Integrating Appearance and Spatial-Temporal Information for Multi-Camera People Tracking

Wenjie Yang[1*§] Zhenyu Xie[1*] Yaoming Wang[1] Yang Zhang[2] Xiao Ma[2] Bing Hao[2]
[1]Shanghai Jiao Tong University [2]AI Lab, Lenovo Research
{13633491388,sp.sat,wang_yaoming}@sjtu.edu.cn, {zhangyang20,maxiao3,haobing1}@lenovo.com

Abstract

Multi-Camera People Tracking (MCPT) is a crucial task in intelligent surveillance systems. However, it presents significant challenges due to issues such as heavy occlusion and variations in appearance that arise from multiple camera perspectives and congested scenarios. In this paper, we propose an effective system that integrates both appearance and spatial-temporal information to address these problems, consisting of three specially designed modules: (1) A Multi-Object Tracking (MOT) method that minimizes ID-switch errors and generates accurate trajectory appearance features for MCPT. (2) A robust intra-camera association method that leverages both appearance and spatial-temporal information. (3) An effective post-processing module comprising multi-step processing. Our proposed system is evaluated on the test set of Track1 for the 2023 AI CITY CHALLENGE, and the experimental results demonstrate its effectiveness, achieving an IDF1 score of 93.31% and ranking 3rd on the leaderboard.

1. Introduction

Multi-Camera People Tracking (MCPT) is an emerging research area that aims to develop advanced computer vision algorithms and systems for tracking individuals across multiple cameras. The goal of MCPT is to accurately locate and track people in a given scene by integrating information from multiple cameras, which can provide different viewpoints and angles of the scene. The application of MCPT is wide-ranging and can include video surveillance, crowd management, social behavior analysis, and more.

Traditional single-camera people tracking methods often face various limitations, such as occlusion and camera viewpoint dependency. In contrast, MCPT overcomes these limitations by fusing information from multiple cameras, enabling the tracking of individuals across different camera views. A typical MCPT pipeline consists of pedestrian detection, re-identification (ReID), Multiple Object Tracking (MOT) and Intra-camera Association (ICA). First, the pedestrian detector outputs pedestrian locations and feature vectors are extracted via ReID module. Based on the detection results and ReID features, single-view trajectories are generated by MOT module for each camera. Finally, ICA module associate trajectories from different views and generates global identities.

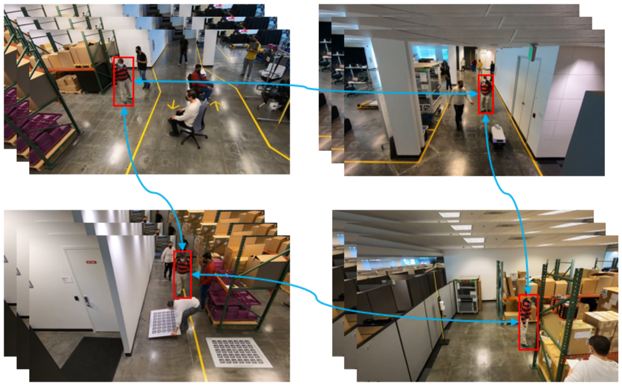

Figure 1. Illustration of MCPT task. People with the same identity in different cameras should be matchted.

In recent years, MCPT has gained significant attention from both academic and industrial communities. Many researchers have proposed various MCPT methods and systems that can achieve competitive accuracy and robustness in challenging scenarios. However, MCPT is still a challenging research area, and we observe following problems that need to be addressed in this challenge.

1. Existing Multiple Object Tracking methods can hardly distinguish mutually occluded human bodies, which often leads to frequent ID-switches.

[*] Equal contribution.

[§] This work was done when Wenjie Yang was a research intern at AI Lab, Lenovo Research.

2. In intra-camera matching, it is difficult to distinguish people with similar appearance or occluded people via ReID features, thus results in ID-switches or ID antinomy (will be explained in 3.2.2).
3. Most tracking-by-detection methods rely heavily on the quality of detection. However, false positive detection results may appear in complex scenes and influence the tracking results.

Faced with the above-mentioned problems, we design a MCPT system that minimizes ID-switches leveraging spatial-temporal and appearance information and removes false positive detection results by multi-body-level detection. Specifically, to prevent ID-switches in single-view trajectories, we propose a new MOT method specially tailored for scenarios where severe occlusion between different people is present. Besides, we adopt a new strategy to calculate more effective appearance (ReID) features of trajectories based on detection confidence, which can also help the correction of identities.

To overvome the limitation of ReID features when similar people appears in the same scene, we leverage multiple sources of spatial-temporal information during intra-camera association to reduce ID-switches between multi-camera trajectories and ID antinomy in single-camera trajectories. Finally, we design an effective post-processing module to further remove ID antinomy and false negative detection results. False negative detection results possibly caused by occlusion are also corrected by trajectory compensation and interpolation.

The main contributions of this paper are summarized as follows:

- We design a MOT method for MCPT task. Our MOT could minimize the ID-switch errors and obtain more accurate appearance feature for trajectories.
- We propose an intra-camera association approach which leverages both appearance information and multiple sources of spatial-temporal information. Our association approach requires only one single step as it can simultaneously handle broken trajectories from single camera and multi-camera trajectories.
- We propose a post-processing module to further refine the MCPT results. The designed module can remove ID antinomy and false positive detections, as well as compensating for missing trajectories.

