An Enhanced Surveillance System Integrating Pedestrian Attribute Recognition and Multi-Target Multi-Camera Tracking Technologies.

Yi-Cheng Lai[1], Chih-Yuan Huang[2*], Yu-Sheng Hsu[3]

[1] Master of Science Student, Department of Civil Engineering, National Central University, Taiwan.

[2] Associate professor, Center for Space and Remote Sensing Research, National Central University, Taiwan.

[3] Research Assistant, Center for Space and Remote Sensing Research, National Central University, Taiwan.

*cyhuang@csrsr.ncu.edu.tw

Abstract

Surveillance systems usually connect to multiple cameras for real-time video feeds. Traditionally, finding certain objects from these video feeds is a time-consuming and labor-intensive task, requiring manual monitoring or review by human operators. Such manual processes can lead to inefficiencies, especially when processing large amounts of footage in a short period. With the rapid development of AI-based computer vision, this research tries to enhance surveillance systems by integrating Pedestrian Attribute Recognition (PAR) and Multi-Target Multi-Camera Tracking (MTMCT) technologies, creating a more efficient and automated solution. The PAR is to identify distinct appearance characteristics of a targeted person from images, enabling surveillance systems to identify and classify various attributes of pedestrians in a textual format, such as gender, age, hairstyle, types of clothing, accessories, etc. These attributes may also include more detailed features, such as posture and movement patterns, allowing for more precise search capabilities. On the other hand, the MTMCT is for identifying and tracking moving objects across multiple cameras, which requires a series of AI-based processes, including object detection and tracking on images from individual cameras, re-identify objects when moving from one camera to another, etc. This tracking process involves various complexities, including ensuring accurate re-identification of objects when they leave one camera's view and enter another, which is key to maintaining continuous tracking. By integrating PAR and MTMCT technologies, a surveillance system can quickly locate target individuals from a large amount of image data based on space, time, and appearance information, which is extremely useful for various applications, including crime investigations, searching for missing persons, etc. In addition, as PAR models can extract local and global appearance features, these features can also help improve the re-identification accuracy in the MTMCT. As a result, we have constructed a surveillance system that provides an effective application for finding pedestrians with an improved MTMCT.

Keywords: Pedestrian Attribute Recognition, Multi-Target Multi-Camera Tracking, Surveillance, Person Retrieval, Re-identification.

Introduction

In recent years, with the development of accelerated computing software and hardware, computer vision has made significant leaps forward (Zaki & Sayed, 2014). Among these advancements, one of the critical concerns for public safety is the ability to quickly and

accurately search for individuals, particularly in surveillance and law enforcement contexts (Zhang et al., 2023). This issue can be explored through the integration of multi-target multi-camera tracking (MTMCT) technology and pedestrian attribute recognition (PAR). MTMCT allows for continuous tracking of individuals across multiple cameras, while PAR assists in tracking and re-identifying individuals by recognizing their appearance characteristics.

PAR primarily focuses on analyzing the appearance characteristics of individuals, including aspects such as age, gender, clothing style, clothing color, and accessories (Li et al., 2016). These attribute features are helpful in re-identifying and matching pedestrians across multiple cameras. Most pedestrian attribute recognition methods rely on deep learning, employing neural network models trained on large datasets to extract and identify pedestrian attribute features, which facilitates subsequent analysis and application. This approach helps improve recognition accuracy, enabling the system to effectively handle complex real-world scenarios.

On the other hand, MTMCT involves tracking pedestrian targets across multiple cameras and ensuring their identity consistency, critical for accurate cross-camera re-identification. This capability is of significant value in crowd management and security surveillance applications (Zhang et al., 2018).

This study aims to enhance the pedestrian search and re-identification capabilities of surveillance systems by exploring and integrating PAR with MTMCT technologies. A PAR and query system is designed to accurately extract and identify various pedestrian attributes, thereby improving surveillance efficiency, enabling rapid target location and tracking, and strengthening public safety management. By improving the accuracy of target identification and tracking, this system directly supports public safety management through quicker response times and more efficient resource allocation. Through the proposed methods and system, this study aims to enhance overall monitoring capabilities to offer a more effective and efficient solution for public safety.

Literature Review

a. Pedestrian Attribute Recognition (PAR)

Early PAR models were primarily based on handcrafted feature extraction techniques and used traditional machine learning models for classification. These methods typically rely on low-level features, such as color histograms, texture features, and Histogram of Oriented

Gradients (HOG) (Dalal & Triggs, 2005), to represent pedestrian attributes (Layne et al., 2012; Liu et al., 2017). While these handcrafted features capture basic visual characteristics, they struggle to handle complex and diverse pedestrian attributes in real-world scenarios. The descriptive power of these handcrafted features is relatively weak, making it challenging to meet the requirements for precise attribute recognition. This limitation drives the shift towards deep learning methods, which can extract more robust and high-level features, significantly improving the recognition of multiple pedestrian attributes in complex environments. However, even deep learning methods face challenges in handling complex scenarios, particularly when identifying multiple attributes simultaneously in highly varied conditions.

