TESIS DE LA UNIVERSIDAD DE ZARAGOZA
2024
338

Sara Casao Martínez

Scene Understanding with Multi-Camera Systems

Director/es
Murillo Arnal, Ana Cristina
Montijano Muñoz, Eduardo

ISSN 2254-7606

Prensas de la Universidad
Universidad Zaragoza

ISSN 2254-7606

Universidad Zaragoza
1542

Tesis Doctoral

SCENE UNDERSTANDING WITH MULTI-CAMERA SYSTEMS

Autor

Sara Casao Martínez

Director/es

Murillo Arnal, Ana Cristina
Montijano Muñoz, Eduardo

UNIVERSIDAD DE ZARAGOZA
Escuela de Doctorado

Programa de Doctorado en Ingeniería de Sistemas e Informática

2024

Departamento de
Informática e Ingeniería
de Sistemas
Universidad Zaragoza

PhD Thesis

Scene Understanding with Multi-Camera Systems

Autor

Sara Casao Martinez

Directores

Eduardo Montijano Muñoz

Ana Cristina Murillo Arnal

Departamento de Informática e Ingeniería de Sistemas
Escuela de Ingeniería y Arquitectura
Universidad de Zaragoza, Abril 2024

Acknowledgments

I would like to begin this manuscript by expressing my gratitude to my supervisors, who gave me the opportunity to embark on this incredible journey. *Ana C. Murillo*, thank you for your understanding and kindness, qualities that proved priceless to me many times throughout these years. *Eduardo Montijano* thank you for constantly encouraging me to strive for improvement and never settle. I am grateful to both of you for your invaluable guidance, advice, and countless hours dedicated to teaching and guiding me.

#TeamI3A, mere words cannot express my gratitude. It has been a privilege to work alongside such exceptional researchers but also with people whom I consider my friends. Thank you very much for the discussions about science (and not so science), for the endless coffee sessions, for the unforgettable trips and shared moments, and for the traditions that disappeared because *this is a workplace*. This experience would not have been the same without all of you. Special thanks to *Pablo*. My friend, we embarked on this journey together and crossed the finish line side by side. Thanks for your good vibes, the legendary "turras miticas" and the support in challenging times. Wherever life takes me, you will have a cross on your map.

Thanks to all the people I met during my research stay in Delft, who warmly welcomed me. *Anurag and Modesto* thank you for being fantastic roommates and a constant source of support. You made me feel at home. *Kyle* thank you for leading the way during our visit to the Netherlands, for all those crazy plans and for making my adaptation to a new environment a breeze.

While this manuscript summarizes the work of four years, this journey began much earlier. The fact that I have shared this path with so many friends fills me with pride and joy. *Sheila, Sofia and Eva* thank you for being my lifelong friends with whom I have shared some of my most cherished memories and I have no doubt, that many more await us. Even when time stretches between gatherings, it feels like no time has passed. *Natalia*, thank you for always being my YES to everything and my partner-in-crime in all our adventures. Your kindness, joy and sense of humor make me grateful to have you in my life.

A mi familia, gracias por todo el apoyo incodicional que siempre he recibido de vosotros, por lanzarme a experimentar, vivir y conocer sitios, personas y situaciones nuevas que sobrepasan mi zona de comfort. Lanzarse al vacío siempre es mucho más fácil teniendo la certeza de que siempre estais ahi.

It makes me very proud to write these lines, to look back and realize all the people who have walked by my side. In one way or another you all have helped me to get here. You are all part of this work, to the person I am, and to the person I will be tomorrow.

Abstract

Scene understanding is an essential problem in computer vision aiming to gain a deeper knowledge of the elements and entities in a scene. This process involves localizing and identifying the elements of interest, analyzing their temporal evolution, and understanding their context. All of these tasks are essential abilities for many AI applications, like smart surveillance systems, autonomous robots, or process automation. The complexity of these applications often surpasses the ability of single cameras to acquire a comprehensive understanding of the scene, prompting the use of multi-camera setups to capture richer data from multiple viewpoints.

Multi-camera systems offer advantages over single-view setups, including enhanced coverage in large areas and leveraging diverse information when using heterogeneous camera sets. However, deploying these systems also poses several challenges. Managing a large number of cameras requires substantial processing power and network bandwidth. Adapting to changing environments or integrating new knowledge online, across multiple data streams, necessitates careful data selection and memory management strategies. Furthermore, collaboration between heterogeneous systems introduces complexities in data fusion, synchronization, and coordination, demanding sophisticated methods to ensure seamless operation. Balancing the benefits of multi-camera systems with their associated challenges is crucial for their effective deployment in real-world applications. This thesis addresses these challenges by delving into three crucial tasks for multi-camera scene understanding:

Distributed multi-target tracking aims to understand dynamic element trajectories by processing the collected information in a distributed fashion. Most of the works centered on multi-target tracking, either process the data in a single central device, thereby limiting the scalability of the system, or assume previous knowledge. In contrast, our method processes visual and non-visual data on each node, without prior knowledge assumptions. Therefore, this thesis provides a fully distributed multi-camera multi-target tracking approach that offers a flexible solution requiring minimal effort to integrate new cameras into the system.

Open-world person re-identification focuses on matching an observed individual against a gallery of known people. The main challenge lies in distinguishing new people from known ones while still correctly matching previously identified individuals despite variations in perspective or lighting. In real-world applications, person re-identification should efficiently adapt to the temporal evolution of the environment. For this purpose, this thesis presents a novel algorithm for building a self-adaptive gallery able to dynamically expand to identify new individuals and update existing information based on new people's observations.

Heterogeneous sensors collaboration combining static RGB cameras with other sensor types, enhance functionality and information gathering. This thesis explores the benefits of two heterogeneous collaborations. First, we analyze the association of static and mobile cameras for monitoring applications, leveraging mobile cameras to capture higher-quality images and improve perception tasks. Second, we explore combining RGB and hyperspectral imaging for object identification in waste material sorting. Hyperspectral sensors capture a spectral signature of the material boosting object identification.

Overall, this thesis contributes to enhancing the scalability of perception solutions, improving monitoring system adaptability to environmental changes, and fostering collaboration among different types of cameras for acquiring complementary knowledge.

Resumen

La comprensión de escena es un problema fundamental en el campo de la visión por computador que tiene como objetivo obtener un conocimiento profundo de los elementos y entidades dentro de una escena. Este proceso implica localizar e identificar los elementos de interés, analizar su evolución temporal y comprender su contexto. Todas estas tareas son capacidades esenciales para muchas aplicaciones de IA como sistemas de vigilancia inteligente, robots autónomos o la automatización de procesos. La gran complejidad de estas aplicaciones a menudo hace que los sistemás de una única cámara no capturen suficiente información para un entendimiento preciso. Por ello, los sistemas tienden a estar compuestos por multiples cámaras capaces de adquirir información más completa desde diferentes puntos de vista.

Los sistemas multi-cámara ofrecen ventajas sobre las configuraciones de monocámaras, incluyendo una mejor cobertura en grandes áreas y el aprovechamiento de datos diversos cuando se combinan cámaras heterogéneas. Sin embargo, la implementación de estos sistemas plantea múltiples desafíos. La gestión de un alto número de cámaras requiere gran potencia de procesamiento y ancho de bancha. Adaptarse a entornos cambiantes o integrar nuevos conocimientos procedentes de varios flujos de datos requiere cuidadosas estrategias de selección de la información así como una gestión de memoria eficiente. Además, la colaboración entre sistemas heterogéneos introduce complejidades en la fusión de datos, su sincronización y su coordinación, lo que exije el desarrollo de sofisticados métodos para garantizar un funcionamiento robusto. De esta forma, equilibrar los beneficios proporcionados por los sitemas multi-cámara con sus desafíos asociados, es crucial para su efectiva implementación en aplicaciones del mundo real. Esta tesis, aborda estos desafíos profundizando en tres tareas cruciales para la comprensión de escenas con múltiples cámaras:

El seguimiento multi-objetivo distribuido tiene como finalidad conocer las trayectorias de los elementos dinámicos de la escena con un procesamiento distribuido de la información recopilada. La mayoría de los trabajos centrados en seguimiento multi-objetivo procesan los datos en un único dispositivo, lo que limita la escalabilidad del sistema, o asumen conocimentos previos de los elementos dinámicos. Por el contrario, nuestra propuesta procesa localmente en cada nodo tanto datos visuales como no visuales, sin suposiciones previas. Por lo tanto, la investigación realizada en esta tesis proporciona un enfoque de seguimineto multi-objetivo con multiples cámaras completamente distribuido ofreciendo una solución flexible que requiere un esfuerzo mínimo en la integración de nuevas cámaras.

La re-identificación de personas en entornos abiertos se centrar en encontrar un individuo comparándo su imagen con una galería de personas conocidas. El principal desafío de las configuraciones en entornos abiertos radica en distinguir a las personas nuevas de las ya conocidas por el sistema y, al mismo tiempo, identificar correctamente las personas previamente identificadas a pesar de las variaciones de perspectiva o iluminación. En aplicaciones del mundo real, la re-identificación de personas debe adaptarse de manera eficiente a la evolución temporal del entorno. Para ello, esta tesis presenta un algoritmo que construye una galería auto-adaptable capaz de expandirse dinámicamente identificando nuevas personas y actualizando los datos existentes con las nuevas observaciones de personas adquiridas.

La colaboración de sensores heterogéneos combina cámaras RGB estáticas con otro tipo de sensores para mejorar la funcionalidad del sistema y la recopilación de información. Esta tesis explora los beneficios aportados por dos tipos de colaboraciones. En primer lugar, analizamos la asociación de cámaras estáticas y móviles para monitorización, explotando las cámaras móviles para capturar imágenes de mayor calidad y así, mejorar las tareas de percepción. En segundo lugar, estudiamos la combinación de imágenes RGB e hiperespectrales para la identificación de objetos en la clasificación de residuos. Los sensores hiperespectrales capturan una firma espectral del material, lo que mejora la identificación final del objeto.

En general, esta tesis contribuye a mejorar la escalabilidad de los métodos de percepción, la capacidad de adaptación de los sistemas de monitorización ante cambios temporales y por último, confirma los beneficios de combinar diferentes tipos de cámaras en la adquisición de conocimientos complementarios.

Contents

Chapter 1

Introduction

Computer vision focuses on developing methods that enable machines to interpret and understand visual information from the world, typically in the form of images or videos. One of the essential problems studied in computer vision is **scene understanding**, which attempts to gain deep knowledge of the existing elements in a scene, including the fundamental tasks of identification and localization but also more complex analysis of context, interactions, and temporal evolution of the entities. The acquisition of these capabilities is of paramount importance to unlock significant advancements in a wide range of practical applications. Figure 1.1 illustrates some examples, such as smart surveillance systems which benefit from detecting and tracking objects and people for monitoring public areas; robots that are able to estimate movement intentions can perform complex tasks more efficiently, like navigating through dynamic environments or adapting to changes in the environment; and process automation empowered by object recognition and localization can detect early faults, increase efficiency, and guarantee quality control.

(a) (b) (c)

Figure 1.1: Different applications that benefit from advanced visual scene understanding (a) surveillance system monitoring a public area, (b) autonomous robot moving among people in a mall and (c) automatic assembly process in a factory. *(Images generated with Microsoft Bing).*

1.1 Scene understanding with multi-camera systems

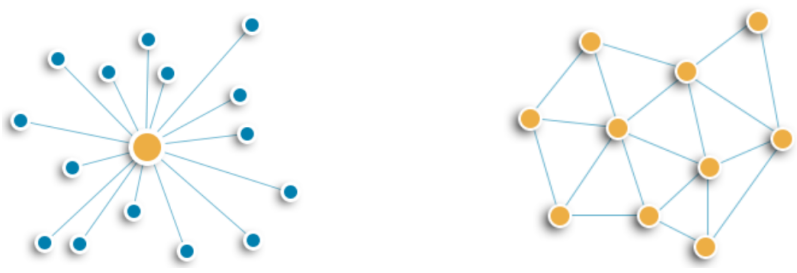
In many scenarios, where the above-mentioned practical applications are developed, the use of a single camera is not sufficient to obtain a thorough understanding of the scene. Consequently, the use of **multiple cameras** represents an extended setup for gathering more

comprehensive information from the environment. In large areas, the spreading of multiple homogeneous cameras enhances the perspective and coverage, thus making it easier to monitor the entire space. Moreover, combining **heterogeneous sets of cameras** provides diverse and valuable information for fulfilling specific tasks. For instance, depth cameras capture the relative position of objects in the scene, making it straightforward to locate elements in space, while hyperspectral cameras capture a wide range of electromagnetic spectrum very useful for material identification. Therefore, multi-camera systems provide significant advantages over single-camera setups for improving visual scene understanding.

Traditionally, all the cameras are connected through a network to one central device (node) that collects and processes all the data. The centralized setup offers certain benefits like ensuring uniformity in data processing and system behavior, simple maintenance avoiding the complexity of coordinating actions across nodes, and the reduction of redundant equipment in different sites. However, the drawbacks inherent to centralized systems, mainly focused on having a single point of failure and the high processing cost associated with incorporating new nodes (scalability), naturally lead to the development of **distributed** solutions. Figure 1.2 illustrates the architecture of each configuration.

Distributed systems comprise independent and interconnected nodes that work together to achieve a common goal through communication and coordination. Among the main characteristics of distributed setups, two are particularly noteworthy. First, each node of the network operates independently, processing its own information and the one received from nodes it communicates with. Thus, the failure of one node does not necessarily lead to the entire system's failure, contributing to enhanced fault tolerance. Second, distributed systems can scale horizontally by incorporating new nodes, enabling them to handle increased workloads and accommodate growing data volumes. These advantages render distributed approaches more appealing than centralized setups. Nevertheless, the challenges arising from the local data processing and the need for information exchange, make the development of robust and efficient methods not trivial, requiring solutions for issues related to bandwidth usage, consensus between nodes, and selection of accurate information to share.

Despite the potential benefits of applying distributed setups to visual scene understanding, research and development in this area remain underrepresented in the scientific community. Most works addressing distributed issues are focused on multi-robot control for accomplishing specific tasks and assume as resolved the visual understanding of the



(a) Centralized systems (b) Distributed systems

Figure 1.2: Comparison between (a) centralized systems where all the nodes communicate their information to a central device for data processing, and (b) distributed systems where every node processed its local information and the one received from its neighbors.

scene (Aldana-López et al., 2023; Shorinwa et al., 2020). This underscores a critical gap in research, where there is a need for more distributed perception solutions capable of understanding complex visual scenes by leveraging the benefits of distributed setups. Closing this gap would foster advancements in fields such as multi-robot systems, surveillance, or autonomous driving.

A sought-after feature for real-world applications within these fields is the implementation of algorithms that operate online. The ability to process data online, incrementally as it arrives to the system, not only ensures up-to-date insights but also allows **dynamic adaptation to the evolving data**. Online methods are particularly suited for scenarios where data streams continuously, and decisions or predictions must be made promptly.

1.2 Challenges and Contributions

This doctoral thesis studies and develops novel computer vision approaches for scene understanding, based on image processing, with a strong focus on their applicability to real-world use cases. For this purpose, the presented research is focused on three main challenges:

Develop distributed solutions for multi-target tracking algorithms. The development of multi-target tracking solutions in distributed setups is currently lacking attention in the community. Consequently, visual scene understanding methods miss out on the inherent benefits of distributed systems, such as the ability to scale the camera network at minimal processing costs. Our research aims to bridge this gap by proposing methods that address the challenge of achieving consistent information in visual data processed in a distributed fashion.

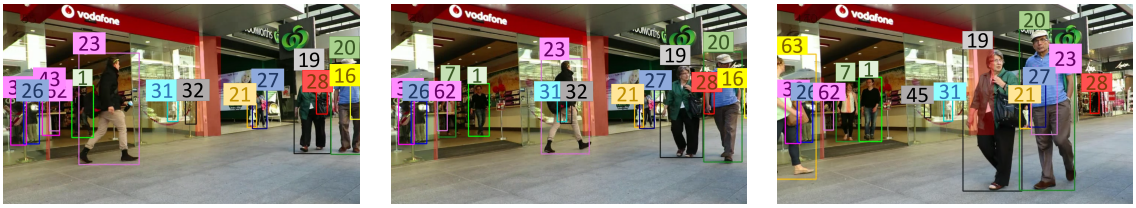
Online adaptation of the scene understanding to evolving data. In a world that is constantly changing and evolving, the demand for automatic visual understanding needs adaptive approaches that efficiently accommodate this continuous transformation. Our focus is on proposing techniques that enable the system to dynamically acquire new knowledge about the elements present in the scene. This involves not only the identification of novel elements but also the adaptive enhancement of existing knowledge whenever new data arrives.

Benefits of heterogeneous systems. While static RGB multi-camera systems are widespread in real-world applications, the combination of these cameras with other types of sensors is a powerful path to improve functionality or information gathering. In our research, we explore the benefits of two heterogeneous aspects in multi-camera systems. First, the association of static and mobile cameras for monitoring public areas. Second, the synergies resulting from combining RGB and hyperspectral imaging for object identification in the context of recycling plants.

Among the possible scene understanding topics, this thesis delves into three crucial tasks: distributed multi-target tracking, open-world person re-identification, and heterogeneous sensor collaboration. In the following subsections, we provide an overview of each topic and highlight their main challenge. Finally, we present our contributions.

1.2.1 Distributed multi-target tracking

Multi-target tracking is the process of knowing the positions of the elements presented in the environment over time, i.e., the trajectories of the objects as they move through the scene. Figure 1.3 depicts multi-target tracking in a single camera at different time instants of the sequence. As observed, the smart camera discerns the trajectory of each person in the scene maintaining their position over time. This process is essential for many real-world applications, including surveillance, autonomous vehicles, and robotics, where the ability to monitor and predict the movements of multiple objects is crucial for effective decision-making and system performance.



t *t+1* *t+2*

Figure 1.3: Sampled frames from a complete sequence where the task of multi-target tracking is performed in a single camera. The system localizes every person in the scene at the instant t and assigns a unique identification to them. In the subsequent frames *t+1* and *t+2*, the tracking maintains the same identification to the same individuals, thus knowing the path of people in the environment. *(Frames from MOT Challenge records Milan et al., 2016).*

The process overview of multi-target tracking algorithms is illustrated in Figure 1.4. Initially, each element of interest presented in the scene, people in this case, is assigned a tracker with a unique identification (*Trackers (t)*). In the subsequent iteration, $t+1$, people present in the frame at that moment are located as bounding boxes defined by their image coordinates (*Detection (t+1)*). The final step consists of associating both sets of data to determine the position of each tracker at instant $t+1$ (*Trackers (t+1)*). The main challenges of the multi-target tracking task are concerned with a correct data association process, occlusions between objects, and the re-identification of elements when they reappear on the scene after being temporarily out of view. In the occlusion scenario, the system does not detect the object and must estimate its position without current reference. Generally, this issue is approached using movement estimators that predict the position based on the historical states of the analyzed element (Rajasegaran et al., 2022). Regarding the re-identification problem, the most common technique is to gather key features of each tracked element to recognize it when it returns to the field of view (Wojke et al., 2017).

To overcome these challenges, the use of multiple cameras with overlapping views is a widespread setup, providing more coverage and diverse perspectives from the same scenario. In centralized systems, all cameras send the data to a common device for joint processing, ensuring consistency in the information. However, developing a distributed system for multi-target tracking is not a trivial task, introducing new challenges primarily related to the association of trackers across cameras and achieving consensus in their position. This complexity arises due to the tracking process conducted locally by each camera and the limited information shared with the network. The vast majority of papers consider both the association of trackers through the network sensors and the association of trackers with

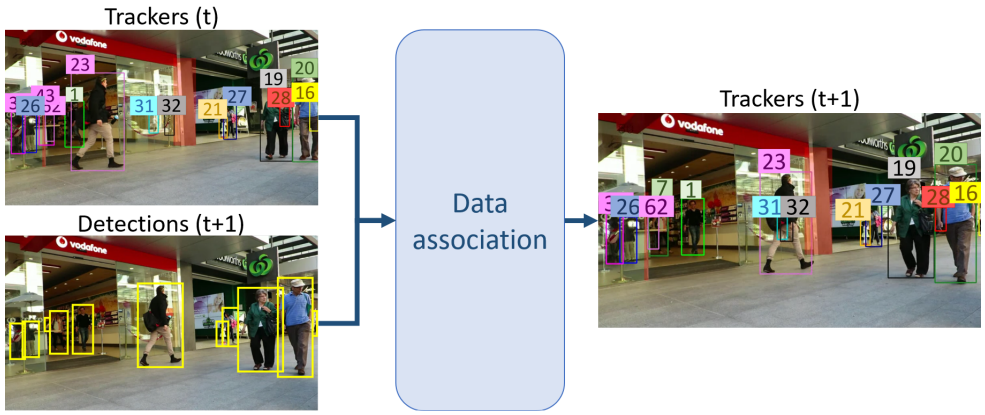


Figure 1.4: Multi-target tracking overview. To obtain the current position of each identity (Trackers (t+1)), the previous known position (Trackers (t)) and the actual people detections (Detections (t+1)) are related through a data association process.

observations as known, centering their studies on theoretical topics, such as how to locally fuse collected information (Kamal et al., 2012; Shorinwa & Schwager, 2023) or when to communicate data (Battistelli et al., 2018).

In contrast to these works, we include the distributed visual understanding of the scene in the problem formulation, addressing data association challenges through the use of visual information (Casao et al., 2021, 2022). The proposed framework expands the Distributed Kalman Filter (DKF) consensus algorithm, which reduces the disagreement on the targets' position between cameras, with an automatic data association. This data association, based on geometry cues and appearance information, manages both the local association of detections and trackers and the global association of trackers across cameras, using only partial information in local nodes. Thereby, the consensus algorithm is able to identify which information received from other cameras must be merged with its local targets' data.

More concretely, the main contributions of this thesis in the field of distributed multi-target tracking are: (1) A Distributed Kalman Filter implementation augmented with fully automatic local data association using geometry and appearance information. Unlike existing integrated approaches, data communication between cameras occurs only as needed based on the density of visible targets, thus lightening the process and increasing applicability. (2) A novel distributed strategy to manage global tracker information across cameras, relying solely on local processing. In contrast to centralized systems where a central node unifies the information, our framework performs a distributed data association across cameras that manages the partial information in local nodes.

The proposed method is tested in several multi-target tracking scenarios with overlapping cameras. These sequences involve a variable number of people over time, different numbers of cameras, and both outdoor and indoor scenes. The study demonstrates the benefits of using a distributed tracking system over a centralized counterpart, obtaining comparable or even better results in certain scenarios. Consequently, this research contributes to narrowing the gap between distributed systems and real-world monitoring applications.

1.2.2 Open-world person re-identification

The great difficulty posed by the aforementioned challenge of re-identification targets when they reappear on the scene after being out of view has led the computer vision community to focus part of its efforts on addressing it. Hence, the person re-identification (re-id) problem is formalized as follows: given a query person-of-interest, the goal is to determine whether this person has appeared in another location at a different time captured by a different non-overlapping camera, or even the same camera at a distinct moment. Figure 1.5 illustrates the most common approach to this problem which consists of comparing the image of an unknown person of interest (query) with a pool of images of known people (gallery), assigning to the query the identity of the most similar person in the gallery.

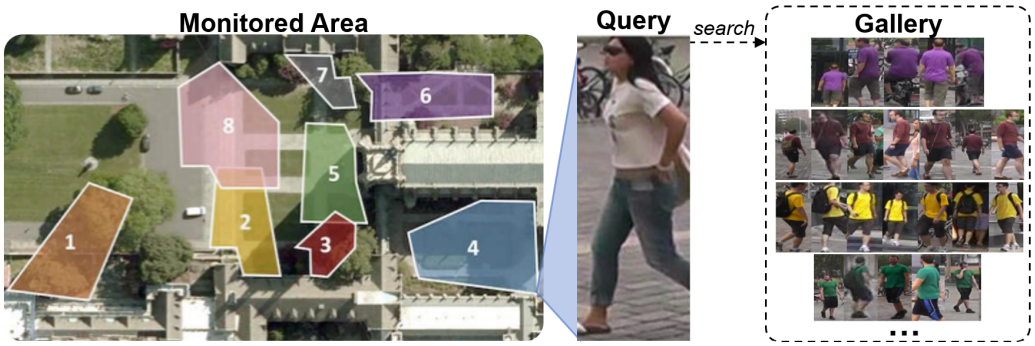


Figure 1.5: Person re-identification overview. In an area monitored with multiple cameras, a query person-of-interest appears on the scene and is compared with a gallery containing the system's knowledge of the environment in an attempt to identify them.