2. Related work

2.1. Pedestrian Detection

Object detection is one of the most fundamental task in computer vision, which aims at localizing and classifying accurately objects of some categories in images and videos. Pedestrian detection, a special branch of object detection, is also an important and challenging task, especially in human-centric assignments. Generally, recent object detection methods are often based on CNNs and can be classified into two categories: single-stage detectors and two-stage detectors.

Single-stage detectors predict object position and category after a single feature sampling and extraction using predefined box scales. You Only Look Once (YOLO) [21] and its follow-up work including YOLOv3 [22], YOLOv5 [12] and YOLOX [8], are among the most popular single-stage detectors. YOLO utilize a backbone inspired from GoogleLeNet [27] and transform the task into a classification problem. It achieves real-time performance while maintaining high prediction precision.

Two-stage detectors generate arbitrary region proposals in the first stage and regress the location and category probability in the second stage. One of the most representative two-stage object detection methods is Faster-RCNN [23]. Faster-RCNN introduces a region proposal network (RPN) to generate region of interest (ROI) with greatly improved efficiency. Afterwards, ROI pooling outputs a series of fixed-size feature maps according to the image feature and the ROIs. Finally, Faster-RCNN outputs refined bounding box location and object category after regression and softmax operation. Mask-RCNN [9] further extends Faster-RCNN for instance segmentation by adding a parallel mask prediction head.

In addition, recent works also propose many transformer-based object detection methods, such as DETR [44], YOLOS [6], Swin Transformer [19] and VitDet [18]. These methods propose new strategies on object query, attention mechanisms, label assignment, feature matching, etc. The main advantage is that global image feature is better captured compared to CNN-based methods.

2.2. Re-identification

As an indispensable component of understanding human behaviours, re-identification (ReID) targets at retrieval of same person regardless of spatial variance. A typical ReID system is composed of three parts: Feature Representation Learning, Deep Metric Learning and Ranking Optimization. Previous studies exploits a variety of feature learning strategies, including global feature [41], local feature [35], auxiliary feature [26], video feature [31], etc. As for Deep Metric Learning in ReID, different loss functions and training strategies are developed to guide the feature representation learning. For example, identity loss, verification loss and triplet loss are three widely adopted loss functions, along with their variants. Ranking optimization aims at the optimization of the ranking order in a ranking list, typically via re-ranking [30, 42, 43] or rank fusion [36, 40] methods.

2.3. Multiple Object Tracking

Multiple Object Tracking (MOT) plays an important role in understanding videos. It is often an indispensable module in Multi-Camera Multi-Target tracking (MCMT). Existing methods can be partitioned into tracking-by-detection methods and joint-detection-tracking methods.

Tracking-by-detection is the mainstream paradigm to achieve high MOT performance. Tracking-by-detection methods first detect multiple objects and then associate objects from different timestamps. SORT [2] is a simple yet efficient framework for MOT task. In SORT, a Kalman Filter is employed for each trajectory and to predict object position. Then SORT calculate a cost matrix based on the overlap of predicted box and detected box and solve the assigning problem using Hungarian algorithm. DeepSORT [34] further introduces a matching cascade to leverage appearance information and motion clues. However, real-time performance can hardly be accomplished due to two-stage feature extraction. Bytetrack [37] removes the background from low-scoring detection results to uncover true objects, thereby improving the tracking performance on occluded and small target. As appearance features are not involved in association, the tracking of Bytetrack is highly dependent on the detection performance.

Joint-detection-tracking is also a feasible solution to improve inference speed. For example, FairMOT [39] simultaneously estimates pixel-wise Re-ID features and objects location based on a single backbone.

2.4. Multi-Camera Multi-Target Tracking

Multi-Camera Multi-Target Tracking (MCMT) has gained increasing attention in research field for its prospect of application for surveillance, health and ecological purpose. Previous works [4, 11, 33] adopt graph-based approaches to associate detections from multiple image flows. Some approaches [24, 28] also integrate Re-ID features in the intra-camera association and can be applied in scenarios without overlapping areas. Other methods leverage camera calibration to provide accurate spatial information during intra-camera association. For instance, Ran and Yael [5] adopt homography matrix to transform head locations and associate head detections across different camera views. Recent approaches [3, 38] utilize 3D pose estimation techniques and camera intrinsic and extrinsic parameters to precisely locate human joints in 3D space, which is considered as high quality spatial information in the association step. However, there methods heavily rely on the performance of 3D pose estimators and the accuracy of camera calibration.

3. Method

The proposed MCPT system is shown in Fig 2, which includes pedestrian detection, Re-ID module, multi-object tracking (MOT), intra-camera trajectory association (ICA) and post-processing module. In this section, we mainly elaborate the MOT, intra-camera trajectory association and post-process in detail, while the rest is documented in Implementation Details.

3.1. MOT

Once obtained detection results and corresponding ReID features, we perform MOT to associate targets throughout the video frames. Considering the final MCPT performance is highly related to the results of MOT, we first analyze the effect of two error cases that often occur in MOT: fragments and ID-switches [16]. Since the fragments could be effectively merged via later intra-camera association but ID-switches is difficult to be corrected once determined, we argue that ID-switches have severer impact on the final MCPT performance than fragments. Meanwhile, it's also non-trival to obtain robust appearance features of a trajectory considering occlusions and various poses. Simply applying single frame ReID feature or averaged ReID features of all frames may not reflect the appearance of trajectory accurately. Thus in this paper, we design a Multi-object tracking (MOT) method for MCPT task to minimize ID-switches and obtain better trajectory appearance features.

