Complementary-View Multiple Human Tracking

Ruize Han,[1,2] **Wei Feng,**[1,2*] **Jiewen Zhao,**[1,2] **Zicheng Niu,**[1] **Yujun Zhang,**[1,2]
Liang Wan,[1,2] **Song Wang**[1,2,3*]
[1]College of Intelligence and Computing, Tianjin University, Tianjin 300350, China
[2]Key Research Center for Surface Monitoring and Analysis of Cultural Relics, SACH, China
[3]Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29208, USA
{han_ruize, wfeng, zhaojw, niuzchina, yujunzhang, lwan}@tju.edu.cn, songwang@cec.sc.edu

Abstract

The global trajectories of targets on ground can be well captured from a top view in a high altitude, e.g., by a drone-mounted camera, while their local detailed appearances can be better recorded from horizontal views, e.g., by a helmet camera worn by a person. This paper studies a new problem of multiple human tracking from a pair of top- and horizontal-view videos taken at the same time. Our goal is to track the humans in both views and identify the same person across the two complementary views frame by frame, which is very challenging due to very large field of view difference. In this paper, we model the data similarity in each view using appearance and motion reasoning and across views using appearance and spatial reasoning. Combing them, we formulate the proposed multiple human tracking as a joint optimization problem, which can be solved by constrained integer programming. We collect a new dataset consisting of top- and horizontal-view video pairs for performance evaluation and the experimental results show the effectiveness of the proposed method.

1 Introduction

Multiple object tracking especially human tracking is one of the most crucial problems in AI and vision (Luo et al. 2015; Wen et al. 2019), with many applications such as video surveillance and environmental monitoring. Although many advanced tracking algorithms are proposed and break the performance records on public benchmarks in every year, there are still many challenges e.g., occlusions and out-of-view problems, which are far from being well addressed in tracking (Han, Guo, and Feng 2018; Feng et al. 2019).

While tracking has many applications (Guo et al. 2020), typically two pieces of information can be provided by tracking results: *accurate trajectories and appearances of the targets over time*. This clearly introduces a conflict – if the camera is too close to the targets, limited coverage and frequent mutual occlusions prevent the accurate detection of their trajectories; if the camera is too far away from the targets, it is difficult to capture the detailed appearance of targets that are important for many applications such as person identification, action recognition, etc. In this paper, we present a new camera setting to address this problem. To track a group of people, which we refer to as subjects in this paper, on the ground, we use two cameras with different views and synchronized clock: A top-view camera at a high altitude, e.g, mounted to flying drone, provides a global birds-eye view of the subjects and the whole scene as shown in Fig. 1(a). A horizontal-view camera on the ground, e.g., mounted to a helmet worn by one person, which is static or moves/rotates smoothly without drastic visual field changes, captures the detailed appearances of subjects of interest, as shown in Fig. 1(b). We expect the collaborative tracking on these two complementary views produces both global motion trajectories and local appearance details of the subjects.

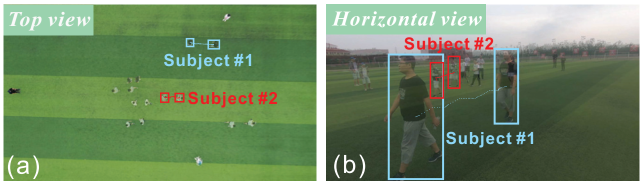

Figure 1: An illustration of the top-view (a) and horizontal-view (b) videos. The former is taken by a camera mounted to a drone in the air and the latter is taken by a GoPro worn by a wearer who walked on the ground. The proposed method jointly tracking multiple subjects, indicated by identical color boxes, across the two videos. Note that the global motion trajectory and local appearance are well presented in these two complementary views.

In this paper, we tackle this collaborative tracking by tracking multiple subjects on each view, together with identifying the same persons across the two complementary views frame by frame, as shown in Fig 1. For this we basically need to accurately associate subjects between the two views at any time and between different frames along each video. Prior works (Xu et al. 2016; 2017) on multi-view object tracking usually assume the multiple horizontal views from different angles, e.g., frontal, side and back views and appearance consistency can still be applied for associating subjects. Differently, in this paper we adopt a top view from

[*]Co-corresponding authors.

a high altitude where view direction is largely perpendicular to the ground, as shown in Fig. 1(a), to capture the subjects' global trajectories. This leads to completely inconsistent appearance and motion features between the two complementary views: on the top view, each subject is largely a small dark region with only the head top and two shoulders visible.

We propose a new joint optimization model to address the proposed collaborative tracking problem. Specifically, we split the video in each view into short clips with equal length and extract the tracklets in each clip, which we refer to as *single-view tracklets*. The tracklets from adjacent clips in both two views are then used for establishing the spatial-temporal data association, resulting in the *cross-view short trajectories*. We formulate the multi-clip cross-view data association as a joint optimization model and solve it by a constrained integer programming algorithm. In this model, the single-view data similarity on each video is based on subjects' appearance and motion consistency while the cross-view data similarity is built by the spatial and appearance reasoning. We finally stitch the short trajectories over time to get the *cross-view long trajectories* as the final tracking results. In the experiments, we collect a new dataset for performance evaluation and the experimental results verify the effectiveness of the proposed method.

The main contributions of this paper are: 1) This is the first work to address the multiple human tracking by combining the complementary top and horizontal views, which can simultaneously capture the global trajectories and local detailed appearances of the subjects. 2) We build a new joint optimization model to associate the subjects across views as well as over time and solve it by constrained integer programming. 3) We collect a new dataset of top-view and horizontal-view videos for performance evaluation[1].