An extensive number of studies have focused on obtaining the most representative features from the appearance of individuals, considering factors such as clothing, body shape, and other visual cues, to correctly rank the known people (Hou et al., 2021; H. Wang et al., 2022). To simplify the analysis, these studies assume labeled all the collected data and a closed-set matching setup, which ensures that the query reappears in the gallery. Both assumptions are unfeasible in a real-world environment, and removing these simplifications is one of the main challenges of the re-identification problem. For this reason, there is a growing tendency to propose solutions for training models using exclusively people bounding boxes, thus avoiding the need for labeling (unsupervised) (Sridhar Raj S & Balakrishnan, 2022) or allowing the possibility that the query may not find a match in the gallery set (open-set) (Huang et al., 2020). Regarding the testing phase, the vast majority of these works employ a static and preset gallery in their development that limits the dynamic nature of the open world. This existing gallery represents the system's knowledge gained from the scene, identifying all the people who have passed through the non-overlapping camera network and maintaining an exhaustive visual representation of each of them. In the final real-world application, where raw data from cameras collect new people, detection errors, and junk data, the re-id system should automatically evolve its understanding of the environment, acquiring new identities and updating information on already known individuals.

To overcome these limitations and implement systems closer to open-world setups, this research focuses on the problem of building a self-adaptive gallery (Casao, Azagra, et al.,

2023) that, to the best of our knowledge, existing approaches in person re-identification have not yet considered. In the developed framework, the gallery is able to dynamically expand to identify new individuals and select representative visual data for modeling their appearance information. This data selection maintains the gallery small and efficient, preventing it from growing uncontrollably by accumulating all the information coming from the scene. Unlike the extended static galleries used for the testing phase, our proposed gallery starts empty and updates its structure as new unlabeled people bounding boxes arrive in the system.

Our work on a self-adaptive gallery seeks a balance between identifying all the people in the scene, minimizing the error of identifying an already known person as new, and maintaining high precision of the appearance models defined for each individual over time. The conducted experiments compare the proposed data selection algorithm with others traditionally used in incremental learning, confirming that ours is the one that best balances the gallery goals sought. Moreover, we conclude that traditional static galleries employed as references for re-identifying a query person-of-interest have redundancy in the data saved, making their use in real-world systems inefficient.

In summary, the main contributions of this thesis to the field of person re-identification in open-world scenarios are: (1) A novel approach for constructing a self-adaptive gallery. The proposed method ensures the appearance model for each person remains compact and representative by employing information theory concepts. Specifically, we leverage uncertainty and diversity metrics to select the most representative samples. (2) A thorough evaluation of the posed problem. This thesis includes a metric based on standard precision and recall to assess the quality of the constructed gallery structure. This metric offers insights into the final quality of the gallery structure, particularly in complex scenarios where identifying the total number of classes is highly challenging.

1.2.3 Heterogeneous sensor collaboration

Static RGB multi-camera systems are widely used sensors in real-world applications. These sensors are highly available, cost-effective, and easy to deploy with the capability of capturing color information with high fidelity. However, their restricted coverage and sensitivity relying solely on visible light, impact their effectiveness in acquiring deeper knowledge from certain scenarios. Thus, there is an increasing interest in exploring collaboration between heterogeneous sensors which not only addresses the limitations of static RGB cameras but also unlocks new possibilities for improved overall scene understanding in diverse and challenging conditions.

This thesis considers two types of heterogeneous systems. First, we remove positional constraints on a few RGB cameras in the network, allowing them to move freely in the environment to acquire new knowledge from the scene (*static-mobile*). The second sensor cooperation analyzed, studies the benefits of capturing additional non-visible light intensity from multiple spectra (*multi-modal*).

Static-mobile cameras. The collaboration between static and mobile cameras presents a compelling approach for monitoring open and large spaces, offering synergies that enhance the capabilities of surveillance systems. Static cameras provide stable and fixed viewpoints contributing to a global understanding of the scene whereas mobile cameras can obtain new perspectives and adapt to changing scenarios. Hence, only by enabling the free movement

of cameras to gain knowledge of what other sensors are currently observing, the set of cameras will be able to maximize the information captured from the environment.

In hybrid surveillance systems, multiple tasks can be addressed. For instance, active tracking consists of transforming visual information into motion decisions for following a dynamic target (Mekonnen et al., 2013); coverage optimization focuses on deploying sensors to maximize the monitored area and targets (Bisagno et al., 2018); and active perception attempts to find the camera location that returns the most informative perspective (Serra-Gómez et al., 2023). Related to the latter task, our proposal allows free movement for some cameras to obtain a close-up view of the individuals in the scene and acquire additional semantic information, e.g., whether or not they are wearing a sanitary mask or a backpack. Traditionally, the approaches that tackle this problem work in closed environments (Kent & Chernova, 2020), or require prior knowledge of where the information is visible from (Alcántara et al., 2021).

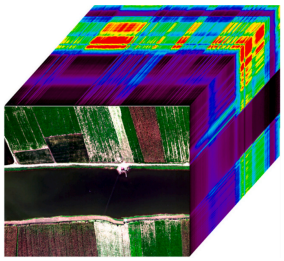
In contrast to these works, we perform an actual distributed tracking of dynamic targets, providing an estimation of the position and orientation of the targets to multiple mobile cameras. These mobile cameras, carried by drones, have to make a decision on their next viewpoint acquired in an open space to maximize the collected data for a specific classification task (Casao et al., 2024). The experimentation is conducted in a photo-realistic environment ("Epic Games Incorporated. Unreal engine, 2022." n.d.) that lets us simulate the processing of visual information and control the mobile cameras in every iteration. Addressing both components, processing of visual information and control of mobile cameras, brings the proposed solution closer to real-world application. The required tools for designing the pedestrian scenarios, including animated people models, trajectory generation, and collecting the environment information are built on the photo-realistic environment for fast and easy prototyping of pedestrian cenes (Casao, Otero, et al., 2023).

Specifically, our main contributions are: (1) A novel hybrid multi-camera framework where static and mobile nodes collaborate for people monitoring. The framework combines distributed tracking with active perception to acquire semantic knowledge from the scene. Mobile camera control is based on real perception, and nodes communicate through the Robot Operating System (ROS), which runs each node as an independent device and manages message exchange. (2) An evaluation in a photo-realistic simulator, thus contributing to bridging the gap with real-world applications. Within this simulator, we develop essential tools to facilitate the design and creation of pedestrian scenarios. These tools include a trajectory plugin for user-friendly path creation, a compilation of pedestrian models ready to use by simply dragging and dropping them, and finally, a Python API to extract environment metadata from cameras and pedestrians.

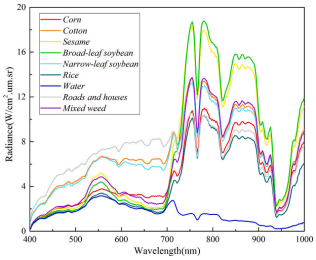
Multi-modal sensors. Secondly, we explore the benefits of combining RGB with hyperspectral sensors (HSI). While traditional cameras capture three color bands (red, green, and blue), to produce an RGB image, hyperspectral cameras capture hundreds of narrow and contiguous bands across the spectrum. Figure 1.6b) illustrates the spectrum of light intensity captured by each band in each pixel. Specifically, Figure 1.6c) represents the values of intensity (y-axis) captured at different wavelengths (x-axis), directly related to the image's bands. Thus, each plot depicts the spectrum acquired by different pixels. These spectral signatures enable the identification and characterization of materials, making hy-

perspectral cameras valuable in applications where a deeper understanding of materials in the scene is critical like medical imaging, agriculture, or environmental tasks.

(a) Hyperspectral camera (Specimen FX17)



(b) Hyperspectral image from HU-Hi Dataset Zhong et al., 2020



(c) Spectral curve from HU-Hi Dataset Zhong et al., 2020

Figure 1.6: Hyperspectral cameras. a) Hyperspectral camera sensor. b) Hyperspectral image of the countryside showing the spectral information collected in every channel. c) The value captured by each of the channels (or wavelengths) for different materials, the whole spectrum gives reference values for the classification of materials pixel-wise.

Among all of them, this thesis focuses on a critical application field nowadays, recycling, where automation and precise separation of objects can significantly improve the efficiency of waste management. Hyperspectral data has been broadly used in this context for material identification. In fact, some commercial cameras already identify automatically several types of materials[1]. However, the automatic identification and localization of particular objects in complex real-world waste industrial settings is still an outstanding challenge. During the last few years, some works have released benchmarks to facilitate research and advances on this topic. For example, *Floating Waste (FloW)* dataset (Cheng et al., 2021) collects information to enable the study of cleaning inland water areas with autonomous boats and *ZeroWaste* dataset (Bashkirova, Abdelfattah, et al., 2022) gather RGB images for industrial waste object localization from a real sorting plant.

Different from previous benchmarks, we capture a multi-modal dataset derived from an operational waste sorting facility, featuring actual industrial data from both HSI and RGB cameras (Casao, Peña, et al., 2023). This dataset was captured in a true-to-life prototype of the equipment setup, where waste travels along a conveyor belt and two synchronized multi-modal cameras collect information from the scene. To enhance the efficiency of waste processing in the facility, the gathered data aims to identify and localize specific objects that can either pose significant issues by clogging the machinery or add value if recovered through redirection to a different recycling sector. Thus, the collected dataset contributes to improving the sorting capacity of the recycling plant by opening the opportunity to automate the removal of problematic elements or recover valuable objects for another sector. Furthermore, the proposed dataset serves as the foundation for evaluating a set of baselines for multi-modal object localization. These baselines consist of different deep neural networks that process either one or both types of images to assign a class label to every pixel of the image. This problem is formally named semantic segmentation and is crucial for understanding the content of an image, enabling machines to interpret visual information more effectively.

[1]Commercial camera used for identification of materials in plastics among others: *iberoptics*

In particular, the main contributions of this thesis to the field of multi-modal waste management segmentation are: (1) The SpectralWaste dataset, the first multimodal (RGB-HSI) dataset specifically designed for real waste management segmentation. This dataset addresses the identification of critical objects that frequently appear in real trash flows and significantly impacts sorting efficiency by causing complete stoppages of the waste separation or belonging to another recycling process. (2) A comprehensive object segmentation analysis that highlights the performance improvement achieved by combining both modalities and, for the first time, explores the suitability of using HSI for object localization in waste sorting scenarios. Additionally, to ensure consistency between the annotated segmentation for both modalities and minimize labeling effort, a novel label transfer algorithm is introduced. This algorithm automatically adapts RGB-annotated segmentation to HSI, relying exclusively on both images without any calibration needed.

1.3 Summary of Results

The results of all the work developed during this thesis have been published and presented in different international conferences and journals.

S. Casao, A. Naya, A. C. Murillo, E. Montijano. **Distributed multi-target tracking in camera networks**. Proccedings of the IEEE/RSJ International Conference on Robotics and Automation. ICRA 2021. Acceptance rate: 49%. CORE: B.

Author Contributions: Conceptualization, S.C, A.C.M, E.M; methodology, S.C, A.C.M, E.M; implementation, S.C, A.N; evaluation, S.C; writing, S.C; reviewing and editing, S.C, A.N, A.C.M, E.M.

S. Casao, A. C. Murillo., E. Montijano. **Data association tools for target identification in distributed multi-target tracking systems**. Iberian Robotics Conference 2022. ROBOT 2022.

Author Contributions: Conceptualization, S.C, A.C.M, E.M; methodology, S.C, A.C.M, E.M; implementation, S.C; evaluation, S.C; writing, S.C; reviewing and editing, S.C, A.C.M, E.M.

S. Casao, P. Azagra, A. C. Murillo, E. Montijano. **Self-Adaptive Gallery Construction Method for Open-World Person Re-Identification**. Sensors 2023. Quartile: Q2. Impact Factor: 3.9.

Author Contributions: Conceptualization, S.C, P.A, A.C.M, E.M; methodology, S.C, P.A, A.C.M, E.M; implementation, S.C; evaluation, S.C; writing, S.C; reviewing and editing, S.C, P.A, A.C.M, E.M.

S. Casao, A. Otero, Á. Serra-Gómez, A. C. Murillo, J. Alonso-Mora, E. Montijano. **A Framework for Fast Prototyping of Photo-realistic Environments with Multiple Pedestrians**. Proccedings of the IEEE/RSJ International Conference on Robotics and Automation. ICRA 2023. Acceptance rate: 43%. CORE: A+. Framework: https://github.com/saracasao/Pedestrian_Environment

Author Contributions: Conceptualization, S.C, A.S, A.C.M, J.A, E.M; methodology, S.C, A.O; implementation, S.C, A.O; evaluation, S.C; writing, S.C, A.O; reviewing and editing, S.C, A.O, A.S, A.C.M, J.A, E.M.

S. Casao, Á. Serra-Gómez, A. C. Murillo, W. Böhmer, J. Alonso-Mora, E. Montijano. **Distributed multi-target tracking and active perception with mobile camera networks**. Computer Vision and Image Understanding. CVIU 2024. Quartile: Q2. Impact Factor: 4.5. Simulated data and photo-realistic environment used available at https://sites.google.com/unizar.es/poc-team/research/hlunderstanding/collaborativecameras

Author Contributions: S.C and A.S.G contributed equally in this work. Conceptualization, S.C, A.S, A.C.M, W.B, J.A.M, E.M; methodology, S.C, A.S, A.C.M, E.M; implementation, S.C, A.S; evaluation, S.C; writing, S.C, A.S; reviewing and editing, S.C, A.S, A.C.M, W.B, J.A.M, E.M.

S. Casao, F.Peña, A. Sabater, R. Castillón, D. Suárez , E. Montijano, A. C. Murillo. **SpectralWaste Dataset: Multimodal data for waste segmentation**. Under Review. Dataset website: https://sites.google.com/unizar.es/spectralwaste/home

Author Contributions: S.C and F.P contributed equally in this work. Conceptualization, S.C, F.P, A.S, R.C, D.S, E.M, A.C.M; methodology, S.C, F.P, D.S, E.M, A.C.M; implementation, S.C, F.P; evaluation, S.C, F.P; writing, S.C; reviewing and editing, S.C, F.P, D.S, E.M, A.C.M.

This work has been supported by the Office of Naval Research Global project ONRG-NICOP-N62909-19-1-2027 and the Spanish projects PGC2018-098817-A-I00 and PGC2018-098719-B-I00 and MCIN/AEI/ERDF/European Union NextGeneration EU/PRTR projects PID2021-125514NB-I00 and RTC-2017-6421-7, DGA T04-FSE and T45_23R.

Along with the presented publications, I also collaborated as **reviewer for different journals and conferences**: BMVC (2020, 2021), CVPR (2020, 2024), ICCV (2023), ICRA (2023), IROS (2022, 2023), MRS (2023), RA-L (2020), TR-O (2023, 2024). Besides, I worked as a volunteer in the organization of the *Fifth Iberian Robotics Conference (ROBOT 2022)* and *Jornadas Automática 2023*.

In this period, I also had the chance to do one **research visit** at the *Technical University in Delft (TU Delft)*. More concretely, I did a 4-month visit at the *Autonomous Multi-Robots Lab*, which resulted in the publication of *A Framework for Fast Prototyping of Photo-realistic Environments with Multiple Pedestrians* and in a subsequent collaboration with the publication of *Distributed multi-target tracking and active perception with mobile camera networks.*

During this thesis, I also collaborated in the **supervision of the Bachelor Thesis** *Reidentificación automática de personas en sistemas multi-cámara* and as **Teaching Assistant** in the *Bachelor's Degree in Informatics Engineering* (artificial intelligence course) from the University of Zaragoza (18h).

1.4 Manuscript Organization

The following chapters describe the four main contributions to the field of computer vision and scene understanding introduced above. Chapter 2 introduces our distributed multi-target tracking approach. Chapter 3 presents the developed adaptive gallery construction method for person re-identification in the open world. Chapter 4 provides detailed insights into our implemented framework where static and mobile cameras collaborate for monitoring the environment and acquiring new information from people in the scene. Chapter

5 describes the collected multi-modal dataset for object segmentation in a waste sorting facility and the conducted analysis on the benefits of using multi-modal sensors on recycling tasks. Finally, Chapter 6, summarizes the conclusions of the presented thesis and outlines our vision for future work.

Chapter 2

Distributed multi-target tracking in camera networks

Despite the significant advantages of distributed systems over centralized setups like scalability, fault tolerance, and lighter communication management, their application in tracking tasks remains largely underexplored. This Chapter addresses this challenge by incorporating the distributed processing of visual information into the problem formulation of distributed multi-target tracking systems. Our focus lies on tackling the inherent problems of traditional multi-target tracking, such as correct data association, handling occlusions, and re-identification, while processing all the information locally and achieving global consensus on both target positions and identities.

The presented research aims to bridge the gap between the established benefits of distributed systems and the limitations of current centralized multi-target tracking approaches. In the following, we introduce our novel distributed multi-target tracking system. The proposed algorithm boosts the benefits of the Distributed-Consensus Kalman Filter with the support of a re-identification network and a distributed tracker manager module to facilitate consistent information. These techniques complement each other and facilitate the cross-camera data association simply and effectively. The proposed system achieves comparable or even better results than certain centralized tracking methods and the conducted ablation studies demonstrate the individual contributions of the developed modules to the overall tracking performance.

2.1 Introduction

Multi-target tracking systems have a broad range of applications in real-world scenarios such as security or crowd behavior analysis (Ardo & Nilsson, 2019; Robin & Lacroix, 2016). Approaching these problems with multi-camera setups brings additional features and benefits with respect to single-camera systems, including more complete information in large spaces or crowded scenes. Besides, multi-camera systems naturally lead to the development of distributed solutions, which are lighter in communication demands, more robust to failures and faster in processing time than centralized ones.

However, the implementation of multi-target multi-camera tracking in a distributed setup poses several challenges. These challenges involve the construction of a robust sys-

tem in terms of bandwidth usage, consensus between nodes and association of high-level information. In the presence of multiple targets, each camera must independently solve a data association problem between measurements and trackers. Moreover, in contrast with centralized systems, where the central node unifies the high-level information returning it labeled to the nodes, in a distributed setup the data association across cameras can only be performed locally and with partial information.

This chapter tackles the challenges discussed above by exploiting synergies between complementary modules, typically studied independently in prior work. We develop a novel approach for multi-target tracking in distributed camera networks boosting the implementation of a Distributed Kalman Filter (DKF) with a local data association process based on geometry and appearance cues, and a novel distributed method for high-level information management. The overall idea is illustrated in Figure 2.1. The DKF exploits the local data association to decide which information needs to be merged into the tracker state estimation consensus. Similarly, the data association is enhanced with the uncertainty estimations given by the DKF. Finally, the tracker manager makes sure of the correct association during the initialization and synchronized deletion of the lost trackers.

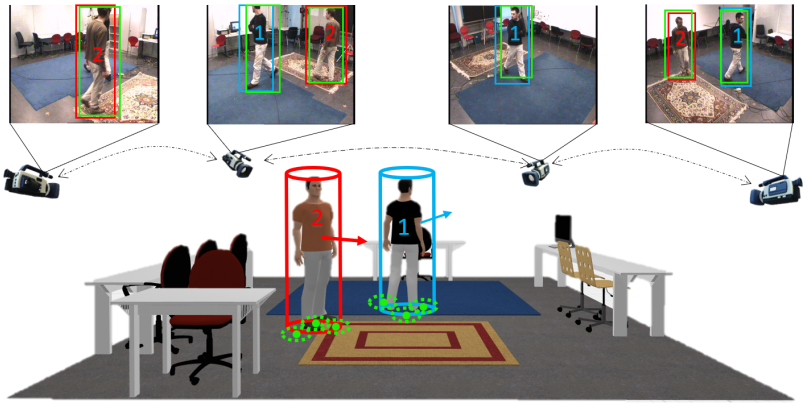


Figure 2.1: Distributed multi-target tracking scenario. The proposed system maintains in each node the 3D tracking information of all targets (seen or not by the current camera) thanks to the information shared among connected cameras. Each camera solves an independent data association problem. Green boxes correspond to people detections while blue and red ones refer to the two existing trackers in the scene, ID1 and ID2 respectively (Best viewed in color).

In order to increase the efficiency of bandwidth usage and strengthen the continual identification of targets within our distributed approach, we introduce an event-trigger mechanism to control the communications between cameras, and a quality-based appearance feature selector to create a robust appearance model for each target. One of the main trade-offs of distributed systems is bandwidth usage vs accuracy, a concern widely addressed in the literature from a theoretical perspective (Battistelli et al., 2018; Sebastián et al., 2021), but often lacking real-world implementations with multiple targets. In this research, the well-established event-triggered mechanism is used to decide when to communicate data based on each target's tracking error. Notable, our implemented event-trigger mechanism incorporates a term that modulates communication frequency according to the complexity of correctly associating measurements and trackers. This term ensures reduced communications in simple scenarios with few targets and straightforward data association.

In contrast, complex scenarios like crowded scenes trigger more frequent data sharing to gather comprehensive information from the environment. Furthermore, this thesis strengthens continual target identification through robust target appearance representation, i.e., creating a reliable appearance model. The quality-based appearance feature selector discards data when the model reaches saturation based on its usefulness and reliability. To measure both features, we proposed a strategy to iteratively score each element of the appearance model and discard the one with the lowest score.

In summary, the main contributions of this thesis to the field of distributed multi-target tracking are: (1) A DKF implementation augmented with a fully automatic data association process that leverages both geometric cues and robust appearance information. Differently from existing approaches, our system triggers communication between cameras based on the complexity of data association, streamlining the process and increasing the efficiency of bandwidth usage. (2) A novel distributed strategy to manage the trackers' information. In contrast to centralized systems that rely on a single node to unify the information, we propose a distributed data association across cameras that manages the partial information in local nodes.

The proposed distributed multi-target tracking method is evaluated in public benchmarks, demonstrating the benefits of its different components with respect to a naive DKF, as well as established centralized algorithms. Furthermore, we discuss the advantage resulting from integrating the event-trigger mechanism and our quality-based appearance feature selector into the complete distributed tracking approach.

2.2 Related work

This section summarizes the related work on core multi-target multi-camera tracking aspects and the proposed approach to tackle this task in a distributed setup.

Multi-view multi-target tracking in centralized systems. Multi-target multi-camera tracking in centralized systems sends the information from each camera to a common location where all the data is processed together (Tesfaye et al., 2019). These implementations, which stand out for their accuracy, are typically used in safety applications (J.-H. Chen & Song, 2018; Ferraguti et al., 2020). Obtaining the complete trajectories of several targets is normally formulated as an optimization problem in a graph. The nodes represent short trajectories, known as tracklets, obtained by the association of detected bounding boxes. There are different variations to compute the weights of the graph. The combination of the similarity measure given by a triplet loss with a linear motion model is proposed in Ristani and Tomasi, 2018. The proposal in L. Wen et al., 2017 is the association of tracklets across views based on correlations in motion, appearance, and smoothness of the resulting 3D trajectory. A method is presented in Le et al., 2018 to select a subset of tracklets from the graph and associate them based on geometry and motion cues. Other works such as Y. Xu et al., 2016 model the tracklet association problem as a hierarchical structure optimization. The main disadvantages of centralized methods are the excessive bandwidth usage required to send all the information to the central computer and the lack of robustness with respect to a single point of failure.

Multi-view multi-target tracking in distributed systems. Distributed implementations typically focus on improving the robustness and efficiency through the study of the bandwidth, the consensus between nodes, and the accuracy of the information shared. Several theoretical algorithms are proposed in Olfati-Saber, 2007 to achieve a consensus in a distributed heterogeneous sensor network performing one communication per estimation cycle. One of those algorithms is implemented in Soto et al., 2009 for a Pan-Tilt-Zoom (PTZ) camera network to track people of interest, although the data association problem is not addressed there. In Kamal et al., 2015, the Information-weighted Consensus Filter (ICF) method Kamal et al., 2012 is selected as the consensus algorithm, filling the gap of data association with the Joint Probabilistic Data Association algorithm (B. Zhou & Bose, 1993), which uses the previous target states to relate measurements and trackers. Using the same consensus algorithm, He et al., 2019 proposes a tracking approach in a distributed camera network. They address data association within each camera using a metric that merges appearance and geometry cues, and across-view data association through the euclidean distance between the 3D position of the targets. Unlike He et al., 2019, we include a specific strategy to manage the high-level information across cameras to improve the consistency of the trackers in the network and the global data association. Furthermore, our approach enhances bandwidth efficiency through event-triggered communications, a common approach in consensus methods (Ge et al., 2019). Some works, such as C. Zhang and Jia, 2017, present an event-triggered Kalman consensus filter for multi-target tracking in a sensor network, where communication events rely on a constant threshold. This chapter introduces a state-dependent event-triggered mechanism similar to L. Wang et al., 2017. However, different from traditional methods, our approach simultaneously deals with multiple targets, making decisions based on the complexity of their identification.