Following Bytetrack [37], we set a low confidence threshold (0.1) for detection results to preserve some low-confidence detections (e.g. occluded targets and small targets). We use the Kalman-filter [13] to predict motion of the tracked targets then adopts Hungarian algorithm [14] to associate detection results to trajectories based on their location.

3.1.1 Two-Step Matching

We adopt a two-step matching strategy in our MOT module. The first step is to match the tracked trajectories with detections in current frame based on IOU. In the second step, we match the untracked trajectories (which may be lost due to occlusions) with the unmatched detections after the first step.

Considering the error of motion estimation, IOU may not fully reflect the relationship between the predicted position by Kalman-filter of an untracked trajectory and its current position (the detection in current frame). Its value remains zero when there is no overlap between two bounding-boxes. To better represents the relation of location in this case, we proposed a new distance metric $MIOU$: given two bounding-boxes B_1 and B_2, we find their smallest enclosed bounding-box B_e, $MIOU$ is calculated by

$$MIOU = 1 - \frac{|B_1| + |B_2|}{2\,|B_e|} \tag{1}$$

When B_1 and B_2 totally coincide, their $MIOU$ equals to 0. When there is no overlap between B_1 and B_2, their $MIOU$

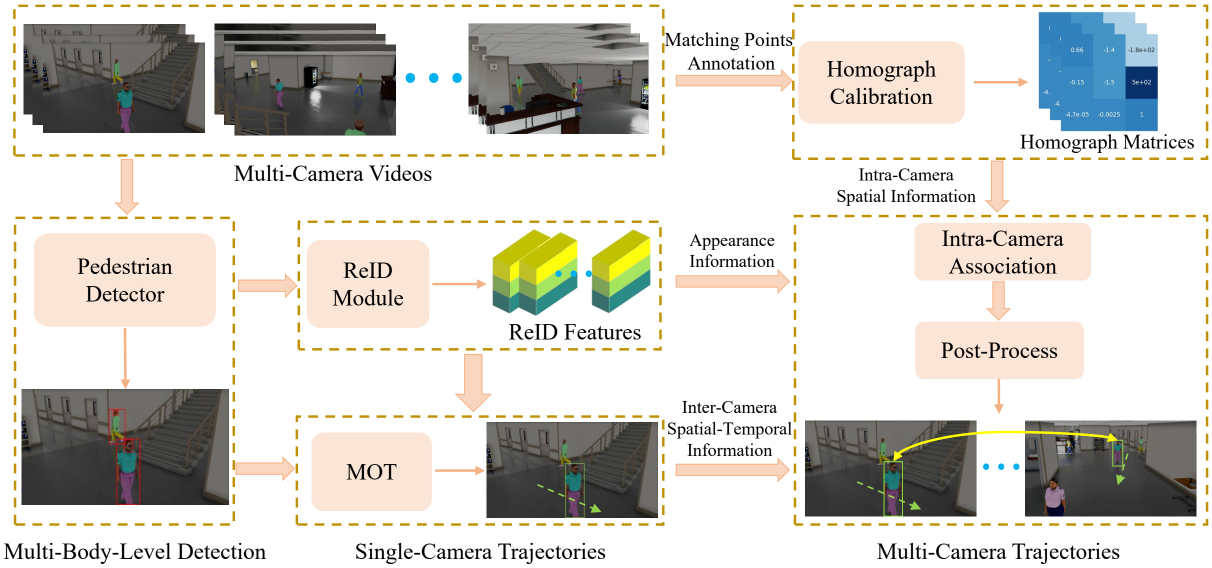

Figure 2. The pipeline of our MCPT system. The MCPT system first performs multi-body-level detection on multi-camera videos, including full-body detection, visible-body detection and head detection. Then the ReID module extracts ReID features, which are fed into our MOT algorithm along with detected full-body bounding boxes. Afterwards, single-camera trajectories are generated and are associated across multiple cameras in intra-camera association via spatial-temporal and appearance information. Finally, the post-processing module further refines the output results by handling with ID antinomy, false positive detection results and missing trajectories.

will be larger 0.5. As the distance between B_1 and B_2 increases, their $MIOU$ will approach to 1. Besides, we adopt an adaptive threshold for Hungarian algorithm in the second step based on the lost time of trajectory. We argue that longer lost time of a trajectory results in an increase of uncertainty of its current position. Thus the threshold will be settled adaptively according to its lost time.

During experiments, we found that if two people nearly coincide, there will be only one common detection box due to non maximum suppression (NMS), and it may been the enclosing bounding box of two people, which mixes their information. Using this inaccurate bounding box for matching may lead to wrong updating of kalman filter state, which will increase the risk of ID-switches. Thus, we discard these coinciding detections and mark both trajectories as untracked until they separate.

Finally, we initialize a new trajectory for each unmatched detection with confidence larger than the initializing threshold (i.e.,0.5) and remove the untracked trajectories that have been lost for a long period (i.e.,>3s) after the second step of association.

3.1.2 Confidence-Aware Trajectory Appearance

During the tracking, we design an online scheme to update the appearance of trajectory via ReID feature of matched detection, which is shown in Algorithm 1. Note that we only consider detections with high confidence (i.e.,larger than a threshold α) in trajectory to extract ReID feature. Since detections with low confidence are often small or occluded targets, their ReID features are not reliable enough. In cases where there are no detections with high confidence in a trajectory, we select the one with highest confidence to extract ReID feature and consider it as the appearance of trajectory.

