Complementary-View Co-Interest Person Detection

Ruize Han*
Jiewen Zhao*
College of Intelligence and Computing, Tianjin University
{han_ruize,zhaojw}@tju.edu.cn

Wei Feng†
College of Intelligence and Computing, Tianjin University
wfeng@tju.edu.cn

Yiyang Gan
College of Intelligence and Computing, Tianjin University
realgump@tju.edu.cn

Liang Wan
College of Intelligence and Computing, Tianjin University
lwan@tju.edu.cn

Song Wang†
Tianjin University
University of South Carolina
songwang@cec.sc.edu

ABSTRACT

Fast and accurate identification of the co-interest persons, who draw joint interest of the surrounding people, plays an important role in social scene understanding and surveillance. Previous study mainly focuses on detecting co-interest persons from a single-view video. In this paper, we study a much more realistic and challenging problem, namely co-interest person (CIP) detection from multiple temporally-synchronized videos taken by the complementary and time-varying views. Specifically, we use a top-view camera, mounted on a flying drone at a high altitude to obtain a global view of the whole scene and all subjects on the ground, and multiple horizontal-view cameras, worn by selected subjects, to obtain a local view of their nearby persons and environment details. We present an efficient top- and horizontal-view data fusion strategy to map multiple horizontal views into the global top view. We then propose a spatial-temporal CIP potential energy function that jointly considers both intra-frame confidence and inter-frame consistency, thus leading to an effective Conditional Random Field (CRF) formulation. We also construct a complementary-view video dataset, which provides a benchmark for the study of multi-view co-interest person detection. Extensive experiments validate the effectiveness and superiority of the proposed method.

CCS CONCEPTS

• **Computing methodologies → Scene understanding**.

KEYWORDS

co-interest person; top view; horizontal view; multi-camera; video surveillance

ACM Reference Format:
Ruize Han, Jiewen Zhao, Wei Feng, Yiyang Gan, Liang Wan, and Song Wang. 2020. Complementary-View Co-Interest Person Detection. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413659

1 INTRODUCTION

Accurate detection of *co-interest person(s) (CIPs)*, i.e., the person(s) that draw interest of multiple surrounding people to look at him/her in a scene, has many important applications in video surveillance. For example, co-interest persons usually play central roles in the ongoing group activities and accurate CIP detections can facilitate the video-based group activity analysis. In many cases, persons with inappropriate behaviors are more likely to draw attention of many other people and the detection of such CIPs can help video-based anomaly detection, which is at the core of many video surveillance tasks. In this paper, our goal is to develop new camera settings and video-analysis algorithms for better CIP detection.

Using fixed cameras for video collection suffer from the problem of limited coverage and pre-specified view angle. While the use of moving cameras, especially wearable cameras, can extend the coverage along with the moving of the camera wearers, they usually capture the people, which we refer to as subjects in this paper, from a horizontal view angle. As illustrated in Figure 1(a-b)[1], in a horizontal view, the limited field of view (FOV) and lack of depth information prevent the reliable judgment of one subject looking at another, which is the foundation of CIP detection. There is also a self conflict in setting the distance (or zoom) of the camera: if the camera is too close, as shown in Figure 1(a-b), some subjects of interest in the scene may not be seen because of the mutual occlusions of subjects or the limited FOV; if the camera is too far, the covered subjects is too small, it will be difficult to identify them and their look-at directions. In either case, the accuracy of CIP detection will be compromised.

*Both authors contributed equally to this research.
†Co-corresponding authors.
Authors are also with the Key Research Center for Surface Monitoring and Analysis of Cultural Relics, SACH, China.

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413659

[1]The two example images of Figure 1(a) and (b) are drawn from the dataset proposed in [11] and [23], respectively.

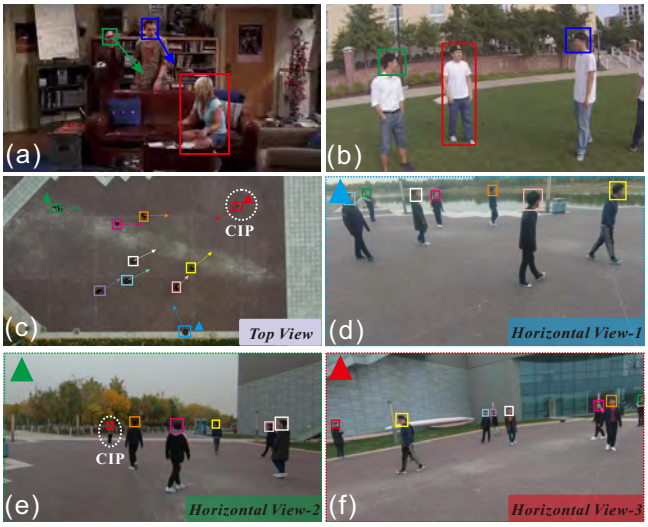

Figure 1: An illustration of CIP detection from (a) single-view video, (b) an egocentric-view video, and (c-f) the proposed combination of a top-view video and multiple (three here) horizontal-view videos. CIPs are denoted by red bounding boxes. As shown in (c-f), we use *multiple* horizontal-view cameras to help alleviate the problem of limited FOV of a single camera, which *may not* 1) cover the CIP at all time (d), 2) provide a good view to capture some subjects' faces (e), and handle the case where camera-wearer himself/herself is the CIP (f). We use a top-view camera (c) to provide a global picture of all or most of the subjects without occlusions.

In this paper, we propose a new camera setting to address this dilemma. We use two types of moving cameras with synchronized clock: Multiple horizontal-view cameras on the ground, e.g., wearable cameras mounted to the head of several involved subjects, and a top-view camera at a high altitude, e.g, a camera mounted to a flying drone. For practicability, fortunately, the above setting *becomes more and more available due to the spread of drones and various wearable cameras*, e.g., GoPro, Google glass. Also, the setting of combining top and horizontal-view cameras for facilitating collaborative video analysis has been researched in many latest works [1–3, 16, 17]. More importantly, this setting can be applied to many different real-world scenarios. For example, to maintain the security in an outdoor social event, we can select several law-enforcement officials as horizontal-view camera wearers on the ground. For fine-grained sport scene analysis of a football game, we can select referees as the horizontal-view camera wearers. In these scenarios, *the collected top- and horizontal-view videos can well complement each other for CIP detection*: The top view provides a global birds-eye view of the subjects (including the camera wearers) on the ground with large FOV and no subject mutual occlusions, as shown in Figure 1(c); the horizontal views capture the detailed appearance of the subjects in their limited FOVs, as shown in Figure 1(d-f). By combining these complementary-view videos, we focus on developing new algorithms to improve the state-of-the-art of CIP detection. To achieve this goal, we need to address several challenging problems: 1) person identification across multiple horizontal views and the top view; 2) the estimation of look-at directions of all the subjects, including camera wearers, in both horizontal and top views; 3) a unified model for CIP detection by fusing all the subjects' look-at directions with both spatial and temporal consistencies.