With the development of deep learning technology, convolutional neural networks (CNNs) (LeCun et al., 1998) have gradually replaced handcrafted feature-based PAR methods. For example, Zhao et al. (2018) proposed the DeepMAR model, which transforms the multi-label classification problem into multiple binary classification tasks. By using a deep learning network to learn multi-level attribute representations in pedestrian images, the model significantly improves attribute recognition accuracy (Zhao et al., 2018). Li et al. (2016) proposed the Hydraplus Net model, which utilizes a multi-scale feature learning architecture to effectively combine global and local features, thereby enhancing the capability of fine-grained attribute recognition.

Additionally, Sudowe & Tsogkas (2015) introduced a multi-task learning framework that jointly trains PAR and pedestrian re-identification tasks, leveraging the inter-task correlations to improve attribute recognition performance. This multi-task learning strategy demonstrates advantages in fully exploiting the relationships between different attributes.

In recent years, the application of the self-attention mechanism and Transformer architecture in PAR has further improved model performance. The self-attention mechanism enhances attribute recognition accuracy by learning the importance of different regions within an image, adaptively focusing on areas relevant to attributes while ignoring irrelevant background information. For example, Tang et al. (2019) proposed the Attribute Context Network (ACN) model, which integrates the self-attention mechanism into CNNs, allowing the model to better focus on key areas within pedestrian images and significantly improve attribute recognition performance. The Transformer architecture has also begun to demonstrate its advantages in PAR. Unlike traditional convolutional operations,

Transformers rely on a global self-attention mechanism to capture long-range dependencies within an image, making them particularly well-suited for handling correlations between attributes (Liu et al., 2021).

b. Multi-Target Multi-Camera Tracking (MTMCT)

Before the emergence of MTMCT technology, Multi-Target Tracking (MTT) primarily focused on tracking targets within a single camera's field of view. Combining detection and tracker techniques allowed for identifying and tracking multiple targets in these methods. Kalman filters (Kalman, 1960) and the Hungarian algorithm (Kuhn, 1955) were commonly used to solve the multi-target association problem. However, single-camera systems were limited by their fixed viewpoints and issues such as occlusions, making them inadequate for addressing the challenges posed by collaborative tracking across multiple cameras. For example, Breitenstein et al. (2009) proposed a multi-target tracking method based on object detection and optical flow, which tracks target trajectories through dynamic modeling of multiple objects. While these methods perform well within a single camera's field of view, they face issues when there are no overlapping areas between cameras or when targets move out of one camera's range into another.

To accurately re-identify individuals across multiple cameras, the system must match appearance features and synchronize spatio-temporal information, as variations in timing and perspectives across cameras can complicate the process. With the introduction of MTMCT, several new challenges emerged. One of the most critical issues is cross-camera re-identification (ReID). When targets move between different cameras, a system must accurately re-identify them based on their appearance features. Despite changes in lighting or viewpoints, the system should maintain consistent identification of the target (Zheng et al., 2011). Additionally, spatio-temporal synchronization between cameras poses another challenge in multi-camera systems. Due to variations in the perspectives and timing of different cameras, the system must handle spatio-temporal transformations to ensure that the trajectories of targets can be continuously tracked and accurately connected across different cameras (Chen et al., 2017). Furthermore, occlusions and environmental changes present additional difficulties for tracking systems. In real-world scenarios, targets may be occluded by other objects or pedestrians, and variations in camera positions and lighting conditions can also affect tracking. This requires the system to possess a high level of robustness, ensuring stable tracking performance across diverse environments (Wang et al.,

2014).

Early MTMCT systems often employed color and appearance feature matching for cross-camera tracking. For example, Zheng et al. (2011) proposed an appearance-based cross-camera matching method that used color histograms and texture features to associate and match targets. As these methods are relatively simple, their performance tends to be unstable under changing lighting conditions and occlusions.

Traditional appearance-based methods, such as color and texture matching, struggle in dynamic environments with varying lighting and occlusions. Deep learning techniques, particularly CNNs, have been increasingly adopted in MTMCT to address these limitations by learning more robust appearance features. ReID technology has become a core component of MTMCT, playing a crucial role in solving the target-matching problem between different cameras. Wang et al. (2014) proposed a deep learning-based cross-camera ReID method that significantly improved the accuracy of cross-camera matching by learning the deep appearance features of targets. Additionally, Chen et al. (2017) developed a CNN-based model that integrates spatial and temporal information to enhance the precision of cross-camera tracking. These deep-learning techniques effectively address issues such as limited overlapping areas between cameras and significant variations in target appearance.

Methodology

a. Data Used

The data used in this study primarily includes the PAR dataset and the MTMCT dataset.

1. PAR dataset

   The PAR dataset used in this study is the publicly available UPAR2024 (Cormier et al.,2023) dataset, which is composed of the PA100K (Liu et al., 2017), PETA (Deng et al., 2014), RAPv2 (Li et al., 2018), and Market1501 (Zheng et al., 2015) datasets. This dataset contains many annotated pedestrian images, each accompanied by labeled attribute information such as age, clothing, and accessories.

