Overlap Suppression Clustering for Offline Multi-Camera People Tracking

Ryuto Yoshida Junichi Okubo Junichiro Fujii
Masazumi Amakata Takayoshi Yamashita
Yachiyo Engineering Co., Ltd. Chubu University
ry-yoshida@yachiyo-eng.co.jp

Abstract

Multi-Camera People Tracking is a multifaceted issue that requires the integration of several computer vision tasks, such as Object Detection, Multiple Object Tracking, and Person Re-identification. This study presents a multi-camera people tracking method that comprises four main processes: (1) single camera people tracking based on overlap suppression clustering, (2) representative image extraction using pose estimation for re-identification, (3) re-identification using hierarchical clustering with average linkage, and (4) low-identifiability tracklets assignment.

Our RIIPS team achieved the highest Higher Order Tracking Accuracy (HOTA) of 71.9446% in the 2024 AI City Challenge Track 1.

1. Introduction

Multi-Camera People Tracking (MCPT) is a complex recognition task that involves tracking people's trajectories and identifying individuals who appear in multiple cameras. This task requires simultaneously solving multiple computer vision tasks, including Object Detection, Multiple Object Tracking (MOT) [34], and Re-identification (Re-ID) [20]. The quantification of people's trajectories through MCPT can contribute to the realization of digital twins, smart cities, and various other entities.

MCPT has two main components: Single Camera People Tracking (SCPT) and Re-ID. Re-ID has traditionally been considered a challenging task; however, recent training methods, such as Triplet Loss [17], have established a technical foundation for models with high identification capabilities for individuals. Therefore, if images with highly identifiable characteristics are provided, individuals can be identified accurately. However, it is possible that the decrease in the accuracy of the MCPT is due to either the significantly different characteristics of the inputs or the inaccuracy of the SCPT, which is the basis for the MCPT. Therefore, to improve the performance of MCPT, it may be beneficial to improve the accuracy of SCPT and use identifiable images for the Re-ID model.

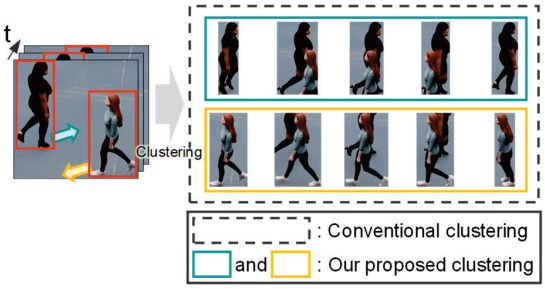


Figure 1: Example of the clustering result. The conventional clustering may result in misclustering, as shown by the dashed rectangle. However, our proposed method can accurately classify individuals, as shown by the solid rectangles. It is suggested that using the proposed method, MCPT performance may improve.

SCPT is often implemented using an online tracking algorithm. Online SCPT uses bipartite matching to associate data in the current frame with tracklets measured in previous frames using motion or appearance features. By relying solely on past information, online tracking often leads to mistracking in complex states, particularly when paths cross. However, in many cases, the trajectory returns to a simple state after crossing, making tracking easy. Thus, if the information before and after the complex states can be utilized for tracking, an improvement in accuracy can be expected. Therefore, offline tracking, which utilizes future information, is considered an approach that contributes to a high tracking performance. Tracking complex trajectories using only motion features is not suitable for identification such as the trajectory represented by x or χ, so it is particularly desirable to effectively utilize appearance features.

Considering these ideas, this study proposes an offline SCPT framework that utilizes information other than complex states based on hierarchical clustering. The correct tracklets are obtained using this method, as shown in Fig. 1. In addition, MCPT consists of three processes: high-identifiability image extraction, Re-ID using hierarchical clustering with average linkage, and low-identifiability image assignment, which together achieve high performance. This is a simple approach based on the idea that the Re-ID performance improves by using only highly identifiable images. The contribution of this study

can be summarized as follows:

- Our original clustering method called Overlap Suppression Clustering improves the accuracy of Single Camera People Tracking.
- Our offline MCPT method, which consists of three processes — (1) high-identifiability image extraction, (2) Re-ID using hierarchical clustering with average linkage, and (3) low-identifiability image assignment — demonstrates high performance.
- This framework achieved the highest HOTA of 71.9446% in the 2024 AI City Challenge Track1.

2. Related Work

2.1. Object Detection for People Tracking

The selection of an object detection model is an important process for SCPT and MCPT, because the localization of a person's position depends on the detection. Typically, object detection model selection considers the trade-off between floating-point operations per second and performance, such as Average Precision. For SCPT, the anchor-free model [7, 36, 37, 38] is more suitable than the anchor-based model [39, 40, 41, 42] because it is less susceptible to occlusion effects [35].

Recently, the Joint Detection and Embedding (JDE) model has been proposed in various studies [27, 28]. In traditional two-stage approaches consisting of detection and embedding, an embedding is required for each bounding box (bbox) in the frame. By contrast, JDE, which can perform this in one shot, including detection, offers superior computational efficiency. On the other hand, detection, which requires costly annotation, and Re-ID, which requires a large training dataset, are combined, leading to an increase in training costs. In addition, the usefulness of the JDE models in Re-ID is not clear because they are evaluated using single camera MOT datasets represented by MOT17 [29] and MOT20 [30].