To address the above challenges, we adopt a recent approach to identify unified subject detections by associating the same subjects across the multiple horizontal-view and the top-view videos and then propose a multi-view fusion approach to estimate the look-at direction of each subject in the top view. In each frame, we take all the detected subjects as a set of CIP candidates and build a Conditional Random Field (CRF) model by treating each frame as a node and the CIP candidates in this frame as its states. In the CRF, we define an intra-frame energy to reflect the confidence of a candidate to be the CIP in each frame, based on the estimated look-at direction information, and an inter-frame energy to reflect the spatial-temporal consistency between frames. The minimization of the combined energy will generate a CIP on each frame of each video. We also introduce an non-CIP frame detection method to handle the case where there is no CIP on a frame. We construct a complementary-view video dataset with annotated ground-truth CIPs and conduct extensive experiments to validate the effectiveness of the proposed method.

The main contributions of this paper are:

- This is the first work to study the CIP detection by combining a top- and multiple horizontal-view videos and it can benefit many application scenarios.
- We introduce an effective CRF formulation and solution for the proposed multi-view CIP detection, for which we define new CRF energies for measuring the intra-frame confidence and inter-frame consistency of the CIPs.
- We construct a new multi-view CIP detection benchmark and use it to evaluate the proposed method. We will release this new benchmark to public[2].

2 RELATED WORK

Video co-segmentation and saliency detection. Prior works on video co-segmentation and saliency detection are related to our work. Video co-segmentation is to simultaneously segment a common category of objects from two or more videos. In practice, almost all the existing video co-segmentation methods are based on the object appearance information [5, 19, 30, 31]. Some methods also use the supplementary motion feature and consider the temporal information for the object co-segmentation [6, 15, 40]. Different from the proposed CIP detection, video co-segmentation finds all the common objects, including the human subjects, across different videos, without considering whether the detected subject(s) draws much interest from other surrounding subjects.

Video saliency detection aims to identify the objects or regions with highest perceptual saliency and it is another well-studied problem [34]. Many machine learning based methods, e.g., sparsity-based reconstruction [8], low-rank consistency [4] and deep learning based models [18, 35] have been developed for video saliency detection. Most of them use the appearance and motion features to

[2]https://github.com/RuizeHan/CIP

separate the salient objects and non-salient regions and are usually developed for a single video. Recently, Xie et al. [36] proposed to detect co-salient objects in multiple videos by fusing the temporal and spatial saliency. However, saliency detection is general-purpose for identifying perceptually salient objects or regions, and our method identifies the co-interest person (CIP) with joint visual attention of surrounding people. Salient detection is usually based on the target's appearance and motion features, while our method detect the CIP based on the view relations with the surrounding people.

Visual attention by gaze estimation. Gaze estimation aims to predict the gaze of a human [20, 22, 25, 29, 41]. Many existing methods aim to model the human visual attention mechanism to identify the salient regions/objects in a natural or social environment [7, 21, 32, 33]. Early work [28] tried to determine a person's look-at direction under unconstrained motion. Recently, Recasens et al. [21] tried to estimate the gaze-focus position of each person present in the image. In [22], a calibrated camera and a scene-based eye tracking method are used to further identify whether a person in the video is looking at the camera. Chong et al. [7] made further extensions to address the problem of estimating the gaze-focus position that are located outside the image view. Besides, Fan et al. [11] tackled the problem of inferring shared attention in the third-person social scene video and in [12], they proposed to understand human gaze communication in social videos. However, almost all these methods handle the scene near to the cameras, with only few subjects in its field of view. In this paper, we address this limitation by combining a top-view video and multiple horizontal-view videos.

Multi-view video analysis. Collaborative analysis of multi-view videos has drawn the interest of many researchers [10, 23, 24, 27, 28, 37, 38]. In [27], face features from multiple views are extracted and then applied to estimate the facing direction. It uses the videos captured by horizontal-view cameras near the scene. Accordingly, the field of view is quite limited and the accuracy of facing direction estimation becomes very poor when the subject's face is not fully captured by the camera. There were a couple of existing works [23, 24] that studied the co-interest person/region detection by analyzing the videos recorded by multiple egocentric-view cameras, which requires all involved people to equip a wearable camera. This setting limits their usability in a real video-surveillance system. Besides, the approach in Park et al. [24] requires the prior 3D reconstruction of the scene via SfM (structure from motion) by using the collected videos, which limits its application in practice. Lin et al. [23] detected the co-interest person from multiple egocentric wearable cameras and required that camera wearers are looking at the co-interest person. More recently, the setting of combining top and horizontal-view cameras has been used for facilitating collaborative video analysis [1–3, 16, 17, 42]. In this paper, we combine the top and horizontal views and take a forward step to the real-world application, i.e., the important person detection for video surveillance.

3 PROPOSED METHOD

3.1 Overview

To detect the CIP in multiple videos over time, we capture $N+1$ temporally synchronized videos with the length of $\tilde{T}$ that are taken by N horizontal-view wearable cameras and a top-view drone-mounted camera. We first use a sliding window technique to consistently divide all $N+1$ videos into non-overlapped short-time clips with the length of T. We next perform a human detection algorithm and the multi-view data fusion strategy to get the CIP candidates in each frame. The proposed algorithm then computes an energy value for each candidate, which negatively reflects its confidence value to be the CIP. We also propose a simple non-CIP frame detection strategy to identify the clips without CIP. Finally, we merge the CIP detection results over all the clips to achieve a CIP detection on all frames over time $\tilde{T}$, as illustrated in Figure 2.

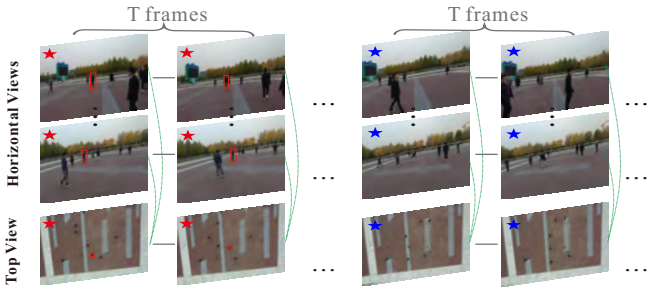

Figure 2: An illustration of the proposed multi-view CIP detection. A red star on the top-left corner indicates that a CIP is detected (with a red bounding box) on this frame, while a blue star indicates that no CIP is detected on this frame.