2 Related Work

Multiple object tracking. In general, multiple object tracking (MOT) can be divided into offline tracking and online tracking. The former takes MOT as an optimization problem of global data association (Tang et al. 2016; 2017; Wang et al. 2016) and is limited to offline applications. The latter only uses the information on the current frame and previous frames (Xiang, Alahi, and Savarese 2015; Zhu et al. 2018), which is suitable for real-time applications but may not handle well long-term occlusions or mis-detections. Different features are used for data association in MOT. Most widely used are appearance and motion features. Color histogram is a popular feature representation for appearance in MOT (Dehghan, Assari, and Shah 2015; Tang et al. 2016). Recently, deeply learned appearance features are also used in tracking (Lealtaixe, Cantonferrer, and Schindler 2016; Chu et al. 2017; Zhu et al. 2018). Both linear and nonlinear models have been used for representing motion features in MOT. While linear motion models assume a linear movement with constant velocity across frames (Zamir, Dehghan, and Shah 2012; Dehghan, Assari, and Shah 2015; Ristani and Tomasi 2018), nonlinear motion models may lead to more accurate predictions (Yang and Nevatia 2012a; 2012b). However, in this paper the consistency of either appearance or motion features are poor across the top and horizontal views.

Multi-view multi-object tracking. Also related to our work is the previous research on multi-view MOT. Some of them focus on excavating more information from multiple views for tracking, such as geometrical relations based tracking methods (Ayazoglu et al. 2011) and reconstruction-assistant trackers (Hofmann, Wolf, and Rigoll 2013). Many others focus on new problem formulations and solutions. For example, Fleuret et al. (2008) propose a generative model with dynamic programming for tracking. Liu (2016) proposes a multi-view method for tracking people in crowded 3D scene. More recently, Xu et al. (2016; 2017) integrate more semantic attributes, e.g., human postures and actions besides appearance and motion features, for cross-view object tracking. However, multi-views in these methods are actually all horizontal views, but with different view angles. This way, most of these multi-view MOTs still use appearance and motion matching to infer the cross-view data association. As mentioned above, this paper aims to track and associate subjects across top and horizontal views, where cross-view appearance and motion consistency are poor.

Top view and horizontal view. Top view and horizontal view are two complementary views and the conjoint analysis of them have drawn much attention recently. Ardeshir and Borji (2016; 2018a) propose a method to identify camera wearers on a top-view video given the horizontal-view videos recorded by these wearable cameras. Similarly, given a horizontal video and a top-view video, Ardeshir and Borji (2018b) study how to identify the horizontal-view camera holder in the top-view video, and re-identify the subjects present in both the horizontal- and top-view videos. Han et al. (2019) exploit the spatial distribution of the subjects to match all the subjects between the top- and horizontal-views, and builds the subject association across the two views. To the best of our knowledge, this paper is the first to track multiple subjects across the top and horizontal views by considering subject association both across views and over time.

3 The Proposed Method

3.1 Overview

Given a pair of temporally-aligned videos that are taken from the top view and horizontal view, respectively, we first synchronously split these two videos into short clips with the same length, e.g., 10 frames. In each video clip, we extract a set of subject tracklets using a simple overlap criteria: detected subjects (in the form of bounding boxes) with good overlap, e.g., higher than 50%, between two adjacent frames, are connected to form tracklets. We discard the tracklets that are overly short, e.g., traversing less than 3 frames. We refer to the resulting tracklets as *single-view tracklets* since they are extracted from the top-view and horizontal-view videos, respectively. We then conduct cross-view cross-clip data association for these single-view tracklets. More specifically, as shown in Fig. 2, the single-view tracklets from two adjacent clips in two views are fed into corresponding cluster

[1] https://github.com/HanRuize/CVMHT

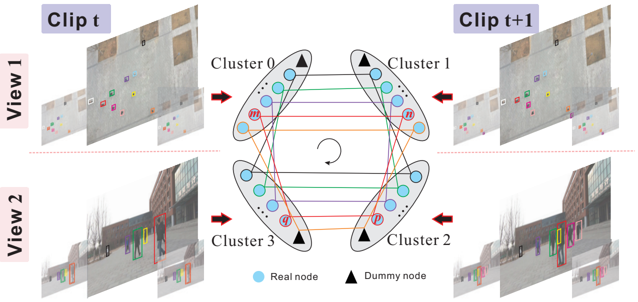

Figure 2: An illustration of subject association between consecutive clips and across two views. Solid triangle in each cluster represent its dummy node.

as nodes. We then establish the subject association between clips and across views by using a joint optimization function. We refer to the generated tracking trajectories between two clips and across two views as *cross-view short trajectories*. Finally, we stitch the short trajectories by considering the frame overlap over time and obtain the *cross-view long trajectories* as the final tracking results.

3.2 Problem Formulation

The single-view tracklets from two adjacent clips in two views constitute four clusters 0, 1, 2, and 3, as shown in Fig 2. Each (real) node in a cluster represents a single-view tracklet in respective clip/view. We define e_{mn}^{i} as the binary variable for the edge between the node m in cluster i and the node n in cluster $(i+1)|4$, where $|$ is the modulo operation. For the four clusters shown in Fig 2, cluster $(i+1)|4$ is adjacent to cluster i in a clockwise order. This way, edge connection is only considered between tracklets from the same view or from the same clip. Besides, the edge connection is intended to be unidirectional and we define c_{mn}^{i} as the weight of edge e_{mn}^{i}. Edge weight reflects the similarity of two tracklets over time or across views and we will elaborate on its constructions later.