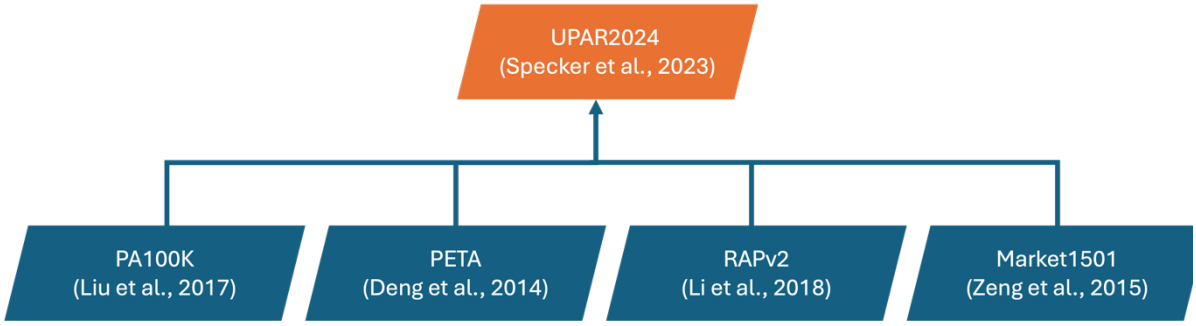

Figure 1: Composition of UPAR2024.

2. MTMCT dataset

   Based on the seventh AI City Challenge (Naphade et al., 2023), the MTMCT dataset includes images captured by multiple cameras simultaneously, covering various real-world scenarios. These images come from both real and synthetic environments, showcasing different pedestrian postures, lighting conditions, and perspectives. After processing and cropping, the dataset faces challenges in pedestrian re-identification (ReID) and pedestrian attribute recognition (PAR), highlighting the difficulties in maintaining accurate detection and tracking under diverse environmental conditions.

The PAR dataset is diverse, featuring pedestrians from different age groups, genders, and clothing styles, with detailed attribute annotations. The MTMCT dataset, derived from the seventh AI City Challenge, captures temporally and spatially continuous pedestrian activities through footage from cameras placed in various positions across real and synthetic environments. The combination of different environments introduces variations in lighting conditions, perspectives, and pedestrian postures. These datasets present significant challenges for image processing and pedestrian recognition tasks, with the MTMCT dataset further complicated by the lower resolution of some footage, adding difficulty to tasks such as re-identification (ReID) and pedestrian attribute recognition (PAR).

b. Research Design

This study designs a system that integrates PAR with MTMCT technologies to enhance surveillance capabilities. The system consists of two main components: PAR, which focuses on identifying and extracting pedestrian attributes, and MTMCT, which handles the continuous tracking of multiple individuals across different camera views. The model is trained and tested using both publicly available datasets and real-world surveillance footage to ensure robustness and adaptability. Finally, the information from PAR and MTMCT is integrated to create a unified surveillance system, enabling precise identification and

continuous tracking of individuals across multiple cameras, thereby improving overall system effectiveness.

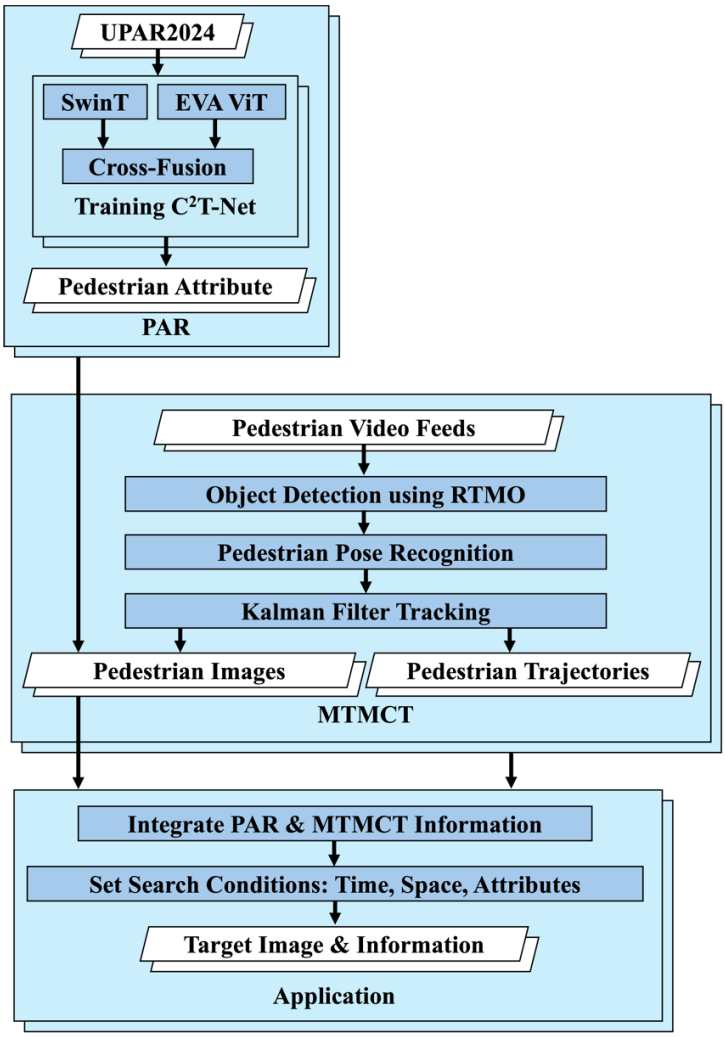

Figure 2: Workflow.