Data association and re-identification. A wide variety of techniques have been proposed to tackle the problem of data association. One of the most popular techniques is the use of motion models to compute similarity based on geometric constraints, and features such as histograms for appearance criteria (Chandra et al., 2019; de Langis & Sattar, 2020). Commonly, a global similarity function is defined based on both metrics. Other works extract body poses and relate them by the nearest observation statistically consistent with the distribution of positions (Virgona et al., 2018). Taking into account the factors included in the data association problem, previous work defines several costs related to geometry, shape, appearance, pose and coordinate transformation to obtain a complete similarity function (Sharma et al., 2018). The work proposed in Tsai et al., 2019 uses inertial sensing and RGB-D cameras to capture the skeleton data and perform a short-term pairing. A long-term pairing process adds the color histogram to the similarity function to increase robustness. The data association process in our approach is similar to de Langis and Sattar, 2020, but our implementation takes advantage of re-identification strategies supported by deep learning techniques. A generalized strategy is based on comparing feature vectors, obtained from a network output, to measure the similarity between a query image and a global gallery, i.e., appearance model of individuals (T. Chen et al., 2019; P. Liu et al., 2018; Quan et al., 2019). To be efficient and effective in extracting these appearance feature vectors, we use the architecture proposed in K. Zhou et al., 2019, which mixes global and local features in a lightweight network, changing traditional convolution operations with depth-wise operations.

2.3 Distributed Tracking Approach

This section describes the main approach of our multi-target tracking system, which incorporates both low-level and high-level information from the scene in a distributed fashion. Figure 2.2 summarizes the proposed architecture. First, image target positions are obtained with a people detector (Y. Wu et al., 2019) in each camera. Then, the association between the current detections and existing trackers is performed locally in the Local Data Association process (LDA). This association is based on a global score computed from the geometric cues, provided by the local tracking filter, and the appearance similarity with respect to the target's appearance model stored in a local gallery. To continue the cycle, each camera exchanges a communication message with the neighboring cameras, the trackers get into the DKF where the new state of the target is updated and sent to the Distributed Tracker Manager (DTM) block. Finally, the DTM manages the initialization of new trackers and the deletion of unobserved ones uniformly over the network.

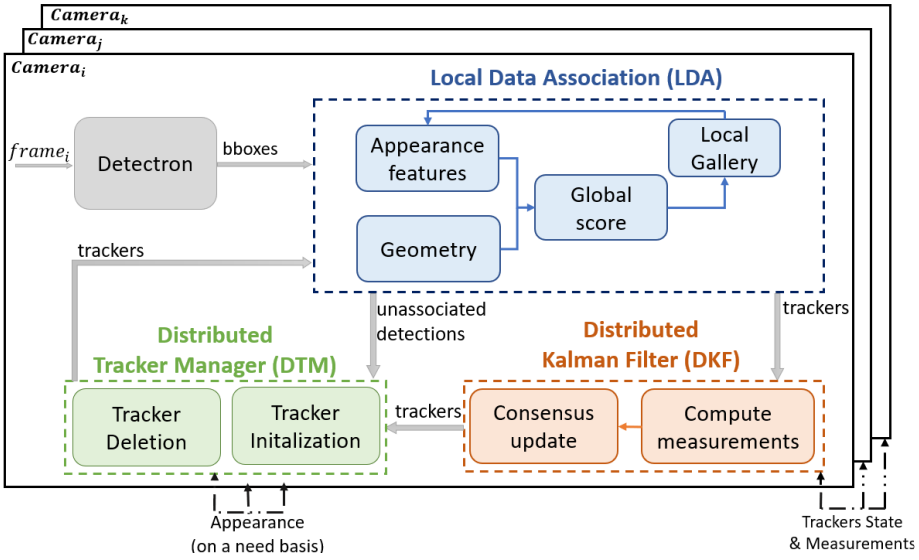


Figure 2.2: Overall architecture of our distributed multi-target multi-camera tracking system.

2.3.1 Distributed Kalman Filter

For simplicity in the exposition, throughout this sub-section we will consider the distributed tracking of a single target. The target model, $\mathbf{x}(k) = (x(k), y(k), w(k), h(k), \dot{x}(k), \dot{y}(k))$, is represented as a 3D cylinder moving on a ground plane, where $(x(k), y(k))$ are the ground plane coordinates of the cylinder center, $(w(k), h(k))$ are the width and height of the cylinder and $(\dot{x}(k), \dot{y}(k))$ are the velocity of the target in the x and y directions. The filter models the motion of the target considering a discrete-time linear dynamical system,

$$\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \mathbf{w}(k), \quad \mathbf{z}(k) = \mathbf{H}\mathbf{x}(k) + \mathbf{v}(k), \tag{2.1}$$

where $\mathbf{w}(k)$ and $\mathbf{v}(k)$ are zero mean Gaussian noise ($\mathbf{w}(k) \sim \mathcal{N}(0, \mathbf{Q}(k)), \mathbf{v}(k) \sim \mathcal{N}(0, \mathbf{R}(k))$), being $\mathbf{Q}(k)$ and $\mathbf{R}(k)$ the model and measurement covariance matrices respectively. The change of the target model in each step depends directly on the transition

matrix $\mathbf{A}$, which considers a constant velocity model on the target position and no dynamics on the rest of the state. The noisy measurement $\mathbf{z}(k)$ is the 3D cylinder obtained as the projection of the bounding box given by the detector, i.e., $\mathbf{z}(k) = (x(k), y(k), w(k), h(k))$ defined with the output matrix $\mathbf{H}$ plus the noise. To obtain this projection we assume known homographies for each camera to map the image plane to ground plane coordinates.

The independent execution of the filter in each camera, C_i, produces a local estimation of the target, $\hat{\mathbf{x}}_i(k)$, possibly different to other cameras' estimation. A Distributed Kalman-Consensus filter is used to mitigate these differences. The consensus algorithm works with known data association between the local measurement, $\mathbf{z}_i(k)$, and target prediction for all the cameras. From this association, each camera computes its sensor data information, $\mathbf{u}_i(k)$, and its inverse-covariance matrix, $\mathbf{U}_i(k)$, defined as the vector and matrix information respectively and obtained as

$$\mathbf{u}_i(k) = \mathbf{H}^T {\mathbf{R}_i}^{-1}(k)\mathbf{z}_i(k), \;\; \mathbf{U}_i(k) = \mathbf{H}^T {\mathbf{R}_i}^{-1}(k)\mathbf{H}. \tag{2.2}$$

The values in (2.2) are exchanged with neighboring cameras in the network, $C_j \in C_i^n$, together with the prediction of the target state, $\bar{\mathbf{x}}_i(k)$, obtained as $\bar{\mathbf{x}}_i(k) = \mathbf{A}\hat{\mathbf{x}}_i(k-1)$. Assuming the measurement noises of the sensors are uncorrelated, the representation in information form allows the cameras to combine all the received measurements with the acquired one by simply adding them,

$$\mathbf{y}_i(k) = \sum_{C_j \in C_i^n} \mathbf{u}_j(k), \qquad \mathbf{S}_i(k) = \sum_{C_j \in C_i^n} \mathbf{U}_j(k). \tag{2.3}$$

The state is then updated by the correction in the prediction of the target state with the merged information and the predictions from the neighboring cameras,

$$\hat{\mathbf{x}}_i(k) = \bar{\mathbf{x}}_i(k) + \mathbf{M}_i(k)\left[\mathbf{y}_i(k) - \mathbf{S}_i(k)\bar{\mathbf{x}}_i(k)\right] + \gamma \mathbf{M}_i(k) \sum_{C_j \in C_i^n} (\bar{\mathbf{x}}_j(k) - \bar{\mathbf{x}}_i(k)), \tag{2.4}$$

where $\mathbf{M}_i(k) = (\mathbf{P}_i(k)^{-1} + \mathbf{S}_i(k))^{-1}$ is the Kalman Gain in the information form, $\mathbf{P}_i(k)$ is the covariance of the target state and $\gamma = 1/\|\mathbf{M}_i(k) + 1\|$. Finally, the covariance matrix is updated according to $\mathbf{P}_i(k+1) = \mathbf{A}\mathbf{M}_i(k)\mathbf{A}^T + \mathbf{Q}_i(k)$.

2.3.2 Local Data Association

In our method, the local data association required for a correct update of the filter is made merging two constraints based on geometry and appearance. Let us consider now that in a particular estimation cycle, the set of measurements $\mathcal{Z} = \{\mathbf{z}_j\}^1$ is provided by the detector to the LDA module. Since the DKF updates the uncertainty of the tracker state, we take advantage of this information to calculate the Mahalanobis distance between the (x, y) position on the ground of each measurement, $\mathbf{z}_j$, and the predicted position, $\bar{\mathbf{x}}_i$,

$$d(\mathbf{z}_j, \bar{\mathbf{x}}_i) = \sqrt{(\mathbf{z}_j - \mathbf{H}\bar{\mathbf{x}}_i)\mathbf{V}^{-1}(\mathbf{z}_j - \mathbf{H}\bar{\mathbf{x}}_i)^T}, \tag{2.5}$$

being $\mathbf{V} = \mathbf{P}_{xy} + \mathbf{R}_{xy}$, with $\mathbf{P}_{xy}$ and $\mathbf{R}_{xy}$ the sub-matrices of $\mathbf{P}_i$ and $\mathbf{R}_j$ that encode the position covariance of the prediction and the measurement respectively. Then, the

[1] We use now the index j to denote different measurements observed in a single camera instead of neighbors in the camera network.

similarity value in geometry is computed as

$$s_d(\mathbf{z}_j, \bar{\mathbf{x}}_i) = \begin{cases} \frac{1}{\alpha}\, d(\mathbf{z}_j, \bar{\mathbf{x}}_i) & \text{if } \; d(\mathbf{z}_j, \bar{\mathbf{x}}_i) < \tau \\ 1 & \text{otherwise,} \end{cases} \tag{2.6}$$

where α is a configuration parameter and τ a threshold applied to ignore highly unlikely candidates.

The candidates selected by the geometry constraint are then evaluated in appearance. Instead of using traditional hand-crafted descriptors to measure the appearance similarity, we employ a network designed for people re-identification (K. Zhou et al., 2019), whose weights have been pre-trained with the MSMT17 Benchmark (Wei et al., 2018). Inspired by the re-identification task evaluation methodology, each tracker is associated with a local gallery, $\mathcal{A}_i = \{\mathbf{a}_i^\nu\}_{\nu=0}^N$, of limited size N, that models the appearance of target i. Then, the appearance similarity between a query and the tracker gallery is obtained with the minimum cosine distance,

$$s_a(\mathbf{a}_j, \mathcal{A}_i) = \min_{\mathbf{a}_i^\nu \in \mathcal{A}_i} \left(1 - \frac{\mathbf{a}_j^T\, \mathbf{a}_i^\nu}{\|\mathbf{a}_j\| \|\mathbf{a}_i^\nu\|} \right), \tag{2.7}$$

where $\mathbf{a}_j$, the query, is the appearance feature vector associated to detection $\mathbf{z}_j$. The construction of the gallery, $\mathcal{A}_i$, is done by locally storing observed patches for each target with periodic updates every Φ frames. This chapter presents two types of appearance feature selectors to determine which features compose the appearance model:

1. **Temporal feature selector (TFS).** The naive version of the local gallery keeps all the features periodically captured. When the gallery saturates, the temporal feature selector discards the oldest feature to make room for the newest one.

2. **Quality-based feature selector (QFS).** The quality-based appearance feature selector relies on metrics that estimate the usefulness and reliability of the features composing the local gallery. The patches arriving at the gallery are saved if and only if they comply with a minimum quality criterion, $q(\mathbf{a}_i^\nu) \rightarrow \mathbb{R} \geq 0$. Specifically, we used the ratio of visible skeleton joints, $q(\mathbf{a}_i^\nu) = q_i^\nu / q_t$, being q_i^ν the number of skeleton joints that represent the feature $\mathbf{a}_i^\nu$ and, q_t the number of joints in a complete skeleton. To avoid introducing low-quality appearance features such as strong occlusions or partial observations, every saved component must comply with $q(\mathbf{a}_i^\nu) \geq 0.5$. Then, we define a scoring system to discard the least valuable features when the gallery saturates. Each component of the appearance model is assigned a score that varies based on two factors. The first factor rewards the usefulness of the feature by increasing the score value in one of the ν component that provides the minimum cosine distance in $s_a(\mathbf{a}_j, \mathcal{A}_i)$. The second factor sorts the features of the gallery based on their spatial distribution in the descriptor space. The closest feature to the gallery centroid increases its score by one, while the farthest feature decreases it by one. When the gallery is full, the feature with the lowest score is discarded.

The final decision is based on selecting the tracker $\bar{\mathbf{x}}_\mathbf{i}$ with the minimum cost c_i to be associated with the measure $\mathbf{z}_j$. The cost function models how far away target i and measurement j are from each other and is defined by

$$c_i = s_d(\mathbf{z}_j, \bar{\mathbf{x}}_\mathbf{i}) s_a(\mathbf{a}_j, \mathcal{A}_i). \tag{2.8}$$

Furthermore, as an improved version of our distributed multi-target tracking algorithm, we extend the dimension of this assignment problem taking into account the cost of associating every combination, c_{ji}, of the set of measurements $\mathcal{Z} = \{\mathbf{z}_j\}$ with the set of candidate trackers $\mathcal{X} = \{\bar{\mathbf{x}}_\mathbf{i}\}$ through the optimal assignment problem

$$\begin{aligned} \underset{p_{ji}}{\text{minimize}} \quad & \sum_j \sum_i p_{ji}(k) c_{ji}(k), \\ \text{subject to} \quad & \sum_i p_{ji}(k) = \sum_j p_{ji}(k) = 1,\ p_{ji}(k) \in \{0,1\}, \forall j, i, \end{aligned} \tag{2.9}$$

This linear assignment problem can be efficiently solved using the Hungarian algorithm (Kuhn, 1955).

2.3.3 Distributed Tracker Manager

Another important issue to address in a practical implementation of the DKF is the management of the trackers through the full distributed system. This requires a correct data association of trackers across different cameras to guarantee that the information mixed in (2.3) and (2.4) corresponds to the same target. Similarly, the cameras need to agree upon when a particular tracker is no longer relevant and should be dropped. We propose how to address these problems in a distributed fashion.

Distributed Global Data Association. Trackers are identified locally by a two-dimensional unique identifier, ID_i, described by the camera id in the network, i, and a local counter, n. New trackers can either be initialized because of a new local observation or because of a transmission from neighboring cameras. Local initialization is done whenever a new target generates two observations in its local gallery. This helps filtering spurious measurements from the detector, giving enough time for a new tracker to ensure that it corresponds to a valid target. Once this happens, we attach to the DKF data the appearance model, $\mathcal{A}_{\text{ID}_i}$, in the message sent to neighbors. It is important to highlight that this is the only moment when appearance is transmitted through the network in our algorithm, consisting in the two descriptors available in the local gallery at that time. The second case that can trigger new tracker initializations in our system is the reception of messages from neighbor cameras. Our algorithm considers three situations for this case: 1) A single neighbor camera sends a new tracker. Then, the camera creates a new tracker and associates it to the received one for the future DKF consensus updates. The local gallery is initialized with the appearance model received. 2) The camera receives new trackers from several neighboring cameras. 3) A new local tracker is initialized at the same time that new trackers from other cameras are received. In situations 2) and 3), it is necessary to check whether the new trackers from the different cameras are of the same target or not. We perform a similar process to the one for local data association described in Section 2.3.2, replacing in (2.5) the measurement by the other camera's estimation, $d(\hat{\mathbf{x}}_i, \bar{\mathbf{x}}_j)$, and the Mahalanobis distance with the Euclidean distance, $\mathbf{V} = \mathbf{I}$, since the covariance matrices associated to the trackers are not part of the communication messages. This also requires a different threshold in (2.6). If two or more trackers are similar enough, they are merged locally into a single one.

Consensus-based Tracker Drop and Re-initialization. The other main task of the Distributed Tracker Manager is to decide when to drop a tracker. Instead of letting each camera to decide this process individually, we have opted for a consensus-based solution that reduces, as much as possible, the number of iterations that different cameras carry out the tracking individually. We let ℓ_i be the local estimation that camera i has on the number of iterations gone since the last local data association of a measurement to the target made by any camera of the network. Since this is a global parameter that involves the whole network, its estimation is sent as part of the tracker message. If the camera achieves the local data association of the tracker, the new value of this parameter is set to zero. Otherwise, the camera chooses as new value the minimum among all the values received, including its own, and adds one unit,

$$\ell_i(k+1) = \begin{cases} 0 & \text{if detected} \\ \min\limits_{j \in \mathcal{C}_i^n} (\ell_i(k), \ell_j(k)) + 1 & \text{otherwise} \end{cases} \tag{2.10}$$

The camera drops the target when $\ell_i(k+1)$ is higher than a threshold κ. The tracker ID is saved together with its gallery as an *old tracker*, to be recovered if the same target is back in the camera field of view or some other camera re-activates it. This process is checked during the local initialization. Before assigning a new ID, a re-identification score is computed between the new tracker gallery and the galleries from *old trackers* through (2.7). The re-initialization is accomplished if s_a is lower than ϵ.

2.3.4 Bandwidth Requirements and Event-triggered Communication.

The information shared between cameras consists, for each active tracker, of the tracker ID, its predicted state, the measurements obtained in (2.2) and the counter of the last observation, (ID_i, $\bar{\mathbf{x}}_\mathbf{i}, \mathbf{u}_i, \mathbf{U}_i$, ℓ_i). This message is encoded using a total of 51 elements, representing less than 1kB of bandwidth information per tracker. In order to carry out the process explained in the *Distributed Global Data Association*, the message with the information related to the new trackers is sent together with appearance information. The appearance information exchange between cameras is composed of two appearance feature vectors required in the trackers' initialization. Each one has 512 elements, representing 16.38kB of bandwidth information that is sent only once per tracker.

The consensus nature of our algorithm makes it amenable to a fully asynchronous data transfer between cameras. Thus, we aim to minimize the standard DKF's requirement of communicating data between cameras once per cycle through an event-triggered communication function that shares the complete message only when deemed relevant. The implemented event-triggered mechanism tackles the inherent trade-off in distributed systems of balancing reduced bandwidth usage between cameras while maintaining tracking accuracy. Following the existing event-triggered method L. Wang et al., 2017, a camera will send the information of target i at time k when the difference of its estimated state, $\hat{\mathbf{x}}_i(k)$, with respect to the last communication is greater than a certain quantity,

$$(\hat{\mathbf{x}}_i(k) - \hat{\mathbf{x}}_i(k_c))^T(\hat{\mathbf{x}}_i(k) - \hat{\mathbf{x}}_i(k_c)) > \beta \hat{\mathbf{x}}_i(k_c)^T \hat{\mathbf{x}}_i(k_c), \tag{2.11}$$

where k_c is the time of the last communication for that target. Differently from other works, we propose to design β as an adaptive threshold that takes into account the influence of

other targets on the tracking. For each measurement, $\mathbf{z}_j \in \mathcal{Z}$, we count the number of targets within a threshold distance, τ,

$$\mathcal{X}_j(\mathbf{z}_j(k)) = \{\bar{\mathbf{x}}_i \in \mathcal{X} \mid d(\mathbf{z}_j(k), \bar{\mathbf{x}}_i(k)) \leq \tau\} \tag{2.12}$$

where $d(\mathbf{z}_j, \bar{\mathbf{x}}_i)$ is the Mahalannobis from (2.5). Let $|\mathcal{X}_j(\mathbf{z}_j(k))|$ be the cardinality of the set and define the target density of the scene at time k as

$$\rho(k) = \max_{\mathbf{z}_j \in \mathcal{Z}} |\mathcal{X}_j(\mathbf{z}_j(k))|. \tag{2.13}$$

This parameter indicates how hard (or easy) is going to be the data association of measurement $\mathbf{z}_j$ to a target. When the target density is high, i.e., two or more targets are close (Figure 2.3), the matching is challenging, and therefore exchanging information with other cameras is potentially beneficial for the system. On the other hand, if the target density is low, i.e., only one or distant targets, there is no need for gathering information frequently. Based on this, our adaptive threshold in (2.11) is defined by

$$\beta(k) = 0.1/\rho^2(k). \tag{2.14}$$

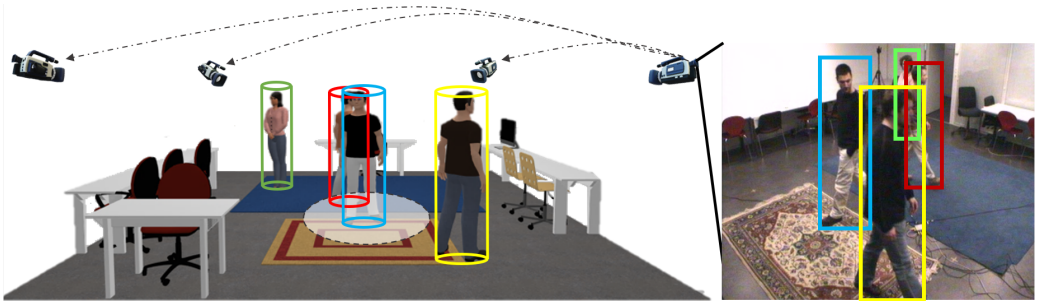


Figure 2.3: Density computation. The red and blue measurements in the right image have two potential tracker candidates represented by the same color in the figure on the left. In this scenario, the data association is more challenging due to the crossing between targets. The density parameter softens the event triggering that sends information to neighbors.

Furthermore, to ensure consistency on the consensus-based tracker drop explained in Section 2.3.3, which determines when a tracker becomes inactive and is defined as *old tracker*, we introduce a second event generator function,

$$k - k_c \geqslant \kappa, \tag{2.15}$$

where κ is the maximum number of iterations since the last time the tracker was associated with a measurement. Thereby, the tracker's information will be sent just before any camera deletes or stores it, ensuring that if another camera of the network is observing the target, it will resume tracking with up-to-date information.

2.4 Experiments

This section details the common experimental setup used throughout all evaluations. Then, it discusses the study conducted on our distributed tracking approach. This analysis focuses on the influence of the key modules essential to our method: the Distributed Kalman Filter (DKF), the Local Data Association (LDA), and the Distributed Tracker Manager (DTM). We compare the results obtained with our approach to those of some centralized tracking setups. Finally, the section assesses how the extended modules, i.e., the optimal data association assignment, the quality-based feature selector, and the event-trigger mechanism, impact the performance of the main method.

2.4.1 Experimental Setup.

Datasets. Our experiments are run on three sequences from the EPFL dataset (Berclaz et al., 2011): *Terrace* set with 4 outdoor cameras, *Laboratory*, with 4 indoor cameras and *Campus* set with 3 outdoor cameras. Additionally, we include qualitative results on a subset of the more complex WildTrack dataset (Chavdarova et al., 2018), captured with 7 outdoor cameras with partial overlap.