Fine-tuning the confidence level of the threshold is also important, given the delicate balance between precision and recall. In the context of SCPT, low recall can lead to underestimation of the number of people, whereas low precision can lead to overestimation.

Non-Maximum Suppression (NMS) [14] requires fine-tuning of the threshold, similar to confidence. NMS, which suppresses multiple bboxes without the highest-confidence bbox, is used to solve the problem of multiple bboxes assigned to the same object. In the context of SCPT, NMS can be critically affected when the foreground person occludes the background person. Although it is important to suppress overlapping boxes, using an overly strict NMS threshold can increase the risk of missing people in the background during occlusion. In SCPT, the behavior of the tracking algorithm should be considered when tuning the threshold.

2.2. Person Re-Identification

Person Re-ID [20] aims to identify the same individual from several images using only visual information. Re-ID uses any kind of method to measure similarity as a metric. Methods such as descriptor learning [21, 22] and color calibration [23] have been widely used [24]. However, in recent years, dominant methods have used embedding features extracted by Re-ID models trained using representation learning [16, 17, 18].

Re-ID models are trained using a loss function designed to maximize the similarity of identical pairs and minimize the similarity of non-identical pairs. Recently, Re-ID models, such as the model-trained LUPerson [19] and OSNet [5], have demonstrated high individual identifiability. In fact, they achieved a mean Average Precision of over 0.85 on the Market1501 dataset [25], which is widely recognized as one of the benchmark datasets for Re-ID.

2.3. People Tracking

SCPT is performed using an association between the bboxes assigned to the same individual in each frame. MCPT is performed by an association between each tracklet assigned to the same individual. Therefore, although the terminology for each task differs, the commonality lies in the associated data. Regardless of whether SCPT or MCPT is used, data association approaches can be broadly categorized as motion-feature tracking or appearance-feature tracking.

2.3.1 Motion-Feature Tracking

Motion-feature tracking is based on the principle that a person's positional information exhibits spatiotemporal continuity. Therefore, the positions of the bboxes are used to associate the individuals in each frame. To associate the same individual in SCPT, common methods, such as the Kalman filter or particle filter, evaluate the difference between the current position estimated from the past position and the current position observed by object detection.

SORT [1] based on the Kalman filter, estimates the position and size of the bboxes from the last two frames and evaluates the distance using Intersection over Union (IoU). SORT has inspired various derivative methods such as SimpleTrack [3] which evaluates the distance using a generalized IoU [26] to measure the distance between non-overlapping bboxes, and StrongSORT [2], which incorporates a momentum term to estimate the current position from more than the last two frames.

In the context of MCPT, there are limitations to associating tracklets based on motion features. Motion features can only be used when the spatial regions captured by the camera overlap with one another. Furthermore, it is impossible to identify the same individual based on camera coordinates alone, which requires an aligning camera

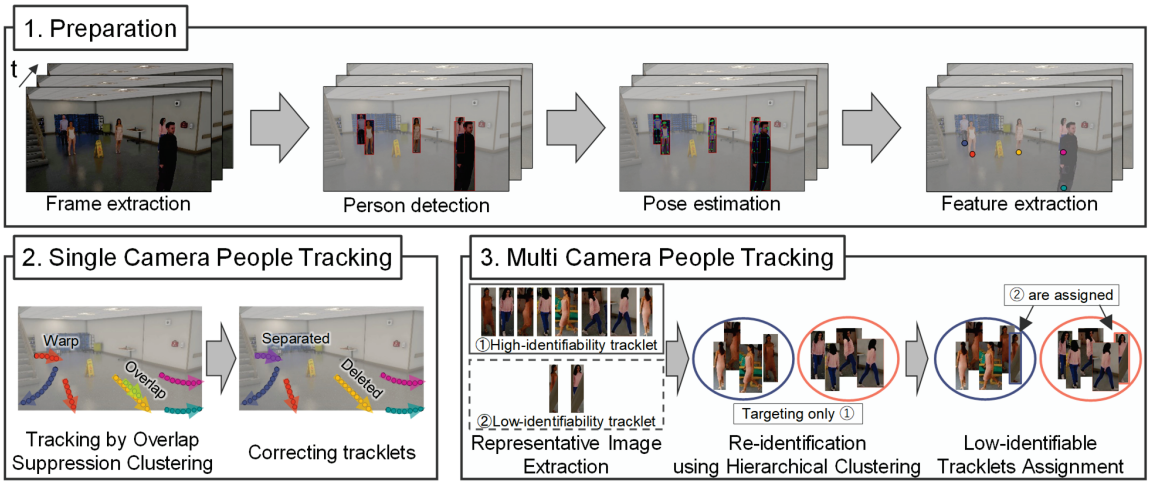


Figure 2: Overview of our proposed framework. First, some inference is performed on each frame as a preparation. Based on these results, Single Camera People Tracking is performed as shown in Fig. 3 and Multi-Camera People tracking is performed as shown in Fig. 5.

positions and computing world coordinates.