3.1.3 ReID-Based ID Correction

As the ID-switches usually happen during matching occluded targets, we adopt a delayed strategy to correct identities. In details, we dynamically maintain a temporal overlap matrix C during tracking, in which $C(i, j) > 0$ means that track i and track j both appears in some frames, thus they must belongs to different people. As shown in Figure 3, once an occluded trajectory i is matched again, we calculate the ReID similarity between its matched detection and trajectories $i \cap \{j, C(i, j) > 0\}$ (i.e., trajectories i and the trajectories that have temporal overlaps with i). If the trajectory with maximal ReID similarity to matched detection is not i, indicating that the trajectory i is matched to the wrong detection, we will thus initialize a new trajectory for the wrongly matched detection to avoid ID-switches.

3.2. Intra-Camera Association

After MOT, we acquire all the single-camera trajectories $T_{all} = \bigcup_{c_i=1}^{c_n} T_{c_i}$, where $T_{c_i} = [T_{c_i}^1, \ldots, T_{c_i}^{n_{c_i}}]$, c_n the total number of camera views and n_{c_i} the number of trajectories

Algorithm 1: Confidence-Aware Trajectory Appearance

```
Input: previous trajectory appearance $A^{t-1}$, ReID
       feature of matched detection $F^t$, confidence
       of matched detection $c_F^t$, previous
       confidence of trajectory $c_A^{t-1}$, number of
       previous high-confidence detections in
       trajectory $n^{t-1}$
Output: updated trajectory appearance $A^t$, updated
        number of previous high-confidence
        detections in trajectory $n^t$, updated
        confidence of trajectory $c_A^t$
if $c^t > \alpha$ then
    $A^t \leftarrow \frac{n^{t-1}}{n^{t-1}+1} A^{t-1} + \frac{1}{n^{t-1}+1} F^t$;
    $n^t \leftarrow n^{t-1}$+1 ;
    if $c_F^t > c_A^{t-1}$ then
        $c_A^t \leftarrow c_F^t$ ;
    end
else
    if $c_F^t > c_A^{t-1}$ then
        $c_A^t \leftarrow c_F^t$ ;
        $A^t \leftarrow F^t$ ;
    end
end
```

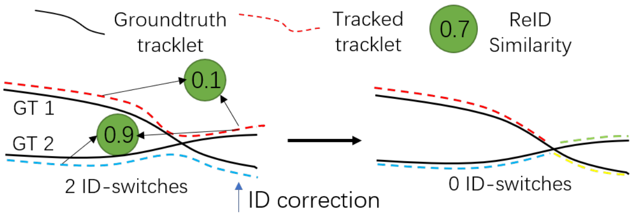

Figure 3. Illustration of ID correction. Two tracked trajectories switch their ID due to occlusion in the left image. After occlusion, we calculate the ReID similarity between matched detection and trajectories to judge if there is ID-switches and assign new IDs for the ID-switched trajectories after occlusion in the right image

from camera c_i. The confidence score c and appearance feature A of each trajectory are also stored to guide the following association step. Similar to [15], we perform hierarchical clustering to associate trajectories. But in addition to appearance information, we also leverage spatial-temporal constraints and integrate them into a distance matrix. Compared to appearance-based methods, our distance matrix can better reflect both the relation between single-camera trajectories and between trajectories from different cameras, thus can achieve more accurate association.

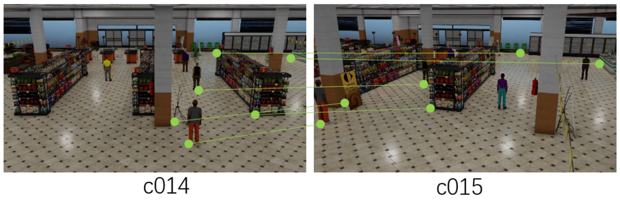

Figure 4. Example of matching points annotation between 2 cameras

3.2.1 Appearance Distance Matrix

We construct a matrix D^A to represent the distance between the appearance features of trajectories. It is taken into consideration of the formulation of distance matrix to help associate trajectories with similar appearance. For two arbitrary trajectories T_i and T_j from T_{all}, the appearance distance can be calculated using the cosine score:

$$D^A(T_i, T_j) = 1 - cos(A_i, A_j) \tag{2}$$

where A_i and A_j are appearance feature vectors of T_i and T_j, respectively.

3.2.2 Spatial-Temporal Distance Matrices

Simply using trajectory appearance matrix as final matrix is insufficient to tackle problems such as severe occlusions, variance of view angles and people with similar clothes. To cope with this issue, we construct another distance matrix that incorporates both intra-camera spatial information (homograph distance) and inter-camera spatial-temporal information (ID antinomy, speed constraint).