c. Pedestrian Attribute Recognition (PAR)

The goal of pedestrian attribute recognition (PAR) is to capture the unique appearance characteristics of target pedestrians from images. Since everyone's appearance is generally distinct, this feature can be applied in various scenarios such as personnel searches, criminal investigations, and unmanned stores. This study utilizes the UPAR2024 datasets to train the Channel-Aware Cross-Fused Transformer-Style Networks (C^2T-Net) (Bui et al.,2024) model. C^2T-Net is designed to better capture both local and global appearance features by combining two transformers for image recognition: the Shifted Window Transformer (SwinT) (Liu et al., 2021) and the EVA Vanilla Transformer (EVA ViT) (Fang et al., 2023).

SwinT employs shifted windows to more effectively identify different areas of the human body and uses the Channel-Aware Self-Attention mechanism (Vaswani et al., 2017) to focus on specific information within each channel, such as color or texture features, while EVA ViT helps reconstruct occluded parts of the human body.

During PAR, SwinT and EVA ViT perform feature extraction simultaneously. SwinT focuses on capturing detailed local features, while EVA ViT is dedicated to extracting global appearance features. The extracted features are then integrated through a cross-fusion mechanism (Chen et al., 2021), enhancing the model's ability to recognize multi-scale features. Once feature extraction and integration are complete, these features are converted into feature vectors and compared with pre-labeled textual attribute tags. By calculating the similarity between these feature vectors, the relative probability of each attribute label is determined, thereby identifying the pedestrian's attributes.

d. Multi-Target Multi-Camera Tracking (MTMCT)

The objective of multi-target multi-camera tracking (MTMCT) is to continuously track multiple targets across different cameras, enabling cross-camera re-identification. This technology can be applied in various scenarios such as crowd management, security surveillance, and smart cities. In this study, we utilize the RTMO multi-person pose estimation technology (Lu et al., 2024) combined with human pose recognition (Maji et al., 2022) and Kalman filtering (Kalmen, 1960) for multi-target multi-camera tracking. First, RTMO is used for pedestrian detection and pose estimation in surveillance footage, identifying both their locations and postures. Next, human pose recognition is employed to extract detailed motion features, enhancing the system's ability to differentiate and track pedestrians based on their pose. Finally, Kalman filtering is applied to predict and continuously update each pedestrian's movement trajectory across camera frames.

e. Integrated System

To achieve cross-camera pedestrian re-identification, we integrate the pedestrian attribute recognition module to extract appearance features, including age, clothing color, and gender, which serve as auxiliary information for multi-camera tracking. Additionally, this study has developed an integrated system that allows for the search of pedestrian targets by setting parameters such as time, space, and appearance feature descriptions, enabling more precise pedestrian target search and identification.

Results and Discussion

First, this study employs C^{2}T-Net for pedestrian attribute recognition and trains the model on the UPAR2024 dataset. During the training process, metrics such as loss function value, accuracy, and precision were evaluated (as shown in Figures 3 and 4). The results indicate that the model achieved optimal performance at the fourth epoch.

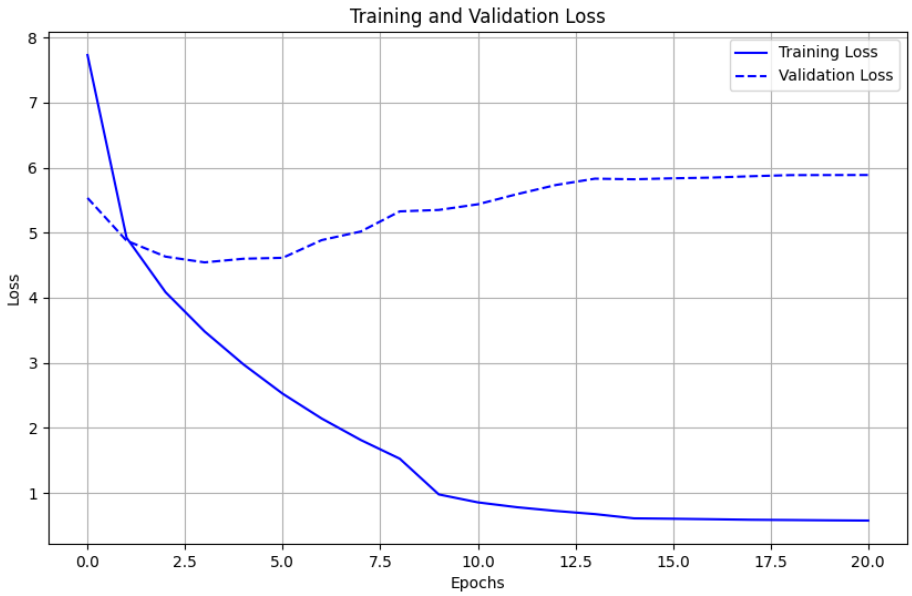

Figure 3: the loss of Training and Validation.