For the 2024 AI City Challenge Track1, the multi-camera views are overlapped and each camera provides a camera matrix Mat_c, that satisfies the relationship between the camera coordinates C_C and world coordinates C_w as shown in Eq.(1). This makes it possible to associate the same individual captured by different cameras using motion features.

$$C_C = Mat_c C_w \tag{1}$$

2.3.2 Appearance-Feature Tracking

The basic principle of appearance-feature tracking is the premise that images of the same individual have a high degree of visual similarity. Consequently, if the similarity of the features extracted from two images exceeds a certain threshold, they represent the same individual. This tracking method requires attention that ignores the relationships of positions between frames. Therefore, it is preferable to use the motion features simultaneously, such as in DeepSORT [15]. Some offline tracking algorithms that use appearance features have also been proposed. Some of these approaches, such as the Appearance Evaluation Network [32] and Tracklet-Plane Matching [33], aim to correct mistracking by considering the appearance features after the initial tracking.

Appearance-feature tracking is applicable to both SCPT and MCPT. However, it would probably be better to distinguish the similarity parameter ε that identifies the same individual. In SCPT, there are continuous variations in the images of the same individual, which makes it easier to show high similarity in nearby frames. However, in MCPT, there are discontinuous variations that make it easier to show a lower similarity than SCPT.

2.4. Multi-Camera People Tracking

MCPT competed in the 2023 AI City Challenge Track 1 [8]. This competition is similar to the 2024 competition but differs in that IDF1 [52] is used as the evaluation metric and real-world data are also used as test data. The top teams use appearance features in the MCPT framework [9, 10, 11, 12, 13]. In particular, all of their approaches use clustering that classifies clusters based on the similarity of neighboring data, such as hierarchical clustering and DBSCAN. In addition, most of the approaches address the problem caused by poorly person-identifiable images, such as occlusion with a person or a unique foreground.

Compared to the MOT17 dataset, the 2023 AI City Challenge Track1 dataset was qualitatively confirmed to contain higher-resolution images. Although the highly accurate approach in MOT17 uses motion-feature tracking [43, 44], in the 2023 and 2024 AI City Challenge Track1, appearance-feature tracking will be useful because there are highly identifiable appearance features.

3. Method

Fig. 2 shows an overview of the proposed framework. First, this method performs preparation, that is, frame extraction from videos and inference of detection, feature extraction, and pose estimation. Considering that the accuracy of each process required in MCPT is guaranteed by previous research, a two-stage model is adopted.

In this study, the point representing the location and the image, and embedding feature information obtained from the bbox is defined as a node. After preparation, SCPT is performed using an original method based on hierarchical clustering. In our framework, SCPT does not identify re-entered individuals on the screen. Our MCPT algorithm is

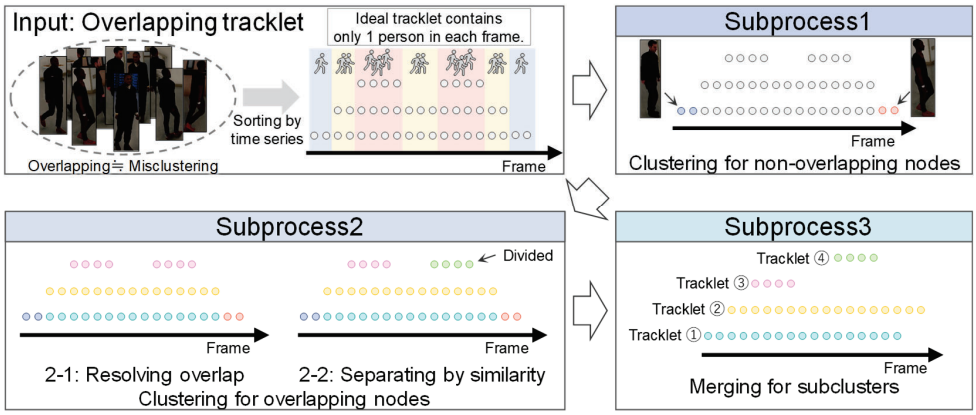


Figure 3: Processing flow of Overlap Suppression Clustering. The goal of this process is to obtain non-overlapping tracklets from the overlapping tracklets. In this figure, the colors of the nodes represent each cluster. In subprocess 1, non-overlapping nodes are clustered. In subprocess 2-1, overlapping nodes are clustered while avoiding overlap. Subprocess 2-2 separates non-identical individuals by similarity, because non-identical individuals belong to the same cluster in subprocess2-1. Subprocess 3 merges each cluster. Overlap Suppression Clustering provides tracklets without overlap.

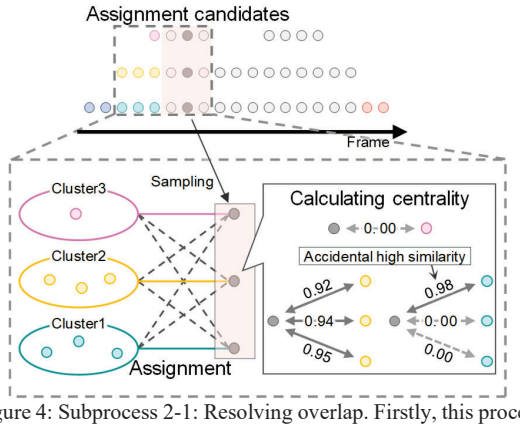


Figure 4: Subprocess 2-1: Resolving overlap. Firstly, this process assigns any overlapping nodes to the initial cluster. By solving the assignment problem, each overlapping node is clustered separately, step by step.

also based on hierarchical clustering. Re-ID is performed using only representative nodes from each tracklet.

Further details of the processing steps are presented in the following section.