We denote $\mathcal{F}_k^n$ as the k-th clip from the n-th horizontal-view video, $n = 1, 2 \cdots, N$, and $\hat{\mathcal{F}}_k$ as the k-th clip from the top-view video, $k = 1, 2, \cdots, K$. For each window W_k, we use multi-view video clips, i.e., $\mathcal{F}_k^1, \mathcal{F}_k^2, ..., \mathcal{F}_k^N$ and $\hat{\mathcal{F}}_k$ to detect the CIP $\hat{P}_k$ in the top view and its corresponding subjects P_k^n in the horizontal views, $n = 1, 2, \cdots, N$. In the following, we elaborate on the proposed CIP detection algorithm in window W_k.

3.2 Problem Formulation

As mentioned above, in window W_k, we use $N+1$ multi-view video clips $\{\mathcal{F}_k^n | n = 1, 2, ..., N\}$ and $\hat{\mathcal{F}}_k$ for CIP detection. In the following, we simplify the notation of $\mathcal{F}_k^n$ as $\mathcal{F}^n$ and $\hat{\mathcal{F}}_k$ as $\hat{\mathcal{F}}$ by dropping the window number k when there is no confusion. Each clip contains T frames, i.e., $\hat{\mathcal{F}} = \{\hat{F}_t | t = 1, 2, \cdots, T\}$ and $\mathcal{F}^n = \{F_t^n | n = 1, 2, \cdots, N; t = 1, 2, \cdots, T\}$, where $\hat{F}_t$ (F_t^n) is the t-th frame of the top-view video clip (n-th horizontal-view video clip).

We first perform the human detection [26] on each frame of $N+1$ videos. We use a multi-view human association algorithm to unify the subjects detected in different videos, i.e., identifying and matching the same persons across different videos. As shown in Figure 1(c-f), all the subjects with identical color bounding boxes in different-view videos denote the same person. We will discuss the details of unifying subject detections[3] in Section 3.3. We take the unified detections $\mathcal{P} = \{P_t | t = 1, 2, \cdots, T\}$ as the CIP candidate set, where P_t is the CIP candidate, i.e., one of the unified subjects in

[3]We map all the subjects in multiple horizontal-view videos into the top-view video and get the unified subject detections as shown in Figure 1(c).

frame t. We introduce the conditional random field (CRF) [9] model for CIP detection, where each frame is treated as a node in CRF and each candidate as a state of this node. Through the CRF model, our goal is to seek a candidate u_t in each frame t as the detected CIP. The CIP candidates $\mathcal{P}$ have a posterior probability

$$\mathrm{p}(\mathcal{P}|\mathcal{F}) \propto \exp(-\mathrm{E}(\mathcal{P}|\mathcal{F})) \tag{1}$$

$$\text{with} \quad \mathrm{E}(\mathcal{P}|\mathcal{F}) = \sum_{t,s} \Phi(P_t, P_s|F_t, F_s), \tag{2}$$

where $\mathcal{F} = \{\hat{\mathcal{F}}, \mathcal{F}^1, \mathcal{F}^2, \cdots, \mathcal{F}^n\}$ denotes the multi-view video clips and P_t denotes the CIP candidate in frame t. We use the energy function $\Phi(P_t, P_s|F_t, F_s)$ to estimate the confidence of P_t, P_s as the same person and take it as the CIP. This way, the CIP detection in a window is transformed to a problem of finding the optimal state of $\mathcal{P}$ that minimizes the energy $\mathrm{E}(\mathcal{P}|\mathcal{F})$. In the following, we simplify the energy function $\mathrm{E}(\mathcal{P}|\mathcal{F})$ as $\mathrm{E}(\mathcal{P})$ and $\Phi(P_t, P_s|F_t, F_s)$ as $\Phi(P_t, P_s)$ when there is no ambiguity.

The remaining problem is the definition of the energy function $\Phi(P_t, P_s)$, which should reflect the confidence of a candidate to be the CIP and the correspondence of the CIP across different frames and views. In this paper, we consider two factors: 1) each candidate's confidence to be the CIP in each frame (**intra-frame factor**), 2) the temporal consistency of the candidates between two adjacent frames (**inter-frame factor**). For factor 1), the CIP in the same frame shows two typical properties: i) it should appear in the FOV of *most* people in the scene, and ii) it usually appears at the *center* of a subjects' FOV, if he/she is looked at by this subject. For factor 2), the CIP shows two properties between adjacent frames: i) the level of interest by other people should *increase* when the CIP appears; ii) the CIP detection should keep good *consistency* over a short time. Besides, at a time the CIP should be the same person across different-view videos. This way, the first step is to match all the subjects and calibrate the subjects' look-at directions across different views, which is a very challenging problem given the cross-view appearance/motion variations and uncalibrated mobile cameras. In the following, we first discuss the multi-view look-at direction estimation in Section 3.3. We then define the CRF energy functions in Section 3.4 by considering the above two factors.

3.3 Multi-View Look-at Direction Estimation

We use FOV to estimate the target that draws interest of each subject. To more reliably estimate a subject's FOV, we combine multiple-view videos to handle the cases shown in Figure 1(d-f). A subject's FOV consists of his/her look-at direction and the field-of-view angle. As shown in Figure 3(a), since the top-view video provides the global picture of the whole scene from a high altitude, we can associate the subjects in different horizontal views by mapping them to the top view. This way, we can identify a set of the unified detections of subjects $\mathcal{P}$ in the top-view video and transform the subjects' look-at directions to the top view for fusion. Without loss of generality, we consider the mapping from one horizontal-view video to the top-view video as shown in Figure 3.