Homography distance We manually annotate several matching points on the ground for each camera pair with overlapping areas to compute homography matrix $H_{c_i \to c_j}$ from camera c_i to camera c_j. The annotation is illustrated in Figure 4. In details, we find more than 5 pairs of match points for each camera pair with overlapping areas and record their coordinates. The selected points should meet the following restrictions: (1) All the points should be on the ground plane, since homography transform could only represents transform between 2 planes. (2) The points in the same camera should be as far away from each other as possible. (3) It's not allowed that more than 3 points in one camera are on the same line. With the coordinates of matched points pair $\bigcup_{k=1}^{m}(P_{c_i}^k, P_{c_j}^k)$, m the number of pairs, we estimate the homography matrix $H_{c_i \to c_j}$ from camera c_i to camera c_j by minimizing the L2 error of distance between the transformed points from c_i and their matched points from c_j.

$$H_{c_i \to c_j} = \underset{H \in R^{3\times3}}{\text{argmin}} \sum_{k=1}^{m} \left\| H P_{c_i}^k - P_{c_j}^k \right\|_2^2 \tag{3}$$

Given all trajectories from different cameras $T_{all} = \bigcup_{c_i=1}^{c_n} T_{c_i}$, where $T_{c_i} = [T_{c_i}^1, \dots, T_{c_i}^{n_{c_i}}]$, n_{c_i} the number of trajectories from camera c_i and c_n the number of camera views. For $T_{c_i}^a = \bigcup_{t_f \in [t_1,\dots,t_{\left|T_{c_i}^a\right|}]} B_{c_i,a}^{t_f}$ and $T_{c_j}^b = \bigcup_{t'_f \in [t'_1,\dots,t_{\left|T_{c_j}^b\right|}]} B_{c_j,b}^{t'_f}$ two trajectories from different cameras $(c_i \neq c_j)$, we can obtain the normalized homograph distance between $T_{c_i}^a$ and $T_{c_j}^b$:

$$D^H(T_{c_i}^a, T_{c_j}^b) = \begin{cases} \sum\limits_{t_s \in I_{T_{c_i}^a, T_{c_j}^b}} D_{t_s}^H(T_{c_i}^a, T_{c_j}^b) & \text{if } I_{T_{c_i}^a, T_{c_j}^b} \neq \emptyset \\ 0 & \text{else} \end{cases} \tag{4}$$

$$D_{t_s}^H(T_{c_i}^a, T_{c_j}^b) = \frac{\left\| \Delta P_{(c_i,a;c_j,b)}^{t_s} \right\|_2}{\bar{h}_{(c_i,a;c_j,b)}^{t_s} \left| I_{T_{c_i}^a, T_{c_j}^b} \right|} \tag{5}$$

$$\Delta P_{(c_i,a;c_j,b)}^{t_s} = H_{c_i \to c_j} P_{c_i,a}^{t_s} - P_{c_j,b}^{t_s} \tag{6}$$

$$\bar{h}_{(c_i,a;c_j,b)}^{t_s} = \frac{h_{c_i,a}^{t_s} + h_{c_j,b}^{t_s}}{2} \tag{7}$$

where $I_{T_{c_i}^a, T_{c_j}^b} = [t_1, ..., t_{\left|T_{c_i}^a\right|}] \bigcap [t'_1, ..., t_{\left|T_{c_j}^b\right|}]$, $n_{T_{c_i}^a, T_{c_j}^b} = \left| I_{T_{c_i}^a, T_{c_j}^b} \right|$, $P_{c_i,a}^{t_s}$ the midpoint of the bottom of bounding box $B_{c_i,a}^{t_s}$ and $h_{c_i,a}^{t_s}$ the height of bounding box $B_{c_i,a}^{t_s}$. We utilize the midpoint of the bottom of bounding box to conduct homography transform because it is usually the closest point to the ground and can be regarded as the best point to reflect the location of a person. For trajectories from the same camera, we assign zero values for their homograph distances.

ID antinomy To prevent the phenomenon that multiple people in the same frame of the same camera are assigned with the same identity (we called this as ID antinomy, which will also be used in latter post-processing), we penalize it according to the number of overlapping frames of two trajectories in the same camera:

$$D^T(T_{c_i}^a, T_{c_i}^b) = n_{T_{c_i}^a, T_{c_i}^b} \tag{8}$$

where $n_{T_{c_i}^a, T_{c_i}^b}$ denotes the number of overlapping frames. Similarly, trajectories from different cameras are assigned with zero values.

Speed constraint Since there is an upper bound for human speed in the real life, two trajectories from the same camera cannot belong to the same person if the interval between one's ending time and the other one's beginning time is too short for the distance between one's endpoint and the other one's beginning point. We penalize the distance of such pairs via another distance matrix D_S. Without losing generality, we assume $T_{c_i}^a$ ends before the beginning of $T_{c_i}^b$.

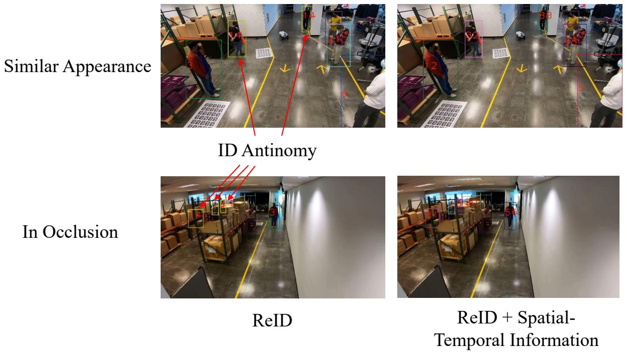

Figure 5. Example of ID antinomy caused by the limitation of appearance (ie., ReID) features. The left column show that people with similar appearance or in occlusion can be assigned with the same identity when only appearance features are considered. As illustrated in the right column, this problem can be resolved by introducing our distance matrix that leverages spatial-temporal information.

$$D^S(T_{c_i}^a, T_{c_i}^b) = \begin{cases} N & \text{if } v_{(c_i,a;c_i,b)} > \alpha_s \\ 0 & else \end{cases} \tag{9}$$

$$v_{(c_i,a;c_i,b)} = \frac{2 \left\| P_{c_i,a}^{t_{\left|T_{c_i}^a\right|}} - P_{c_i,b}^{t'_1} \right\|_2}{\left| t_{\left|T_{c_i}^a\right|} - t'_1 \right| (h_{c_i,a}^{t_{\left|T_{c_i}^a\right|}} + h_{c_i,b}^{t'_1})} \tag{10}$$