3.1. SCPT using appearance feature

3.1.1 Tracking by Overlap Suppression Clustering

In our proposed SCPT, the tracklet of each individual is tracked by clustering its nodes over several periods. Basically, if the distance between two embedding features is less than ε, they are considered to be the same individual. In this study, $1 - cosine\ similarity$ is defined as a distance, and cosine similarity is defined as a similarity. As ε increases, the likelihood of associating the same individual as the same individual increases. However, there is also a greater risk of misclustering another individual.

When analyzing high-frame-rate videos typically recorded at approximately 30 fps, images within nearby frames exhibit high similarity as the inter-frame variance decreases significantly. Therefore, agglomerative hierarchical clustering using a single linkage with a small ε would improve the performance of SCPT. However, the personal identifiability of the embedding features is not necessarily high. Occasionally, non-identical feature pairs lead to misclustering. If the target persons are similar, the original images share the same unique foreground, or there is a significant intersection, misclustering can occur.

Owing to the impossibility of a person existing in two different places simultaneously, it becomes obvious that misclustering occurs when there are multiple nodes with the same frames. In this study, the state of the tracklet when there are multiple nodes at the same frame is defined as an overlap, and an original method called Overlap Suppression Clustering is implemented to resolve this state, which occurs through ordinary hierarchical clustering.

Fig. 3 shows the processing flow of overlap suppression clustering. To obtain non-overlapping tracklets, this process re-clusters the overlapping tracklets. There are three subprocesses: subprocess 1, clustering for non-overlapping nodes in tracklets with overlap, subprocess 2, clustering for overlapping nodes, and subprocess 3, merging for subclusters.

Subprocess 1. Subprocess 1 performs clustering of non-overlapping nodes. Subclusters are assigned to each connected component based on similarity.

Subprocess 2. Subprocess 2 performs clustering for the overlapping data. Fig. 4 shows the details of subprocess 2-1. In the tracklet with overlap, the maximum number of nodes in the same frame can be considered the maximum number of simultaneous appearances. Therefore, the

groups of overlapping nodes in each frame are classified into subclusters according to the number of maximum overlaps in the tracklet.

In this subprocess, initial nodes in the same frame are assigned to subclusters, and then other unclustered nodes are assigned by step-by-step bipartite matching. The initial nodes were adopted based on the minimum similarity between high-similarity pairs of overlapping nodes. This criterion implies that the adopted initial nodes have higher identifiability than other groups of overlapping nodes.

Unclustered nodes are assigned to each subcluster using bipartite matching. Matching candidates are selected based on their similarity to existing clustered nodes, independent of frames. This process allows highly identifiable nodes to be preferentially assigned to subclusters over less identifiable nodes. The cost function used in bipartite matching is the weighted degree centrality. Weighted degree centrality is the importance criterion of a node that calculates the sum of the edges from the node used in graph theory. In this study, weighted degree centrality is referred to as centrality. The centrality between un-clustered nodes x and the cluster C is represented as follows:

$$Centrality_{WD}(x, C) = \sum_{i=1}^{n_C} w_{x,x_{C_i}} \quad (2)$$

where w denotes the edge between nodes and $x_{C_i} \in C$. The edge is similarity if $similarity > \varepsilon$, otherwise 0. Centrality prevents accidental misclustering because the sum of the edges between the nodes of identical individuals tends to be larger than that between the nodes of non-identical individuals. When all overlapping nodes are assigned to clusters, the unconnected components in each cluster are separated, as shown in Fig.3 subprocess2-2.

Subprocess3. Subprocess3 merges the subclusters into a new tracklet. Subclusters are merged using bipartite matching, which maximizes the centrality between clusters unless there is an overlap or the centrality is less than zero. This process can also be used to merge the clusters between periods.

3.1.2 Correcting tracklets

Sequential NMS. NMS with a high threshold will result in missing detections for individuals with occlusions. The optimal NMS parameter helps to reduce the overlapping boxes of an identical individual and avoids excessive reduction of the overlapping boxes of a non-identical individual. However, it is impossible to reject only the overlapping boxes of an identical individual from a single image.

This study performed an original process called Sequential NMS (SNMS), which aims to reject only overlapping boxes given identical individuals by the sequential overlapping state, instead of NMS. This method considers tracklets with high overlap as identical individuals. In general, NMS evaluates IoU as an overlap metric. However, bboxes with identical individuals will receive a low NMS score if they have different sizes. Thus, SNMS evaluates the overlap coefficient shown in Eq.(3) as the overlap criterion between sets X and Y.

$$Overlap\ Coefficient = \frac{|X \cap Y|}{min(|X|, |Y|)} \quad (3)$$

SNMS calculates the overlap coefficient both temporally and spatially. Temporally SNMS is calculated by the intersection of the appearance times between tracklets. Spatially SNMS is calculated by the averaged overlap coefficient of the bboxes present in the same frame. If the overlap coefficient is above a certain threshold, both temporally and spatially, a small tracklet in the pair is deleted.