The energy function of the problem can be formulated as

$$\arg\max_{\mathbf{e},\mathbf{d}} \quad \sum_{i=0}^{3}\Big(\sum_{n=1}^{N_{i'}}\sum_{m=1}^{N_i} c_{mn}^{i}\cdot e_{mn}^{i} + c_0\cdot d^i\Big), \qquad (1)$$

where N_i denotes the number of real nodes (tracklets) in cluster i, and $i' = (i+1)|4$. Other than real nodes, we also add a dummy node (Dehghan, Assari, and Shah 2015) for each cluster, which can be connected to multiple nodes (both real and dummy node) in other clusters[2]. The variable d^i counts the number of edges connected to the dummy node in cluster i, which can take any nonnegative integer value.

This is an mixed integer programming (MIP) problem and we further consider three constraints:

Constraint 1 limits that there is at most one edge between 1) nodes in cluster i and a (real) node in cluster i', 2) a (real) node in cluster i and nodes in cluster i', $i' = (i+1)|4$:

$$\sum_{m=1}^{N_i} e_{mn}^{i} \le 1, \sum_{n=1}^{N_{i'}} e_{mn}^{i} \le 1. \qquad (2)$$

Constraint 2 ensures that resulting edge connections form loops among four clusters:

$$e_{mn}^{i} + e_{np}^{i'} + e_{pq}^{i''} \le 2 + e_{qm}^{i'''}, \qquad (3)$$

where $i' = (i+1)|4$, $i'' = (i'+1)|4$, and $i''' = (i''+1)|4$. **Constraint 3** ensures that a same number of K nodes (including the number of real nodes and the value of dummy node) are selected for association from each cluster:

$$\sum_{q=1}^{N_{i'''}}\sum_{m=1}^{N_i} e_{qm}^{i'''} + \sum_{m=1}^{N_i}\sum_{n=1}^{N_{i'}} e_{mn}^{i} + d^i = 2K. \qquad (4)$$

With above three constraints, we solve the integer programming to obtain edge sets between different clusters, which provide the desired subject association across two views and two clips. Next we define the edge weights, i.e., the similarity between tracklets.

3.3 Cross-View Data Association

Spatial reasoning. For a pair of synchronized clips from the top view and horizontal view, respectively, we first use a spatial distribution based method (Han et al. 2019) to get the association results of the subjects frame by frame. Specifically, let $\mathcal{T} = \{T_m\}_{m=1}^{M}$ be the collection of M subjects detected on a frame in one view, and $\mathcal{H} = \{H_q\}_{q=1}^{Q}$ be the collection of Q subjects detected on the corresponding frame in the other view. The result of cross-view subject association is to identify all the matched subjects between $\mathcal{T}$ and $\mathcal{H}$. The association results can be expressed as $\mathrm{A}(T_m, H_q) = 1$ if T_m, H_q represent the same person, while $\mathrm{A}(T_m, H_q) = 0$ otherwise. Given two single-view tracklets $\mathbf{T}_m, \mathbf{H}_q$ extracted from synchronized clips in two views clips, respectively, we identify their temporally overlapped L frames, on each of which we have subjects T_m^l and H_q^l, $l = 1, \cdots, L$. The similarity between $\mathbf{T}_m, \mathbf{H}_q$ can be calculated by

$$\hat{c}_{qm}^{i} = \frac{\sum_{l=1}^{L}\mathrm{A}(T_m^l, H_q^l)}{\max(|\mathbf{T}_m|, |\mathbf{H}_q|)}, \quad i = 1, 3 \qquad (5)$$

where $|\mathbf{T}_m|$ denotes the length of the tracklet $\mathbf{T}_m$ (number of contained subjects) and $i = 1, 3$ indicate the two sets of edges in Fig. 2 that reflect cross-view data association.

Appearance reasoning. In addition to the spatial reasoning, we also consider the appearance similarity. In consideration of the difficulty of data annotation and algorithm efficiency, we design a one-short learning based Siamese network (Koch, Zemel, and Salakhutdinov 2015) to measure the similarity between two tracklets. The detailed Siamese network structure is shown in Fig. 3. Given two tracklets, each of which consisting of a sequence of image patches, we

[2] We add dummy nodes to handle the cases of misdetection, occlusion and out of view. For example, a match between a real node in cluster 1 and the dummy node in cluster 2 indicates that the tracked subject underlying the real node in cluster 1 is not detected/tracked in cluster 2.

first compute the mean image of these patches in each tracklet. As shown in Fig. 3, the mean images of two tracklet are fed into the Siamese network, which is composed of three convolution layers and two connection layers, and uses the pair-based contrastive loss function (Hadsell, Chopra, and Lecun 2006). By calculating the Euclidean distance of the two streams' output vectors, we obtain the similarity score $\overline{c}^i_{qm}$ of the two tracklets $\mathbf{T}_m, \mathbf{H}_q$ for $i = 1, 3$.