Metrics. In order to evaluate our contributions, we use standard Multiple Object Tracking (MOT) metrics defined in Bernardin and Stiefelhagen, 2008 and the continual identification metrics presented in Ristani et al., 2016a. The first metric obtained, Multi-Object Tracking Accuracy (**MOTA**), measures failures during the tracking taking into account the number of misses, m_t, false positive, fp_t, and mismatches, mme_t, over the total number of targets, g_t, per frame, t,

$$\text{MOTA} = 1 - \frac{\sum_t (m_t + fp_t + mme_t)}{\sum_t g_t}. \tag{2.16}$$

Multiple Object Tracking Precision (**MOTP**) shows the ability of the tracker to estimate precise object positions through the error in estimated position for matched object-hypothesis pairs, d_t^i, over the total number of matches made, c_t,

$$\text{MOTP} = 1 - \frac{\sum_{i,t} d_t^i}{\sum_t c_t}. \tag{2.17}$$

Then, we evaluate the capability of the system for preserving the identities with Identification Precision (**IDP**), Identification Recall (**IDR**) and their F1 Score (**IDF1**),

$$\text{IDP} = \frac{IDTP}{IDTP + IDFP}, \ \text{IDR} = \frac{IDTP}{IDTP + IDFN}, \tag{2.18}$$

$$\text{IDF}_1 = \frac{2IDTP}{2IDTP + IDFP + IDFN}, \tag{2.19}$$

being IDTP, IDFP and IDFN the true positives, the false positives and the false negatives, respectively, during the tracking process. All the evaluations follow the same process explained in Y. Xu et al., 2016, giving as a final result the median of each metric for all the cameras available.

Topology effect. Moreover, the performed experiments study the effects of different network topologies (complete, ring, chain, and disconnect). For chain and ring topologies we have considered several network alternatives to make the evaluation independent of the individual quality of particular cameras in the tracking. The considered combinations are shown in Figure 2.4.

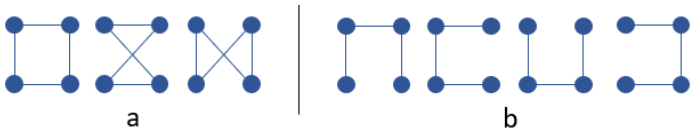


Figure 2.4: Alternatives considered for (a) ring graph topologies and (b) chain graph topologies.

Image Processing. For the person detection, all experiments run the official Detectron implementation (Y. Wu et al., 2019), filtering the output with a process equivalent to non-maximal suppression. Regarding the appearance features, these are obtained with the person re-identification neural network K. Zhou et al., 2019 pre-trained in the MSMT17 Benchmark (Wei et al., 2018).

2.4.2 Distributed Multi-target Tracking Evaluation

Implementation details. The configuration parameters from the proposed distributed tracking algorithm are set to $\alpha_{LDA} = 2000$, $\alpha_{GDA} = 50$, $\tau = 0.5$, $\kappa = 15$, $\epsilon = 0.25$, $N = 20$ and $\Phi = 20$. These parameters were defined in Sections 2.3.2 and 2.3.3.

Ablation study. This first experiment analyzes three configurations of the distributed tracking approach to evaluate the effects of the novel modules included in our architecture. The simplest configuration is our implementation of the Distributed Kalman Filter using only geometric information in the data association process (DKF). A second version adds the appearance information from the local gallery built with the temporal feature selector (DKF + LDA). The latest version includes the distributed tracker management module in the system. This version represents the main algorithm presented in the paper (DKF + LDA + DTM). In all the configurations the communication between nodes is performed once per cycle. Furthermore, we evaluate the influence of the appearance in a disconnected graph, i.e., without information shared between cameras and thus running four independent Kalman filters.

Figure 2.5 summarizes the results of the conducted tests. Each row shows the results obtained using certain dataset, and each column corresponds to the results with a different connectivity graph. Overall, we can see how both proposed modules (LDA and DTM) bring significant improvements in all the metrics and datasets, with more positive influence in the identification metrics (IDF1, IDP and IDR) than the tracking metrics (MOTA, MOTP). For a few cases, it would be slightly better to apply only LDA, but the penalization for including DTM in those few cases (decrease of less than 5%) is not nearly as significant as the benefits it brings in the rest of the cases (up to 20% increase in some network topologies with respect to the DKF+LDA configuration). Figure 2.6 shows some qualitative results for each of the datasets [2].

[2]More qualitative results in: https://www.youtube.com/watch?v=syg4EYXBWX0

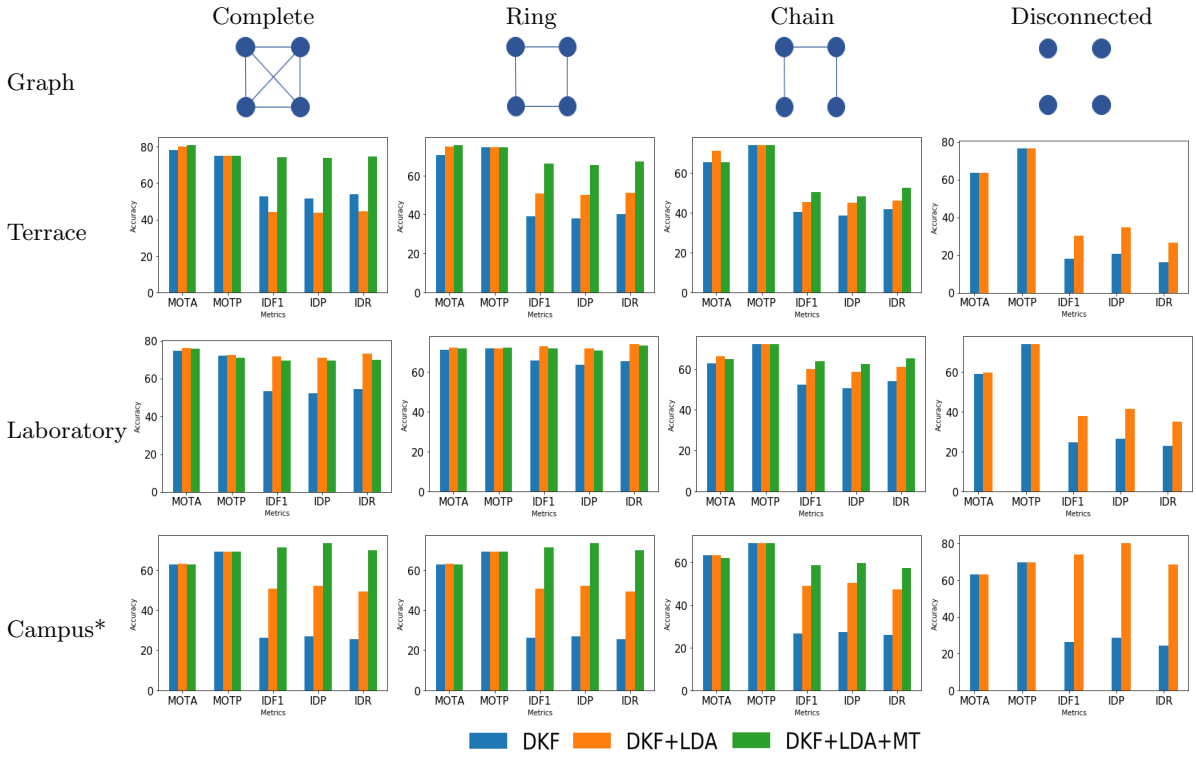


Figure 2.5: Ablation study for our approach considering different graph topologies (complete, ring, chain, disconnected). The variations are run on three sets of data (Terrace, Laboratory and Campus) and evaluated using CLEAR MOT metrics. Running the main steps of the proposed approach (DKF+LDA+MT) achieves the best results in the majority of cases. *The Campus dataset only has three cameras, obtaining the same results in the complete and ring graphs.

Comparison with other algorithms. The results obtained with our distributed multi-target tracking algorithm are compared with centralized methods evaluated in Y. Xu et al., 2016. They publish results on public datasets running their proposed approach, Hierarchical Trajectory Composition (HTC), as well as two other baselines, Probabilistic Occupancy Map (POM) Fleuret et al., 2007 and K-shorted Path (KSP) Berclaz et al., 2011. To make a fair comparison, the evaluation parameters are those described in Y. Xu et al., 2016. Among the datasets used there that provide accurate camera-ground plane calibration (required to run our algorithm), we picked the most challenging sequence of *Terrace*, where a high number of crossings and occlusions occur between targets. Note this does not intend to be a thorough evaluation, but rather an experiment to see where the proposed distributed approach gets in comparison with existing centralized methods.

Previously discussed distributed approaches Kamal et al., 2015; Soto et al., 2009 can not be included in this study because up to our knowledge they do not provide MOT metric results in available benchmarks, or focus on different goals and metrics than us. In Kamal et al., 2015, the experiments focused on the consensus algorithm analysis, whereas Soto et al., 2009 uses its own camera network to test the proposed approach, showing as result the trajectories of the individuals on the ground. Section 2.2 already highlighted our approach advantages with respect to these systems.

The results are summarized in Table 2.1, where we see better performance of our approach with respect to the centralized methods in two of the three graph topologies. The

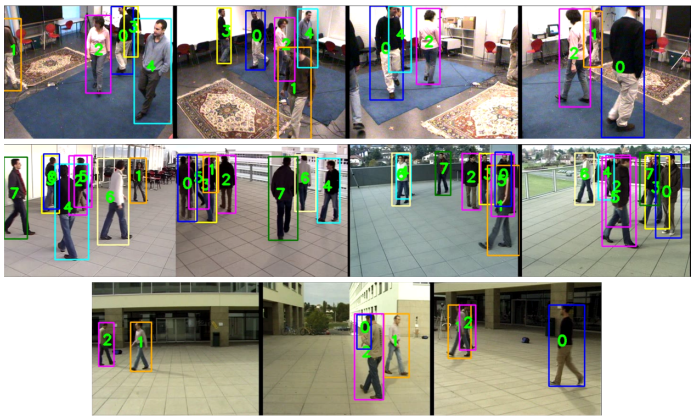


Figure 2.6: Qualitative tracking results of our approach in three datasets: *Laboratory* (first row), *Terrace* (second row) and *Campus* (third row).

Algorithm	MOTA	MOTP	Bandwidth [kB/frame]^o
Centralized baselines			
HTC (Y. Xu et al., 2016)*	71.84	71.15	1215
KSP (Berclaz et al., 2011)*	65.75	57.82	1215
POM (Fleuret et al., 2007)*	56.9	61.33	1215
Distributed proposed approach			
DKF+LDA+DTM (Complete)	80.95	75.2	28.23
DKF+LDA+DTM (Ring)	75.98	74.42	28.23
DKF+LDA+DTM (Chain)	69.74	73.96	28.23

* Results interpolated Y. Xu et al., 2016, only shown graphically.

^{o} Bandwidth calculated analytically more details in the text.

Table 2.1: MOT metrics and bandwidth requirements per frame of our distributed approach communicating once per cycle and existing centralized methods on *Terrace* dataset.

complete graph, being the closest version to the centralized system, gets an improvement of 9.11% in MOTA while the ring graph improves the results by 4.14%. Finally the chain graph obtains lower MOTA, but still comparable results to those of the HTC algorithm. Regarding the bandwidth, for the centralized systems we have computed it considering that every camera is at one hop communication to the central server. Since there are four cameras, the server needs to receive 4 images of 288×360 RGB pixels each one, giving a total of 414,720 pixels. Each pixel is encoded with 3 Bytes (R+G+B), so the total bandwidth required is 1,215kB/frame. For the distributed system, we compute an upper bound of the bandwidth, considering that in every iteration the information of 9 active trackers, which is the maximum number of trackers active during the whole execution. Each tracker requires 0.78kB/frame (Section 2.3.4), so when multiplied by the 4 cameras and the 9 trackers results 28.08kB. The total number of trackers initialized in our algorithm is 24, that over all the iterations average a total of 0.15kB/frame to send the appearance information of the new trackers. The sum of these two quantities results in the 28.23kB/frame of Table 2.1.

2.4.3 Extended Distributed Multi-target Tracking Evaluation.

This subsection analyzes the impact on the performance of the main method by including the extended modules presented in this chapter, the optimal data association assignment, the quality-based feature selector and the event-trigger mechanism.

Implementation details. Due to the decrease in communications caused by the event-triggered communication and the greater influence of local galleries when incorporating the quality-based feature selector, the configuration parameters using any of them have been set to $\alpha_{LDA} = 1200$, $\alpha_{GDA} = 100$, $\tau = 1$, $\kappa = 15$ and the gallery size to 20 samples with $\Phi = 10$. These parameters were defined in Sections 2.3.2 and 2.3.3.

Ablation Study. This study demonstrates the benefits the proposed extended modules bring to the main Distributed Multi-Target Tracking (DMTT) system previously evaluated. Since preserving the assigned identity to each target over time appears to be the most influential metric, IDF1 is assessed in this analysis. We compare four algorithms in terms of accuracy and bandwidth requirements: 1) DMTT, 2) our event-triggered mechanism using the optimal data association assignment (DMTT + E), 3) the quality-based feature selector combined with the optimal data association assignment (DMTT + QFS), and finally, 4) the complete system (DMTT + QFS + E).

Figure 2.7 shows the performance across different datasets and the impact of the connectivity graph. Overall, integrating our quality-based feature selector (DMTT + QFS) enhances the IDF1 metric while the continual identification is not compromised when communications are controlled by the event-trigger mechanism (DMTT + QFS + E). Figure 2.8 illustrates examples of appearance models resulting from the QFS where each row represents the final gallery of a target. The diverse and representative data selected for the gallery provides a more robust appearance assessment compared to storing only the most recent information.

Table 2.2 presents the percentage of communications using our event-triggered mechanism with respect to making one per iteration as in DMTT. The results indicate that our

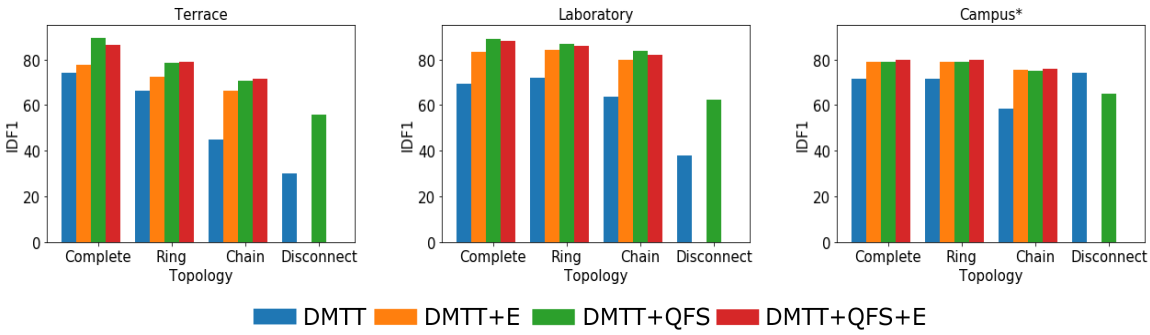


Figure 2.7: Ablation study for our approach considering different graph topologies (complete, ring, chain, disconnected). We compare the DMTT approach (blue), the integration of our event-triggered using the optimal data association (orange), the integration of our quality-based feature selector combined with the optimal data association (green), and finally, the complete integrated system (red). The variations are run on three sets of data (Terrace, Laboratory and Campus) and evaluated using the IDF1 MOT metric. *The Campus dataset only has three cameras, so the Ring and Complete topologies are the same.

adaptive event-triggered mechanism is able to identify the complexity of the scene increasing the communication rate accordingly. Thus, in sequences where the scene is crowded, i.e., Terrace and Laboratory, communications are reduced by about 50%, while in simpler scenarios where just a few targets are tracked, i.e., Campus, the system exchanges information in only 22% of the estimation cycles.

Finally, analyzing both Table 2.2 and Figure 2.7, we can confirm that the proposed event-triggered mechanism (DMTT + QFS + E) offers meaningful savings in communications, without sacrificing accuracy, i.e., achieving comparable results to the DMTT + QFS algorithm.

Figure 2.8: Appearance models obtained with our quality-based feature selector. Each row shows a target model from each dataset evaluated: Terrace, Laboratory and Campus, respectively.

	% Communications		
Dataset	*Complete*	*Ring*	*Chain*
Terrace	55.5	52.7	54.4
Laboratory	52.5	49	50.12
Campus	22	22	21.76

Table 2.2: Percentage of communications with our event-triggered mechanism with respect to making one communication per cycle.

Tracking in the Wild. The last experiment is a proof of concept focused on identifying and tracking a subset of 20 people in the WILDTRACK dataset (Chavdarova et al., 2018). This is a more challenging dataset than the previous ones, especially for the re-identification, than those considered in the previous experiments. It includes 7 cameras and strong light changes from different perspectives.

In order to carry out the re-identification process, it has been necessary to manually build an initial common gallery of each person to store in each camera (global gallery), to indicate the subset of targets of interest for this experiment. We selected the 20 people who appear more frequently and store an average of 6 images per person, selecting one image per existing point of view to form a global gallery. Since the calibration matrices only contemplate the transformation between the image plane and the main ground plane, the other areas of the image are ignored. For this experiment, because the scene conditions and the accuracy of the homographies are different, our algorithm is configured with $\tau_{LDA} =$

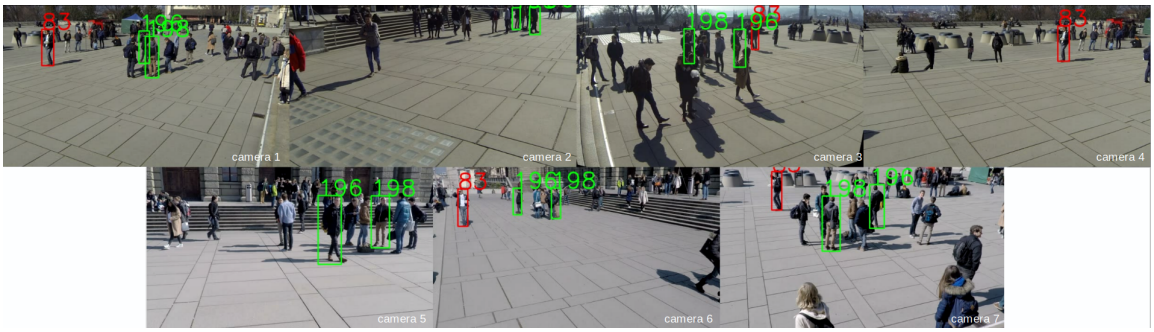


Figure 2.9: Example of the people identified from the Wildtrack subset at a certain time (the seven images correspond to the seven dataset cameras). A few correct identifications are highlighted (with complete or partial occlusions) with green boxes. The failure of not identifying a person-of-interest is marked with red boxes. (Best viewed in color).

2000 and $\tau_{GDA} = 100$. The local gallery size is set to 5 with $\phi = 1$, $\kappa = 1$. The re-identification process is performed with the minimum cosine distance between the targets and the global gallery matching when the distance is lower than 0.15.

Figure 2.9 shows the images of the 7 cameras with the tracking results at a particular iteration, illustrating the benefits and issues of the method. For example, cameras can correctly locate targets even if they are partially occluded or far away, thanks to information shared by other cameras. This happens with camera 5 about ID196, which helps camera 1 to identify the same target, even though it is occluded by ID198. In this dataset, the major problems are caused by the occlusions and people crossings, which generate mismatches between close and similar targets, and the difficulty in obtaining a correct re-identification[3].

The only change in the evaluation process with respect to the previous section is the threshold of the Intersection over Union (IoU) between the tracker and the ground truth to consider the predicted position correct. In this case, as the dataset is more challenging and considers smaller bounding boxes, we set the threshold to 0.3. In order to analyze the re-identification process, we include in this experiment the results of precision and recall used to compute IDF_1. The results obtained for the complete and the ring connected graphs are the following:

- Complete Graph. IDF1: 50.1 IDP: 43.9 IDR: 57.2
- Ring Graph. IDF1: 43.4 IDP: 42.4 IDR: 48.3

As we can see, the recall is higher than the precision in both cases. This tendency is probably caused by false positives in the re-identification, due to high similarity (specially in appearance) between people included and excluded from the tracked subset.

[3]More qualitative results in: https://youtu.be/t6MwklDrwE4

2.5 Conclusions

This chapter has presented a novel distributed approach to multi-target tracking problems. Our multi-target tracking algorithm, designed for distributed camera networks, provides a complete method that deals with the distributed fusion of low- and high-level information. In this thesis, the challenges of a distributed system have been addressed boosting the DKF with a fully automatic data association, based on geometric and appearance constraints, along with a novel tracker manager to handle the misalignment of the high-level information. The distributed tracker manager takes care of the global data association and each tracker's consistency in the network to reduce the number of iterations that different cameras carry out the tracking individually. Regarding the problem of bandwidth usage, we propose an event-triggered mechanism that dynamically adjusts the frequency of communications based on the complexity of the scene. Additionally, the use of appearance information in the data association process is harnessed by selecting valuable appearance features for composing the local galleries of each target, resulting in a more robust appearance assessment. The proposed approach is evaluated in challenging public benchmarks and achieves comparable or even better results than centralized systems. Furthermore, the conducted ablation studies analyze the influence of each presented module, demonstrating the benefits of their integration into the distributed algorithm.

Chapter 3

Self-adaptive gallery for open-world person re-identification

Effective monitoring systems need to be able to handle the changing nature of real-world environments. When monitoring people using multiple cameras with non-overlapping views, the ability to re-identify individuals across time and space is crucial. This task, known as person re-identification (re-id), involves recognizing a person previously captured by another camera in a different location or even by the same camera at a distinctive time. Traditional re-id approaches commonly rely on a pre-built *gallery* with relevant information about the people already observed. However, constructing and maintaining such a gallery is a costly and tedious process due to the challenges associated with labeling and storing continuous incoming data. Consequently, this gallery is typically created in an offline and one-time process, making it static and unable to include new information, thus limiting its applicability in open-world scenarios.

This Chapter introduces a novel unsupervised approach that tackles the aforementioned limitations. Our method automatically identifies new individuals and incrementally constructs a gallery for open-world re-id. The proposed approach continuously adapts the gallery knowledge by incorporating new information as it arrives while ensuring efficient storage by saving only the most representative data for each individual and the overall gallery. Exploiting concepts of information theory, we use the uncertainty and diversity of the new observations to define which ones should be incorporated into the model of each person. Experimental evaluation in challenging benchmarks demonstrates the benefits of our approach against other data selection algorithms. Furthermore, we include a discussion comparing the obtained results with other unsupervised and semi-supervised re-id methods.

3.1 Introduction

Person re-identification, or simply re-id, addresses the problem of matching people across non-overlapping views in a multi-camera system (D. Wu et al., 2019; Ye et al., 2021). Solutions to this problem benefit many robotic applications where people are involved, such as tracking (Chang et al., 2015; de Langis & Sattar, 2020), navigation (Truong & Ngo, 2017) or searching (Mohamed et al., 2019; Shree et al., 2020). An extensive number of studies have focused on obtaining the best feature representation in supervised close-

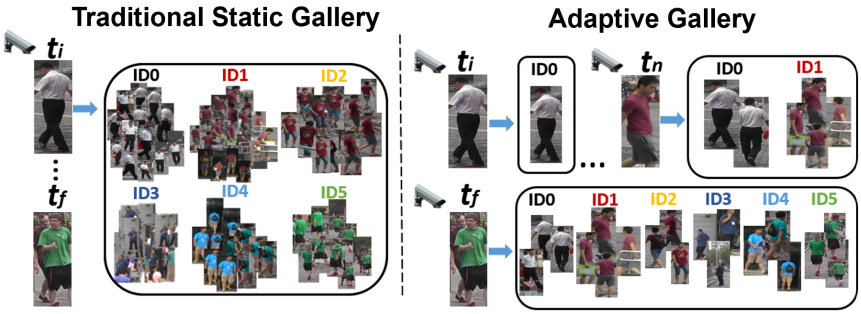


Figure 3.1: Simplified comparison between a large static gallery, traditionally used, and our small self-adaptive gallery. Both have a set of images representing each identity (ID0, ID1, ...), i.e., each person. The traditional gallery is the same for every person query that arrives at different times (t_i, t_f). However, because the adaptive gallery is being built and updated as new data arrives, we can appreciate a more comprehensive gallery for later times ($t_i < t_n < t_f$).

world scenarios (e.g., Hou et al., 2021; H. Luo et al., 2019; Z. Wen et al., 2019; K. Zhou et al., 2021) where the problem is narrowed to seek a query person from an existing pool of labeled people images, generally called *gallery*. While they obtain high performance in commonly used benchmarks, from the viewpoint of practical re-id systems, people identity annotation to obtain sufficient ground truth data could be extremely inefficient (Leng et al., 2019). Hence, there is a tendency in the research community to address other alternatives and still open problems in re-identification, such as unsupervised (Lin et al., 2020; Sridhar Raj S & Balakrishnan, 2022; D. Wang & Zhang, 2020), domain adaptation (Feng et al., 2021; G. Wang et al., 2020; Y. Zheng et al., 2021) or open-set in open-world (Huang et al., 2020; Martini et al., 2020; Y. Zhao et al., 2019). The vast majority of these works use a static and preset gallery in their development that restrains the dynamic nature of the open-world, where raw data from camera systems collect new people, detection errors, or junk data. In order to solve problems related to open-world recognition, the system needs to deal with unknown classes but also be able to incrementally self-adapt by acquiring new knowledge (Bendale & Boult, 2015; Fontanel et al., 2020). Therefore, an open-world re-identification system should automatically evolve its gallery, be able to identify new identities and update known people's data. To the best of our knowledge, existing approaches in person re-identification have not yet considered this fundamental problem of building a fully unsupervised self-adaptive gallery. Thus, the lack of methods that address this problem motivates our research to propose a re-identification framework that focuses on the applicability of re-id approaches in open-world settings without any human assistance.

