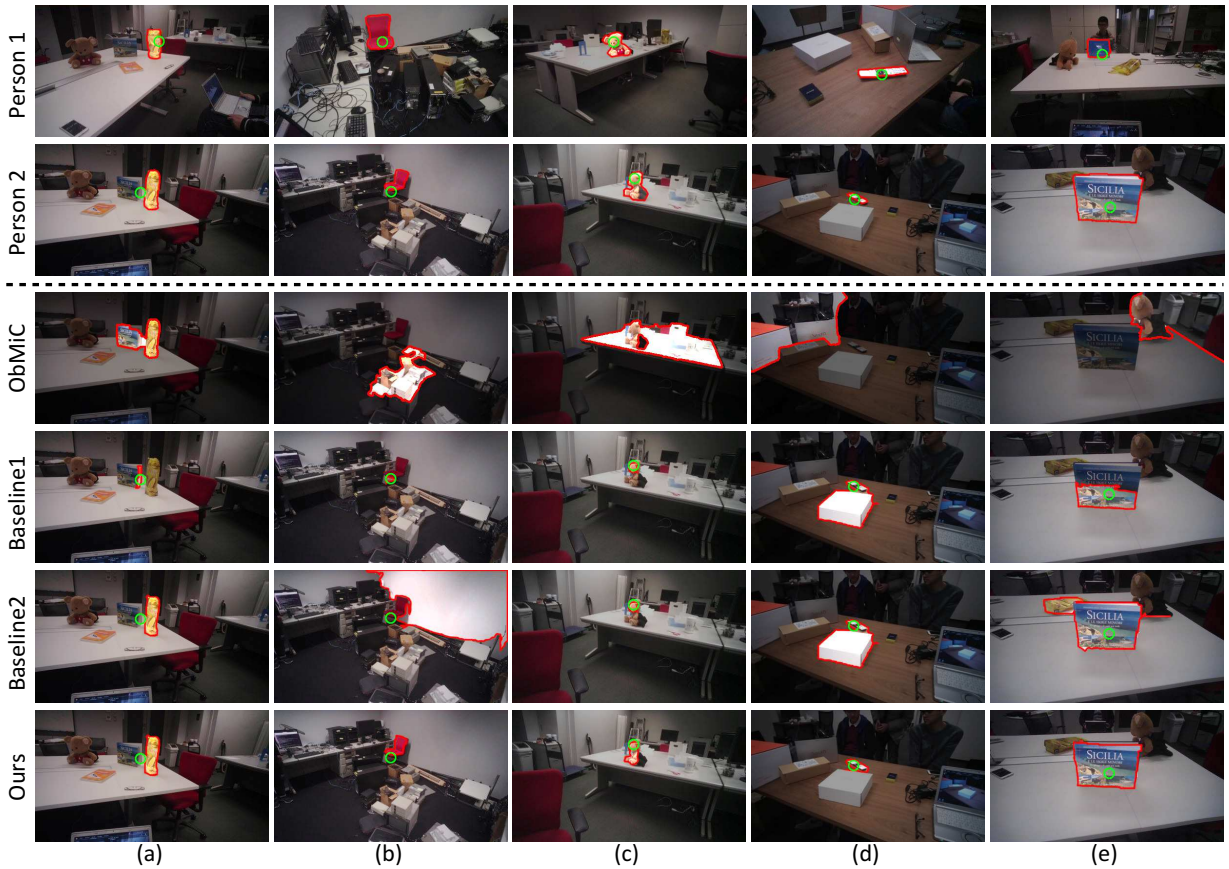

Figure 4. **Segmenting Objects of Joint Attention: Examples**. Red boundaries indicate jointly-attended object segments and green circles describe points of gaze. The first two rows describe the ground truth segments for the two input videos. The remaining rows show segmentation results in the second video.

Method	FtF-large (%)		FtF-small (%)		SbS-large (%)		SbS-small (%)		Avg. (%)
	P	R	P	R	P	R	P	R	F1 score
Kera *et al*. [11]	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 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.

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 and false positive results are found in (b) and (c), respectively. These failure cases imply some potential limitations of our approach, which we will discuss in the next section.

3.4. Limitations

While the proposed approach outperforms existing co-segmentation [9] and joint-attention discovery [11] 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 5 (b) and degrades the performance in the FtF conditions in Table 1. In addition, different objects with similar appearance, like the cards in Figure 5 (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. For instance, object regions near the intersection of two viewing directions are more likely to correspond to an object of joint attention. We leave this for our future work.