For this purpose, there are three problems to be solved: i) Subject matching between the top and multiple horizontal views. ii) look-at direction estimation of the horizontal-view camera wearer in the top view. iii) look-at direction estimation of all the other subjects

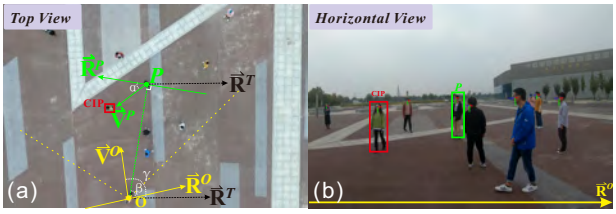

Figure 3: An illustration of the look-at direction transformation from a horizontal-view video to the top-view video.

without wearing cameras in the top view. We extend a recent cross-view subject association algorithm [17] to address the problems i) and ii). Specifically, this algorithm provides the subject association across two complementary views, e.g., the subject P in Figure 3(a) and (b) as well as the look-at direction $\vec{V}^{\mathrm{O}}$ of the horizontal-view camera in the top-view video as shown in Figure 3(a). We can apply this algorithm to associate the subjects in each horizontal view to the top view and then use the top view as a bridge to construct unified detections of subjects across the top and multiple horizontal views. For problem iii), as shown in Figure 3, we define three direction vectors: 1) $\vec{R}^T$: the horizontal vector pointing to the right in the top view, 2) $\vec{R}^{\mathrm{O}}$: the horizontal vector pointing to the right of the horizontal-view camera, and its vertical direction is denoted as $\vec{V}^{\mathrm{O}}$ in the top view, and 3) $\vec{R}^P$: the vector that is vertical to vector $\overrightarrow{\mathrm{PO}}$ in the top view. We also define three angles as

$$\begin{cases} \alpha = \delta^{\mathrm{ac}}(\vec{V}^P, \vec{R}^P) \in [0, \pi] \\ \beta = \delta^{\mathrm{ac}}(\overrightarrow{\mathrm{OP}}, \vec{R}^O) \in [0, \pi] \\ \gamma = \delta^{\mathrm{ac}}(\vec{V}^O, \vec{R}^T) \in [0, 2\pi), \end{cases} \tag{3}$$

where $\delta^{\mathrm{ac}}(\vec{x}, \vec{y})$ denotes the counterclockwise angle from $\vec{y}$ to $\vec{x}$. With the definition of above direction vectors, the angles β and γ can be computed in the top view. The angle α can be derived by given vector $\vec{V}^P$, i.e., the look-at direction of subject P (paralleling to the ground plane). This way, $\vec{V}^P$ can be transformed as a look-at direction estimation in the horizontal-view video, as shown in Figure 3(b), which can be solved by a 3D head pose estimation algorithm[4], e.g., [39]. Through the plane geometric transformation, we calculate the view angle of the subject P in the top view as

$$\delta^{\mathrm{ac}}(\vec{V}^P, \vec{R}^P) = \mod(\alpha + \beta + \gamma, 2\pi). \tag{4}$$

Moreover, since we use multiple horizontal-view videos, it is common that a subject in the top view is also captured in multiple horizontal views. We find that the look-at direction estimation is more accurate when the subject is facing to and near the horizontal-view camera. On the contrary, the look-at direction estimation will get worse or even fail when the horizontal-view camera captures the side or back of the subjects. Therefore, for each subject, we fuse his/her look-at directions estimated from multiple horizontal views for the final look-at direction in the top view. Specifically, for the unified detection of a subject in multiple horizontal views, we pick the view that leads to the highest face-detection confidence of this

[4]We use the head pose estimation result to estimate the look-at direction, which ignores the squint and assumes the look-at direction is same as the facing direction.

subject to estimate his/her look-at direction. We will further show the ablation study of different data fusion methods in Section 4.4.

3.4 Spatial-Temporal CIP Potential Energy

In this section, we consider the intra-frame and inter-frame factors of CIP discussed in Section 3.2 and reformulate the energy function in Eq. (1) as

$$\mathrm{E}(\mathcal{P}) = \sum_{t} \Phi_1(P_t) + \sum_{t,s} \Phi_2(P_t, P_s), \tag{5}$$

where Φ_1 and Φ_2 are two terms for the intra-frame and inter-frame energies, respectively. Different from many previous works [14, 23], we do not define the inter-view energy term, i.e., the relationship of the candidates from different-view videos, since we unify them by a multi-view data fusion strategy as discussed in Section 3.3.

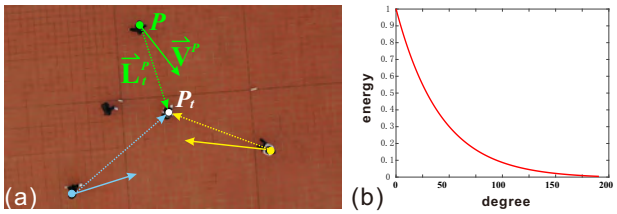

Figure 4: An illustration of the (a) deviation angle $\langle \vec{V}^P, \vec{L}_t^P \rangle$ and (b) damping function.

Intra-frame confidence. Base on the **intra-frame factor** discussed in Section 3.2, a CIP that draws the subjects' interest usually stays in the view center of these subjects. We define the view-deviation level of a candidate based on its position in the FOV of other subjects. We also calculate the focus-ratio level of a candidate, i.e., the ratio between the number of subjects whose FOV covers the candidate to the total number of subjects in the scene. This way, we construct the intra-frame energy as

$$\Phi_1(P_t) = -\mathrm{Sig}(\mathrm{V}(P_t) + \mathrm{F}(P_t)), \tag{6}$$

where $\mathrm{Sig}(\cdot)$ denotes the sigmoid function, i.e., $\mathrm{Sig}(x) = \frac{1}{1+e^{-x}}$ and $\mathrm{V}(\cdot)$ and $\mathrm{F}(\cdot)$ are two energies to reflect the view-deviation and focus-ratio level of a candidate P_t, respectively. Specifically, $\mathrm{V}(\cdot)$ is defined as

$$\mathrm{V}(P_t) = \frac{1}{|\mathbb{P}_t| - 1} \sum_{P \in \mathbb{P}_t \setminus P_t} \mathrm{Damp}\langle \vec{V}^P, \vec{L}_t^P \rangle, \tag{7}$$

where $\mathbb{P}_t$ denotes all the detected subjects with total number $|\mathbb{P}_t|$ at frame t, $\vec{V}^P$ denotes the look-at direction of a subject P and $\vec{L}_t^P$ denotes the vector from the subject P to the candidate P_t, P traverses all the subjects $\mathbb{P}$ except for P_t . The brackets $\langle \cdot \rangle$ represent the angle between two vectors. As shown in Figure 4(a), we use the angle between $\vec{V}^P$ and $\vec{L}_t^P$ to estimate the view-deviation level of P_t in the view of P. We further use a damping function to modify it as shown in Figure 4(b). We also define $\mathrm{F}(\cdot)$ as

$$\mathrm{F}(P_t) = \frac{1}{|\mathbb{P}_t| - 1} \sum_{P \in \mathbb{P}_t \setminus P_t} \mathrm{Ind}(\langle \vec{V}^P, \vec{L}_t^P \rangle \le \theta), \tag{8}$$

where $\vec{V}^P$ and $\vec{L}_t^P$ are the same as those in Eq. (7) and $\mathrm{Ind}(\cdot)$ is the indicator function. This energy term is to compute the proportion of subjects whose FOV contains the candidate P_t, where 2θ is the subject's field-of-view angle.