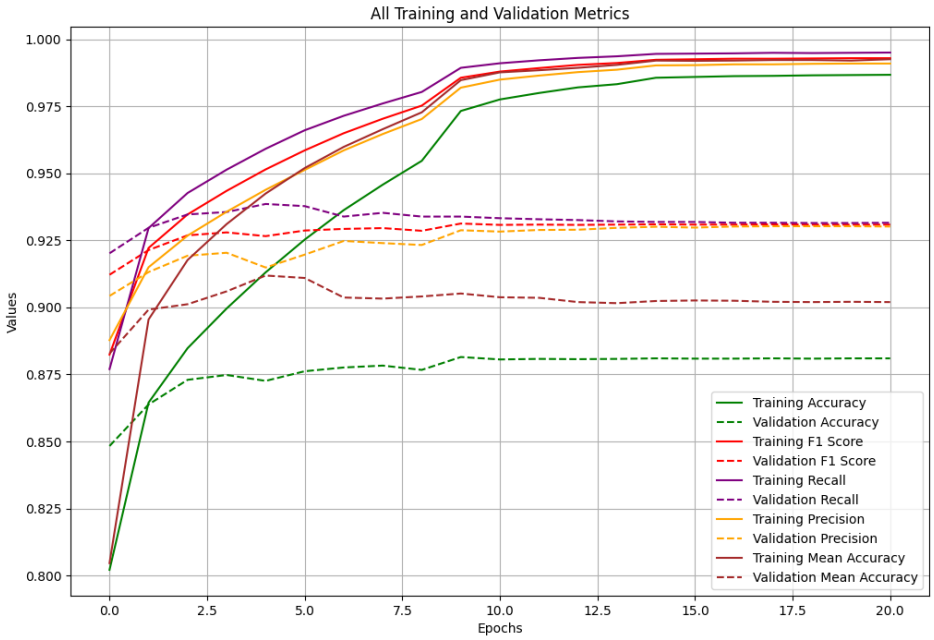

Figure 4: All Training and Validation Metrics.

However, based on the trends shown in the figures, as training progressed, the training loss continued to decrease, while the validation loss, after an initial drop, began to rise and eventually stabilized, indicating that the model experienced overfitting. Therefore, this study selected the model from the fourth epoch, where the best performance was achieved, to avoid the negative impact of overfitting on the model's generalization ability.

Figure 5 illustrates the variation in average accuracy for all attributes across different threshold values. According to the chart, the average accuracy shows little variation under different thresholds, consistently remaining around 73%. This indicates that even with changes in the threshold, the overall predictive performance of the model does not significantly change. This could be because the model has already achieved a high level of stability in predicting all attributes, and minor adjustments to the threshold do not have a significant impact on the overall accuracy.

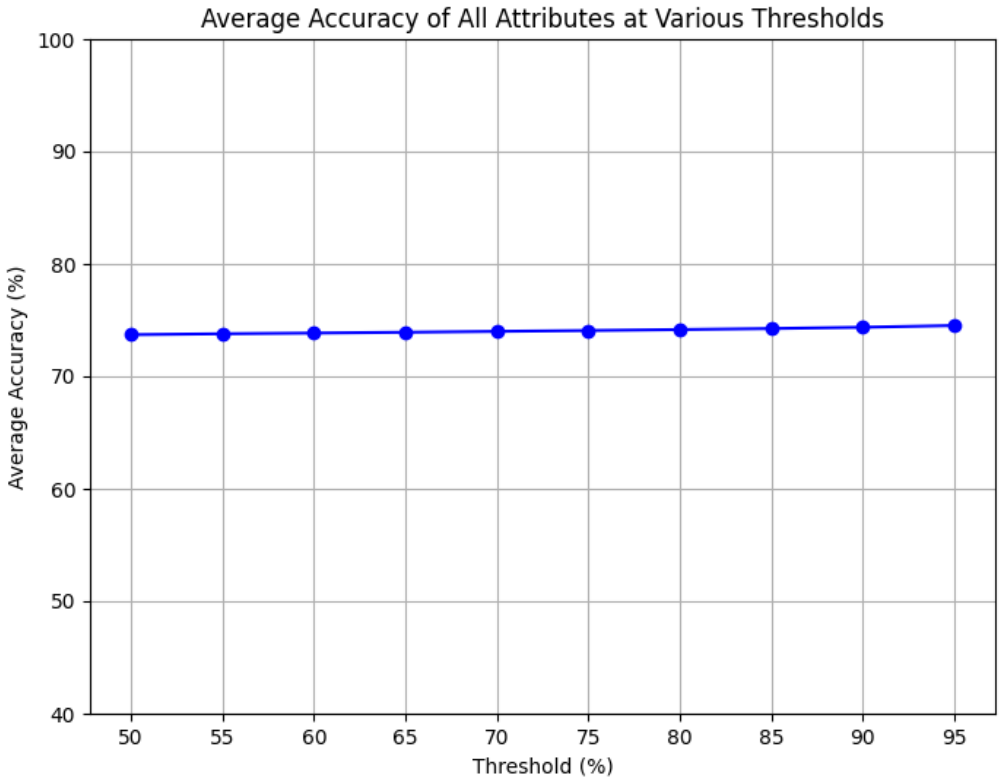

Figure 5: The Average Accuracy of All Attribute at Various Thresholds.