This chapter presents a novel framework for person re-identification focusing on a self-adaptive gallery that evolves over time in an unsupervised fashion. The presented framework is able to dynamically expand to identify new individuals and build their appearance models with representative information. Figure 3.1 gives an overview of the differences between a labeled and static gallery traditionally used and our proposed adaptive gallery. Unlike the static gallery, we start with an empty gallery and update its structure as new samples arrive (unlabeled person images) to acquire new knowledge. The samples that provide the most representative appearance description of each person are selected to be included in the gallery. This selection is fully unsupervised and assembled using concepts

of active learning techniques. Specifically, we analyze the uncertainty and diversity of each sample to evaluate its informativeness, keeping only those that present a good balance between low uncertainty and high diversity, i.e., less likely to be failures but not redundant with the rest.

More concretely, the main contributions of this chapter to the field of open-world person re-id are two-fold: (1) A novel approach for building a self-adaptive gallery for person re-identification in open-world scenarios in an unsupervised fashion. The appearance model of each person is kept small and representative by selecting those samples that are most representative using information theory concepts. (2) A thorough evaluation of the posed problem. We include a metric based on the standard precision and recall to evaluate the quality of the gallery structure. This metric provides an intuition of the final quality of the gallery structure when the problem is complex and identifying the total number of classes is highly challenging.

The experiments section provides a detailed analysis of the main parameters defined in the method, along with a comparison of different data selection algorithms commonly used in incremental settings. A comparison with other unsupervised and semi-supervised re-id methods is also discussed.

3.2 Related Work

The problem of person re-identification has been widely studied through time, as shown in L. Zheng, Yang, and Hauptmann, 2016. Early works defined the problem as tracking (Zajdel et al., 2005), then moved to image-based classification (Gheissari et al., 2006) and video-based classification (Bazzani et al., 2010). With the success of deep learning, works have shifted from hand-crafted descriptors (R. Zhao et al., 2013) to deep learning methods (Yi et al., 2014). The next step in person re-identification research was the shift from close-world settings, assuming only known classes and correctly annotated data, to an open-world approach with multiple modalities, limited noisy annotations and an undefined number of people. This shift has raised interesting new research challenges (Bendale & Boult, 2015) relating the problem to other fields.

Unsupervised and Semi-Supervised Re-Id Methods. Several works attempt to tackle the re-id problem by building the re-id models in an unsupervised or semi-supervised manner. For example, Panda et al., 2019 presents a method to add a new camera to a multi-camera re-id system using unsupervised transfer learning from the knowledge obtained on the other cameras. Unsupervised algorithms typically focus on modeling the spatiotemporal information to match people images between them (Lin et al., 2019; Sridhar Raj S & Balakrishnan, 2022), generate new data from unlabeled samples similar to data augmentation techniques (H. Chen et al., 2022; X. Zhang et al., 2022), or reduce the error in hard pseudo-labels using softer adaptable pseudo-labels (Lin et al., 2020). Semi-supervised methods leverage the available annotated information by gradually refining the descriptors with the unlabeled data most similar to the labeled one (Y. Wu et al., 2018) or by generating virtual samples based on the annotated data (Han et al., 2020). Different from these works, we propose a method that focuses on creating a gallery that incrementally adds new unsupervised data with no retraining of the feature descriptors.

Incremental Person Re-Id. Incremental person re-identification has been approached from two main perspectives. First, the incremental adaptation of the learned model as new data arrives at the system (Martinel et al., 2016). This perspective trains the model in the same domain as the queries that will be analyzed later and uses a human in the loop to label the most representative data for the model adaptation through active learning techniques. Second, instead of adapting the feature representation, the goal is to perform a re-ranking in the gallery as new queries are matched with the labeled images (Z. Wang et al., 2019). Both perspectives use a static large gallery that ensures a match for the query person.

Gallery Construction. The construction of the gallery is based on the principle that instances of the same class are close in the feature space. This problem is often solved using clustering algorithms (Lin et al., 2019), which have been studied thoroughly in the literature (D. Xu & Tian, 2015; R. Xu & Wunsch, 2005) and applied in many fields. Close to our approach, DeCann and Ross, 2015 and (De Feyter et al., 2019) present works that update the reference database (gallery) if the new data is not similar to any user by adding new users. However, they focus on the face recognition problem and an unlimited amount of stored data. To deal with the gallery construction problem in incremental scenarios, the available system resources should be taken into account since storing all the information received in a limitless fashion is not feasible. Therefore, the imposition of a bounded memory is commonly applied in many of these incremental approaches (Castro et al., 2018; Rebuffi et al., 2017). Some works address the dynamical expansion of the classes aided by the manual labeling of the novel samples (Mancini et al., 2019; Valipour et al., 2017), while others also consider receiving new instances of already known classes, facing the challenges related to the update of existing class models (Azagra et al., 2020; Hayes et al., 2019). They perform the update of each class model using a scoring system and controlling the size limit of each class by merging the most similar elements. This scenario is the most similar to our approach, but different from these existing works, the proposed framework updates the model by analyzing not only the diversity of the samples but also the global uncertainty of the gallery. The result sought by combining both properties of obtaining a more varied model, is similar to that of prior work in Bang et al., 2021, which selects data with different levels of uncertainty from a set of labeled images. Unlike all these methods, our approach deals with incremental and unlabeled information in an open-world scenario.

3.3 Method

This section describes in detail the problem addressed, the method overview, and the main stages of the proposed system.

3.3.1 Problem Description

We define the gallery as a set of classes, $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_N\}$, where each class, $\mathcal{C}_i \in \mathcal{C}$, represents one person. Each class is represented by a set of at most m features $\mathcal{C}_i = \{f_i^1, \ldots, f_i^m\}$ with f_i^j the jth feature of the class, respectively. The features are extracted from sample images, named samples for simplicity, and comprise an appearance descriptor, x_i^j, obtained from a generic re-id neural network, and the skeleton joints visible in the sample, s_i^j, thus $f_i^j = (x_i^j, s_i^j)$. Specifically, we select the re-identification OsNet model (K.

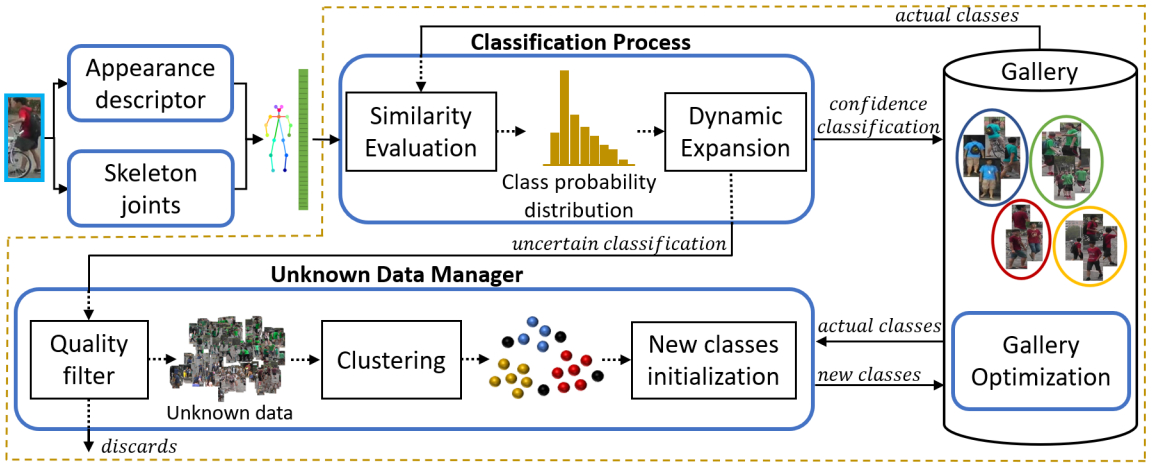


Figure 3.2: Overview of the self-adaptive gallery construction method. The person bounding box undergoes a pre-processing where the sample features are obtained with existing deep neural network encoders. Then, the proposed method analyzes the features obtained to decide which ones are used to adapt and evolve the gallery with the new information.

Zhou et al., 2021) to extract the appearance descriptors, and the OpenPose network (Cao et al., 2019) to obtain the skeleton joints.

The problem is to devise a method able to incrementally create the gallery of observed people from an empty initialization as new samples arrive in the system, considering an unknown (possibly unlimited) number of classes, N.

3.3.2 Method Overview

The overall idea of the proposed method is represented in Figure 3.2. First, whenever a new sample is acquired, the associated feature, f_q, is obtained. Then, the method performs a classification process by computing the class probability distribution of the new sample through a similarity evaluation. Based on the confidence of the classification, the system decides whether to conduct a dynamic expansion or not. Samples with high confidence enter the gallery, while samples with low confidence are sent to the unknown data manager for further analysis. The set of unknown data is periodically clustered to generate new potential classes that are compared with the existing ones to identify and initialize new classes. Finally, since there is a limit in the memory budget of m features per class, the gallery optimization handles the efficient use of memory resources by deciding the relevant data to keep.

3.3.3 Classification Process

Initialization Stage. In the initial phase of the gallery construction, the low number of classes initialized does not allow to work properly with probability distributions. Therefore, the proposed system runs a short initialization stage. In order to perform this initialization following the incremental setup, a set of candidate-classes, $\mathcal{B} = \{\mathcal{B}_1, \ldots, \mathcal{B}_k\}$, is defined, where the first candidate-class is created with the arrival of the first sample $\mathcal{B}_1 = \{f_1^1\}$. Then, the cosine similarity of incoming samples is computed between x_q and those appearance descriptors already included in $\mathcal{B}$. If the maximum cosine similarity is greater than

a threshold, ε, the sample is included in the corresponding candidate-class set; otherwise, a new candidate-class is initialized. As soon as a candidate-class reaches a minimum size of l, it becomes a person-class, i.e., a real class, belonging to the gallery $\mathcal{C} = \{\mathcal{C}_1\}$. Once the gallery reaches a minimum number of person-classes, Q, the proposed decision-making based on the class probabilistic distribution of the samples is run as detailed next.

General Regime. Once the gallery is initialized, the system *evaluates the similarity* of each new sample with the current gallery to obtain a probability distribution over the set of existing classes. This is accomplished using the softmax operator

$$p(x_q \in \mathcal{C}_i) \equiv p_i(x_q) = \frac{\exp(\bar{x}_i^\top x_q/\upsilon)}{\sum_{j=1}^{N} \exp(\bar{x}_j^\top x_q/\upsilon)}, \tag{3.1}$$

where υ is a temperature parameter that controls the softness of probability distribution over classes (Lin et al., 2019), x_q is the normalized appearance descriptor of the new sample, and $\bar{x}_i$ is the weighted centroid of $\mathcal{C}_i$. Working with normalized vectors, the product of both descriptors, $\bar{x}_i^\top x_q$, is equivalent to the cosine similarity between them. In our framework, the weighted centroid $\bar{x}_i$ is defined as

$$\bar{x}_i = \frac{\sum_{j=1}^{m} r_i^j x_i^j}{\sum_{j=1}^{m} r_i^j}, \tag{3.2}$$

being $r_i^j = s_i^j/s_T$ the ratio of visible joints in the person image bounding box with s_i^j the number of detected joints and s_T the total number of joints in a complete skeleton. By weighting the samples according to the number of joints, we favor the selection of samples with more body parts shown.

In a similar fashion to existing techniques for incremental learning (Fontanel et al., 2020; Rao et al., 2019), a threshold is used to control the *dynamic expansion* of the classes identified in the current gallery. More concretely, a simple and intuitive condition is used to measure the classification confidence of x_q through its class probability distribution,

$$\frac{\max_i p_i(x_q)}{\max_{j \neq i} p_j(x_q)} \geq \tau, \tag{3.3}$$

where τ is the expansion threshold. Samples whose probability distribution does not comply with the condition (3.3) are considered doubtful and go into the pool of unknown data. Conversely, if the confidence of the classification obtained with (3.1) is higher or equal than τ, the pseudo-label assigned to the sample corresponds to the class with maximum probability, $i^* = \arg\max_i p_i(x_q)$, and will be considered to be part of its representation model, $\mathcal{C}_{i^*}$.

3.3.4 Unknown Data Manager

Samples that do not satisfy the classification confidence criteria (3.3) are defined as unknown. The role of the Unknown Data Manager is to identify new identities as well as to recover samples that could not be previously classified with enough certainty. To avoid the initialization of new classes with sets of poorly-explained features, i.e., images showing

only one arm or one leg, all the unknown samples first undergo a quality filter to ensure that the appearance descriptors represent at least half of a person, formally $r \geq 0.5$, being r the ratio of joints.

The identification of new classes is tackled through the periodic *clustering* of the unknown data. In open-world scenarios, the number of classes is unbounded, making the use of clustering methods such as K-Means unfeasible. Thus, to partition the set of unknown data, we use a DBSCAN algorithm (Ester et al., 1996) based on sample density and able to deal with noisy information. The resulting clusters that reach the minimum size l, are compared with the current classes in the gallery to check whether they belong to an existing class or represent a new one. Following the analysis performed in Lin et al., 2019 on criteria methods to decide which pair of clusters to merge, the minimum distance criterion is used to verify if a potential new class, $\mathcal{C}_w$, shares identity with any of the existing in the gallery. The minimum distance criterion takes the shortest distance between samples from the new cluster, $\mathcal{C}_w$, and all elements of the gallery, $\mathcal{C}$,

$$D(\mathcal{C}_w, \mathcal{C}) = \min_{\mathcal{C}_i \in \mathcal{C}} \left(\min_{x_j \in \mathcal{C}_i, x \in \mathcal{C}_w} \left(1 - x^\top x_j\right) \right). \tag{3.4}$$

Since the computational cost of this process is considerably high, we compute an approximation limiting the number of existing classes that are compared with $\mathcal{C}_w$ from the set N to a subset of k. To select which classes are analyzed, for each $x \in \mathcal{C}_w$, we compute the k-Nearest centroids of the gallery and then select the k most frequent classes among all of them. Using only these classes in the first minimum of (3.4), the computational cost remains constant with the size of the gallery.

Finally, if the approximated minimum distance is higher than α, the cluster $\mathcal{C}_w$ is *initialized in the gallery as a new class*. Otherwise, the new cluster and the class with the closest sample represent the same identity and are merged, complying with the memory budget by means of the gallery optimization process.

3.3.5 Gallery Optimization

Our approach performs an intelligent decision-making process with the goal of storing the most representative features of each existing class and making efficient use of memory resources. In order to address this goal, we use two metrics that describe the relationship of each appearance descriptor with those in the same class and with all the rest.

The first metric is the *intra-class diversity* of the samples. For a descriptor, x, that belongs to class $\mathcal{C}_i$, we define its diversity through the minimum cosine distance among all the other descriptors that belong to the same class,

$$D_i(x) = \min_{x_j \in \mathcal{C}_i \setminus x} \left(1 - x^T x_j\right). \tag{3.5}$$

Then, the diversity of the whole class is defined as the minimum diversity among all its features,

$$D(\mathcal{C}_i) = \min_{x_j, x_k \in \mathcal{C}_i, x_j \neq x_k} \left(1 - x_j^T x_k\right). \tag{3.6}$$

This metric aids to identify redundant information, i.e., similar samples within a class. Leveraging this information, when a new sample is classified and assigned to an existing

class of the gallery, $\mathcal{C}_i$, it is only added to the representation model if its diversity is greater than the current diversity of the class,

$$D_i(x_q) \geq D(\mathcal{C}_i). \tag{3.7}$$

The second metric is the *uncertainty* of the sample with respect to the whole gallery, which is measured through Shannon's entropy by

$$H(x) = -\sum_{i=1}^{N} p_i(x) \log\left(p_i(x)\right), \tag{3.8}$$

where N is the number of classes at the moment in the gallery, and $p_i(x)$ is the probability described in (3.1). High entropy values stand for appearance descriptors that can be easily confused with those of other classes. In contrast, a feature with low entropy indicates high confidence in belonging to a certain class. Therefore, this metric provides an intuition of the relative distance between the feature and the rest of the classes of the gallery (inter-class).

The dependency on all the classes in (3.8), together with the constant evolution of the class centroids required for (3.1), makes the computation of this metric very heavy. For efficient computation, we keep a matrix, $\mathbf{R}_i$, for each class, i, with the cosine similarity between its samples, x_i^j, and all the weighted centroids of the gallery,

$$\mathbf{R}_i = \begin{bmatrix} \bar{x}_1^\top x_i^1 & \bar{x}_2^\top x_i^1 & \cdots & \bar{x}_N^\top x_i^1 \\ \bar{x}_1^\top x_i^2 & \bar{x}_2^\top x_i^2 & \cdots & \bar{x}_N^\top x_i^2 \\ \vdots & \vdots & \ddots & \vdots \\ \bar{x}_1^\top x_i^m & \bar{x}_2^\top x_i^m & \cdots & \bar{x}_N^\top x_i^m \end{bmatrix}, \tag{3.9}$$

as well as a list of the classes that have changed since the last gallery optimization of $\mathcal{C}_i$. This list is used to update only the columns associated with classes with changes, noting that the other distances have not changed and can be reused. Note that the $\mathbf{R}_i$ matrix is the changing element of (3.1) since υ is a constant value. Once we compute the update of the probability distribution of the samples belonging to $\mathcal{C}_i$, obtaining entropy with (3.8) is straightforward.

When the memory budget of a class is exceeded, because of a merge caused by the Unknown Data Manager or the insertion of a new sample, an optimization process using both metrics decides which sample to drop. In particular, the sample to drop is

$$x^* = \arg\max_{x \in \mathcal{C}_i} \left(\gamma \frac{H\left(x\right)}{\log\left(1/N\right)} - (1-\gamma) D_i(x) \right), \tag{3.10}$$

where $\gamma \in [0, 1]$ is a parameter to weigh the relevance of the uncertainty and the diversity terms, and the logarithm, $\log\left(1/N\right)$, normalizes the entropy to a value between zero and one, equivalent to the diversity. The proposed optimization function seeks a balance between how much a feature mixes the different classes (entropy) and how distinctive it is from the rest of the features within the same class (diversity). Figure 3.3 shows a simplified optimization process with only two clusters, $\mathcal{C}_1$ and $\mathcal{C}_2$, where $\mathcal{C}_1$ has exceeded its size constraint $m = 3$. Additionally, we include two examples of the final appearance models obtained with the proposed process to illustrate the balance achieved between uncertainty and diversity even though the two identities look very similar.

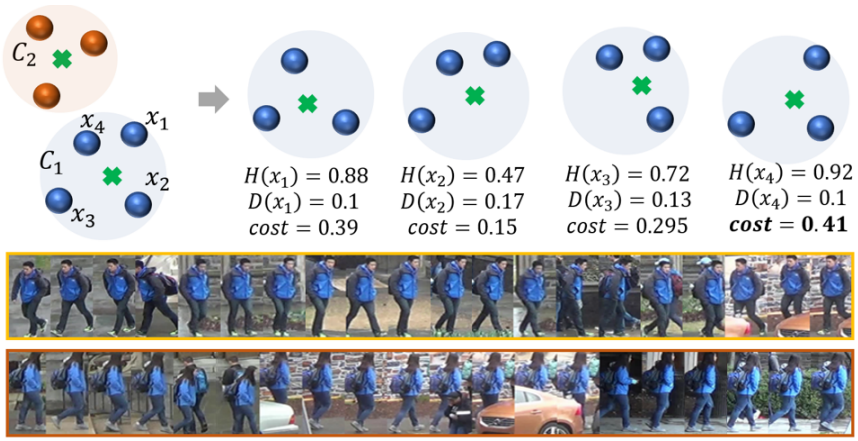


Figure 3.3: Gallery optimization. Upper area: example simplified where C_1 exceeds the memory budget and the gallery optimization selects the feature with the maximum cost to be dropped, x_4. Lower area: visual sample of two appearance models from similar identities that are correctly separated in the *DukeMTMC-VideoReID* dataset. The yellow edge corresponds to identity 86 and the orange edge to identity 194, both ground truth identities.

3.4 Experiments

This section analyzes the influence of the main parameters defined in the system, the algorithm selected to model the person's appearance and compares the performance of the proposed framework with other unsupervised and semi-supervised re-id approaches.

3.4.1 Experimental Setup

Datasets. The evaluation is performed with two challenging public benchmarks, MARS (L. Zheng, Bie, et al., 2016) and DukeMTMC-VideoReID (Y. Wu et al., 2018). In both of them, we use the official test set split into the *query set* and the *gallery set*.

Experimental Settings and Metrics. Two experiments are performed in this section. First, several analysis of the *gallery construction* process assesses the key aspects of our approach. The second experiment, *query re-identification*, runs a conventional evaluation for re-id methods in order to compare the proposed framework with other unsupervised and semi-supervised approaches. The details and metrics employed in each experiment are described next.

1. Gallery Construction. The *gallery set* from both datasets is used to evaluate the self-adaptive gallery construction process. As in traditional incremental settings, the tracklets are randomly shuffled, and then, the images from each tracklet are provided one by one to simulate an incremental input to the self-adaptive gallery. In order to evaluate the global performance of the proposed approach, we consider the following three metrics based on the classic precision, recall, and $F1$ score:

   - **Gallery Structure**: The perfect gallery structure has one (and only one) class per ground truth identity (GT-ID). This GT-ID is set for each class with the mode of all the sample identities present at the class initialization. In order to evaluate the

quality of the final gallery structure, we compute the precision (P), recall (R), and F1 score metrics as

$$P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN} \quad \text{and} \quad F1 = \frac{2 \cdot (P \cdot R)}{P + R}, \tag{3.11}$$

where we define the false negatives (FN) as those GT-ID not associated with any class, i.e., identities not found, the true positives (TP) as all GT-IDs associated with at least one class, i.e., identities found, and the false positives (FP) as the additional classes with the same GT-ID associated, i.e, two classes associated to the same GT-ID count as one FP and one TP.

- **Class Precision**: This metric assesses the precision of the samples that enter the gallery over time. The true positives (TP) are the samples whose identity matches the GT-ID of the class they have been assigned, and the false positives (FP) are the samples that do not.
- **Sample Classification F1**: This metric evaluates the pseudo-label assigned to every sample that arrives to the system. Considering that the gallery structure often has redundancy due to the unsupervised nature of the method, we deem a limited number of redundant classes for each identity. In particular, for a given GT-ID, we only consider the K classes with the highest number of samples associated with them, discarding the rest. Hence, the true positives (TP) are the samples that match the GT-ID with the assigned class. The false positives (FP) are the samples with mismatching GT-IDs, and the false negatives (FN) include samples classified as unknown or assigned to the discarded classes.

2. Query Re-Identification. In order to compare the proposed framework with other unsupervised and semi-supervised approaches, we use the *query set* to evaluate the gallery obtained at the end of the *gallery construction* process. Thus, the *query set* is matched with the limited size gallery created in the previous experiment, which remains static during this evaluation. The conventional evaluation for re-identification (K. Zhou et al., 2021) is performed including the **Rank-1** and **Rank-5** metrics.

Implementation details. For both experiments, the settings for our approach configuration are: similarity threshold in the initialization stage $\varepsilon = 0.9$, temperature parameter in the softness operator $\upsilon = 0.1$, the k-Nearest centroids with $k = 3$ used by the Unknown Data Manager, distance threshold to initialize a new cluster $\alpha = 0.1$, gallery size to run the probabilistic decision making $Q = 20$, the re-identification network used in cross-domain is an OsNet model (K. Zhou et al., 2021) trained with the MSMT17 Benchmark (Wei et al., 2018), and the OpenPose network (Cao et al., 2019) is used to obtain the skeleton joints.