Where N is a large constant and α_s is an empirical speed threshold. t'_1,$t_{\left|T_{c_i}^a\right|}$ are the beginning time of $T_{c_i}^b$ and the end time of $T_{c_i}^a$, respectively. $P_{c_i,a}^{t_{\left|T_{c_i}^a\right|}}$ and $P_{c_i,b}^{t'_1}$ are the midpoints of the bottom of boundding boxes $B_{c_i,a}^{t_{\left|T_{c_i}^a\right|}}$ and $B_{c_i,b}^{t'_1}$, respectively. $h_{c_i,a}^{t_{\left|T_{c_i}^a\right|}}$ and $h_{c_i,b}^{t'_1}$ are the heights of $B_{c_i,a}^{t_{\left|T_{c_i}^a\right|}}$ and $B_{c_i,b}^{t'_1}$, respectively.

3.2.3 Final Association Distance Matrix

Adaptive weight matrix To better combine different distance matrices, we first construct an adaptive weight matrix W based on the confidence of trajectories (i.e., the maximal detection confidence of trajectories).

$$W(T^a, T^b) = \frac{\lambda min(c^a, c^b) + \beta}{w_{a,b}^H + \beta} \tag{11}$$

$$w_{a,b}^H = \lambda min(c^a, c^b) + (1 - \lambda)\theta(D^H(T^a, T^b)) \tag{12}$$

Where c is the trajectory confidence, $\theta(\cdot)$ is a function and $\theta(x) = \begin{cases} 0 & x \leq 0 \\ 1 & x > 0 \end{cases}$. β is a small value to prevent the denominator from zero value. The construction of weight matrix is based on the following observation: The confidence of detection and the reliability of ReID feature are both highly correlated to the completeness of the target in detection box. The final distance matrix can be represented as a weighted sum of the four distance matrices:

$$D = WD^A + (1 - W)D^H + w_T D^T + w_S D^S \tag{13}$$

3.2.4 One-Step Trajectory Association

Once the distance matrix of all trajectories from different cameras is constructed, we associate the trajectories via hierarchical clustering. Unlike previous work [15] utilizing two steps to cluster single-camera trajectories and multi-camera trajectories respectively, we tackle the problem by bridging the above two types of trajectories adopting an adaptive weighting strategy in Equation 13, thus perform one-step clustering.

Due to difficulties caused by possible false positive detections and severe occlusions, the detection noise and ReID noise would have a great impact on clustering results. Therefore, strict restrictions have to be applied on the initialization of clustering centers. First, clusters with a total detection number above a certain threshold is selected as clustering center candidates. The distance between each pair of candidates is examined to ensure that each candidate corresponds to a distinct person, otherwise two clusters will be merged. Then clusters possessing large distance values to existing centers are appended as new center candidates to avoid possible omission. Finally, all the rest clusters will re-clustered to the center candidate to which the distance is the minimum among all candidates.

3.3. Post-Processing

The multi-camera trajectories after inter-camera association still suffer from ID antinomy and false negative detection. Besides, the full-body bounding boxes that we adopted in intra-camera association is not totally consistent with the official standard of labeling. Thus we design a post-processing module to further refine the multi-camera trajectories.

Removal of ID antinomy As mentioned earlier, it is not appropriate to assign the same identity to multiple people in the same frame of the same camera, thus we remove the part of multi-camera trajectories that have ID antinomy.

Trajectory compensation from multiple cameras Due to removal of ID antinomy and low detection confidence (occluded or small targets), there will be some missing trajectories although with the existence of detection bounding boxes. To address this problem, we design a trajectory compensation procedure that leverages trajectories from other camera views to compensate missing trajectories. We perform homography from the trajectory in other camera views (if exists) to the camera view where the person is possibly missing. if there is a detection box nearly coinciding the transformed box (i.e, with a high IOU value), it will be regarded as a part of the missing trajectory and will be appended to the multi-camera trajectory of this person.

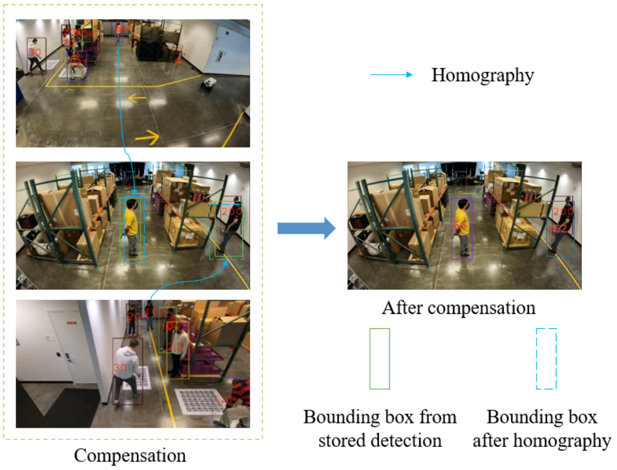
Homography
After compensation
Bounding box from stored detection
Bounding box after homography
Compensation

Figure 6. Illustration of compensation. Missing trajectories are compemsated from other camera views by homography and the criteria based on IOU with detection boxes.