Warp tracklets separating. Ideal tracklets contain spatiotemporal continuity. If the tracklet contains spatiotemporal discontinuities, it is likely to contain misclustering. In this study, this condition is defined as a warp. This process detects a warp using motion features and separates the tracklets after the warp as a new tracklet. The warp is detected using the Euclidean distance between the detected bbox position (x_{obs}, y_{obs}) at $t+1$ and the predicted bbox position (x_{pred}, y_{pred}) at $t+1$ calculated from the position information up to t. The predicted bbox position was computed as follows:

$$\begin{Bmatrix} x_{t+1,pred} \\ y_{t+1,pred} \end{Bmatrix} = \begin{Bmatrix} x_{t,obs} \\ y_{t,obs} \end{Bmatrix} + \begin{Bmatrix} e_{x,t} \\ e_{y,t} \end{Bmatrix} \quad (4)$$

$$e_{i,t} = \alpha e_{i,t-1} + (1-\alpha)\left(i_{t,obs} - i_{t-1,obs}\right) \quad (5)$$

where α denotes a momentum term. The prediction of the bbox position uses the weighted cumulative sum of past tracklets, such as StrongSORT.

3.2. Multi-Camera People Tracking

Fig. 5 shows an overview of our MCPT process. Our method uses a hierarchical clustering-based method as well as SCPT. Before performing clustering, this process extracts representative nodes from each tracklet. Clustering is then performed on nodes that are considered highly identifiable among the representative nodes. Tracklets containing only low-identifiability nodes are assigned to each cluster by solving the assignment problem. This process provides stable tracking results, because misclustering is caused by low-identifiable nodes in many cases.

3.2.1 Representative Image Extraction

To address the computational complexity of Re-ID, our framework clusters only representative nodes extracted from each tracklet. Incidentally, to improve the interpretability of Fig.5, Fig.5 and the title of this section are referred to as Representative “Image” Extraction. The Re-ID model can accurately identify individuals if the input nodes contain the entire body and are of high resolution. Therefore, this process extracts representative nodes based on the confidence of pose estimation. Nodes are classified into three levels based on the status of the

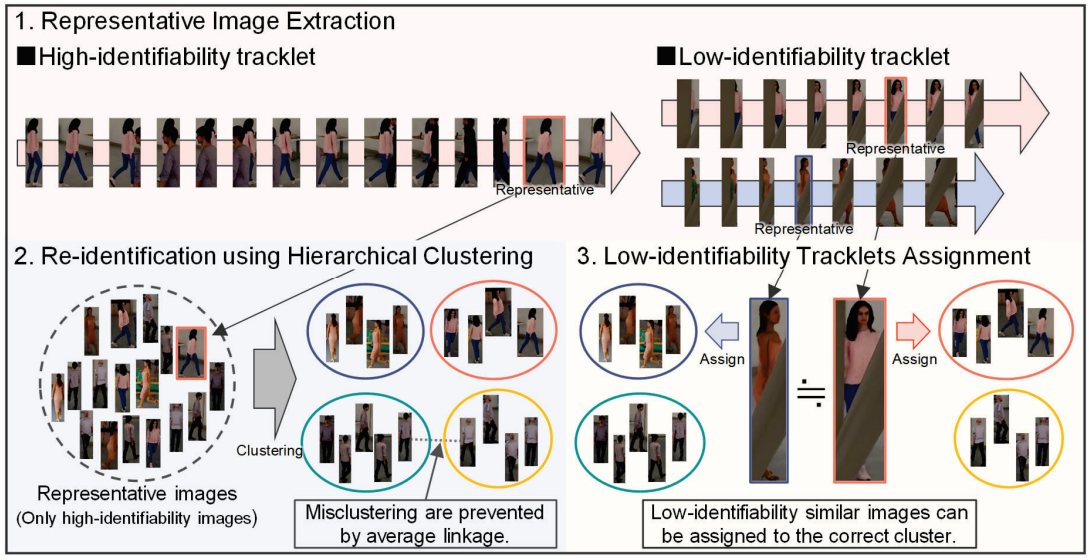


Figure 5: Overview of Multi Camera People Tracking. (1) Representative images are extracted from each tracklet. (2) Re-identification performs only high-identifiability images considered by key points of pose estimation. (3) Tracklets containing only low-identifiability images are assigned to clusters separated in the Re-identification process.

key points, as shown in the table below:

Level	status
Lv.3	All of the confidence of key points over a certain threshold.
Lv.2	At least one of the confidences with symmetry key points of the body is above a certain threshold.
Lv.1	Without Lv.3 or Lv.2.

Table 1: Node status considering key points

When multiple nodes exist at the same level in the tracklet, the nodes with the largest bbox areas are selected from the highest level. Furthermore, the use of high-level images in Re-ID is recommended because low-level nodes may contain occlusions or a unique foreground, which can cause misclustering. In this study, a tracklet containing only low-level nodes is defined as a low-identifiability tracklet. To achieve a highly accurate Re-ID, this process clusters only high-identifiability tracklets and assigns low-identifiability tracklets to each cluster composed of high-identifiability tracklets.

3.2.2 Re-Identification using Hierarchical Clustering

Creating similarity matrix. A similarity matrix is created as the input data for clustering, which represent the similarity between representative nodes. Elements of the similarity matrix less than threshold $1-\varepsilon$ are replaced by zero. In addition, to ensure accurate tracking, the similarities of the matrix are replaced by considering the Euclidean distances measured using world coordinates. If there is no time intersection between the tracklets, the Euclidean distance is not measured.

If the world distance between tracklets is small, these tracklets will be the same individual, and the elements of the matrix are replaced with 1. Conversely, if the world distance is large, these tracklets will be non-identical individuals, and the elements of the matrix are replaced with a negative value. Because the tracklets obtained from SCPT are not always complete, similarities are replaced only when the distances are measured in many frames.