3.4.2 Gallery Construction. Parameter Evaluation

The first study of this chapter analyzes the effect of the three key parameters on the gallery construction process: (1) the weight used in Equation (3.10) to balance the influence of the uncertainty and the diversity, γ, (2) the expansion threshold, τ, in Equation (3.3), and (3) the minimum size required to initialize a class, l, along with the memory budget per identity, m, defined in Section 3.3.1. The goal of this parameter evaluation is to choose the

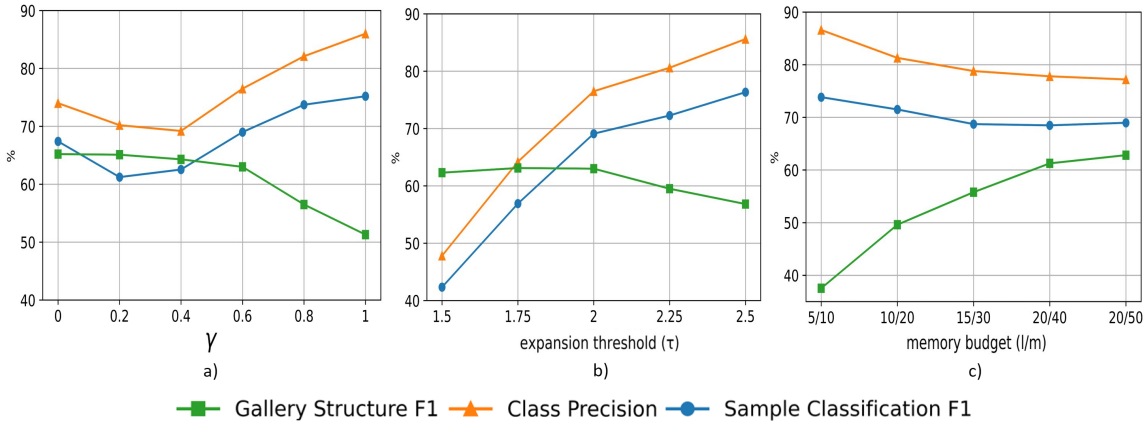


Figure 3.4: Parameter evaluation in the gallery construction process using the MARS dataset: a) effect of the weight assigned to uncertainty and diversity (γ); b) influence of the expansion threshold (τ); c) effect of the minimum size to create a class and the memory budget (l/m).

parameters that yield balanced galleries based on the defined metrics. In this study, we set $K = 4$ for the *sample classification F1*.

The results of the performed analysis are shown in Figure 3.4. The influence of each parameter at the end of the process is examined in Figure 3.4a–c, revealing that the trend of the quality *gallery structure F1* is inverse to the tendencies of *class precision* and *sample classification F1*. Figure 3.4a) shows the effect of **weighting the uncertainty and diversity** with γ, fixing all the other parameters to $\tau = 2$, $l = 20$ and $m = 50$. The increase in γ favors the selection of samples with low entropy but less diverse ones in the appearance models. The balance between uncertainty and diversity in the gallery is attained at $\gamma = 0.6$. Then, the **expansion threshold**, τ, is analyzed in Figure 3.4b). We keep $l = 20$, $m = 50$, and from the former analysis, γ is set to 0.6. When τ increases, more samples are sent to the Unknown Data Manager, resulting in the initialization of more classes. The trade-off between the metrics analyzed is accomplished at $\tau = 2$. Finally, the influence of the minimum size to create a class, l, and the **memory budget** per identity, m, is evaluated in Figure 3.4c). The rest of the parameters are set to $\gamma = 0.6$, $\tau = 2$. The increase in the *gallery structure F1* is caused by the reduction in the initialization, leading to fewer redundant classes. This implies greater confidence in the classification of the samples as m increases. Therefore, the selected memory budget configuration is the one that generates the highest *gallery structure F1*, $l = 20$ and $m = 50$, with an influence not highly significant in the other metrics analyzed.

To verify the stability of the obtained results and the construction process, Figure 3.5 presents the evolution of the metrics over time with the final parameters set, $\gamma = 0.6$, $\tau = 2$, $l = 20$, and $m = 50$. Since it is an evaluation over time, in this particular case we consider $K = \infty$ for the *sample classification F1*. All the metrics settle after processing 20% of the samples. Then, it can be fairly assumed that the method's behavior is stable beyond that stage.

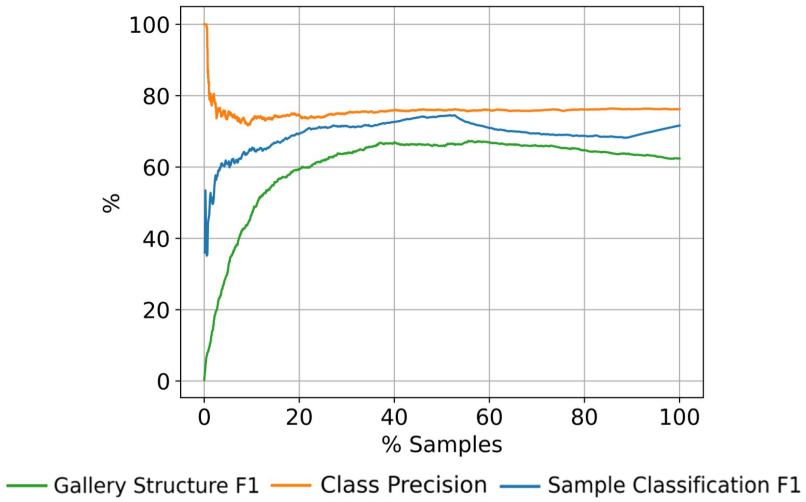


Figure 3.5: Evolution over time of the metrics with the final parameters set in the gallery construction process using the MARS dataset .

3.4.3 Gallery Construction. Data Selection Method Comparison

Following the analysis from the previous section, this experiment sets $\gamma = 0.6$, $\tau = 2$, $l = 20$, and $m = 50$. This study evaluates different gallery optimization processes for deciding which sample to remove from the appearance model when the memory budget is exceeded. We compare existing algorithms commonly used in incremental clustering works that have to deal with memory budget requirements. These algorithms are assessed at the end of the gallery construction process. The first method is uniform sampling (Uniform) which saves a new feature for every $U = 5$ instance. When the size limit is exceeded, the oldest data is dropped to save a newer one. Another typical process is random decision-making (Random) which removes a random index when the memory reaches its budget. Regarding more sophisticated methods, we compare the two closest approaches in the literature, the method proposed in Azagra et al., 2020, called Incremental Object Model (IOM), and the ExStream method present in Hayes et al., 2019. In both cases, we use the implementation provided by the authors to evaluate the effect of the data dropped in the gallery in our overall method. Moreover, due to the influence on the final results of the data arrival order in incremental setups, three different iterations are run (i.e., three different random data arrival orders). To make a fair comparison, all five methods use the same features extracted from OsNet (K. Zhou et al., 2021).

First, a comprehensive analysis of the final quality of the *gallery structure* is performed. The number of classes created per GT-ID and the *gallery structure* metrics are shown in Figure 3.6a) and Figure 3.6b), respectively. The results in Figure 3.6a) indicate that the ExStream and the Uniform algorithms create a high number of redundant classes in the gallery. This means that the appearance models resulting from these methods are significantly less representative, leading to more uncertain classifications. Thus, they send a high number of samples to the unknown pool and create new classes for already existing identities. The proposed optimization process (Ours) creates only one class for the same number of GT-IDs as IOM while identifying more people in the scene, represented by a smaller number of GT-IDs with 0 classes created. Then, derived from this analysis and

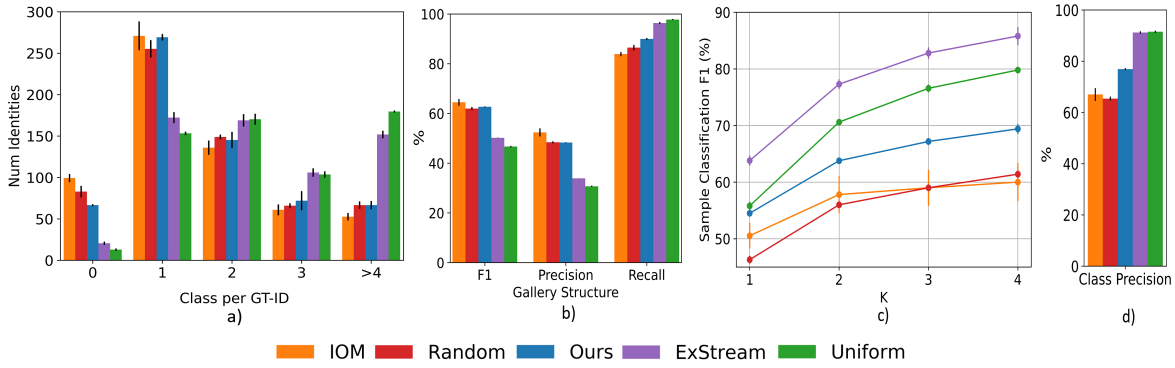


Figure 3.6: Data selection method comparison with the MARS dataset. We analyze a) the number of classes created (x-axis) per GT-ID in the dataset (y-axis) showing the number of GT-ID with more than one class associated or those GT-ID that have not been correctly found, i.e., 0 classes associated; b) *gallery structure* metrics: $F1$, precision, and recall; c) *sample classification* $F1$ analyzing the influence of varying K; d) *class precision.*

verified in the $F1$ results on Figure 3.6b), the methods which provide a gallery structure of better quality are IOM, Random, and Ours, being Ours the one that identifies the most people in the scene among them, as measured with the *gallery structure recall.*

Second, Figure 3.6c) present the analysis of varying K in the *sample classification* $F1$, and Figure 3.6d) shows the *class precision* results. As expected, the *sample classification* $F1$ improves in all algorithms with the increment of K. Comparing the methods that generate a gallery with a suitable structure, i.e., IOM, Random, and Ours, the results shown in Figure 3.6c) and Figure 3.6d) demonstrate that the proposed gallery optimization process (Ours) outperforms IOM and Random in both metrics. Our approach is able to create more reliable people models without losing diversity, thus enhancing the classification of the samples. The ExStream and Uniform methods obtain high scores in these metrics because of a large number of redundant classes, limiting their practical effectiveness in the actual ability to re-identify known people.

In summary, our algorithm achieves the best balance between building a well-structured gallery and achieving reliable classification metrics of the individual samples. While ExStream and Uniform methods generate galleries with worse quality structure, IOM and Random approaches obtain lower *class precision* and *sample classification F1* results.

3.4.4 Gallery Construction. Final Results

A detailed evaluation on MARS and DukeMTMC-VideoReID is provided using the same parameter values from the previous section for both benchmarks.

Table 3.1 shows the final results of the complete self-adaptive gallery construction framework on both datasets. In the *gallery structure* analysis, the table includes the number of GT-IDs, classes created and the gallery structure $F1$, the precision, and the recall. The larger number of people in DukeMTMC-VideoReID makes it more challenging to identify most of them, causing lower recall metrics than in the MARS dataset, i.e., the 80.06% of the people have been correctly identified in DukeMTMC-VideoReID against the 89.43% in MARS (*gallery structure recall*). In terms of *class precision*, note that the proposed framework obtains similar and consistent results for both datasets, 76.69% in MARS and 80.1%

Figure 3.7: Visualization of the evolution of appearance models in the gallery. Each row corresponds to gallery samples at a certain time. The different colors represent the time stamp of the samples included in the gallery (best viewed in color).

Metrics	Dataset MARS	Dataset DukeMTMC-VideoReID
Gallery Structure		
Total IDs (GT)	620	1110
Classes Created	1147.6 (±2.5)	1337.33 (±16)
F1	62.67 (±0.19)	72.62 (±0.26)
Precision	48.24 (±0.12)	66.45 (±0.51)
Recall	89.43 (±0.4)	80.06 (±0.47)
Class Precision	76.9 (±0.36)	80.1 (±0.60)
Sample Classification		
F1	69.4 (±0.86)	62.6 (±0.87)
Precision	72.23 (±0.12)	72.43 (±1.12)
Recall	66.8 (±1.69)	55.21 (±0.69)

Table 3.1: Detailed results of the proposed framework on MARS and DukeMTMC-VideoReID datasets. The results show the mean and the standard deviation of three performed iterations, mean (±std).

in DukeMTMC-VideoReID. Thus, the method creates robust appearance models, being able to correctly distinguish the people in the scene, which in turn helps in the *sample classification* obtaining precision results of 72%.

Finally, Figure 3.7 includes samples of the gallery for one identity per dataset at three different times during their construction, showing in each row the person model at different times. The left identity includes an example of corruption that the gallery can suffer remarked by a discontinuous red line. In both cases, the third row shows how our resulting gallery presents a high variability of samples, resulting in a representative model for each identity[1].

3.4.5 Query Re-Identification

This final experiment performs the traditional evaluation of person re-id, i.e., obtains the expectation that the true match is found within the first R ranks (Hirzer et al., 2011).

[1]More qualitative results in https://youtu.be/m6sD0GZxZfw

Method	Setting	DukeMTMC-VideoReID			MARS		
		GS (%)	Rank-1	Rank-5	GS (%)	Rank-1	Rank-5
Full-gallery	Cross-Domain	100	63.2	72	100	66.4	73.3
EUG	OneEx	100	72.7	84.1	100	62.67	74.94
SCLU	OneEx	100	72.7	85	100	63.74	78.44
BUC	Unsp. (None)	100	76.2	88.3	100	57.9	72.3
Softened	Unsp. (None)	100	76.4	88.7	100	62.7	77.2
GLC+	Unsp. (None)	100	80.9	91.5	100	66.5	78.7
Ours	IUCD	18.4	59.5 (±1.2)	69.1 (±1.04)	8.1	60.1 (±0.78)	69.8 (±0.38)

Table 3.2: Comparison with re-id approaches in DukeMTMC-VideoReID and MARS *query set*.

However, instead of matching the *query set* with a completely labeled gallery, the *query set* is matched with the resulting gallery from the *gallery construction* process. In this experiment, the gallery remains static. The proposed method obtains its results in an incremental unsupervised cross-domain setting (IUCD). Table 3.2 shows the results of this experiment, including the setting in which the different methods operate. Our offline baseline is the Full-gallery method, which has the whole gallery available and manually labeled using the same descriptors as our approach. This method is our upper bound result in the cross-domain setting. Moreover, due to the unsupervised component of our approach, we present the results of unsupervised and semi-supervised systems that perform offline training in the same domain as the *query set*. The unsupervised methods that included BUC (Lin et al., 2019), Softened (Lin et al., 2020) and GLC+ (H. Chen et al., 2022) do not use any labeled data in the whole process (None). Concerning the semi-supervised approaches, they use one tracklet labeled per identity (OneEx). Note that we are the only algorithm working on the incremental unsupervised cross-domain (IUCD) setting, while the rest perform the entire process offline. Thus, although Table 3.2 is not a fair or direct comparison for our approach, we believe that it is interesting to see how close the proposed approach results are with respect to existing methods, despite the much more challenging and realistic scenario of our approach. The resulting values for our approach are the average and the standard deviation for the three random iterations performed previously, i.e., mean (±std). Besides, since the proposed gallery deals with memory requirements, the percentage of the gallery size used with respect to the total (GS) is shown. In this case, the standard deviation is not included, but we remark that it is lower than 0.01 in all cases.

The DukeMTMC-VideoREID results show the impact of the different goals sought. In our case, the correct identification of the 1110 people that compose the gallery is a really challenging task, where some of the queries analyzed in this evaluation do not have corresponding models in the gallery. In contrast, the methods that focus on improving the feature representation obtain better results than in the MARS dataset due to the lack of distractors in the gallery. Regarding the MARS dataset, which is closer to an open-world scenario, the results with our approach are close to the unsupervised or semi-supervised approaches using two orders of magnitude less in the amount of data stored in the gallery. Finally, considering the difference between the Full-gallery baseline and our approach, we see how the proposed approach achieves comparable performance despite a much smaller (one or two orders of magnitude less) and unsupervised built gallery.

3.5 Conclusions

This Chapter has addressed the challenge of developing methods able to adapt to temporal variations while maintaining a balance between available resources and the volume of information processed. More concretely, the presented research introduces a novel framework designed to tackle critical issues of person re-id within open-world scenarios. Our approach focuses on automatically detecting new identities and updating the appearance model of already known people for the system through a self-adaptive gallery. This is accomplished while efficiently managing limited memory resources and removing the need for manual intervention. Therefore, we present a fully unsupervised self-adaptive gallery for person re-identification that constructs efficient appearance models of the individuals saving only the most representative samples. In the short-term person re-identification problem, our framework can identify more than 80% of the people presented in the challenging scenarios evaluated by comparing the new unlabeled data and the existing classes in the gallery. The existing classes in the gallery are modeled with an optimization process that selects the most representative information to represent each class, balancing the uncertainty (inter-class) and the diversity (intra-class) of the samples. Experimental results demonstrate that the proposed optimization process returns a *class precision* of about 80% while encouraging the variability inside the classes, thus generating well-balanced and better-structured galleries than those resulting from similar existing methods. The high *class precision* maintained over time aids the continuous person re-id by obtaining an *F1 sample classification* of 62.6% and 69.4% in the Mars and DukeMTMC-VideoReID datasets, respectively. Furthermore, compared to existing re-id algorithms, our method obtains similar results to the fully labeled galleries with one or two orders of magnitude less data.

Chapter 4

Collaborative hybrid surveillance systems in a photo-realistic environment

Previous chapters addressed the development of methods for static RGB multi-camera systems, commonly used in real-world surveillance and monitoring applications for their affordability and low maintenance. However, their fixed viewpoint and limited coverage can hinder performance in open and large scenarios with blind spots. To overcome these limitations, our research explores the benefits of removing positional constraints on a few cameras mounted on drones that can actively reposition themselves to gather optimal information at any given time.

In this Chapter, we introduce a hybrid camera system that combines static and mobile cameras to collaboratively observe the environment and efficiently capture specific attributes from each person. The proposed solution integrates a multi-camera distributed tracking system for precise localization of all individuals with a control scheme that moves the mobile cameras to the best viewpoint for a specific classification task. Our main contribution is the novel framework that leverages the synergies resulting from the cooperation of the tracking and control modules, obtaining a system closer to the real-world application and capable of high-level scene understanding. Evaluating the proposed framework with freely moving people while controlling drones is nearly unfeasible in real-world scenarios due to security constraints. Thus, we develop the required components to easily generate scenarios with pedestrians on the photo-realistic simulator Unreal Engine and AirSim. The simulated environment is able to create random and customized trajectories for each person and provides ready-to-use animated people models along with an API for their metadata retrieval. Finally, we prove the usefulness of the developed environment and leverage it to evaluate the proposed hybrid framework. Our experiments demonstrate the advantages of using collaborative cameras for surveillance applications compared to static camera setups.

4.1 Introduction

The use of multiple cameras increases the coverage and the amount of information collected from large-scale scenes. Although the most frequent configuration in surveillance applica-

tions is a network of static cameras, including mobile cameras brings plenty of potential benefits. In addition to the improved coverage capabilities of such a hybrid system, mobile cameras can be guided to acquire more detailed information and particular viewpoints when needed. Enhancing collaborative behavior among them is then essential to achieve an efficient mutual scene understanding (X. Li et al., 2018; Mekonnen et al., 2013; Miller et al., 2022).

One of the main challenges of collaborative camera network systems is to attain robustness and efficiency. Hence, there has been a tendency to transition from centralized to distributed setups that can easily scale and are more robust against individual node failures (Yu et al., 2022; Y. Zhou et al., 2022). Another common challenge in multi-camera systems is finding a suitable viewpoint that maximizes gaining new knowledge for a given recognition task. For instance, solving tasks such as person identification or clothing brand recognition requires a specific viewpoint. Active perception enables the capability of moving a camera to the location of the most informative perspective. Developing and evaluating distributed solutions, where mobile cameras with autonomous decision-making are involved, is not a trivial task.

To address all of these challenges, we propose a novel active and distributed framework. Our *collaborative hybrid system for surveillance* has static cameras to monitor the scene and mobile cameras to strengthen the visualization of certain attributes with high-resolution close-up target images, as summarized in Figure 4.1. The mobile cameras, drones in our case, are guided by a control policy built upon Serra-Gómez et al., 2023. This policy continuously determines the cameras' next position and orientation to capture viewpoints that maximize the acquisition of relevant information for certain people's attributes class. Differently from this prior work, here multiple drones are considered working together with a network of static cameras that provide information about the targets' position and orientation using real data, taking into account the challenges associated with the use of a real tracking system. The distributed tracking process in charge of this task is based on (Casao et al., 2021) detailed in Chapter 2. In this chapter, the implementation of this module is improved making the transition to a real system easier thanks to the integration with ROS to handle communications.

The assessment of the proposed framework is performed with a photo-realistic simulator since testing autonomous drones in pedestrian environments is almost impractical in real-world scenarios. In particular, the presented research uses the open-source Unreal Engine together with the AirSim simulator (Shah et al., 2018), both of which provide a photo-realistic environment to simulate drones and static camera data generation. Unfortunately, the generation of complex pedestrian scenarios can be a time-consuming solution with a steep learning curve, in addition to the tedious work of collecting and blending the suitable actors (avatars, movements, textures...). Therefore, we develop the core *tools for fast prototyping of environments with multiple pedestrians* and easily generate realistic dynamic scenes[1]. First, we implement a trajectory plugin for a user-friendly definition of paths directly on the environment map. The plugin offers the capability to traverse the created paths in a random or customized fashion. Then, a compilation of animated pedestrian models, created using open-source tools, is ready to be used by simply dragging and

[1]Simulated data and photo-realistic environment used available at https://sites.google.com/unizar.es/poc-team/research/hlunderstanding/collaborativecameras.

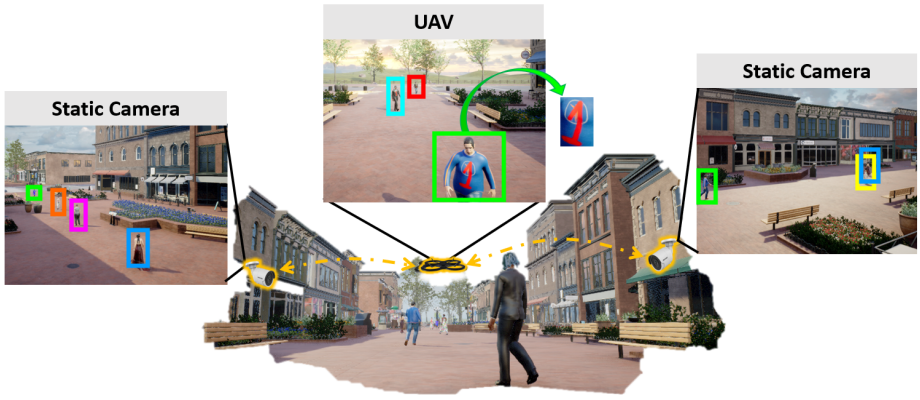


Figure 4.1: Overview of our multi-camera collaborative system. The system comprises a camera network that performs a distributed multi-target tracking process. The static cameras monitor the scene and the mobile cameras are guided by a control policy to capture close-up images of viewpoints likely to strengthen the classification of certain attributes.

dropping them on the map. Finally, a Python API obtains all the environment metadata from AirSim, providing automatic annotations of all the pedestrians present in the scene.

The experiments conducted in this chapter validate the suitability of using photo-realistic environments to evaluate multi-target tracking methods. Then, the performed studies focus on analyzing the performance of our hybrid multi-camera system under different conditions, demonstrating the benefits of a collaborative setup over static or individual cameras.

4.2 Related Work

Active Perception for Class Recognition. The active perception problem of recognizing certain classes is commonly addressed by defining a set of viewpoints in advance, which are then used to plan trajectories for gathering new information. One-step greedy planners select specific viewpoints to objects based on factors such as class uncertainty and observation occlusions (Patten et al., 2016). Instead, non-myopic methods such as Popović et al., 2017 consider both, movement costs and information gained between the object's viewpoints. Alternatively, some approaches formulate the problem as a partially observable Markov Decision Process (POMDP) and design paths over viewpoints by accounting for costs associated with measurements, occlusions, and potential misclassifications (Atanasov et al., 2014). Likewise, Patten et al., 2018 employs a modified version of Monte-Carlo tree search to generate plans. However, these techniques typically rely on a priori access to the black-box model for estimating the usefulness of viewpoints. More recent works use non-myopic learning methods like Deep Reinforcement Learning (DRL) for static multi-target pose estimation and active perception. They optimize camera movements to reduce observation uncertainty (Sock et al., 2020) or maximize information gain (Q. Xu et al., 2021). Nevertheless, these approaches either assume static targets, are limited to closed environments (Kent & Chernova, 2020), or require prior knowledge of where the information is visible from (Alcántara et al., 2021; Jeon et al., 2020). Our work leverages an attention-based neural network architecture to encode dynamic targets and to provide viewpoint

recommendations that are traced with a low-level controller. In addition, we enable the use of multiple drones and overcome the assumption of possessing prior knowledge about the positions and orientations of the targets by exploiting the collaboration with static cameras responsible for the global understanding of the scene.