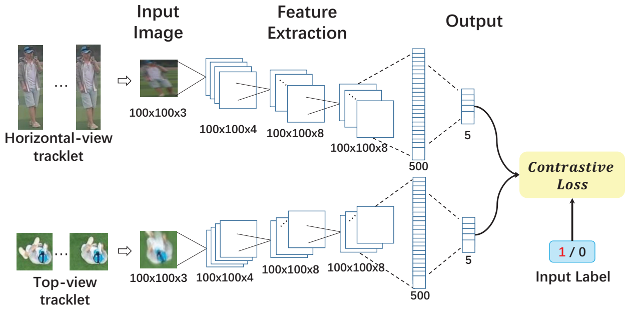

Figure 3: Siamese neural network structure for measuring the cross-view appearance similarity.

3.4 Cross-Clip Data Association

Appearance consistency. To measure the appearance similarity of single-view subjects, we use color histogram as the representation. We first compute the histogram for all the subjects of a single-view tracklet. Then the median of all the histogram is selected as the appearance descriptor of the tracklet (Dehghan, Assari, and Shah 2015). Let $\varphi(\mathbf{T}_m)$ and $\varphi(\mathbf{T}_n)$ be the appearance descriptor of the tracklets $\mathbf{T}_m$ and $\mathbf{T}_n$ respectively. We use Histogram Intersection (Grauman 2005) to calculate the appearance similarity between them:

$$\hat{c}^i_{mn} = \mathrm{K}(\varphi(\mathbf{T}_m), \varphi(\mathbf{T}_n)), \quad i = 0, 2 \tag{6}$$

where K denotes a kernel function (Grauman 2005) and $i = 0, 2$ indicate the two sets of edges in Fig. 2 that reflect cross-clip data association.

Motion consistency. We also consider the motion consistency for single-view subject association. We use the constant velocity motion model to predict motion consistency as in most previous MOT tracking methods. Given two tracklets, the forward and backward deviation error δ_f and δ_b can be computed by the motion model. The deviation error $\delta = \alpha(\delta_f + \delta_b)$ is to measure the difference of the two tracklets $\mathbf{T}_m$ and $\mathbf{T}_n$, where α is a scaling factor. We convert the errors into similarity by $\overline{c}^i_{mn} = e^{-\delta}$, which takes the value in $[0, 1]$ for $i = 0, 2$.

3.5 Tracking Framework

Given two cross-view tracklets $\mathbf{T}_m, \mathbf{H}_q$, we can use the above methods to get the spatial similarity score $\hat{c}^i_{qm}$ and appearance similarity score $\overline{c}^i_{qm}$, respectively ($i = 1, 3$). We use the linear combination to compute the final edge weight c^i_{qm} as

$$c^i_{qm} = w_1 \hat{c}^i_{qm} + (1 - w_1)\overline{c}^i_{qm}, \quad i = 1, 3 \tag{7}$$

where w_1 is a pre-set parameter. Similarly, given two single-view tracklets $\mathbf{T}_m, \mathbf{T}_n$ from different clips, we calculate the edge weight c^i_{mn} $(i = 0, 2)$ by

$$c^i_{mn} = w_2 \hat{c}^i_{mn} + (1 - w_2)\overline{c}^i_{mn}, \quad i = 0, 2 \tag{8}$$

where w_2 is a pre-set parameter, and $\hat{c}^i_{mn}, \overline{c}^i_{mn}$ are the tracklet similarity scores calculated by the appearance color histogram feature and motion model, respectively.

The proposed MOT tracking method can be summarized in Algorithm 1.

Algorithm 1: Complementary-View MOT:

```
Input: $V_T$, $V_H$: Top-view and horizontal-view videos;
       parameters $w_1$, $w_2$, $c_0$.
Output: Tracked subject bounding boxes with ID
        numbers.
Split the cross-view videos into $T$ clips respectively.
for $t = 1 : T$ do
    Detect the subjects then extract the single-view
      tracklets $\mathbf{T}^t_m$ ,$\mathbf{H}^t_q$, $\mathbf{T}^{t+1}_n$, $\mathbf{H}^{t+1}_p$ in clip $t$ and $t + 1$.
    Calculte the tracklets similarity scores then
      compute the edge weight $c$ by Eq. (7) and Eq. (8).
    Solve $\mathbf{e}$ by Eq. (1) to get cross-view
      short-trajectories.
    if $t = 1$ then
        if $e_{qm} = 1$ then
            Assign the same ID numbers to the
              bounding boxes in the tracklets $\mathbf{T}^t_m, \mathbf{H}^t_q$.
        else
            Assign the incremental ID to the other ones.
    else
        if $e_{mn} = 1$ $(e_{pq} = 1)$ then
            Assign the ID number of $\mathbf{T}^t_m$ $(\mathbf{H}^t_q)$ to
              $\mathbf{T}^{t+1}_n$ $(\mathbf{H}^{t+1}_p)$, respectively.
        else if $e_{np} = 1$ then
            Assign the ID number of $\mathbf{T}^{t+1}_n$ to $\mathbf{H}^{t+1}_p$ .
        else
            Assign the incremental ID to the other ones.
return bounding boxes with ID numbers
```

4 Experiments

4.1 Dataset and Metrics

We do not find publicly available dataset with temporally synchronizing top-view and horizontal-view videos with ground-truth labeling for cross-view multiple object tracking. Therefore, we collect a new dataset by flying a drone with a camera to take top-view videos and mounting GoPro over the head of a person to take the horizontal-view videos for performance evaluation. Videos are taken at five sites with different background. The subjects are free to move or stop in the scene without any specific instructions and there may be random mutual occlusions between subjects. We

Table 1: Comparative results of different methods. IDP, IDR, IDF$_1$, IDS, MOTP, MOTA are standard MOT metrics. CVIDF$_1$ and CVMA are the new metrics for evaluating the cross-view MOT.