Inter-frame consistency. Base on the **inter-frame factor** discussed in Section 3.2, the inter-frame energy is dependent on the change of CIP confidence and the temporal consistency across frames. We construct the inter-frame energy as

$$\Phi_2(P_t, P_s) = -\mathrm{Sig}(\Delta\mathrm{V}(P_t, P_s) + \Delta\mathrm{F}(P_t, P_s) + \mathrm{O}(P_t, P_s)), \tag{9}$$

where $\Delta\mathrm{V}(P_t, P_s) = \mathrm{V}(P_s) - \mathrm{V}(P_t)$ represents the change of view-deviation level from P_t to P_s. Similarly, $\Delta\mathrm{F}(P_t, P_s) = \mathrm{F}(P_s) - \mathrm{F}(P_t)$ represents the change of focus-ratio level. In the latter experiments, we simply take $s = t + 1$ to only consider the two adjacent frames for reducing computational cost. The inclusion of these two energy terms encourages the detection of a CIP that appears from frame t to frame s. The last energy term $\mathrm{O}(P_t, P_s)$ denotes the overlap between P_t and P_s, which guarantees the spatial consistency of the CIP from frame t to frame s.

The CRF problem can be treated as the discrete energy minimization problem. In this paper, we use the Viterbi algorithm [13] to minimize the energy function in Eq. (5) for the optimal solution. However, one limitation of the above CRF model is its assumption that there is always a CIP in each frame. This may not be true in practice. Therefore, we propose a simple non-CIP frame detection strategy to identify the frames without CIP. Specifically, we assume $\hat{P}_t$ as the predicted CIP of frame t in a clip. We calculate the average attention level of the clip as $\phi = \frac{1}{T}\sum_{t=1}^{T} \frac{1}{2}(\mathrm{V}(\hat{P}_t) + \mathrm{F}(\hat{P}_t))$, where $\mathrm{V}(\cdot), \mathrm{F}(\cdot)$ are defined in Eqs. (7) and (8), respectively. We use a threshold δ to estimate the confidence of the detected CIP in this clip. We regard that there is no CIP in this clip if $\phi < \delta$. We will discuss the effectiveness of non-CIP frame detection strategy and the parameter selection of δ in Section 4. After that, we get the CIP in the entire top-view video and then we backtrack the CIP in each horizontal-view video based on the cross-view human association discussed in Section 3.3.

4 EXPERIMENTS AND DISCUSSION

4.1 Dataset and Metrics

Dataset collection. We do not find public benchmark with synchronized top-view and horizontal-view video sets with required annotations for complementary-view co-interest persons detection. Therefore, we collect a new dataset by using a camera mounted on the drone to take top-view videos and several GoPros mounted over the heads of some people to take the horizontal-view videos for algorithm evaluation. We capture the videos at three different sites with different backgrounds. At each site, there are 10 subjects moving or standing in the scene, and three of them wear GoPro cameras over the head to collect horizontal-view videos. We combine the top-view video and the corresponding three horizontal-view videos as a set of videos. All the subjects wear dark coats thus sharing very similar appearances. We arrange the video recordings in a way that the 10 subjects alternately play as the CIP in a way that most other subjects look at him/her. There are non-CIP gaps between the co-interest transition from one CIP to another. Note that, all the subjects are free to walk or stand without constraints, and there are

random mutual occlusions among them or out-of-view subjects in horizontal-view videos. We manually label the CIP by a bounding box in each frame and we also label the frames without CIP as non-CIP frames. We also manually align these videos temporally such that corresponding frames from different videos are taken at the same time. We collect 30 sets of video and each set of videos contains temporal synchronized three horizontal-view and one top-view videos. Each video contains both temporal intervals with a CIP and without CIP. In total, we collect 49 minutes 10 seconds of videos containing 88,512 frames.

Evaluation metrics. For each detected CIP, denoted by C, if there is a ground truth box G with an overlap $O = \frac{C \cap G}{C \cup G}$ larger than 0.5, we count this detected CIP C to be a true positive. We can also count the total number of detected and ground-truth CIPs in each video. In this way we calculate the Precision, Recall and F-score $= \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$.

4.2 Setup

Implementation details. We implement the main program in Matlab and run on a desktop computer with an Intel Core i7 3.4 GHz CPU, and the human detection and gaze detection are implemented on RTX 2080Ti GPU. We use the general YOLOv3 [26] detector to detect subjects in the form of bounding boxes in both top-view and horizontal-view videos. For top-view subject detection, we fine-tune the network using 600 top-view human images. Note that all the training data have no overlap with our test dataset. We use a head pose estimation algorithm FSA-Net [39] to estimate the human facing direction in horizontal-view videos. The clip length T is set to 50 in the experiments, and two pre-specified parameters θ and δ are set to 45 degrees and 0.5, respectively.

Baseline methods. • ***Random***: A weak baseline that draws a Gaussian heatmap with random mean and variance. For each ground truth CIP G in a frame, we accumulate the region generated by the heatmap in G, denoted by S. We count the detection result in this frame as the true positive if the ratio $R = \frac{S}{G}$ lager than 0.5. We take the result as non-CIP frame if $R = 0$.

• ***Motion-based***: We use the motion direction of each subject to estimate the look-at direction in top view instead of using our complementary-view data fusion method described in Section 3.3.

• ***Image/Video-saliency***: We choose a salient object detection method [35] for comparison. This approach provides both 1) image based saliency detection results which takes single frames as input, and 2) video based saliency detection results which considers the temporal reasoning. We take a loose condition for evaluating the saliency based method with the same way as the baseline of random and replace the Gaussian heatmap with the saliency map.

• ***Method in*** [23]: One seemingly related work is [23] that uses multiple egocentric cameras for CIP identification. We apply the approach in [23] to the horizontal-view videos in our dataset.