Figure 6 provides a detailed breakdown of the model's accuracy for different attributes (such as age, gender, hair length, upper-body color, lower-body type, etc.) across varying threshold values. In this figure, we can observe that the accuracy for each attribute slightly increases or remains stable as the threshold value rises, but overall changes are minimal. The prediction accuracy for the age attribute consistently stays above 90%, indicating that the model is highly accurate in predicting this attribute. However, the accuracy for the gender and hair length attributes is relatively lower, hovering around 50%, suggesting that the model's performance in predicting these attributes is not ideal.

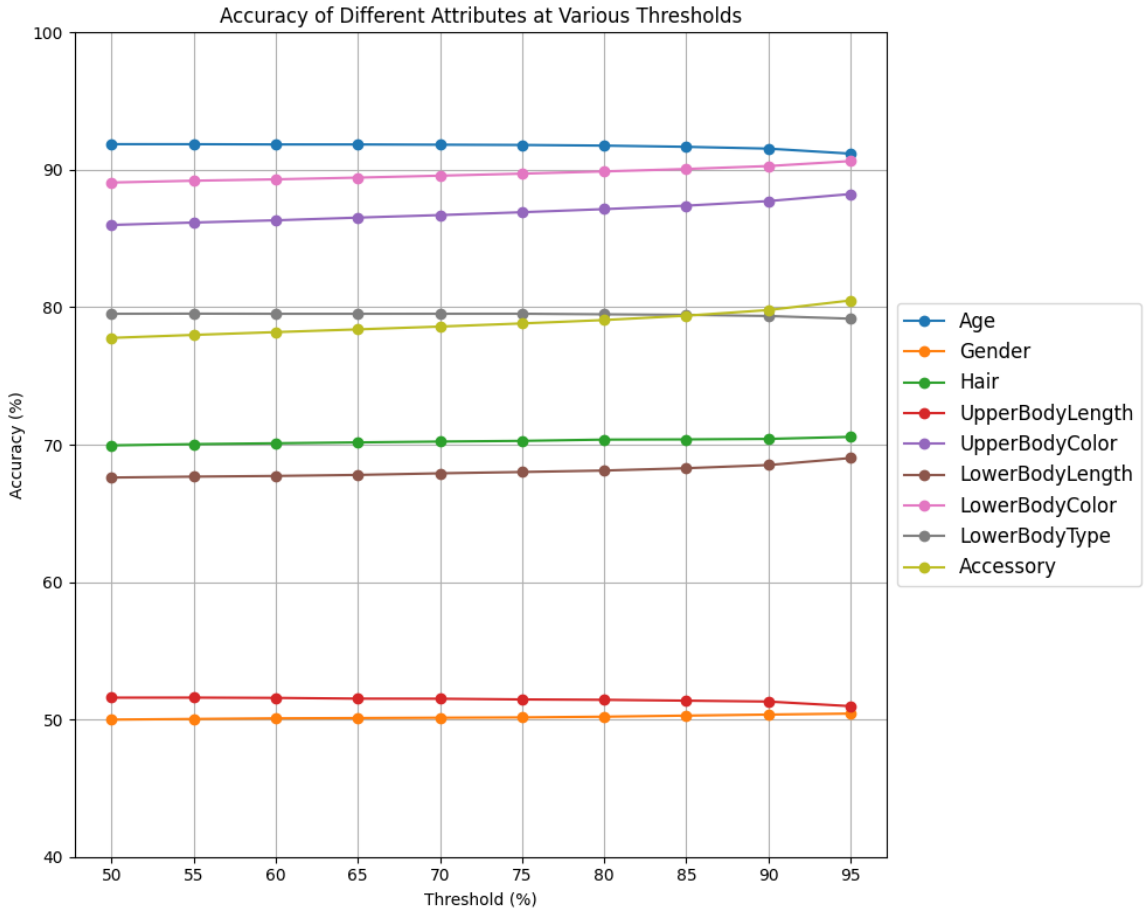

Figure 6: The Accuracy of Different Attributes at Various Thresholds.

The small variation in accuracy shown across both charts could be attributed to several factors:

Stability of model predictions: The model exhibits relative stability in predicting different attributes, suggesting that its decision boundaries are well-defined. This implies that regardless of changes in the threshold value, the model's ability to correctly classify attributes is not significantly affected. When the model's predictions are close to the decision boundary, changes in the threshold would impact the results. However, if the predictions are far from the boundary, threshold adjustments may not lead to notable fluctuations in accuracy.

These charts reveal the model's stability and accuracy when predicting various attributes. Although the effect of different threshold values on the prediction results is minimal, it indicates that the model is highly accurate in predicting certain attributes, such as age, while its performance in predicting other attributes, like gender, is relatively weaker. Overall, the

model demonstrates a high level of stability, but there is still room for improvement in the prediction of certain attributes.

Data quality and attribute characteristics: For certain attributes, like age, the annotations may be highly accurate and consistent, allowing the model to effectively learn these patterns. As a result, threshold adjustments do not significantly impact the accuracy. For other attributes that are more difficult to distinguish, such as gender or hair length, the model might struggle to capture subtle differences between them, leading to lower accuracy and minimal variation across different thresholds.