The similarities are replaced with negative values if the minimum Euclidean distance exceeds a certain threshold. Furthermore, if the average world distance over a certain number of frames is less than another threshold, the similarities are replaced with 1. The purpose of replacing it with a negative value rather than negative infinity is to ensure the formation of a correct cluster, even if misclustering occurs between pairs that have not been distance-measured.

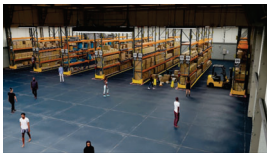
Hierarchical clustering. To perform initial Re-ID, this process uses hierarchical clustering with average linkage as a linkage criterion. Distance d between clusters C_I and C_J is calculated as follows:

$$d(C_I, C_J) = \frac{1}{n_{C_I} \cdot n_{C_J}} \sum_{i=1}^{n_{C_I}} \sum_{j=1}^{n_{C_J}} w_{x_{C_{I,i}} x_{C_{J,j}}} \quad (6)$$

where w denotes the edge between nodes, and x denotes the node in each cluster, as in Eq. (2).

Nodes of non-identical individuals may occasionally have a high similarity; however, in general, nodes of an identical individual tend to have a higher similarity. Therefore, hierarchical clustering using average linkage prevents misclustering because the edges between the nodes of non-identical individuals are averaged by other


(a) Scene01 Camera0008


(b) Scene65 Camera0583


(c) Scene79 Camera0777


(d) Scene88 Camera0917

Figure 6: Examples of dataset

edges, even if there are some edges with occasional high similarities. In particular, negative values in the matrix, replaced by the previous process, help to form a correct cluster.

3.2.3 Low-identifiability Tracklets Assignment

This process assigns low-identifiability tracklets, which are excluded from the representative image extraction and Re-ID processes, to the clusters. By assigning them to restricted clusters, the nodes are clustered more accurately than in ordinary clustering. In addition, if each individual appears several times on the screen, a cluster with small intrinsic nodes is unlikely to be suitable as a representative cluster for an individual. Therefore, tracklets belonging to these clusters are assigned to another cluster if there are sufficient nodes. These tracklets are assigned only if the similarity threshold is exceeded and are assigned to the most frequent cluster that exceeds the threshold.

4. Experiments

4.1. Dataset

The target videos are synthetic animated human data generated using the NVIDIA Omniverse Platform. They have 1080p feeds at 30 fps. Table 2 shows the breakdown of the dataset.

	Scenes	Cameras	Frames	Labels
Train	01-40	360	8,637,840	74,227,868
Validation	41-60	174	4,174,956	26,420,461
Test	61-70	100	2,399,400	-
	71-80	159	3,815,046	-
	81-90	159	3,815,046	-

Table 2: Breakdown of datasets

Fig. 6 displays examples of images in the dataset. These examples are selected from images that capture a wider range in each scene. The training dataset resembles a warehouse space with some shelves along the wall. The validation dataset and scene 61-70 in the test dataset appear to represent the same space as that in the training dataset. However, because of the occlusion caused by shelves placed in the middle of the space, tracking is more difficult than that in the training dataset. Scene 71-80 of the test dataset looks like a supermarket space, and some cameras are placed in another space, such as a backroom. Scene 81-90 of the test dataset looks like a corridor in a hospital. Compared to scene 01-80, cameras are placed in a lower position, and people walk through the nearby cameras.

4.2. Evaluation metrics

Higher Order Tracking Accuracy (HOTA) [4] is an MOT evaluation metric that explicitly balances the effect of performing accurate detection, association, and localization into a single unified metric for comparing the tracking results used in the 2024 AI City Challenge Track1.

Dataset annotations are established according to a set of criteria. The labeling of occluded objects depends on both the height and width visibility requirements. For truncated objects, either the height or the width visibility criterion must be satisfied. Height visibility is assessed based on whether the head is observable, with at least 20% of the object's height visible. Alternatively, if the head is not visible, a minimum of 60% of the object's height should be discernible for labeling. Width visibility requires that more than 60% of the object's body width is visible for labeling purposes.

4.3. Implementation details

In this study, no model training was performed to reduce training costs. Instead, the champion model from the 2023 AI City Challenge Track 1 was used. Evaluating the impact of the model is future work. The detection model is YOLOX-X [7], trained by the MOT17, CrowdHuman [46], CityPersons [47] and ETHZ [48]. This model was utilized in ByteTrack [45]. The feature extractor is OSNet-AIN [5], trained by the Market1501 [49], CUHK03 [50] and MSMT17 [51]. The pose estimation model is HRNet-w48 [6], trained by the COCO [31]. During the preparation process, experiments were performed on one Tesla T4. During the tracking process, experiments were performed on one Xeon Platinum 8175M. The required machine specifications for the tracking process are determined by the number of hard overlapping tracklets. This framework can perform tracking on machines with lower specifications than Platinum 8175M, as long as the camera does not capture a large number of people simultaneously.

To increase the HOTA score, we fine-tune the results using the following process:

Removing noise image. The YOLOX algorithm utilized in this study has a tendency to detect excess body parts, resulting in decreased HOTA scores despite the qualitative accuracy of the tracking. To address this issue,

a subprocess was implemented to remove extraneous images from the tracking results based on the key points of pose estimation. Images with low confidence in keypoints and unusual aspect ratios for full-body shots were removed from each tracklet.