Method	IDP	IDR	IDF$_1$	IDS	MOTP	MOTA	CVIDF$_1$	CVMA
GMMCP	49.4	50.7	50.1	1172	**75.2**	79.3	17.9	27.2
MDP	65.8	68.4	67.1	723	**75.2**	**84.9**	33.7	38.1
DMAN	72.3	**77.2**	74.7	**311**	75.1	82.4	44.7	43.2
Ours	**77.6**	76.6	**77.1**	382	74.9	84.2	**84.0**	**78.3**

Table 2: Comparative results of different methods on the subsets of top-view videos and horizontal-view videos, respectively.

Method	Top view						Horizontal view					
	IDP	IDR	IDF$_1$	IDS	MOTP	MOTA	IDP	IDR	IDF$_1$	IDS	MOTP	MOTA
GMMCP	50.7	50.7	50.7	822	69.3	76.5	47.7	50.8	49.2	350	83.6	83.4
MDP	76.2	77.8	77.0	331	69.6	**86.3**	50.8	54.3	52.5	392	83.6	82.7
DMAN	**85.1**	**87.4**	**86.2**	**65**	**70.0**	85.8	54.8	61.9	58.1	**246**	82.7	77.4
Ours	79.5	80.9	80.2	115	69.2	84.0	**74.3**	**70.1**	**72.2**	267	**83.8**	**84.5**

manually synchronize these videos such that corresponding frames between them are taken at the same time. We then cut out 15 pairs of sequences with length from 600 to 1,200 frames as our dataset. We manually annotate the subjects in the forms of rectangular bounding boxes and ID numbers: the same subject across two views are labeled with the same ID number.

We apply standard MOT metrics for evaluating the tracking performance (Lealtaixé et al. 2015), including multi-object tracking precision (MOTP) and multi-object tracking accuracy (MOTA). One key task of the proposed collaborative MOT is to identify/track the same subject across two views. Therefore, we also select four ID-based metrics, i.e., ID precision (IDP), ID recall (IDR), ID F$_1$ measure (IDF$_1$) and ID switches (IDS) for performance evaluation.

Besides, to fully measure the performance of the proposed cross-view multiple object tracking, we define the following new metrics. First, the cross-view ID F$_1$ metric – CVIDF$_1$ is defined as

$$\mathrm{CVIDF}_1 = \frac{2\mathrm{CVIDP} \times \mathrm{CVIDR}}{\mathrm{CVIDP} + \mathrm{CVIDR}}, \quad (9)$$

where CVIDP and CVIDR denote the cross-view subject matching precision and recall, respectively. We further define the cross-view matching accuracy – CVMA, as

$$\mathrm{CVMA} = 1 - \left(\frac{\sum_t \mathrm{m}_t + \mathrm{fp}_t + 2\mathrm{mme}_t}{\sum_t \mathrm{g}_t} \right), \quad (10)$$

where m_t, fp_t, mme_t are the numbers of misses, false positives, mismatch pairs of cross-view object matching at time t, and g_t is the total number of objects in both top and horizontal views at time t.

4.2 Experiment Setup

We implement the main program in Matlab and on a desktop computer with an Intel Core i5 3.4GHz CPU, and the Siamese network for cross-view appearance similarity measurement is implemented on GPU. We use the general YOLOv3 (Redmon et al. 2016) detector to detect subjects in the form of bounding boxes in both top- and horizontal-view videos. For top-view subject detection, we fine-tune the network using 600 top-view human images. For training the Siamese based network, given a subject detected in the top-view frame, we use it paired with its corresponding subject in horizontal view as a positive sample, and paired with other subjects as a negative training sample. Note that all the training data have no overlap with our test dataset. The pre-specified parameters w_1, w_2 and c_0 are set to 0.3,0.5 and 0.3, respectively. The mixed integer programming problem is solved by the MIP solver of *cplex*.

We choose three multiple object trackers, i.e., GMMCP (Dehghan, Assari, and Shah 2015), MDP (Xiang, Alahi, and Savarese 2015), and DMAN (Zhu et al. 2018) as the compared methods. Among them, GMMCP uses the color histogram based appearance feature and constant velocity model based motion feature for data association, which is same with our single-view data association. Both MDP and DMAN are single object tracker based online MOT approaches, where DMAN learns deep appearance features for data association. All the comparison trackers are implemented to track on the top-view and horizontal-view videos separately, initialized with the ground-truth subjects and labels on the first frame. Note that, for fair comparison, we use the same subject detector for all the methods, including the proposed method and these comparison methods. In practice, we did not find existing methods with code that can handle the proposed cross-view multiple objects tracking. One seemingly related work is Xu et al. (2016; 2017) for cross-view multiple people tracking. However, we could not include it directly into comparison because it assumes multiple horizontal or sloped views and can still use pose/appearance features for data association, which is not applicable to the top-view videos used in this paper.

4.3 Results

We evaluate the proposed method on our dataset. We evaluate the single-view MOT results using the standard MOT metrics. We show the results of different trackers in Table 1(left). We find that although using the same features as

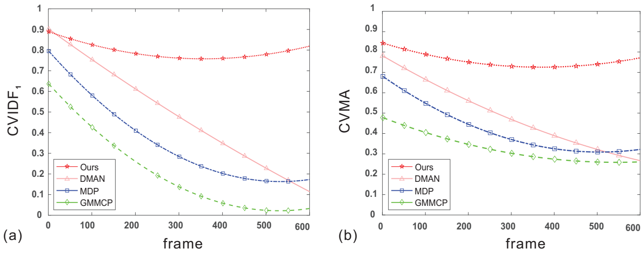

Figure 4: Cross-view subject matching results over time.