Multi-body-level detection matching In our system, three different body-level detectors are employed: full-body detector, visible-body detector and head detector. In intra-camera association, the full-body bounding boxes are essential to provide an approximate foot location. However, we observe that only the visible body parts is labelled in synthetic scenes when the corresponding person's head is visible. For real scenes, A bounding box is annotated if 60% of the body is seen, or the head and shoulder are seen. Thus we first match different body-level detection results via an IOU threshold. Then we judge if a bounding box should be outputted based on whether it is matched with a head bounding box or whether the matched visible bounding box occupies 60% the area of the full-body bounding box in real scenes. During this procedure, false positive full-body detection results can be filtered because there is low probability that multi-body-level false positive detection results occur at the same time. Finally, the full-body bounding boxes in trajectories will be replaced by visible-body bounding boxes.

Interpolation As a commonly used post-processing for tracking, we also perform interpolation to fill the missing frames of trajectory which may been caused by false negative detections in occluded scenarios.

4. Experiments

4.1. Dataset and Evaluation Metrics

The AIC23 MCMT Tracking dataset[1] [20] comprises 22 indoor scenes captured by multiple cameras across various settings. It includes 10 training scenes, 5 validation scenes, and 7 testing scenes. The dataset contains real-world data captured from cameras placed in warehouse buildings, as well as synthetically generated data from multiple indoor settings. The synthetic animated people dataset, which makes up a significant portion of the total dataset, is created using the NVIDIA Omniverse Platform.

In order to improve the performance of the ReID module, we finetune the model on Randperson dataset, which is a public synthetic dataset containing 8,000 virtual characters, 11 scenes, 19 cameras, 38 videos of dense pedestrians, and 1,801,816 cropped pedestrian images.

For metrics of evaluation, the IDF1 score will be used to rank the performance of each team on the leaderboard. IDF1 measures the ratio of correctly identified detections over the average number of ground-truth and computed detections. Other evaluation measures adopted by the MOTChallenge [1, 17], such as IDP and IDR, will be displayed but they will not be used for ranking purposes.

4.2. Implementation Details

Multi-body-level detection. For human full-body detection, we adopted the same detection model of Bytetrack [37], which is a YOLO-v5x [12] trained on several public pedestrian detection datasets. The visible-body detection and head detection model is a YOLO-v5m model pretrained on Crowdhuman dataset [25].

Re-identification. The network structure we adopted for ReID is MGN(R101) [29]. We initialize the model with the pretrained weight from LUPerson [7]. In order to extract accurate ReID features both from real-world data and from synthetic data, we finetune the model with Randperson dataset [32] based on the Fastreid toolbox [10].

Empirical parameter setting. α for high confidence threshold is set as 0.88. N, α_s in Equation 9 are set as 10 and 2, respectively. λ in Equation 11 and Equation 12 is set as 0.65, β in Equation 10 is set as 0.01. w_T and w_S in Equation 13 are both set as 1. The linkage criterion for hierarchical clustering is average distance and the threshold is set as 0.35.

4.3. Results

Table 1 displays the ablation study of the different proposed strategies. Compared to the baseline (ByteTrack + hierarchical clustering based on ReID features), our refined MOT approach yields a 1.75% increase in IDF1, demonstrating the effectiveness of the proposed single-camera tracking method. Integrating spatial-temporal information into distance matrix results in an IDF1 of 93.08%. Implementing a post-processing step to remove ID antinomy, compensate and interpolate trajectories and refine final outputs further enhances the IDF1 by 0.23%, achieving an IDF1 of 93.31% on the final leaderboard.

Method	IDF1	IDP	IDR
Baseline	89.30	88.84	89.76
+our MOT	91.05	90.62	91.50
+spatial-temporal	93.08	92.35	93.82
+post-processing	**93.31**	**93.43**	**93.19**

Table 1. Comparaison of different MCPT methods on AIC23 MCMT Tracking test set.

Rank	Team ID	Team Name	IDF1
1	6	UWIPL_ETRI	95.36
2	9	HCMIU-CVIP	94.17
3	41	AILab (ours)	93.31
4	51	FraunhoferIOSB	92.84
5	10	Skygazer	92.33
6	113	hust432	92.07
7	133	ctcore	91.09

Table 2. Leaderboard of Track 1 in the AI City Challenge 2023.

The result of our proposed system was entered into the evaluation system of the AICity Challenge 2023 Track 1, which achieved an IDF1 score of 93.31% and ranked third among more than 40 teams. The final leaderboard is shown in Table 2.

5. Conclusion

In this paper, we propose an effective Multi-Camera People Tracking system. It mainly contains three modules: MOT, intra-camera association and post-process module. For MOT, we minimize the ID-switches error and obtain more accurate appearance feature for trajectories. Besides, we develop an intra-camera association approach which leverage both appearance information and multiple sources of spatial-temporal information. The post-process module which contains multi-step post processing to eliminate ID antinomy and false positive detections as well as compensate missing trajectories. The experimental results on the public test set of Track1 for the 2023 AI CITY CHALLENGE validate the effectiveness of our method, as it attains an IDF1 score of 93.31%, securing the third place on the leaderboard.