Deleting distant persons. If misclustering occurs in MCPT, the incorrectly assigned node is positioned far from the other nodes. In addition, if a tracklet contains ID switching, where two or more individuals are assigned the same ID, then a person other than the representative nodes is necessarily misclustered because MCPT is performed by only one node of each tracklet.

In this subprocess, the coordinates of the nodes within the same cluster are measured in each frame. If a person is tracked by more than three cameras simultaneously, and the maximum Euclidean distance among these coordinates exceeds a certain threshold, the node with the maximum sum of distances to other nodes is deleted from that cluster as a distant person. Owing to time constraints in the competition, we were unable to experiment with the approach of potentially assigning deleted nodes to existing clusters. It may also be beneficial to repeat the correction process until all nodes are addressed. However, we only submitted the results of a single execution owing to the submission period.

Interpolating missing detection. The objective of this process is to complete the missing detections in each tracklet. However, if all the tracklets are interpolated from start to end, the number of out-of-scope bboxes increases, whereas the number of missing detections decreases. To address this issue, the tracklet is interpolated only if the number of missing detections is below a certain threshold for a continuous number of frames.

4.4. Results

4.4.1 Result on test data

The proposed method recorded the highest HOTA of 71.9446% in the 2024 AI City Challenge Track1. Table 3 displays the results for the top 5 teams.

Rank	Team ID	HOTA
1	**221 (Ours)**	**71.9446**
2	79	67.2175
3	40	60.9261
4	142	60.8792
5	8	57.1445

Table 3: Result on test data

4.4.2 Ablation study on validation data

To demonstrate the usefulness of each process described thus far, our framework evaluated the performance of the validation dataset as the ablation study. SCPT used the same settings in all the studies.

The baseline approach is a simple hierarchical clustering method that employs average linkage with all representative nodes. Ablation study (2) evaluates the influence of the low-identifiability tracklet assignment process on the baseline method. Ablation studies (1) and (2) rely solely on appearance features, whereas ablation study (3) examines the impact of incorporating motion features. As explained in Section 3.2.2, a replacement process based on Euclidean distances, is performed. Ablation Study (4) examines the effects of all processes proposed in this study. Table 4 summarizes the results of the ablation studies.

Method	HOTA (%)	DetA (%)	AssA (%)	LocA (%)
(1) Baseline	57.348	59.023	55.949	87.666
(2) (1) + Tracklets assignment	61.761	63.601	60.082	87.686
(3) (2) + Motion features	67.773	70.654	65.081	87.734
(4) All processes	**68.455**	**71.401**	**65.703**	**87.952**

Table 4: Performance of our framework on validation data

With each process added, the tracking performance improves. Motion features significantly contribute to performance improvements, as shown in Table 4 (3). It has been confirmed that a certain level of performance can be achieved solely through appearance features. In this study, MCPT were completed in about 10~20 minutes. Without using motion features, the processing time decreases significantly. The processing time of SCPT varies depending on the crowding situation. For many cameras, it takes less than five minutes. However, it takes over 30 minutes when there are many people, as shown in Fig. 6.

SCPT errors are sometimes confirmed when the person leaves the screen and another person immediately enters the screen from the same position. These errors can be caused by the similarity of the background.

5. Conclusion

We proposed an MCPT framework based on hierarchical clustering. The framework consists of four main processes: (1) SCPT based on the original clustering method, called Overlap Suppression Clustering; (2) representative image extraction considered by pose estimation; (3) re-identification using hierarchical clustering with average linkage; and (4) low-identifiability tracklet assignment Our framework achieved the highest HOTA of 71.9446% in 2024 AI City Challenge Track1.

Our proposed method presents technical challenges in terms of inference time, since pose estimation is performed additionally. As a future work, we will investigate the correlation between model performance and efficiency.

6. Acknowledgement

To create a program based on our ideas, this research was supported by TQuality Co., Ltd., a technology company. We express our gratitude to the reviewers.

References

[1] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In ICIP, pages 3464–3468, 2016.

[2] Yunhao Du, Yang Song, Bo Yang, and Yanyun Zhao. StrongSORT: Make deepsort great again. In IEEE, pages 8725–8737, 2023.

[3] Jiaxin Li, Yan Ding, and Hualiang Wei. Simpletrack: Rethinking and improving the jde approach for multi-object tracking. Sensors, 22(15), 2022.

[4] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixe, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. International journal of computer vision, pages 548–578, 2021.

[5] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person reidentification. In ICCV, pages 3702–3712, 2019.

[6] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. In TPAMI, pages 3349–3364, 2020.

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

[8] 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 CVPR, 2023.

[9] Hsiang-Wei Huang, Cheng-Yen Yang, Zhongyu Jiang, Pyong-Kun Kim, Kyoungoh Lee, Kwangju Kim, Samartha Ramkumar, Chaitanya Mullapudi, In-Su Jang, Chung-I Huang, and Jenq-Neng Hwang. Enhancing multi-camera people tracking with anchor-guided clustering and spatiotemporal consistency ID re-assignment. In CVPR Workshop, pages 5239–5249, 2023.

[10] Quang Qui-Vinh Nguyen, Huy Dinh-Anh Le, Truc ThiThanh Chau, Duc Trung Luu, Nhat Minh Chung, and Synh Viet-Uyen Ha. Multi-camera people tracking with mixture of realistic and synthetic knowledge. In CVPR Workshop, pages 5495–5505, 2023.