GMMCP in single-view data association, our method outperforms GMMCP by a wide margin in the ID-related metrics. The proposed method achieves the comparable performance with the state-of-the-art DMAN tracker.

Moreover, we divide the dataset into top-view and horizontal-view videos and evaluate the MOT performance, respectively. As shown in Table 2, we can first find that the tracking results in top view shows better performance compared to horizontal view. This is due to the mutual occlusions which frequently appear in the horizontal view but are rare in top view. In this case, the proposed method achieves better tracking results in horizontal view with the assistance of tracking in top view. As for top-view tracking, the proposed method outperforms GMMCP by a large margin with the cross-view joint optimization for data association. The result verifies that the top view and horizontal view are complementary in improving the tracking accuracy.

Besides the standard MOT metrics, we also compare the cross-view MOT performance using CVIDF_1 and CVMA as shown in Table 1(right). While the selected comparison methods can only handle the single-view tracking, we provide the ground-truth ID of each subject on the first frame of both views. This way, the tracking on each view actually propagate the subject IDs to later frames and from the IDs, we can match the subjects across views over time. We can find that all the three compared trackers produce very poor results because the cross-view subject matching will fail once a target is lost in tracking in any one view. The proposed method can produce an acceptable CVIDF_1 and CVMA results of 84.0% and 78.3%, respectively. To better evaluate the cross-view MOT, we show the average CVIDF_1 and CVMA scores over time[3]. As shown in Fig. 4, we can find that the performance of all the single-view trackers show a downward trend. Our method shows a steady scores with no performance decrease over time.

Actually, MOT is expected to maintain the subject ID after the initialization on the first frame. We evaluate the ID accuracy over time based on the consistency of current and initial ID numbers. As shown in Fig. 5, DMAN performs best in the top view. However, in the horizontal view, our approach gets better performance than other trackers as the frame number increases. This is because our approach can re-identify the horizontal-view subjects by associating to the tracked subjects in the top view.

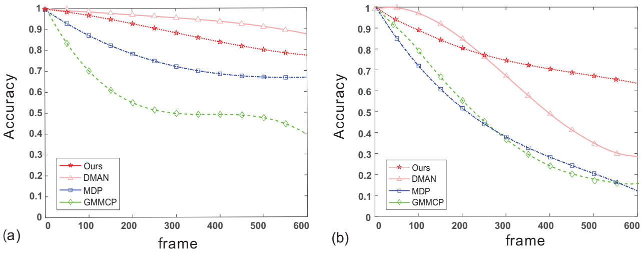

Figure 5: ID consistency accuracy over time in top view (a) and horizontal view (b).

4.4 Ablation Studies

Features for similarity measurements. We study the influence of using different similarity measures. As shown in Table 3, 'w/o sv-App' and 'w/o Motion' denote the proposed method without the appearance and motion features in single-view data association, respectively. And 'w/o cv-App' and 'w/o Spatial' denote the proposed method without the appearance and spatial features in cross-view data association, respectively. The results in the first and second rows show that using only one of the single-view data association features cannot achieve performance as good as the proposed method that combines both two. However, using any one of the two types of features can achieve the acceptable results. By comparing the results in the third and fourth rows with the last row, we can see that the proposed method using only appearance features produces poor results in cross-view subjects matching, which verifies the appearance is not very useful for human association across two views. Fortunately, spatial reasoning provides more accurate information than appearance features in cross-view data association.

Table 3: Comparative study of using different association features. 'sv-App' and 'w/o Motion' denote the appearance and motion features in single views. 'cv-App' and 'Spatial' denote the appearance and spatial features in cross views.

Features	IDF_1	MOTA	CVIDF_1	CVMA
w/o sv-App	71.4	84.0	83.4	77.6
w/o Motion	71.1	83.9	81.9	76.4
w/o cv-App	72.1	84.1	82.6	76.5
w/o Spatial	62.6	81.4	15.5	25.9
Ours	**77.1**	**84.2**	**84.0**	**78.3**

Parameters selection. There are three free parameters in the proposed method: w_1, w_2 in Eq. (8), Eq. (9) and c_0 in Eq. (1). We examine their influence to the final tracking performance. Table 4 reports the results by varying one of these three parameters while fixing the other two. We can see that the final tracking performance, including standard MOT metrics, i.e., IDF_1 and MOTA or cross-view MOT metrics, i.e., CVIDF_1 and CVMA, are not very sensitive to the selected values of these three parameters.

[3] In this experiment, we only consider the first M frames of each video, where M is the minimum length of all the videos. This way, we can compute the average CVIDF_1 and CVMA scores of all the videos frame by frame.