[1] https://www.aicitychallenge.org/2023-data-and-evaluation/

References

[1] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008. 8

[2] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE, 2016. 3

[3] Long Chen, Haizhou Ai, Rui Chen, Zijie Zhuang, and Shuang Liu. Cross-view tracking for multi-human 3d pose estimation at over 100 fps. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3279–3288, 2020. 3

[4] Weihua Chen, Lijun Cao, Xiaotang Chen, and Kaiqi Huang. An equalized global graph model-based approach for multi-camera object tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(11):2367–2381, 2016. 3

[5] Ran Eshel and Yael Moses. Homography based multiple camera detection and tracking of people in a dense crowd. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 3

[6] Yuxin Fang, Bencheng Liao, Xinggang Wang, Jiemin Fang, Jiyang Qi, Rui Wu, Jianwei Niu, and Wenyu Liu. You only look at one sequence: Rethinking transformer in vision through object detection. *Advances in Neural Information Processing Systems*, 34:26183–26197, 2021. 2

[7] Dengpan Fu, Dongdong Chen, Jianmin Bao, Hao Yang, Lu Yuan, Lei Zhang, Houqiang Li, and Dong Chen. Unsupervised pre-training for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14750–14759, 2021. 8

[8] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 2

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2

[10] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. Fastreid: A pytorch toolbox for general instance re-identification. *arXiv preprint arXiv:2006.02631*, 2020. 8

[11] Martin Hofmann, Daniel Wolf, and Gerhard Rigoll. Hypergraphs for joint multi-view reconstruction and multi-object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3650–3657, 2013. 3

[12] Glenn Jocher. yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation. https://github.com/ultralytics/yolov5, Oct. 2020. 2, 8

[13] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. 1960. 3

[14] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 3

[15] Fei Li, Zhen Wang, Ding Nie, Shiyi Zhang, Xingqun Jiang, Xingxing Zhao, and Peng Hu. Multi-camera vehicle tracking system for ai city challenge 2022. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3265–3273, 2022. 5, 7

[16] Yuan Li, Chang Huang, and Ram Nevatia. Learning to associate: Hybridboosted multi-target tracker for crowded scene. In *2009 IEEE conference on computer vision and pattern recognition*, pages 2953–2960. IEEE, 2009. 3

[17] Yuan Li, Chang Huang, and Ram Nevatia. Learning to associate: Hybridboosted multi-target tracker for crowded scene. In *2009 IEEE conference on computer vision and pattern recognition*, pages 2953–2960. IEEE, 2009. 8

[18] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 280–296. Springer, 2022. 2

[19] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[20] Milind Naphade, Shuo Wang, David C. Anastasiu, Zheng Tang, Ming-Ching Chang, Yue Yao, Liang Zheng, Mohammed Shaiqur Rahman, Meenakshi S. Arya, Anuj Sharma, Qi Feng, Vitaly Ablavsky, Stan Sclaroff, Pranamesh Chakraborty, Sanjita Prajapati, Alice Li, Shangru Li, Krishna Kunadharaju, Shenxin Jiang, and Rama Chellappa. The 7th AI City Challenge. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2023. 8

[21] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 2

[22] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 2

[23] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 2

[24] Ergys Ristani and Carlo Tomasi. Features for multi-target multi-camera tracking and re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6036–6046, 2018. 3

[25] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018. 8

[26] Chi Su, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Deep attributes driven multi-camera person re-identification. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 475–491. Springer, 2016. 2

[27] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with

convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 2

[28] Minh Vo, Ersin Yumer, Kalyan Sunkavalli, Sunil Hadap, Yaser Sheikh, and Srinivasa G Narasimhan. Self-supervised multi-view person association and its applications. *IEEE transactions on pattern analysis and machine intelligence*, 43(8):2794–2808, 2020. 3

[29] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 274–282, 2018. 8

[30] Hanxiao Wang, Shaogang Gong, Xiatian Zhu, and Tao Xiang. Human-in-the-loop person re-identification. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 405–422. Springer, 2016. 2

[31] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, pages 688–703. Springer, 2014. 2

[32] Yanan Wang, Shengcai Liao, and Ling Shao. Surpassing real-world source training data: Random 3d characters for generalizable person re-identification. In *Proceedings of the 28th ACM international conference on multimedia*, pages 3422–3430, 2020. 8

[33] Longyin Wen, Zhen Lei, Ming-Ching Chang, Honggang Qi, and Siwei Lyu. Multi-camera multi-target tracking with space-time-view hyper-graph. *International Journal of Computer Vision*, 122:313–333, 2017. 3

[34] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017. 3

[35] Hantao Yao, Shiliang Zhang, Richang Hong, Yongdong Zhang, Changsheng Xu, and Qi Tian. Deep representation learning with part loss for person re-identification. *IEEE Transactions on Image Processing*, 28(6):2860–2871, 2019. 2

[36] Mang Ye, Chao Liang, Yi Yu, Zheng Wang, Qingming Leng, Chunxia Xiao, Jun Chen, and Ruimin Hu. Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing. *IEEE Transactions on Multimedia*, 18(12):2553–2566, 2016. 2

[37] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 1–21. Springer, 2022. 3, 8

[38] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenyu Liu, and Wenjun Zeng. Voxeltrack: Multi-person 3d human pose estimation and tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2613–2626, 2022. 3

[39] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, 129:3069–3087, 2021. 3

[40] Liang Zheng, Shengjin Wang, Lu Tian, Fei He, Ziqiong Liu, and Qi Tian. Query-adaptive late fusion for image search and person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1741–1750, 2015. 2

[41] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1367–1376, 2017. 2

[42] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1318–1327, 2017. 2

[43] Jiahuan Zhou, Pei Yu, Wei Tang, and Ying Wu. Efficient online local metric adaptation via negative samples for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2420–2428, 2017. 2

[44] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2