[11] Wenjie Yang, Zhenyu Xie, Yang Zhang, Hao Bing, and Xiao Ma. Integrating appearance and spatial-temporal information for multi-camera people tracking. In CVPR Workshop, pages 5260–5269, 2023.

[12] Andreas Specker and Jurgen Beyerer. ReidTrack: Reid-Only Multi-Target Multi-Camera Tracking. In CVPR Workshop, pages 5442–5452, 2023.

[13] Zongyi Li, Runsheng Wang, He Li, Yuxuan Shi, Hefei Ling, Jiazhong Chen, Bohao Wei, and Boyuan Liu. Hierarchical clustering and refinement for generalized multi-camera person tracking. In CVPR Workshop, pages 5520–5529, 2023.

[14] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, pages 580–587, 2014.

[15] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In ICIP, pages 3645–3649, 2017.

[16] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In CVPR, pages 1735–1742, 2006.

[17] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In SIMBAD, pages 84–92, 2015.

[18] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In CVPR, pages4690-4699, 2019.

[19] 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 CVPR, pages 14750–14759, 2021.

[20] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person reidentification: A survey and outlook. In TPAMI, pages 2872–2893, 2021.

[21] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In ECCV, pages 262–275, 2008.

[22] W.R. Schwartz and L.S. Davis. Learning Discriminative Appearance-Based Models Using Partial Least Squares. In XXII, 2009.

[23] Omar Javed, Khurram Shafique, and Mubarak Shah. Appearance modeling for tracking in multiple non-overlapping cameras. In CVPR, 2005.

[24] A. Bedagkar-Gala and S. K. Shah. A survey of approaches and trends in person re-identification. In Image and Vision Computing, 32(4):270–286, 2014.

[25] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person reidentification: A benchmark. In ICCV, pages 1116–1124, 2015.

[26] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-IoU loss: Faster and better learning for bounding box regression. In AAAI, pages 12993–13000, 2020.

[27] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. FairMOT: On the fairness of detection and re-identification in multiple object tracking. In IJCV, pages 3069–3087, 2021.

[28] Sisi You, Hantao Yao, Bing-Kun Bao, and Changsheng Xu. UTM: A unified multiple object tracking model with identity-aware feature enhancement. In CVPR, pages 21876–21886, 2023.

[29] Patrick Dendorfer, Aljosa Osep, Anton Milan, Konrad Schindler, Daniel Cremers, Ian Reid, Stefan Roth, and Laura Leal-Taixe. Motchallenge: A benchmark for single-camera multiple target tracking. In IJCV, 129(4):845–881, 2021.

[30] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixe. MOT20: A benchmark for multi object tracking in crowded scenes. arXiv preprint arXiv: 2003.09003, 2020.

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In ECCV, pages 740–755, 2014.

[32] Yang Zhang, Hao Sheng, Yubin Wu, Shuai Wang, Weifeng Lyu, Wei Ke, and Zhang Xiong. Long-term tracking with deep tracklet association. In TIP, 29:6694–6706, 2020.
[33] Jinlong Peng, Tao Wang, Weiyao Lin, Jian Wang, John See, Shilei Wen, and Erui Ding. TPM: Multiple object tracking with tracklet-plane matching. In Pattern Recognition, 2020.
[34] Wenhan Luo, Junliang Xing, Anton Milan, Xiaoqin Zhang, Wei Liu, and Tae-Kyun Kim. Multiple object tracking: A literature review. Artificial Intelligence, 2020.
[35] Jia Chen, Fan Wang, Chunjiang Li, Yingjie Zhang, Yibo Ai and Weidong Zhang. Online Multiple Object Tracking Using a Novel Discriminative Module for Autonomous Driving. In Electronics, 2021.
[36] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully convolutional one-stage object detection. In ICCV, pages 9627–9636, 2019.
[37] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In Proceedings of the IEEE/CVF international conference on computer vision, pages 6569–6578, 2019.
[38] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In ECCV, pages 734-750, 2018.
[39] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In NeurIPS, pages 91–99, 2015.
[40] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single Shot MultiBox Detector. In ECCV, pages 21–37, 2016.
[41] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In ICCV, pages 2980–2988 2017.
[42] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In CVPR, pages 779–788, 2016.
[43] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. Botsort: Robust associations multi-pedestrian tracking. arXiv preprint arXiv:2206.14651, 2022.
[44] Jinkun Cao, Xinshuo Weng, Rawal Khirodkar, Jiangmiao Pang, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. arXiv preprint arXiv:2203.14360, 2022.
[45] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. ByteTrack: Multi-object tracking by associating every detection box. In ECCV, 2022.
[46] 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.
[47] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In CVPR, pages 3213–3221, 2017.
[48] A. Ess, B. Leibe, K. Schindler, and L. Van Gool. A mobile vision system for robust multi-person tracking. In CVPR, pages 1–8, 2008.
[49] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person reidentification: A benchmark. In ICCV, pages 1116–1124, 2015.
[50] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. DeepReID: Deep filter pairing neural network for person reidentification. In CVPR, pages 152–159, 2014.
[51] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person reidentification. In CVPR, pages 79–88, 2018.
[52] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In ECCV, pages 17–35, 2016.