This study integrates PAR with MTMCT to establish a pedestrian attribute-based search system, allowing users to specify attributes such as age group, gender, and clothing. The search attributes are optional, enabling effective searches even when some information is uncertain.

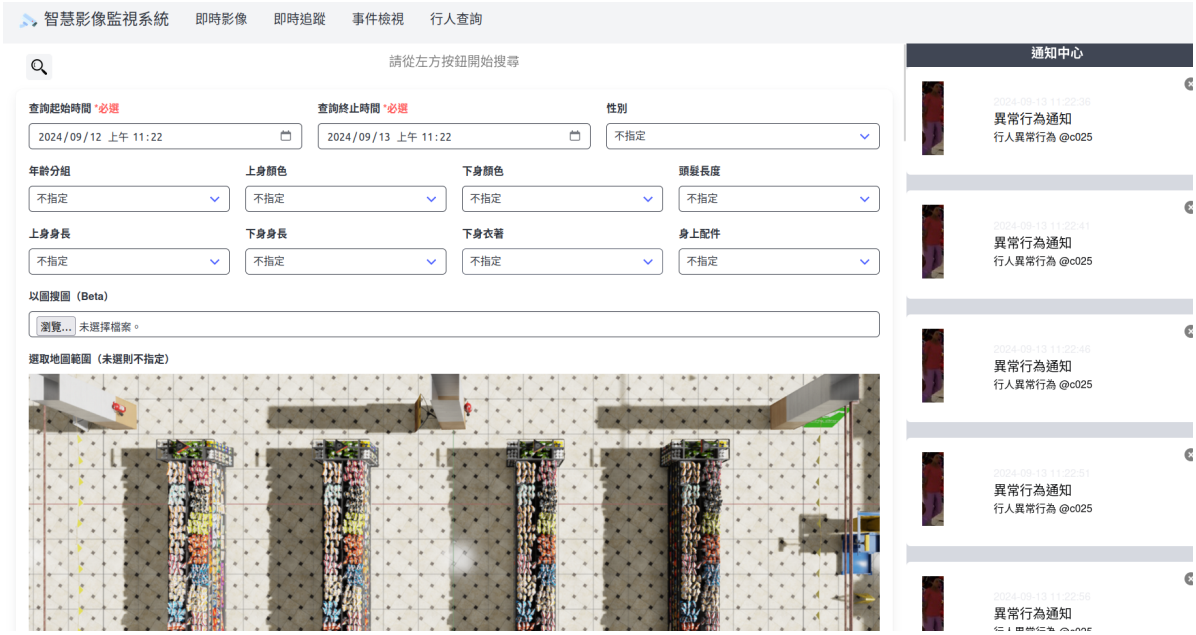

Figure 7: Application of the System.

We also tested pedestrian attribute recognition in actual surveillance footage from the Taoyuan Metro A18 Taoyuan High-Speed Rail Station, as illustrated in Figure 8. Despite the resolution of the surveillance footage, the system was able to clearly analyze the attributes.

Figure 8: Pedestrian in the actual surveillance footage.

Figure 9 shows the detailed feature analysis results from the pedestrian attribute recognition system. According to the system's recognition, the pedestrian was identified as carrying a bag, with an accuracy of 96.40%. The pedestrian was classified as an adult, with an accuracy of 97.11%, and the system identified the pedestrian as female with an accuracy of 99.98%. Furthermore, the pedestrian's hairstyle was recognized as long hair, with an accuracy of 54.93%. In terms of clothing, the system confirmed that the pedestrian's lower body was wearing a long garment, with an accuracy of 88.86%, and it identified that the pedestrian was wearing a skirt or dress with an accuracy of 99.99%. The color of the upper garment was recognized as white, with an accuracy of 98.05%, and the system identified the upper garment as short-sleeved with an accuracy of 99.91%.

Figure 9: Pedestrian Attribute Recognition (1).

Figure 10 shows the results of the pedestrian attribute recognition system. According to the system's analysis, the pedestrian was identified as carrying a backpack with an accuracy of 99.84%. The pedestrian was classified as an adult, with an accuracy of 99.07%, and was recognized as having short hair with an accuracy of 99.49%. In terms of clothing, the system confirmed the pedestrian's lower garment as long with an accuracy of 86.72%, and the accuracy for identifying a skirt or dress was 99.97%. The color of the upper garment was recognized as blue with an accuracy of 99.72%, and the system identified the upper garment as short-sleeved with an accuracy of 99.98%.

Figure 10: Pedestrian Attribute Recognition (2).

Conclusions and Recommendation

This study integrates pedestrian attribute recognition with multi-target multi-camera tracking to develop an enhanced surveillance system. The pedestrian attribute recognition capability effectively extracts various pedestrian attributes, while the multi-target multi-camera tracking technology enables continuous tracking across multiple cameras. The search system allows users to locate pedestrian targets based on appearance attributes and spatiotemporal information. The research results demonstrate that users can perform precise target searches based on pedestrian appearance attributes, providing crucial technical support for the development of future intelligent surveillance systems.