Collaborative Systems for Perception Tasks. Multiple works have developed collaborative systems to address complex perception tasks. One of the most common problems tackled is active object tracking, where visual observations are transformed into a camera control signal to improve the tracking process, e.g., turning left or moving forward (Schranz and Andre, 2018). The combination of a fixed camera, that globally monitors the scene, with a pan-tilt-zoom (PTZ) camera, used to increase the image quality of the target of interest, is proposed in X. Li et al., 2018. In J. Li et al., 2020, this setup is extended to a centralized PTZ camera network, where reinforcement learning techniques are employed to learn the new pose of the cameras for finding the target and tracking it as long as possible. In order to follow an object capable of moving in all directions, Trujillo et al., 2019 develop a cooperative aerial robotic approach with two drones for achieving overlapping images and forming a pseudo-stereo vision system. The collaboration of hybrid systems has been studied for different tasks such as dynamic obstacle avoidance, where the information of the static cameras is leveraged by the mobile robot (Mekonnen et al., 2013), or the localization, planning, and navigation of ground robots using a semantic map created by a high-altitude quadrotor (Miller et al., 2022). Furthermore, some works have focused on distributed collaborative perception tasks. Yu et al., 2022 propose an approach for distributed learning where each robot only shares the weights of the network for privacy protection and Y. Zhou et al., 2022 present a general-purpose graph neural network for fusing node information and obtaining accurate perception tasks. Closer to our work, Bisagno et al., 2018 leverage the collaboration of fixed cameras, PTZ, and UAVs for crowd scene covering in a distributed manner. Different from them (Bisagno et al., 2018), we do not assume as known the target positions, which entail addressing the challenges of a distributed multi-target tracking system.

Multi-camera Multi-target Tracking. Multi-camera centralized setups are commonly used in real-world applications to cover larger areas (Guo et al., 2022; Quach et al., 2021) or acquire a greater amount of information (Byeon et al., 2018; R. Zhang et al., 2020). These centralized approaches process the entire camera network information in one unique node, making it difficult to scale up. Thus, there is a trend toward distributed setups to increase the applicability of multi-camera systems (Xompero & Cavallaro, 2022). While theoretical works have proposed solutions to problems such as consensus algorithms to unify local estimations (Z. Li et al., 2023; Soto et al., 2009), only a few works have addressed the distributed multi-target tracking with real data. For example, Kamal et al., 2015 combine the Information-weighted Consensus Filter (ICF) with the Joint Probabilistic Data Association Filter (JPDAF), which uses the previous target states, to fill the gap of relating measurements and trackers in the consensus algorithm. Based on the same ICF consensus method, He et al., 2019 address the association of measurements and trackers through a global metric that merges appearance and geometry cues. To associate trackers across cameras, they employ the Euclidean distance between the 3D position of the targets.

Different from Kamal et al., 2015 and He et al., 2019, this Chapter tackles the problem of having mobile nodes in the camera network. To this end, we implement the proposed framework within a photo-realistic environment, enabling the integration of both visual information processing and mobile camera control.

Photo-realistic Simulators. The use of photo-realistic simulators has become extremely popular in a wide variety of problems, ranging from navigation (Vorbach et al., 2021) to cinematography (Pueyo et al., 2022). Some studies have leveraged the potential of these simulators to address end-to-end active tracking by training navigation policies based on reinforcement learning with RGB image as the only input (W. Luo et al., 2019; Tallamraju et al., 2019). Regarding the development of methods focused on people within simulated environments, most of the works benefit from these photo-realistic environments to create benchmarks that present new diverse scenes for multi-target tracking algorithms (Fabbri et al., 2018; Kerim et al., 2021). Moreover, several computer vision applications have tapped into the effortlessness of getting labeled data through simulators, with various works proposing domain adaptation algorithms from synthetic to real-world data for action recognition (da Costa et al., 2022) and re-identification (T. Zhang et al., 2021). These approaches often rely on tools such as Blender and Unreal Engine to generate people models, introducing the required animations from Mixamo ("Adobe Systems Incorporated. Mixamo, 2022." n.d.). Previous works rarely release the simulator environment, with the exception of T. Zhang et al., 2021 and Fabbri et al., 2018. Unfortunately, Mixamo no longer allows the use of its software for any machine learning or artificial intelligent tasks[2] and *GTA V* lacks the functionality of working with drones within the simulation. Therefore, this research introduces multiple tools built on Unreal Engine and AirSim to easily create photo-realistic scenarios with moving pedestrians. The developed tools significantly reduce the workload.

4.3 Preliminaries

4.3.1 Problem Formulation.

This thesis addresses the distributed tracking and correct visualization of people's attributes in large-scale environments. We monitor an area populated by a set of I targets, $\{\mathcal{X}_i\}_{i=1}^{I}$, with a system of J cameras, $\{C_j\}_{j=1}^{J}$, where a subset of $Q < J$ cameras can translate and rotate, e.g. they are installed on drones. Each camera in the network captures an RGB image and a depth map to estimate the state of the targets locally by fusing its information with that received from its neighbors, $\mathcal{N}_j$. The state of target i in camera j is defined as $\mathbf{x}_i^j = (x_i^j, y_i^j, z_i^j, w_i^j, h_i^j, \dot{x}_i^j, \dot{y}_i^j)$ represented by a 3D cylinder with (x_i^j, y_i^j, z_i^j) the 3D coordinates of the center cylinder's base, w_i^j the width, h_i^j the height, and $(\dot{x}_i^j, \dot{y}_i^j)$ the velocity of the target in the x and y directions, respectively. The orientation of the target, φ_i^j, is estimated based on their velocities $\dot{x}_i^j$ and $\dot{y}_i^j$. The responsibility for correctly visualizing the attribute's class of the targets lies in the moving cameras (drones). It is important to note that these attributes can only be observed from specific viewpoints, such as determining if the targets are wearing a backpack or glasses. The state of the drones

[2] https://www.adobe.com/legal/terms.html

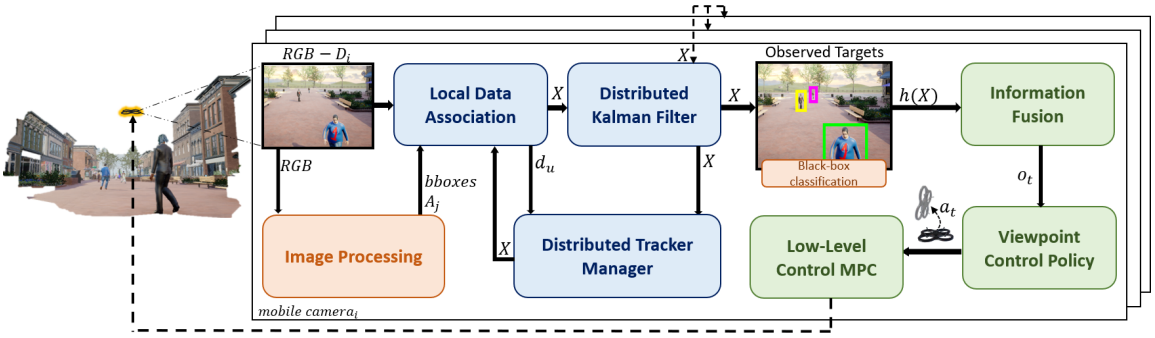


Figure 4.2: Method overview deployed in one mobile camera. The whole system is implemented in ROS, initializing each camera as a node and the image processing module as a service. First, the Local Data Association relates people detection (*bboxes*) with the corresponding trackers ($\mathcal{X}$). Then, the cameras exchange and fuse data with their neighboring cameras to obtain a collaborative distributed tracking system. The knowledge of the environment is provided to the control policy for obtaining a new recommendation of viewpoint (a_t) to improve the gathering people's information.

$\mathbf{y}_q = (\mathbf{u}_q, \psi_q)$, assumed as known, is represented as their position $\mathbf{u}_q$ and their heading ψ_q, being $q \in \{1, \ldots, Q\}$. Each drone is controlled by a hierarchical policy, where a viewpoint control policy operating at $\frac{1}{\tau_h}$ Hz takes as input the knowledge of the scene and outputs a viewpoint recommendation $\mathbf{a}^q$. Next, the recommended viewpoint is traced with a low-level controller operating at $\frac{1}{\tau_l} \gg \frac{1}{\tau_h}$ Hz. The purpose of the policy is to position the targets' attributes within the field of view (FOV) of the drone. We assume that the drones are faster than the targets and fly at a constant height above them, avoiding collisions.

The goal of the presented work is to achieve an accurate estimation of the targets' position and visualize all people's attributes as quickly as possible.

4.3.2 Overview

Figure 4.2 presents an overview of the proposed method to address the problem described in the previous section. The complete framework has been implemented in ROS, with each camera defined as a node of the system and ensuring synchronization between them. Neural networks have been implemented in the image processing module as services to save memory.

First, each camera captures an RGB image and a depth map (D_i) to compute the re-projection between the image plane and the real-world coordinates, different from the homographies used in Chapter 2 for obtaining the conversion between planes. We incorporate depth information to simplify the re-projection but this could be replaced by a network calibration in a more realistic setup. Then, a general detector provides the people bounding boxes (*bboxes*) that are used as measures for the tracking system and that are associated with the current trackers through the Local Data Association module (LDA). Once the cameras in the networks exchange the targets' information ($\mathcal{X}$) with their neighbors, the Distributed Kalman Filter (DKF) implemented attempts to obtain consensus on the targets' state. Finally, the Distributed Tracker Manager (DTM) initializes new trackers and associates them locally with the trackers received from the neighboring cameras. The mobile cameras of the system obtain the output of a black-box CNN, with perception

information about the visible targets ($h(\mathcal{X})$), and update their class beliefs with an efficient information fusion method. Based on the latter and the estimated state of the targets, the viewpoint control policy recommends a new camera pose (a_t) to maximize the information acquired in the next step. The new viewpoints are then tracked with a low-level controller.

4.4 Distributed Tracking

The distributed multi-target tracking module implemented in our hybrid system is based on the method detailed in Chapter 2. Therefore, this section only focuses on explaining the modifications required to integrate mobile cameras with dynamic communications in the multi-camera system. These modifications primarily involve the *Distributed Kalman Filter* (subsection 2.3.1) and the required information exchange between cameras to ensure temporal consistency in the association of trackers across cameras approached in the *Distributed Tracker Manager* (subsection 2.3.3).

For a more in-depth explanation of the distributed multi-target tracking algorithm, please refer to Chapter 2.

Distributed Kalman Filter. The proposed framework defines the target motion model as a discrete-linear dynamic system with constant velocity. Each camera executes a Kalman filter independently producing a local estimation of the target state, $\hat{\mathbf{x}}_i(k)$, and the associated error covariance matrix $\mathbf{P}_i(k)$. Note that local estimations may vary among different cameras. Then, the Distributed Kalman-Consensus filter (Soto et al., 2009) is implemented to mitigate these differences and seek to reach a consensus in $\hat{\mathbf{x}}_i(k)$ for all cameras C_j.

The consensus algorithm assumes knowledge of the data association between the local measurement $\mathbf{z}_i(k)$ and the target prediction $\bar{\mathbf{x}}_i(k)$, which is obtained by applying the linear motion model to the previous target state estimation $\hat{\mathbf{x}}_i(k-1)$. The measurement $\mathbf{z}_i(k)$ is the 3D cylinder obtained as the projection of the bounding box given by the detector and the velocity of the target computed with the last data association, i.e., $\mathbf{z}_i(k) = (x(k), y(k), z(k), w(k), h(k), \dot{x}(k), \dot{y}(k))$. This measurement is coupled in the filter with a zero mean Gaussian noise characterized with $\mathbf{R}_i(k)$ as its covariance matrix. Using mobile cameras requires online updates of the transformation matrix from the image plane to the three spatial global coordinates of the world. The cameras of the photo-realistic environment follow the pinhole model, which combined with the depth information, $d(k)$, enables the conversion of image plane coordinates $v_x(k)$ and $v_y(k)$, to the relative 3D world camera coordinates $x_r(k), y_r(k)$ and $z_r(k)$ by

$$x_r(k) = d(k), \;\; y_r(k) = \frac{d(k)}{f}(v_x(k) - c_x), \;\; z_r(k) = \frac{d(k)}{f}(v_y(k) - c_y) \tag{4.1}$$

where f is the focal length and, c_x and c_y are the image center coordinates in x and y, respectively. Then, the relative camera coordinates are transformed into the common global world system demand by the consensus-filter algorithm following

$$\begin{bmatrix} x(k) \\ y(k) \\ z(k) \\ 1 \end{bmatrix} = \begin{bmatrix} R_j(k) & | & T_j(k) \\ 0 & & 1 \end{bmatrix} \begin{bmatrix} x_r(k) \\ y_r(k) \\ z_r(k) \\ 1 \end{bmatrix} \tag{4.2}$$

being $R_j(k)$ the rotation matrix and $T_j(k)$ the translation vector of the camera at instant k, assumed as known. In a real setup, this information could be computed by offline calibration of the cameras and using onboard sensors such as GPS or IMUs together with SLAM algorithms for the drones. Regarding the velocity, we take advantage of the online tracking to measure the time the target has taken to arrive at the current position at k since the last data association between $\bar{\mathbf{x}}_i$ and $\mathbf{z}_i$.

Distributed Tracker Manager. In the practical implementation of distributed tracking systems, it is essential to obtain a correct association of trackers across the different network cameras. Our distributed tracker manager handles this task for the newly arriving trackers from neighboring cameras, relying on both geometric and appearance cues. Specifically, the algorithm computes the Euclidean distance between the targets' 3D positions in the environment normalized by a maximum distance threshold, τ_d, defined to identify potential association candidates. Additionally, the appearance constraint is assessed using the cosine distance between the trackers' appearance features. Finally, the optimal assignment problem is solved to determine the final tracker associations, with a cost function based on the product of both distance metrics.

Since sharing appearance is limited exclusively to newly initialized trackers for saving bandwidth, the tracker consensus process across cameras occurs only when a new tracker is initialized in any of them. To ensure robustness in dynamic communication scenarios where mobile cameras may exchange information with different cameras over time, we include the cross-camera trackers association in the communication message. This cross-camera trackers association consists of a look-up table where each tracker locally stores the unique identifier, i, assigned to the same target by the rest of the cameras in the network C_j. Consequently, once the message has traversed the entire network, the cameras achieve a global consensus on the association of trackers across all the cameras in the network.

4.5 Active Perception

In addition to collaborating in the distributed tracking of multiple targets, mobile cameras tackle the task of active perception to gain additional knowledge about the people presented in the scene. They leverage shared information to efficiently position themselves for effectively visualizing each target's attribute class. Furthermore, mobile cameras are allowed to communicate between them in order to gather global knowledge of the visualization process's status.

This section is not a direct contribution of this thesis. However, since the active perception approach is a crucial part of the implemented hybrid framework, its components are detailed in the following to be self-contained.

4.5.1 Target Class Observations and Belief Updates.

Every time step τ_h, the drone uses a black-box perception algorithm (e.g., a pre-trained CNN classifier) to compute the class probability distribution for each target visualized from the correct viewpoint. Let $\mathcal{P} = h(\mathcal{X}) = \{\mathbf{p}_i\}_{i=1}^{I}$ be the class probability distribution, where $\mathbf{p}_i$ represents the likelihood of target i belonging to each one of the G classes in the class

set $\mathcal{G}$. To simplify the notation, in this complete Section 4.5, t will denote times periods of τ_h.

The probability distribution over time is modulated by belief vectors $\mathbf{b}_i^t$ for each target i. These vectors contain G belief values b_{ig}^t representing the aggregate likelihood of target i belonging to a class $g \in \mathcal{G}$ up to time t, i.e., combines the historical class probabilities distributions up to time t. The process of aggregating the drone's observations to derive class beliefs for each target is a crucial consideration. Standard Bayesian recursive estimation is not recommended in this case due to the unavailability of the measurement likelihood model, $\mathbb{P}(\mathbf{p}_i^t|\mathbf{b}^{t-1})$, from the black-box sensor. Building a precise pose-dependent likelihood model requires the construction of a dense dataset and considering all targets and occlusions for optimal viewpoint search. This process is expensive and does not scale well due to its computational demands.

Instead, we propose the use of the conflation operator $\zeta(\mathbf{p}_i^{1:t})$, a mathematical method introduced by Hill and Miller, 2011. Conflation enables the aggregation of probability distributions obtained from measurements of the same phenomena under different conditions. It possesses the remarkable property of minimizing the loss of Shannon information when combining multiple independent probability distributions into a single distribution, specifically when computing $\mathbf{b}_i^t$ based on the measurements $\mathbf{p}_i^{0:t}$. The conflation is defined by

$$\mathbf{b}_i^t = \zeta(\mathbf{p}_i^{1:t}) \equiv \zeta(\mathbf{b}_i^{t-1}, \mathbf{p}_i^t) = \frac{\mathbf{b}_i^{t-1} \odot \mathbf{p}_i^t}{(\mathbf{b}_i^{t-1})^\top \mathbf{p}_i^t}, \tag{4.3}$$

where the Hadamard product $\odot$ in the numerator is taken component-wise, whereas the dot product is the normalization factor. Conflation's commutative and associative properties enable efficient recursive computation, making it suitable for onboard and decentralized belief updates in the presence of multiple communicating drones. The beliefs are initialized at $t = 0$ with a uniform prior probability distribution over all possible target classes, formally $b_{ig}^0 = 1/G \ \ \forall g \in \mathcal{G}$.

4.5.2 Viewpoint Control Policy.

The lack of an observation model that maps target relative poses to a probability distribution, i.e., the h function that maps $\mathcal{P} = h(\mathcal{X})$, hinders the direct solution of the active perception for class recognition problem. Therefore, we leverage Reinforcement Learning to train a viewpoint control policy, π_ϕ, that learns to recommend viewpoints $\mathbf{a}_t$ that minimize the accumulated entropy of all targets' beliefs over a given time horizon. The policy is parameterized by ϕ and operates at the perception low-frequency, $\frac{1}{\tau_h}$.

Each drone uses a copy of the same learned viewpoint control policy that solves the viewpoint recommendation problem. The viewpoint recommendation problem is formulated as a Partial Observable Markov Decision Process (POMDP), denoted by $\langle S, A, \mathcal{T}, \Omega, \mathcal{O}, R \rangle$. The state S includes the state of the drones, the targets' pose, their beliefs, and their visualization status (visualized or not). Actions A represent recommended viewpoints within a constrained neighborhood and transitions $\mathcal{T}$ assume timely movement to the next viewpoint. The drone receives partial information Ω about the environment through the observation function $\mathcal{O}$. The observation of each target is defined by $\mathbf{o}_{q,i}^t = [\bar{\mathbf{o}}_{q,i}^p, \bar{\mathbf{o}}_{q,i}^c] \in \Omega$ where $\bar{\mathbf{o}}_{q,i}^p$ is the observation of each target physical attributes (poses and velocities). Each target's attribute information is represented by $\bar{\mathbf{o}}_{q,i}^c$ which includes the entropy of the local

class estimates from the drone q and the entropy of the global class beliefs.We define the joint target observation vector as $\mathbf{o}_q^t = \{\mathbf{o}_{q,i}^t\}_{i=1}^I = [\bar{\mathbf{o}}_q^p, \bar{\mathbf{o}}_q^c]$.

The reward function is based on the formulation of Serra-Gómez et al., 2023. It provides rewards to the agent for successfully classifying each and all targets and reducing the entropy of target class beliefs. Additionally, it penalizes the agent for movement and for each time step in which the task remains incomplete.

Architecture. The generalization ability of the learned policy $\pi_\phi(\mathbf{a}|\mathbf{o}_q)$ depends on the neural network architecture chosen. The main challenge lies in the size and dynamical changes over time of the set $\mathbf{o}_t^q = \{\mathbf{o}_{q,i}^t\}_{i=1}^I$.

Inspired by Relational Graph Convolutional Networks (Schlichtkrull et al., 2018) and self-attention mechanisms (Vaswani et al., 2017) used in static knowledge graphs, we employ a self-attention block (SAB) to capture the relationships among all targets at time t. Note that the focus in this first layer is on spatial features such as poses and velocities $\bar{\mathbf{o}}_p^q$, since the purpose is to encode important information including target visibility, observation perspective, occlusions, and potential simultaneous observations. Therefore, the initial layer is,

$$\begin{aligned}
\tilde{\mathbf{e}}_{i,p}^{1,h} &= F(\bar{\mathbf{o}}_{q,i}^p; \mathbf{W}_{q,h}^1) + \sum_{j\in\mathcal{J}} \lambda_{i,j}^h F(\bar{\mathbf{o}}_{q,j}^p; \mathbf{W}_{v,h}^1)\,, \\
\mathbf{e}_{i,p}^1 &= LN(Res^1(LN(concat(\{\tilde{\mathbf{e}}_{i,p}^{1,h}\}_{h=1\dots H})))\,, \\
\lambda_{i,j}^h &= \mathrm{softmax}\Big(\frac{1}{\sqrt{d_h}} F(\bar{\mathbf{o}}_{q,i}^p; \mathbf{W}_{q,h}^1)^\top F(\bar{\mathbf{o}}_q^p; \mathbf{W}_{k,h}^1)\Big)_j\,,
\end{aligned} \tag{4.4}$$

where $i \in \mathcal{J}$, $Res^l(x) = x + \sigma(F(x; \mathbf{W}^l))$, with σ being a ReLU activation function and F a parametric affine transformation. LN stands for Layer Normalization. $\mathbf{W}^1 \in \mathbb{R}^{d_{enc}\times(d_h H+1)}$ and $\mathbf{W}_{w,h}^1 \in \mathbb{R}^{d_h\times(d_{in}+1)}, w \in \{v,q,k\}$, are learnable parameters. d_{in}, d_h, d_{enc} are the dimensionality of the input, each head h, and the first layer. Note that each head h encodes a different relation $\boldsymbol{\lambda}^h$ between targets. To incorporate the information acquired about each target's class, we concatenate it with the latent representation of each target from the previous layer. Then, we map it back to a latent space of dimension d_{enc} using a learned linear layer. The process can be expressed as $\mathbf{e}_i^1 = F([\mathbf{e}_i^1, \bar{\mathbf{o}}_{q,i}^c]; \mathbf{W}_c)]$, where $\mathbf{e}_i^1$ represents the updated latent representation, $\bar{\mathbf{o}}_{q,i}^c$ is the class information of target i observed by drone q, and $\mathbf{W}_c$ is the learned weight matrix.

Next, we use a pooling multi-head attention mechanism (PMA) that incorporates a learned seed vector per head $\mathbf{v}_s^h \in \mathbb{R}^{d_h}$ to calculate the attention weights for a single query,

$$\begin{aligned}
\tilde{\mathbf{e}}^{2,h} &= \mathbf{v}_s^h + \sum_{j\in\mathcal{J}} \lambda_j^h F(\mathbf{e}_j^1; \mathbf{W}_{v,h}^2)\,, \\
\mathbf{e}^2 &= LN(Res^2(LN(concat(\{\tilde{\mathbf{e}}^{2,h}\}_{h=1\dots H})))\,, \\
\lambda_j^h &= \mathrm{softmax}\Big(\Big\{\frac{1}{\sqrt{d_h}} \mathbf{v}_s^{h,\top} F(\mathbf{e}_j^1; \mathbf{W}_{k,h}^2)\Big\}_{j\in\mathcal{J}}\Big)_j\,.
\end{aligned} \tag{4.5}$$

The output latent vector $\mathbf{e}^2$ is further processed by a fully connected layer to obtain the parameters $\mu_{\mathbf{a}_t}$ and $\log(\sigma_{\mathbf{a}_t})$ of a diagonal Gaussian distribution $\mathcal{N}(\mu_{\mathbf{a}_t}, \sigma_{\mathbf{a}_t})$ over viewpoints. The learned policy π_ϕ then samples recommended viewpoints $\mathbf{a}_t$ from this distribution. We assume that the drone can reach the recommended viewpoint before the next time step.