• ***Face*** [39] + ***Att.*** [11]: We apply the look-at direction estimation approach [39] used in our method and generate the visual attention map proposed in [11] to locate the CIP in horizontal-view videos. In this approach, we use the bounding boxes of all subjects instead of the region proposals used in [11].

• ***Face*** [39] + ***Fus.*** + ***Att.*** [11]: Similarly, we use the method in [39] and the proposed complementary-view data fusion strategy to estimate the subjects' look-at directions in top view. We then generate the attention map in [11] to locate the CIP in top-view videos and backtrack it in horizontal views.

• ***Face*** [39] + ***Fus.*** + ***Voting***: Similar to the above one, after getting the subjects' look-at directions in top view, we replace the attention map with a voting strategy – each subject votes an interesting target by its look-at direction to elect the CIP with the most votes.

Table 1: Comparison of the baseline methods and ours.

Method	Precision	Recall	F-score
Random	7.7%	2.1%	3.3%
Motion-based	6.8%	5.8%	6.3%
Image-saliency [35]	29.8%	52.2%	37.9%
Video-saliency [35]	20.2%	33.9%	25.3 %
Method in [23]	11.9%	20.2%	15.0%
Face [39] + Att. [11]	23.0%	22.7%	22.8%
Face [39] + Fus. + Att. [11]	42.7%	32.6%	37.0%
Face [39] + Fus. + Voting	38.1%	30.8%	34.1%
Ours	**61.5%**	**59.2%**	**60.3%**

4.3 Results

As shown in Table 1, we evaluate the proposed method on the proposed video dataset using the CIP detection Precision, Recall and F-score metrics. Existing related methods are only applied to the horizontal-view videos. For a fair comparison, we only take the results on the horizontal view videos when comparing the proposed method against the other methods. We can see that the motion based method fails in CIP detection. In most cases, the subjects may turn his/her head to look at the CIP, which makes the look-at direction inconsistent to the motion direction. From the third and fourth rows of Table 1, we can see that the saliency detection methods, both image based and video based, provide unsatisfactory performance in our task. It is because that the CIP in a video/image may not have the notable properties used in salient object detection. We can also see that the video based saliency detection provides worse performance than the image based methods. The reason might be that the moving of subjects has a significant impact on video based saliency detection. Also, the method in [23] can not handle our problem well because it assumes that the relative location and size of the CIP does not change in each video, which requires the camera wearers to always keep a steady distance to the CIP and move his/her eyes to follow the CIP. Results generated by 'Face + Att.' and 'Face + Fus. + Att.' do not perform well since the straightforward attention map is not robust enough for handling the complex scene in our problem. Further, using the multi-view fusion strategy, 'Face + Fus. + Att.' performs much better than 'Face + Att.' by a large margin, which demonstrates that the proposed *multi-view fusion strategy can estimate each subject's look-at direction more reliably*. Finally, the comparative results of last three rows in Table 1 show the *superiority of our CRF model for CIP detection* compared to the the attention map in [11] and the simple voting strategy.

Table 2: Comparative study of variants of our method evaluated on the horizontal-view, top-view and all videos.

Types	Horizontal view			Top view			All		
	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
w/o H-viewer	57.9%	59.1%	58.5%	64.8%	70.1%	67.3%	59.6%	61.8%	60.7%
w only H-viewer	54.0%	47.1%	50.4%	59.3%	52.6%	55.8%	55.4%	48.5%	51.7%
w all-view	51.5%	59.9%	55.3%	57.2%	65.9%	61.2%	52.9%	61.4%	56.8%
w single-view	50.4%	13.5%	21.3%	58.2%	20.0%	29.8%	52.3%	15.1%	23.4%
w/o inter-frame	46.9%	47.7%	47.3%	52.4%	54.5%	53.4%	48.3%	49.4%	48.8%
w/o non-CIP	48.1%	**62.8%**	54.5%	54.4%	**75.7%**	63.3%	49.7%	**66.1%**	56.7%
Ours	**61.5%**	59.2%	**60.3%**	**68.6%**	70.1%	**69.3%**	**63.3%**	61.9%	**62.6%**

Table 3: Results by varying values of T, θ and δ.

T	Precision	Recall	F-score	θ	Precision	Recall	F-score	δ	Precision	Recall	F-score
25	60.6%	58.2%	59.4%	40°	64.4%	57.5%	60.8%	0.45	55.7%	65.6%	60.3%
50	63.2%	61.9%	62.6%	45°	63.2%	61.9%	62.6%	0.50	63.2%	61.9%	62.6%
75	62.3%	62.1%	62.2%	50°	60.2%	63.8%	61.9%	0.55	72.4%	48.2%	57.8%

4.4 Ablation Studies

Look-at direction estimation. As shown in Table 2, '***w/o H-viewer***' and '***w only H-viewer***' denote the proposed method without and only using the look-at directions of the horizontal-view camera wearers, respectively. From the first row, we can see that our method is not heavily dependent on the look-at directions of the horizontal-view camera wearer. Moreover, the CIP detection performance of the proposed method is acceptable with only the look-at directions of the horizontal-view camera wearers.

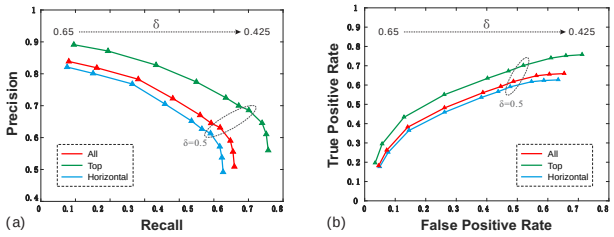

Figure 5: PR (a) and ROC (b) curves under different δ.

Multi-view data fusion. In the proposed method, when computing the subjects' look-at directions in top view, we take the corresponding horizontal-view subject with the highest face detection confidence. Another fusion strategy, denoted by '***w all-view***', is to average a subject's look-at directions estimated from different horizontal views. As shown in Table 2, the latter fusion strategy leads to significant CIP detection performance decrease since the look-at direction estimation can be highly inaccurate when the subject's face is not well visible in some horizontal views. If we only use one horizontal-view video for look-at direction estimation, i.e., '***w single-view***' in Table 2, we achieve an acceptable precision but very poor recall. It is because that a single horizontal-view video has limited FOV and many subjects are not covered.

Effectiveness of the CRF model. To validate the effectiveness of the temporal consistency constraint in our method, we remove the inter-frame energy in the CRF model, i.e., '***w/o inter-frame***' shown in the fifth row. We can see that the precision and recall scores of CIP detection show a significant decrease. In '***w/o non-CIP***', we remove the non-CIP detection strategy and force to detect a CIP in every frame, even if a frame does not contain a CIP. This will raise the recall score, but generate more false positives, leading to a lower precision score.