Table 4: Results by varying values of w_1, w_2 and c_0.

w_1	IDF$_1$	CVIDF$_1$	MOTA	CVMA	w_2	IDF$_1$	CVIDF$_1$	MOTA	CVMA	c_0	IDF$_1$	CVIDF$_1$	MOTA	CVMA
0.2	74.9	82.7	84.2	77.4	0.4	76.2	83.7	84.1	78.1	0.2	77.0	81.4	84.3	77.1
0.3	77.1	84.0	84.2	78.3	0.5	77.1	84.0	84.2	78.3	0.3	77.1	84.0	84.2	78.3
0.4	77.5	82.8	84.3	77.8	0.6	76.8	83.6	84.2	78.2	0.4	73.3	83.2	84.3	77.5

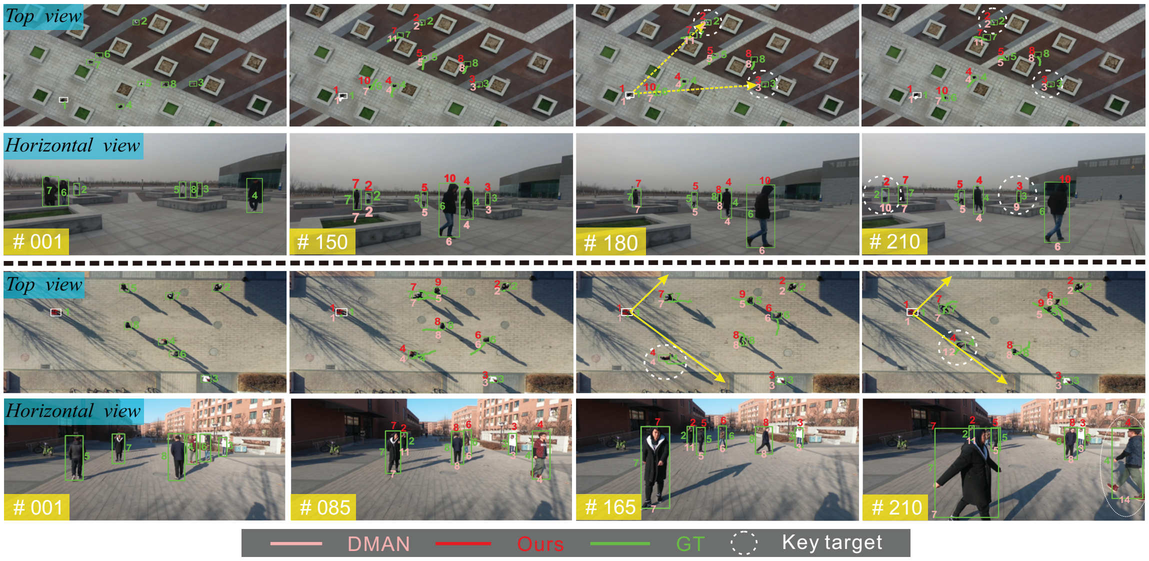

Figure 6: Case analysis of long-term occlusion (top) and out-of-view (bottom) scenario.

4.5 Discussion

Occlusion and out of view. In horizontal-view videos, it is common to have subjects with mutual occlusion and being out-of-view. In this case, existing online trackers, e.g., DMAN can not associate the long-term lost subjects when they reappear in the view. Two examples are shown in Fig. 6. The top two rows show the case of mutual occlusions. From the top view at frame #180, we can find that two subjects (ID number 2, 3) are occluded by others and DMAN switches the ID of them when they reappear in the field of view at frame #210. Our method keeps the original ID number. Similarly, we focus on the key subject (ID number 4) which goes out of view at frame #165 in the horizontal view. We can find that this subject is reassigned to a new ID number by DMAN. Our approach gets the original ID number of the target, which is consistent to its ID number in the top view.

Table 5: Time performance of each component (sec/frame).

Component	tracklet	sv-sim	cv-sim	solution
Time	0.16	0.26	0.06	0.10
Proportion	27.5%	44.7%	9.6%	18.2%

Speed analysis. As shown in Table 5, we record the running time taken by each component of the proposed method. In this table, 'tracklet' denotes the single-view tracklet construction, 'sv-sim' and 'cv-sim' denote the single-view and cross-view data similarity computation, respectively. 'solution' denotes the step of solving the optimization problem. We can find that the single-view similarity computation takes 44.7% of the total running time. The total time taken by cross-view similarity computation and final optimization is similar to that taken by the tracklet construction. This demonstrates the time efficiency of the proposed cross-view similarity computation and optimization. We further compare the speed of our approach with three comparison MOT trackers. From Table 6, we can find that our method runs faster than MDP and DMAN. Note that, our main program is implemented in Matlab with CPU and it can get much faster with the multithreading or GPU acceleration.

Table 6: Running speeds of different methods (frames/sec).

Type	Ours	GMMCP	MDP	DMAN
Speed	1.74	2.44	0.77	0.91

5 Conclusion

In this paper, we have studied a new problem to track multiple subjects in the complementary top and horizontal views. We formulate a joint optimization problem for the subjects association in the time-space domain. In each view, the data similarity over time is represented by appearance and motion features. Across the top and horizontal views, the data similarity is represented by spatial and appearance reasoning. We solve the optimization problem by a constrained mixed integer programming algorithm for final cross-view multiple subject tracking. We collect a new complementary-view dataset, as well as manually label the ground truth, for performance evaluation. Experimental results on this dataset demonstrate that the proposed MOT method can collaboratively track subjects in both views, by providing both the global trajectory and the local details of the subjects.

6 Acknowledgments

We thank Sibo Wang, Shuai Wang, Chenxing Gong for their kind assistance in the collection and annotation of our dataset. This work was supported, in part, by the NSFC under Grants U1803264, 61672376, 61671325, 61572354.

References

Ardeshir, S., and Borji, A. 2016. Ego2top: Matching viewers in egocentric and top-view videos. In *European Conference on Computer Vision*, 253–268.

Ardeshir, S., and Borji, A. 2018a. Egocentric meets top-view. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41(6):1353–1366.