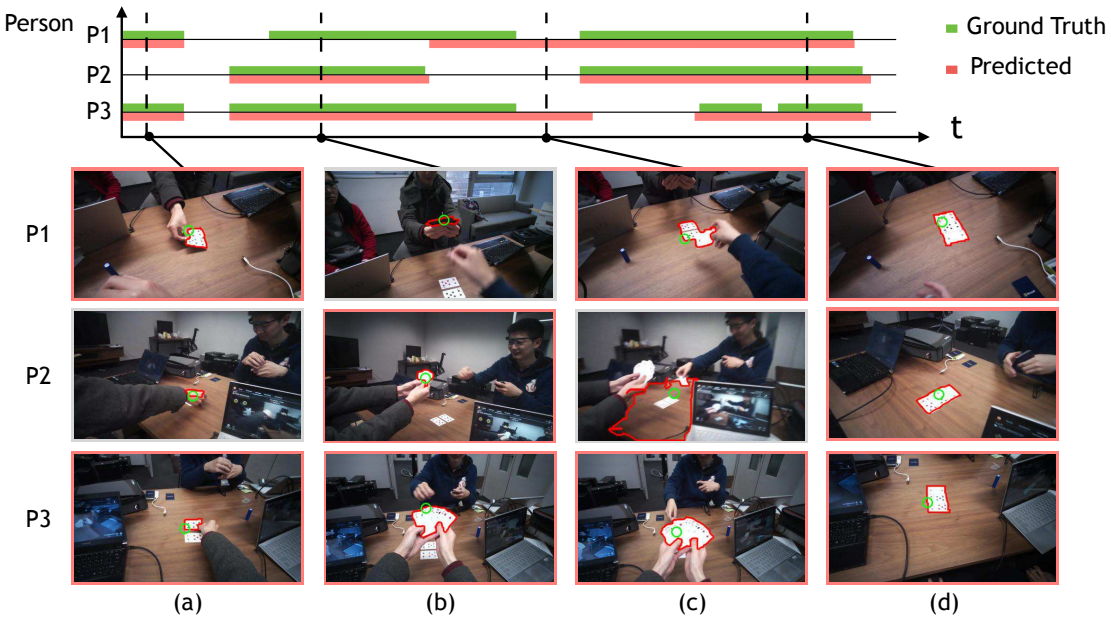

Figure 5. **Joint Attention Discovery for Three Persons Case**. The top half shows the ground truth and predicted results of temporal localization. The bottom half depicts some segmentation results in pink boundaries and points of gaze in green circles. Images highlighted in pink borders are judged as joint attention periods by the proposed model.

4. Conclusions

We proposed a new method for temporally localizing and spatially segmenting objects of joint attention in multiple first person videos recorded with gaze data. 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. Experimental results demonstrate that our approach is able to achieve state-of-the-art performance in both tasks.

In future work, we plan to use the predicted points of gaze instead of the measured gaze data in testing. This would make our method more practical in real world scenarios, as high-performance eye-trackers are expensive and not always available with respect to single wearable cameras. Another interesting extension of our future work is to take advantage of deep learning techniques, which have been proved to achieve substantial higher performance compared with traditional methods in numerous computer vision tasks.

Acknowledgments

This work was supported by JST CREST Grant Number JPMJCR14E1, Japan. We thank Binhua Zuo, Zhenqiang Li, Dailin Li, Ya Wang and Jiehui Wang for helping to collect and annotate our joint attention dataset.

References

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

[2] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen. icoseg: Interactive co-segmentation with intelligent scribble guidance. In *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, pages 3169–3176. IEEE, 2010.

[3] M. Cai, K. M. Kitani, and Y. Sato. An ego-vision system for hand grasp analysis. *IEEE Transactions on Human-Machine Systems*, 37(4):524–535, 2017.

[4] T. Charman, J. Swettenham, S. Baron-Cohen, A. Cox, G. Baird, and A. Drew. Infants with autism: an investigation of epathy, pretend play, joint attention, and imitation. *Developmental psychology*, 33(5):781, 1997.

[5] W.-C. Chiu and M. 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.

[6] J. Dai, Y. Nian Wu, J. Zhou, and S.-C. Zhu. Cosegmentation and cosketch by unsupervised learning. In *Proc. of IEEE Int. Conf. Computer Vision*, pages 1305–1312, 2013.

[7] I. Endres and D. Hoiem. Category independent object proposals. In *Proc. of European Conf. Computer Vision*, pages 575–588, 2010.

[8] A. Fathi, X. Ren, and J. M. Rehg. Learning to recognize objects in egocentric activities. In *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, pages 3281–3288. IEEE, 2011.

[9] H. Fu, D. Xu, B. Zhang, and S. Lin. Object-based multiple foreground video co-segmentation. In *Proc. of IEEE Conf.*

Computer Vision and Pattern Recognition, pages 3166–3173, 2014.

[10] A. Joulin, F. Bach, and J. Ponce. Discriminative clustering for image co-segmentation. In *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, pages 1943–1950, 2010.

[11] H. Kera, R. Yonetani, K. Higuchi, and Y. Sato. Discovering objects of joint attention via first-person sensing. In *Proc. of IEEE Workshop on Egocentric Vision*, pages 7–15, 2016.

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

[13] R. Nicholas and K. M. Kitani. First-person forecasting with online inverse reinforcement learning. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[14] H. S. Park, E. Jain, and Y. Sheikh. 3d social saliency from head-mounted cameras. In *Proc. of Int. Conf. Neural Information Processing Systems (NIPS)*, pages 422–430, 2012.

[15] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *Proc. of European Conf. Computer Vision*, pages 143–156, 2010.

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

[17] 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.

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

[19] H. Soo Park, E. Jain, and Y. Sheikh. Predicting primary gaze behavior using social saliency fields. In *Proc. of IEEE Int. Conf. Computer Vision*, pages 3503–3510, 2013.

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

[21] C. Sutton, A. McCallum, et al. An introduction to conditional random fields. *Foundations and Trends in Machine Learning*, 4(4):267–373, 2012.

[22] K. Tang, A. Joulin, L.-J. Li, and L. Fei-Fei. Co-localization in real-world images. In *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, June 2014.

[23] D. Tsai, M. Flagg, A. Nakazawa, and J. M. Rehg. Motion coherent tracking using multi-label mrf optimization. *International Journal of Computer Vision*, 100(2):190–202, 2012.

[24] 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.

[25] Z. Wang, B. Fan, G. Wang, and F. Wu. Exploring local and overall ordinal information for robust feature description. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 38(11):2198–2211, 2016.

[26] R. Yonetani, K. M. Kitani, and Y. Sato. Recognizing micro-actions and reactions from paired egocentric videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2629–2638, 2016.

[27] B. Zhang, H. Zhao, and X. Cao. Video object segmentation with shortest path. In *Proc. of ACM Int. Conf. Multimedia*, pages 801–804, 2012.

[28] F. Zhou, F. De la Torre, and J. 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.