BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2025
http://excel.fit.vutbr.cz

Multi-Camera People Tracking for AI City Challenge

Dominik Kroupa*

Abstract

This paper focuses on Multi-Camera People Tracking task, which is one of the challenges issued by AI City Challenge. The aim is to process provided camera records from different locations and track pedestrian movement across these cameras. An offline framework consisting of three main stages if proposed: (1) generation of single-camera tracklets through pedestrian detection (along with keypoint estimation) and appearance feature extraction, (2) refinement and completion of tracklets using appearance features and strategies to reduce identity switches, and (3) inter-camera association via global ID assignment leveraging appearance features. Additionally, a model trained on detected body keypoints is employed for ground position estimation. The solution was evaluated in the AI City Challenge MTMCT Track 1 in the previous year, achieving a Higher Order Tracking Accuracy (HOTA) score of 31.52%, thereby establishing a baseline for multi-camera pedestrian tracking that the future work can extend.

*xkroup12@stud.fit.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

Multi-Target Multi-Camera Tracking (MTMCT) is a challenging task in computer vision, where the aim is to track multiple objects across multiple cameras. Algorithms being developed to handle such tasks need to address challenges such as changes of appearance due to light conditions or object occlusions, which may lead to identity switches, resulting in decreased tracking precision. The pedestrian multi-camera tracking result can be further utilized for tasks such as crowd analysis.

This paper takes part in the AI City Challenge [1] benchmark, specifically in the issued challenge *Track 1*, where the researchers are given a dataset consisting of roughly 1,300 camera recordings that capture various environments (called scenes), where each environment is a standalone multi-camera subset, each subset containing from 9 up to 16 camera recordings.

While some teams [2] work on approaches like clustering to refine appearance features, to further improve a common MTMCT pipeline that utilizes detection, re-identification, single-camera tracking and multi-camera matching, other teams experiment with other approaches, like Yoshida et al. [3], that perform single-camera tracking by clustering detected pedestrians based on their appearance and with further refinement of such clusters, this solution achieved highest HOTA score of 71.94%. Other teams, like Wang et al. [4], fully utilize information about camera calibration by projecting multi-camera views into a unified Bird's-Eye View (BEV) representation and perform detection and tracking in the BEV space.

This paper proposes an offline framework consisting of three stages to deal with the MTMC tracking problem. Firstly, we generate single-camera tracklets and extract appearance features. Secondly, we perform refinement extracted tracklets to reduce identity switches. Finally, we perform inter-camera association by using appearance features. For the estimation of the ground position for each person, we trained a regression model with the detected body keypoints as input.

The contributions are as follows. Implementation of an offline multi-stage MTMCT pipeline, integration of keypoint-based ground position estimation method utilizing the YOLOV8x-pose model's output, and exploration of re-identification (ReID) performance on a person ReID dataset, that was created using the train subset of AI City Challenge MTMCT dataset.

2. Proposed Pipeline

Figure 1 illustrates the multi-stage MTMCT pipeline. Videos of a given scene, found in the provided, synthetically generated AI City Challenge (AICC) dataset, are first processed by the YOLOv8x-pose [5] detection model, resulting in detections and keypoints on output. The keypoint information is then used to estimate the position of the pedestrian on the ground. Detection information is used for feature extraction. For feature extraction, we experimented with two main models in the submissions *osnet_x1_0* model by Zhou et al. [6], and the *CLIP* model by Radford et al. [7]. Detections and ReID features are then passed to the tracking algorithm, where we again tried two different trackers, ByteTrack [8] and Bot-SORT [9]. Now that we have single-camera tracklets, we refine them in order to reduce identity switches. For this, we utilize information about the closeness of tracked bounding boxes in each frame, which we store in the single-camera tracking stage. Whenever tracklets were closer than given threshold, implying that identity switch could have occurred, we check the similarity of identity vectors to confirm and correct such cases. The refined single-camera tracklets are then matched together across cameras based on their appearance similarity, producing the MTMCT result.

3. Re-ID Dataset

In both the synthetically generated dataset and real-world environments, there is an occurrence of people of similar appearance, making it difficult to correctly assign identities due to the similar feature distance. To address this issue, we created the pedestrian ReID dataset from the provided AICC dataset. While creating a ReID dataset, we found that same pedestrians have different IDs assigned when they appear in more than one scene (MTMCT subset). Since having the same person labeled with different ID is not correct in ReID datasets, we used the aforementioned *osnet_x1_0* model to group most similar pedestrians and then manually assigned correct IDs to such pedestrians. Figure 2 shows a few samples from the created dataset, which contains 429 unique identities in total.

4. Increasing Inter-Class Separability

Figure 3 shows separability between positive and negative pairs (intra-class and inter-class) of differently trained *osnet_x1_0* models, with (a) being a model pre-trained by the authors on Market-1501 [10] dataset with softmax loss function for 60 epochs. Models (b) and (c) were trained for 60 epochs on the created ReID dataset, where model (b) has been trained with circle loss and (c) with softmax loss. The graphs of these two models show increase in inter-class separability, promising improvement in feature vector matching.

5. Activation Maps

Figure 4 shows an example output with visualized activation maps of the models described in Section 4. The images are from the testing subset of the created ReID dataset, and thus none of the models were trained on them. Based on this output model, (c) shows activation both on person's head and clothing, which often provides crucial details that help with correct recognition.

6. Conclusions

Although this paper does not achieve state-of-the-art performance, our solution provides a baseline of the MTMCT pipeline, and we hope to further improve the resulting score by using the improved ReID models in this years' AI City Challenge that has been delayed for this year.

Acknowledgements

I would like to thank my supervisor prof. Ing. Adam Herout, Ph.D., for his help and Ing. Markéta Juránková, Ph.D., for providing a ground position estimation model.

References

[1] Shuo Wang, David C. Anastasiu, Zheng Tang, Ming-Ching Chang, Yue Yao, Liang Zheng, Mohammed Shaiqur Rahman, Meenakshi S. Arya, Anuj Sharma, Pranamesh Chakraborty, Sanjita Prajapati, Quan Kong, Norimasa Kobori, Munkhjargal Gochoo, Munkh-Erdene Otgonbold, Ganzorig Batnasan, Fady Alnajjar, Ping-Yang Chen, Jun-Wei Hsieh, Xunlei Wu, Sameer Satish Pusegaonkar, Yizhou Wang, Sujit Biswas, and Rama Chellappa. The 8th AI City Challenge. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2024.

[2] Jeongho Kim, Wooksu Shin, Hancheol Park, and Donghyuk Choi. Cluster self-refinement for enhanced online multi-camera people tracking. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 7190–7197, 2024.

[3] Ryuto Yoshida, Junichi Okubo, Junichiro Fujii, Masazumi Amakata, and Takayoshi Yamashita. Overlap suppression clustering for offline multi-camera people tracking. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 7153–7162, 2024.

[4] Yizhou Wang, Tim Meinhardt, Orcun Cetintas, Cheng-Yen Yang, Sameer Satish Pusegaonkar, Benjamin Missaoui, Sujit Biswas, Zheng Tang, and Laura Leal-Taixé. Mcblt: Multi-camera multi-object 3d tracking in long videos, 2025.

[5] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8, 2023.

[6] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[8] 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, 2022.

[9] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. Bot-sort: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651*, 2022.

[10] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *IEEE International Conference on Computer Vision*, 2015.