Ardeshir, S., and Borji, A. 2018b. Integrating egocentric videos in top-view surveillance videos: Joint identification and temporal alignment. In *European Conference on Computer Vision*, 300–317.

Ayazoglu, M.; Li, B.; Dicle, C.; Sznaier, M.; and Camps, O. I. 2011. Dynamic subspace-based coordinated multicamera tracking. In *IEEE International Conference on Computer Vision*, 2462–2469.

Chu, Q.; Ouyang, W.; Li, H.; Wang, X.; Liu, B.; and Yu, N. 2017. Online multi-object tracking using CNN-based single object tracker with spatial-temporal attention mechanism. In *IEEE International Conference on Computer Vision*, 4846–4855.

Dehghan, A.; Assari, S. M.; and Shah, M. 2015. GMMCP Tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, 4091–4099.

Feng, W.; Han, R.; Guo, Q.; Zhu, J.; and Wang, S. 2019. Dynamic saliency-aware regularization for correlation filter-based object tracking. *IEEE Transactions on Image Processing* 28(7):3232–3245.

Fleuret, F.; Berclaz, J.; Lengagne, R.; and Fua, P. 2008. Multicamera people tracking with a probabilistic occupancy map. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30(2):267–282.

Grauman, K. 2005. The pyramid match kernal : discriminative classification with sets of image features. In *IEEE International Conference on Computer Vision*, 1458–1465.

Guo, Q. Han, R.; Feng, W.; Chen, Z.; and Wan, L. 2020. Selective spatial regularization by reinforcement learned decision making for object tracking. *IEEE Transactions on Image Processing*.

Hadsell, R.; Chopra, S.; and Lecun, Y. 2006. Dimensionality reduction by learning an invariant mapping. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1735–1742.

Han, R.; Zhang, Y.; Feng, W.; Gong, C.; Zhang, X.; Zhao, J.; Wan, L.; and Wang, S. 2019. Multiple human association between top and horizontal views by matching subjects' spatial distributions. In *arXiv preprint arxiv:1907.11458*.

Han, R.; Guo, Q.; and Feng, W. 2018. Content-related spatial regularization for visual object tracking. In *IEEE International Conference on Multimedia and Expo*.

Hofmann, M.; Wolf, D.; and Rigoll, G. 2013. Hypergraphs for joint multi-view reconstruction and multi-object tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, 3650–3657.

Koch, G.; Zemel, R.; and Salakhutdinov, R. 2015. Siamese neural networks for one-shot image recognition. In *International Conference on Machine Learning Workshop*.

Lealtaixé, L.; Milan, A.; Reid, I.; Roth, S.; and Schindler, K. 2015. Motchallenge 2015: Towards a benchmark for multi-target tracking. In *arXiv preprint arxiv:1504.01942*.

Lealtaixe, L.; Cantonferrer, C.; and Schindler, K. 2016. Learning by tracking: Siamese CNN for robust target association. In *IEEE Conference on Computer Vision and Pattern Recognition*, 418–425.

Liu, X. 2016. Multi-view 3d human tracking in crowded scenes. In *AAAI Conference on Artificial Intelligence*, 3553–3559.

Luo, W.; Stenger, B.; Zhao, X.; and Kim, T. 2015. Automatic topic discovery for multi-object tracking. In *AAAI Conference on Artificial Intelligence*, 3820–3826.

Redmon, J.; Divvala, S. K.; Girshick, R. B.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 779–788.

Ristani, E., and Tomasi, C. 2018. Features for multi-target multi-camera tracking and re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 6036–6046.

Tang, S.; Andres, B.; Andriluka, M.; and Schiele, B. 2016. Multi-person tracking by multicut and deep matching. In *European Conference on Computer Vision*, 100–111.

Tang, S.; Andriluka, M.; Andres, B.; and Schiele, B. 2017. Multiple people tracking by lifted multicut and person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 3701–3710.

Wang, X.; Turetken, E.; Fleuret, F.; and Fua, P. 2016. Tracking interacting objects using intertwined flows. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38(11):2312–2326.

Wen, L.; Du, D.; Li, S.; Bian, X.; and Lyu, S. 2019. Learning non-uniform hypergraph for multi-object tracking. In *AAAI Conference on Artificial Intelligence*, 8981–8988.

Xiang, Y.; Alahi, A.; and Savarese, S. 2015. Learning to track: Online multi-object tracking by decision making. In *IEEE International Conference on Computer Vision*, 4705–4713.

Xu, Y.; Liu, X.; Liu, Y.; and Zhu, S. 2016. Multi-view people tracking via hierarchical trajectory composition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 4256–4265.

Xu, Y.; Liu, X.; Qin, L.; and Zhu, S. 2017. Cross-view people tracking by scene-centered spatio-temporal parsing. In *AAAI Conference on Artificial Intelligence*, 4299–4305.

Yang, B., and Nevatia, R. 2012a. Multi-target tracking by online learning of non-linear motion patterns and robust appearance models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1918–1925.

Yang, B., and Nevatia, R. 2012b. An online learned crf model for multi-target tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2034–2041.

Zamir, A. R.; Dehghan, A.; and Shah, M. 2012. GMCP-Tracker: Global multi-object tracking using generalized minimum clique graphs. In *European Conference on Computer Vision*, 343–356.

Zhu, J.; Yang, H.; Liu, N.; Kim, M.; Zhang, W.; and Yang, M. 2018. Online multi-object tracking with dual matching attention networks. In *European Conference on Computer Vision*, 379–396.