4.5 Discussion

Parameters selection. Figure 5(a) plots the PR (Precision Recall) curve of the proposed method using different values of δ. We can see that the precision score decreases while recall score increases as δ gets smaller. It is because our method generates less non-CIP frames when δ decreases, resulting in more CIP frame detections with more false positives. Figure 5(b) shows the ROC curve under different δ. There are two additional free parameters in the proposed method: the clip length T and the pre-set angle threshold θ in Eq. (8). We select different values for them and examine their influence on the detection performance. Table 3 reports the results by varying one of these two parameters while fixing the other one. We can see that the final performance, including precision, recall and F-score, is not very sensitive to the selected values of these two parameters.

Qualitative analysis. Figure 6 shows the CIP detection results on sample frames from the top-view and three horizontal-view videos. Red and green boxes indicate the detected CIP and the ground truth, respectively. Frames with a solid blue star on the top-left corner indicate that no CIP is detected by our algorithm, e.g., they are drawn from the CIP's egocentric video or the CIP is occluded in these frames. As shown in Figure 6, the proposed algorithm can detect CIP even if the CIP shows similar appearance and motion characteristics with other people. As shown in the

Figure 6: Qualitative analysis of CIP detection on sample frames from different view videos. Red and green boxes indicate the detected CIP and the ground truth, respectively. Frames with a solid red/blue star on the top-left corner indicate that CIP/non-CIP is detected on this frame by our algorithm. Best viewed in color.

first row of Figure 6, the CIP is fully occluded by other subjects in 'Horizontal View 2', and the proposed method can handle it and detect the CIP in the other two horizontal views. The second row shows a case where the CIP appears in all three horizontal-view videos, and the proposed method can obtain its appearance from different views. Note that, our method can identify the CIP with serious partial occlusion, as shown in the third row of 'Horizontal View 3'. It is because the other two horizontal views can well observe the CIP and the top view provides a good picture of all the subjects. In the last row, the camera wearer for 'Horizontal View 1' is the CIP, who is also out of the FOV of 'Horizontal View 3' video. Moreover, almost all the subjects' faces can not be seen in the FOV of 'Horizontal View 2'. Even so, the proposed method successfully identifies the CIP in the top view and then backtracks him in the three horizontal views. Based on the above analysis, we can see that the proposed method can combine multiple horizontal views and a complementary top view for more comprehensive and accurate CIP detection.

Table 4: Time performance of different components.

Component	Sub. Det.	Head Pose	Data Fus.	CRF
Time (sec/frm)	0.0281	0.7500	0.0285	0.0004
Platform	GPU	GPU	CPU	CPU

Speed analysis. As shown in Table 4, we record the running time taken by each component of the proposed method. In this table, 'Sub. Det.' and 'Head Pose' denote the subject detection and human head pose estimation, respectively. 'Data Fus.' denote the multi-view fusion for look-at direction estimation. 'CRF' denotes the step of building and optimizing the CRF model. We can see that the most time-consuming steps in the proposed algorithm are subject detection and head pose estimation: the latter takes over 90% of the total time consumption. The components of 'Data Fus.' and 'CRF', where our main contribution lies, run over real-time speed on a desktop computer with Intel i7 3.4 GHz CPU.

5 CONCLUSION

In this paper, we studied a new problem of identifying the co-interest person (CIP) by using complementary top- and horizontal-view videos taken by mobile cameras. The subjects' look-at directions were inferred from multiple horizontal-view cameras and then mapped to the top-view video for co-interest person detection. For this, we built an effective Conditional Random Field (CRF) model by considering both the intra-frame confidence and inter-frame consistency. We collected a new complementary-view video dataset, as well as manually labelled ground-truth CIPs, for performance evaluation. Experimental results on this dataset demonstrated that the proposed method can effectively identify CIPs by capturing both global locations in top view and the local appearance details in horizontal views. This work is not an incremental method based on a well-studied problem, and we just hope to open a new door to study the CIP detection in surveillance videos. In the future, we plan to improve the performance by considering more complex cases, auch as the possible presence of multiple CIPs on a frame.

ACKNOWLEDGMENTS

This work was supported, in part, by the NSFC under Grants U1803264, 61672376, 61671325, 61572354.

REFERENCES

[1] Shervin Ardeshir and Ali Borji. 2016. Ego2Top: Matching Viewers in Egocentric and Top-View Videos. In *European Conference on Computer Vision*.

[2] Shervin Ardeshir and Ali Borji. 2018. Egocentric Meets Top-View. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 6 (2018), 1353–1366.

[3] Shervin Ardeshir and Ali Borji. 2018. Integrating Egocentric Videos in Top-View Surveillance Videos: Joint Identification and Temporal Alignment. In *European Conference on Computer Vision*.

[4] Chenglizhao Chen, Shuai Li, Yongguang Wang, Hong Qin, and Aimin Hao. 2017. Video Saliency Detection via Spatial-Temporal Fusion and Low-Rank Coherency Diffusion. *IEEE Transactions on Image Processing* 26, 7 (2017), 3156–3170.

[5] Ding-Jie Chen, Hwann-Tzong Chen, and Long-Wen Chang. 2012. Video object cosegmentation. In *ACM International Conference on Multimedia*.

[6] Weichen Chiu and Mario Fritz. 2013. Multi-class Video Co-segmentation with a Generative Multi-video Model. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[7] Eunji Chong, Nataniel Ruiz, Yongxin Wang, Yun Zhang, Agata Rozga, and James M Rehg. 2018. Connecting Gaze, Scene, and Attention: Generalized Attention Estimation via Joint Modeling of Gaze and Scene Saliency. In *European Conference on Computer Vision*.

[8] Runmin Cong, Jianjun Lei, Huazhu Fu, Fatih Porikli, Qingming Huang, and Chunping Hou. 2019. Video Saliency Detection via Sparsity-Based Reconstruction and Propagation. *IEEE Transactions on Image Processing* 28, 10 (2019), 4819–4831.

[9] Thomas Deselaers, Bogdan Alexe, and Vittorio Ferrari. 2012. Weakly Supervised Localization and Learning with Generic Knowledge. *International Journal of Computer Vision* 100, 3 (2012), 275–293.