In the future, the feature extraction methods used in pedestrian attribute recognition will be incorporated into the re-identification process of multi-target multi-camera tracking. Since pedestrian attribute recognition can accurately extract both local and global features, it is expected to represent an individual’s uniqueness more comprehensively, thereby improving the overall performance of re-identification. Additionally, near-real-time system

implementation will be a future area of exploration for this study. Establishing a deep learning-based image recognition system that is both fast and accurate within certain computational resource constraints remains a significant challenge for practical applications.

References

Zaki, M. H., & Sayed, T. (2014). Using automated walking gait analysis for the identification of pedestrian attributes. *Transportation Research Part C: Emerging Technologies, 48*, 16–36.

Zhang, Y., Zhang, F., Jin, Y., Cen, Y., Voronin, V., & Wan, S. (2023). Local correlation ensemble with GCN based on attention features for cross-domain person Re-ID. *ACM Transactions on Multimedia Computing, Communications, and Applications, 19*(2), 1–22.

Li, D., Zhang, Z., Chen, X., Ling, H., & Huang, K. (2016). A richly annotated dataset for pedestrian attribute recognition. *arXiv preprint arXiv:1603.07054*.

Zhang, W., Li, Y., Lu, W., Xu, X., Liu, Z., & Ji, X. (2018). Learning intra-video difference for person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology, 29*(10), 3028–3036.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 886–893).

Layne, R., Hospedales, T. M., & Gong, S. (2012). Person re-identification by attributes. In *Proceedings of the British Machine Vision Conference (BMVC)* (pp. 1–11).

Liu, X., Zhang, S., Huang, J., & Gao, Z. (2017). Pedestrian attribute recognition in surveillance: Dataset and approach. *Pattern Recognition, 63*, 148–160.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278–2324.

Zhao, L., Tian, Y., Yan, S., & Tang, J. (2018). Deep learning-based pedestrian attribute recognition. *IEEE Transactions on Image Processing, 27*(6), 3071–3084.

Li, W., Zhao, R., & Wang, X. (2016). Hydraplus-net: Attentive deep features for pedestrian analysis. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 350–359).

Sudowe, P., & Tsogkas, S. (2015). Attribute-based person re-identification with multi-task

learning. In *Proceedings of the IEEE International Conference on Computer Vision Workshops* (pp. 573–580).

Tang, Y., Niu, L., Huang, W., & Yang, J. (2019). Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization. *IEEE Transactions on Image Processing, 28*(7), 3443–3455.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 10012–10022).

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering, 82*(1), 35–45.

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly, 2*(1–2), 83–97.

Breitenstein, M. D., Reichlin, F., Leibe, B., Koller-Meier, E., & Van Gool, L. (2009). Robust tracking-by-detection using a detector confidence particle filter. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1515–1522).

Zheng, L., Zhang, L., & Huang, Y. (2011). Person re-identification by multiple camera networks: A comprehensive survey. *IEEE Transactions on Image Processing, 22*(12), 4310–4321.

Chen, Y., Ai, H., & Shang, C. (2017). Multi-camera multi-target tracking with space-time consistency. *IEEE Transactions on Circuits and Systems for Video Technology, 27*(10), 2167–2179.

Wang, X., Zhang, Z., & Zhang, H. (2014). Cross-camera pedestrian identity verification using part-based CNN. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 70–78).

Cormier, M., Specker, A., Junior, J., Jacques, C. S., Moritz, L., Metzler, J., ... & Beyerer, J. (2024). Upar challenge 2024: Pedestrian attribute recognition and attribute-based person retrieval—dataset, design, and results. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 359–367).

Liu, X., Zhang, S., Huang, J., & Gao, Z. (2017). Pedestrian attribute recognition in surveillance: Dataset and approach. *Pattern Recognition, 63*, 148–160.

Deng, Y., Luo, P., Loy, C. C., & Tang, X. (2014, November). Pedestrian attribute recognition at far distance. In *Proceedings of the 22nd ACM International Conference on Multimedia* (pp. 789–792).

Li, D., Zhang, Z., Chen, X., Ling, H., & Huang, K. (2018). Richly annotated pedestrian attribute dataset for person retrieval. *IEEE Transactions on Image Processing, 28*(2), 612–624.

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., & Tian, Q. (2015). Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1116–1124).

Naphade, M., Wang, S., Anastasiu, D. C., Tang, Z., Chang, M.-C., Yao, Y., ... & Chellappa, R. (2023). The 7th AI City Challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5538-5548). IEEE.

Bui, D. C., Le, T. V., & Ngo, B. H. (2024). C2t-net: Channel-aware cross-fused transformer-style networks for pedestrian attribute recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 351–358).

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 10012–10022).

Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., ... & Cao, Y. (2023). Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 19358–19369).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems, 30*.

Chen, C. F. R., Fan, Q., & Panda, R. (2021). Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 357–366).

Lu, P., Jiang, T., Li, Y., Li, X., Chen, K., & Yang, W. (2024). RTMO: Towards high-performance one-stage real-time multi-person pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1491–1500).

Maji, D., Nagori, S., Mathew, M., & Poddar, D. (2022). Yolo-pose: Enhancing YOLO for multi-person pose estimation using object keypoint similarity loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2637–2646).