[10] Junting Dong, Wen Jiang, Qixing Huang, Hujun Bao, and Xiaowei Zhou. 2019. Fast and Robust Multi-Person 3D Pose Estimation from Multiple Views. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[11] Lifeng Fan, Yixin Chen, Ping Wei, Wenguan Wang, and Song-Chun Zhu. 2018. Inferring Shared Attention in Social Scene Videos. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[12] Lifeng Fan, Wenguan Wang, Siyuan Huang, Xinyu Tang, and Song-Chun Zhu. 2019. Understanding Human Gaze Communication by Spatio-Temporal Graph Reasoning. In *International Conference on Computer Vision*.

[13] G David Forney. 1973. The viterbi algorithm. *Proceedings of the IEEE* 61, 3 (1973), 268–278.

[14] Huazhu Fu, Xu Dong, Bao Zhang, and Stephen Lin. 2014. Object-Based Multiple Foreground Video Co-segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[15] Jiaming Guo, Zhuwen Li, Loongfah Cheong, and Steven Zhiying Zhou. 2013. Video Co-segmentation for Meaningful Action Extraction. In *International Conference on Computer Vision*.

[16] Ruize Han, Wei Feng, Jiewen Zhao, Zicheng Niu, Yujun Zhang, Liang Wan, and Song Wang. 2020. Complementary-View Multiple Human Tracking. In *AAAI Conference on Artificial Intelligence*.

[17] Ruize Han, Yujun Zhang, Wei Feng, Chenxing Gong, Xiaoyu Zhang, Jiewen Zhao, Liang Wan, and Song Wang. 2019. Multiple Human Association between Top and Horizontal Views by Matching Subjects' Spatial Distributions. In *arXiv*.

[18] Lai Jiang, Mai Xu, Tie Liu, Minglang Qiao, and Zulin Wang. 2018. DeepVS: A Deep Learning Based Video Saliency Prediction Approach. In *European Conference on Computer Vision*.

[19] Armand Joulin, Kevin Tang, and Li Fei-Fei. 2014. Efficient image and video co-localization with frank-wolfe algorithm. In *European Conference on Computer Vision*.

[20] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. 2019. Gaze360: Physically Unconstrained Gaze Estimation in the Wild. In *International Conference on Computer Vision*.

[21] Aditya Khosla, Carl Vondrick, and Antonio Torralba. 2015. Where are they looking?. In *Advances in Neural Information Processing Systems*.

[22] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra M Bhandarkar, Wojciech Matusik, and Antonio Torralba. 2016. Eye Tracking for Everyone. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[23] Yuewei Lin, Kareem Ezzeldeen, Youjie Zhou, Xiaochuan Fan, Hongkai Yu, Hui Qian, and Song Wang. 2015. Co-Interest Person Detection from Multiple Wearable Camera Videos. In *International Conference on Computer Vision*.

[24] Hyun Soo Park, Eakta Jain, and Yaser Sheikh. 2012. 3D Gaze Concurrences From Head-mounted Cameras. In *Advances in Neural Information Processing Systems*.

[25] Adria Recasens, Carl Vondrick, Aditya Khosla, and Antonio Torralba. 2017. Following Gaze in Video. In *International Conference on Computer Vision*.

[26] Joseph Redmon, Santosh Kumar Divvala, Ross B Girshick, and Ali Farhadi. 2016. You Only Look Once: Unified, Real-Time Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[27] Neil Robertson and Ian Reid. 2006. Estimating gaze direction from low-resolution faces in video. In *European Conference on Computer Vision*.

[28] Kevin Smith, Sileye O Ba, Jeanmarc Odobez, and Daniel Gaticaperez. 2008. Tracking the Visual Focus of Attention for a Varying Number of Wandering People. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, 7 (2008), 1212–1229.

[29] Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. 2014. Learning-by-Synthesis for Appearance-Based 3D Gaze Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[30] Le Wang, Gang Hua, Rahul Sukthankar, Jianru Xue, Zhenxing Niu, and Nanning Zheng. 2014. Video Object Discovery and Co-Segmentation with Extremely Weak Supervision. In *European Conference on Computer Vision*.

[31] Le Wang, Gang Hua, Rahul Sukthankar, Jianru Xue, Zhenxing Niu, and Nanning Zheng. 2016. Video object discovery and co-segmentation with extremely weak supervision. 39, 10 (2016), 2074–2088.

[32] Wenguan Wang, Qiuxia Lai, Huazhu Fu, Jianbing Shen, and Haibin Ling. 2019. Salient Object Detection in the Deep Learning Era: An In-Depth Survey.. In *arXiv*.

[33] Wenguan Wang and Jianbing Shen. 2018. Deep Visual Attention Prediction. *IEEE Transactions on Image Processing* 27, 5 (2018), 2368–2378.

[34] Wenguan Wang, Jianbing Shen, Fang Guo, Ming-Ming Cheng, and Ali Borji. 2018. Revisiting video saliency: A large-scale benchmark and a new model. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[35] Wenguan Wang, Jianbing Shen, and Ling Shao. 2018. Video Salient Object Detection via Fully Convolutional Networks. *IEEE Transactions on Image Processing* 27, 1 (2018), 38–49.

[36] Yufeng Xie, Linwei Ye, Zhi Liu, and Xuemei Zou. 2016. Video co-saliency detection. In *International Conference on Digital Image Processing*.

[37] Mingze Xu, Chenyou Fan, Yuchen Wang, Michael S Ryoo, and David J Crandall. 2018. Joint Person Segmentation and Identification in Synchronized First- and Third-Person Videos. In *European Conference on Computer Vision*.

[38] Yuanlu Xu, Xiaobai Liu, Lei Qin, and Songchun Zhu. 2017. Cross-View People Tracking by Scene-Centered Spatio-Temporal Parsing. In *AAAI Conference on Artificial Intelligence*.

[39] Tsunyi Yang, Yiting Chen, Yenyu Lin, and Yungyu Chuang. 2019. FSA-Net: Learning Fine-Grained Structure Aggregation for Head Pose Estimation From a Single Image. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[40] Dong Zhang, Omar Javed, and Mubarak Shah. 2014. Video Object Co-segmentation by Regulated Maximum Weight Cliques. In *European Conference on Computer Vision*.

[41] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2015. Appearance-based gaze estimation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[42] Jiewen Zhao, Ruize Han, Yiyang Gan, , Liang Wan, Wei Feng, and Song Wang. 2020. Human Identification and Interaction Detection in Cross-View Multi-Person Videos with Wearable Cameras. In *ACM Multimedia*.