For training the network, the Proximal Policy Optimization (PPO) algorithm is employed (Liang et al., 2018; Schulman et al., 2017). PPO requires an estimate of the state-value $V^{\pi_\phi}(\mathbf{s}_t)$, which is approximated by a linear layer predicting $V^{\pi_\phi}(\mathbf{s}_t) \approx \mathbf{v}_v^\top \mathbf{e}^2$. This value estimation is used during training to guide the policy. The training process combines the surrogate loss and KL-divergence term to ensure stability. Additionally, an entropy regularization term is included to promote exploration (Haarnoja et al., 2017).

Low-Level Control MPC. During training, we assume the drone reaches the suggested viewpoint by the next time step. However, at test time we employ a low-level controller operating at a frequency $\frac{1}{\tau_l}$Hz $>>$ $\frac{1}{\tau_h}$Hz, to guide it there while accounting for the drone dynamics. The controller solves the following receding-horizon constrained optimization problem:

$$\begin{aligned} \min_{\mathbf{y}_{1:N}, \rho_{0:N-1}} \quad & \sum_{k=0}^{N-1} w_\rho \left\| \rho_k \right\| + w_g \frac{\left\| \mathbf{y}_N - \mathbf{a}_t \right\|}{\left\| \mathbf{y}^0 - \mathbf{a}_t \right\|} \\ \text{s.t.} \quad & \mathbf{y}_0 = \mathbf{y}_t, \quad \mathbf{y}_{k+1} = f(\mathbf{y}_k, \rho_k) \\ & \rho_k \in \mathcal{P}, \quad 0 \leq k \leq N-1 \end{aligned} \tag{4.6}$$

where ρ_k is the low-level control input sent to the robot, that needs to be inside the possible values $\mathcal{P}$, $f(\mathbf{y}_k, \rho_k)$ the internal dynamics and w_u and w_g are the respective weights of the stage and terminal costs (Serra-Gómez et al., 2023; Zhu & Alonso-Mora, 2019). Although our full method accounts for the drone dynamics using this low-level controller, our formulation is flexible to other low-level controllers as long as they track the recommended viewpoint $\mathbf{a}_t$. This is why during simulation we employ both the in-built drone dynamic model and the controller from AirSim (Shah et al., 2018).

4.6 Photo-realistic Environment

Testing autonomous drones in scenarios with pedestrians is unfeasible in the real world due to multiple safety constraints. Therefore, this thesis harnesses the advantages of photo-realistic simulators for processing visual information while controlling mobile cameras and gathering automatic labeling. Aiming to create and design scenarios with moving people quickly and easily, we develop the essential tools to effortlessly create photo-realistic simulated pedestrian scenarios within Unreal Engine.

Figure 4.3 shows the overall structure of the implemented tools. In the photo-realistic environment managed via Unreal Engine, the trajectory plugin allows users to create paths and control pedestrians' movement along them either randomly or according to customized references. Then, we provide a collection of ready-to-use pedestrian models that can be integrated into the scene by simply dragging and dropping them. These models consist of self-generated meshes integrated with a database of existing open-source animations, and controlled via the trajectory plugin. Finally, a Python API developed in AirSim extracts environment metadata including images, pedestrian information ground truth, sensor states, and timestamps. In the following, a detailed explanation of our tools is provided.

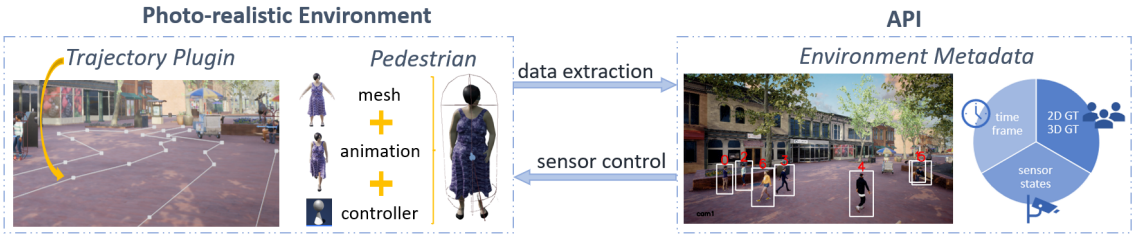


Figure 4.3: Overall structure of our implementation. In the photo-realistic environment, the trajectory plugin provides a user-friendly way to define the path to follow for the pedestrian models (white lines). In addition, pedestrian models are ready to be used, where the appearance (mesh), the movement (animation), and the pedestrian behavior (controller) are integrated. Finally, the API extracts the environment metadata.

4.6.1 Trajectory Plugin

The developed plugin offers several options for defining pedestrian trajectories thus, enabling the design of scenes with varying complexity. First, the *continuous path* tool allows the creation of specific paths that characters are forced to follow. It provides a user-friendly way to define routes and achieve natural-looking behaviors, such as pedestrians walking along sidewalks. This attribute is crucial for complex scenarios with numerous obstacles like urban areas. The second tool, *target points*, defines a set of goal positions that pedestrians can reach by moving freely across the map. This is particularly useful for open environments with minimal restrictions. Both tools are used directly on the simulated world map to define the pedestrian routes. Therefore, this plugin provides a user-friendly method for generating scenarios where pedestrians follow the traced path by simply dragging and dropping the *continuous path* or the *target point* directly onto the simulated world. This avoids the tedious task of collecting the specific coordinates of the desired trajectory. Figure 4.3 shows an example of various routes created using the *continuous path* tool.

Pedestrian models integrate an *AI controller* that makes the character move along these paths in either a random or customized manner. In random mode, the controller searches for goals within a defined area and randomly selects one. Then, the pedestrian moves towards that goal until it is reached. This process is repeated to set the next target location. Regarding the customized mode, this option offers more control but is more time-consuming. The user specifies the exact sequence of goals each pedestrian must follow. The controller loads the goals tagged with the current pedestrian's name and guides the character through them from oldest to newest.

4.6.2 Pedestrians

In order to make multiple agents available to interact with the world, our environment gathers a set of 50 general ready-to-used rigged pedestrian models. These models consist of realistic human characters, encompassing a diverse cast, in terms of gender, ethnicity, height, or clothing, for a truthful simulation of a real-life environment.

The 3D human meshes are produced in *MakeHuman* "Makehuman community. Makehuman, 2022." n.d., an open-source application that allows the mass production of people with random characteristics by considering different adjustable rules. To increase the diversity of human models, we use a community gallery included in the application to download

Creative Commons assets, such as topologies, skin tones or clothes. The final people models that may not be realistic or appropriate have been manually filtered out.

Next, the pedestrians must mimic the action of walking within the environment, hence, the animations associated with this action have to be blended with the previous meshes. In our framework, the animations are from the CMU Graphics Lab Motion Capture Dataset Carnegie Mellon University Graphics Lab, n.d., which consists of 2605 motion capture segments in different formats. Focusing on the walking animations, multiple motion capture segments have been assessed to select those most suitable and with sufficient quality for a photo-realistic environment. The final adaptation of these animations to a pedestrian movement has been done in *Blender*, where we discard the unrealistic frames in the merge of movement and mesh. Additionally, the frames attributable to the motion capture process are removed, such as the first moments of preparation of the motion capture actor and those at the end of the movement. Lastly, the models in Blender are exported to Unreal where the integration of all the components, i.e., mesh, animation and controller, is performed. These Unreal pedestrian models are ready to be used by simply dragging and dropping them on the world. Figure 4.4a shows a scheme of the process.

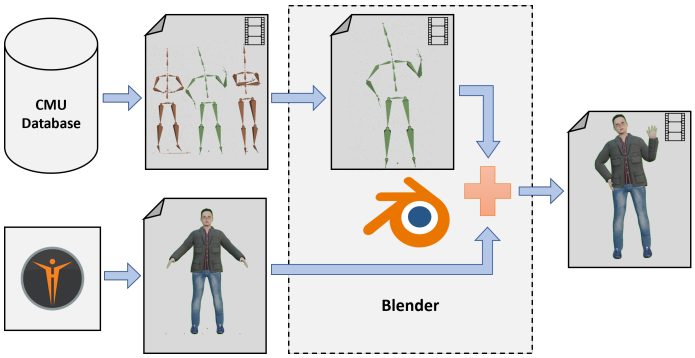


Figure 4.4: Scheme of the characters' models obtention. The animations from the CMU Motion Capture Database are trimmed to only contain valuable frames and then applied on the skeleton of a model via Blender. Finally, the model performing the animation is exported in Unreal Engine.

4.6.3 API for Environment Metadata Capture.

The API implemented in AirSim performs the automatic annotation, the image acquisition, and the control of the cameras at the scene. AirSim is an open-source simulator built on Unreal Engine that aims to narrow the gap between simulation and reality for the development of autonomous vehicles, particularly drones and cars (Shah et al., 2018).

The data annotation is conducted automatically by isolating the pedestrians from the rest of the environment with a unique key structure embedded in their Unreal identity. Every pedestrian model provided shares this key structure to be easily found. Then, pedestrians' ground truth, including instance segmentation, 3D position, and 2D bounding boxes, are gathered in every iteration of metadata acquisition. All the collected information is referenced to the *AirSimLocal* coordinates system, including the dynamic cameras which AirSim references to their initial position.

4.7 Experiments

This section first presents a simple analysis to assess the suitability of photo-realistic environments for evaluating multi-target tracking algorithms. Then, we conduct a thorough evaluation of the novel collaborative framework proposed in this thesis discussing the obtained results.

4.7.1 Simulated Data Analysis

Evaluating existing multi-target tracking algorithms with simulated data based on their previous assessment with real-world datasets offers a manner to validate the performance achieved with photo-realistic images. Achieving similar performance in both real-world and highly realistic simulations supports the use of such engines for developing and assessing perception systems, ultimately streamlining real-world implementation. Therefore, to confirm the suitability of photo-realistic simulations for evaluating our collaborative framework, we offline assess two state-of-the-art single-view tracking methods: PHALP (Rajasegaran et al., 2022) and Tracktor++ (Bergmann et al., 2019).

4.7.1.1 Experimental Setup

Sequences Collected. This experiment uses the well-established single-view MOT Challenge benchmarks (Dendorfer et al., 2021; Leal-Taixé et al., 2015; Milan et al., 2016) as a reference. These benchmarks consist of short sequences captured by static or moving cameras in real-world environments. Leveraging our developed tools, we generate six sequences of varying lengths in two simulated environments: Street and Font. These environments correspond to specific locations within a freely available photo-realistic map from the Unreal community. More concretely, we place dynamic people models in a commercial pedestrian street (Street) and a park (Font). The collected sequences incorporate images from fixed or moving cameras (drones) under various lighting and weather conditions. This design, in addition to validating the results obtained with photo-realistic images, enables us to assess the tracking algorithms' ability to generalize across diverse scenarios. Table 4.1 summarizes the details of the generated sequences.

Dataset	Camera	Length	Resolution	Trajectory	Scene	Description
Day	Static	500	1920x1080	Customized	Street	Commercial street at day time
Night	Static	500	1920x1080	Customized	Street	Commercial street at night time
Fog	Static	900	1920x1080	Random	Font	Fog weather in a small park
Street Moving	Dynamic	500	1920x1080	Customized	Street	Drone flying over people
Midday	Static	900	1920x1080	Random	Font	Small park at midday
Font Moving	Dynamic	600	640x480	Random	Font	Low quality images from a drone

Table 4.1: Details of the sequences acquired to evaluate the state-of-the-art tracking methods.

Metrics. Following the official evaluation of the selected methods, this experiment employs the CLEAR MOT Metrics (Bernardin & Stiefelhagen, 2008) and Identity Metrics (Ristani et al., 2016b). In particular, the Multiple Object Tracking Accuracy (MOTA) and ID F1 Score (IDF1) quantify two of the main aspects, i.e., object coverage and temporal identification, respectively.

4.7.1.2 Results.

This evaluation aims to validate the suitability of photo-realistic images for the multi-target tracking task and thus, further assess our collaborative framework within the simulator. Therefore, note that our goal is not to determine the best tracker, but rather to verify that the algorithms achieve similar performance with simulated data compared to real-world evaluations. Additionally, we analyze their ability to generalize to uncommon scenarios in real-world datasets, such as nighttime or foggy weather, which can isolate more challenging situations for the tracker.

To perform the proposed study, two object tracking methods are implemented, PHALP (Rajasegaran et al., 2022) and Tracktor++ (Bergmann et al., 2019). The obtained results in our collected sequences along with their previous results in real-world datasets are presented in Table 4.2 and Table 4.3. PHALP algorithm exhibits high MOTA and similar IDF1 in the *Day* and *Night* sequence compared to the real-world dataset, likely due to the shared conditions of high-resolution images and pedestrians close to the camera. Similarly, Tracktor++, previously evaluated on the MOT Challenges (Leal-Taixé et al., 2015; Milan et al., 2016) where most of the sequences are in commercial or urban areas, achieves comparable IDF1 in our commercial street sequences (*Day*, *Night* and *Street Moving*). From these results, we can deduce that under similar conditions both algorithms achieve comparable performance on real-world and photo-realistic data.

Finally, our analysis of the performance on the photo-realistic sequences reveals limitations in the generalization capabilities of the evaluated methods. PHALP's performance decreases in wider areas, suggesting a potential limitation in handling large-scale environments. Regarding Tracktor++, the lack of visibility appears to be the major cause of failure generating a performance drop in the foggy sequence compared to the midday scene (both from the same scenario and viewpoint). Figure 4.5 shows qualitative examples of both methods where red bounding boxes are missing trackers.

Dataset	**PHALP**	
	MOTA↑	IDF1↑
Real-world dataset		
PoseTrack	58.9	76.4
MuPoTS	66.2	81.4
AVA	-	62.7
Photo-realistic sequences		
Day	84.4	76.2
Night	83	82
Fog	48.2	52
Street Moving	36.3	52.8
Midday	48.8	51.5
Font Moving*	-	-

Table 4.2: PHALP results in multiple datasets.*PoseTrack*, *MuPoTS* and *AVA* are the benchmarks used for its official evaluation. * *Font Moving*, with poor quality images, makes the algorithm produce unsatisfactory results.

Dataset	**Tracktor++**	
	MOTA ↑	IDF1 ↑
Real-world dataset		
2D MOT2015	44.1	46.7
MOT16	54.4	52.5
MOT17	53.5	52.3
Photo-realistic sequences		
Day	30.6	43.5
Night	37	45
Fog	5.1	22.3
Street Moving	58.2	57.7
Midday	34.8	33.8
Font Moving	3.6	9

Table 4.3: Tracktor++ results in multiple datasets. *2D MOT2015*, *MOT16* and *MOT17* are the benchmarks used for its official evaluation.

(a) Example of qualitative results with the tracker PHALP.

(b) Example of qualitative results with the tracker Tracktor++.

Figure 4.5: Qualitative examples of the multi-target tracking methods analyzed where the failures (missed people) are highlighted with red bounding boxes. (a) Qualitative examples from PHALP. (b) Qualitative examples from Tracktor++.

4.7.2 Collaborative Framework Evaluation

4.7.2.1 Experimental setup

Scenarios designed. This experiment uses the same two high-fidelity virtual environments as above to test the presented framework: Street and Font. The designed setups are presented in Figure 4.6 and their respective dimensions are 97×27m and 97×50m. In both environments, we place three static cameras with overlapping views for global area monitoring and define distant starting points for the drones. The size of the images captured by the camera network is set to 1440 × 900 and the field of view to 90 degrees[3]. Regarding communications, to be as faithful as possible to a real-world scenario, we set the drones to share information with each other as well as with the closest camera to them at the time. Communication between static cameras is limited to their direct neighbor, as shown in Figure 4.6. Finally, the number of pedestrians present on the scene varies between episodes, and their trajectories are randomized.

Regarding the task of correctly visualizing attributes of people visible only from a specific point of view, we devise a marketing study on clothing brands as a use case. Specifically, the appearance (meshes) of some of the pedestrian models created are modified to include a logo on the front of their T-shirts that can be visualized exclusively from the frontal view of the person.

Sequence Evaluated. Several sequences are evaluated in each one of the environments with their corresponding ground truth being automatically obtained from the simulator. More concretely, the number of pedestrians varies to assess the performance for 5, 10, and 15 pedestrians. Thus, the conducted experiments are named as sparse, medium, and busy for 5, 10, and 15 pedestrians respectively, resulting in the following sequences: *Street Sparse*, *Street Medium*, *Street Busy* for *Street* environment, and *Font Sparse*, *Font Medium*, *Font Busy* for *Font* environment. All of them have the same length of 500 frames.

[3]The rest of the camera parameters are those set by default in Unreal Engine and AirSim

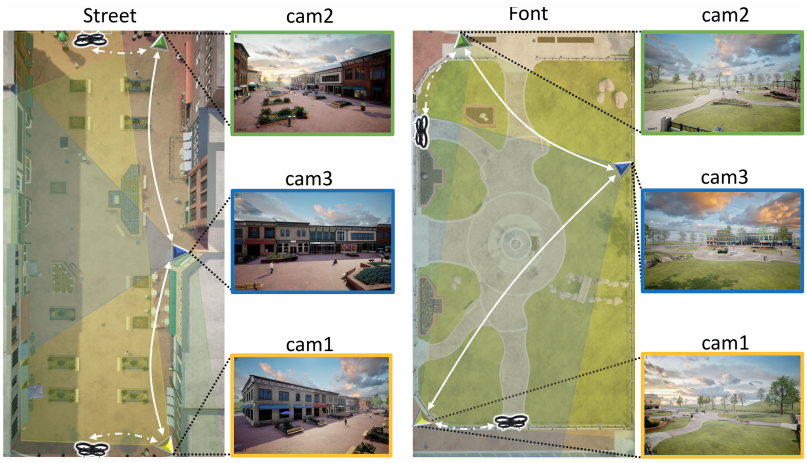


Figure 4.6: Experimental environments used to evaluate the proposed framework. On the left, we show the setup for the experiments performed in the commercial street, *Street*, and on the right the setup for the font area, *Font*. The starting points of the two drones used as mobile cameras are also shown and they always communicate with each other.

Evaluation metrics. To comprehensively evaluate the proposed approach for *distributed multi-target tracking*, the common CLEAR MOT metrics (Bernardin and Stiefelhagen, 2008; Ristani et al., 2016b) adopted are:

- Multiple Object Tracking Accuracy (**MOTA**): measures failures during the tracking taking into account the number of misses, false positives, and mismatches.
- Identity F1 Score (**IDF1**): evaluates the capability of the system for preserving the identities over time.
- Multiple Object Tracking Precision (**MOTP**): shows the ability of the tracker to estimate precise object positions through the error in estimated position.

The above evaluation is performed in the image plane where metrics require setting a threshold between the ground truth and the resulting trackers in order to consider a tracker valid. We evaluate the resulting bounding boxes in the image plane using a minimum intersection over union (IoU) of 0.3 as the threshold to validate the trackers. The final tracking results in each camera are those obtained as output of the Distributed Tracker Manager and the final value presented is the median of all the cameras in the network.

Regarding the acquisition of the correct people's viewpoint obtained from the *active perception* approach, the evaluation is performed using a black-box clothing brands detector. Thus, we employ two metrics:

- Trackers Classified (**TC**): measures the percentage of trackers whose beliefs are higher than 95%.
- Precision (**P**): evaluates the percentage of trackers for which their beliefs exceed 95% and correctly identifies their brand (attribute).

To associate each tracker with a ground truth brand class, we perform a linear sum assignment problem between the trackers and ground truth bounding boxes. The ground truth bounding boxes obtained from the simulator contain the person's attribute class.

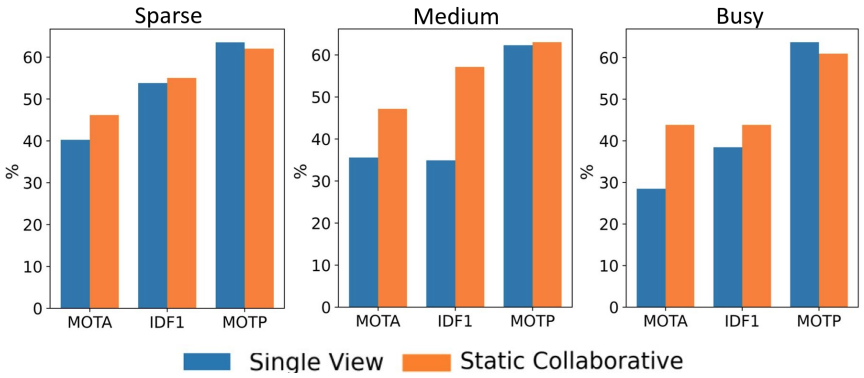


Figure 4.7: Comparison of the cameras responsible for distributed multi-target tracking collaborating with each other with a chain graph of communications (Static Collaborative) and a single view tracking with isolated cameras (Single View)

4.7.2.2 Results and settings

In the following, we explain the baselines selected to compare the proposed method in the *Street* sequences and perform a detailed analysis of the obtained results. To conclude the experiments, we also present the performance of our approach in the metrics described above for both environments, *Street* and *Font*. The parameters defined in the method for all the experiments are set to $\tau_{dLDA} = 1$, $\alpha = 700$, $\tau_{aLDA} = 0.55$, $\tau_{dDTM} = 2$,$\tau_{aLDA} = 1$, $\tau_l = 0.05$, $\tau_h = 0.25$ and the size of the targets' gallery is set to 10.

Collaborative behavior analysis. To demonstrate the benefits of collaborative behavior between nodes in a multi-target tracking network, we gather the three cameras from our system responsible for tracking and assessed their performance with and without communication. The first case (Static Collaborative) follows the initial setup where cameras communicate exclusively with their direct neighbor in a chain graph (Figure 4.6). In the second setup, the different cameras perform individual tracking without any communication between nodes (Single View). The obtained results, present in Figure 4.7, demonstrate the benefits of sharing information once per iteration so that no node in the network is isolated. The Static Collaborative setup achieves up to 21% and 15% of improvement in the IDF1 and MOTA metrics, respectively, in comparison with the tracking in Single View. Therefore, we can conclude that in large scenarios, the use of collaborative cameras with overlapping perspectives enhances tracking performance in comparison to the use of independent cameras.

Mobile cameras analysis. Furthermore, we evaluate the efficiency of our mobile cameras (MC) to correctly visualize the desired people's viewpoint against a baseline of static cameras (SC). The static setup is composed of five cameras, the three already existing in the system and two more located on the other side of the street for more visual coverage of the scene. Communications among the five cameras are defined as a ring graph, i.e., each camera shares information with its two nearest neighbors. As a consequence of the distributed nature of the system, the static cameras collaborate to gain knowledge of the overall scene, and the evaluation of the correct viewpoint visualization is performed individually. The final results of the baseline are the median of all the cameras.

The results obtained of the percentage of trackers classified (TC) and correctly identified their brands (P) with beliefs higher than 95% are presented in Table 4.4. In the sparse scenario, where the occlusions between targets are not frequent, the static camera setup gets better results than the mobile cameras. However, in more crowded scenarios, static cameras struggle to avoid occlusions for obtaining a view with high confidence from the pedestrian. In contrast, mobile cameras can be actively positioned to capture the desired viewpoint, achieving a coverage (TC) of 81.5%, against the 70.83% obtained from the static setup, in the most challenging scenario (*Street Busy*). In addition, the quality of the people data captured by each one of the systems is unmatched. Figure 4.8 shows examples of the same pedestrian captured with the mobile cameras (blue box) and with the static cameras (red dashes box). Every two columns correspond to the same person and we can notice the great difference in quality. The images from mobile cameras revealed much more clear details than the static ones, whose images are of low quality and blurry. These result in better identification of the person's brand in most of the sequences evaluated (P).

Method	**Classification Process (%)**					
	Street Sparse		*Street Medium*		*Street Busy*	
	↑TC	↑P	↑TC	↑P	↑TC	↑P
SC	75	75	80	60	70.83	58.33
MC	71.5	64.3	76.2	66.7	81.5	74.1

Table 4.4: Percentage of trackers classified (TC) and percentage of trackers' brand correctly identified (P) in the *Street* sequences. Results for the baseline static camera network (SC) and our hybrid system with two mobile cameras (MC).

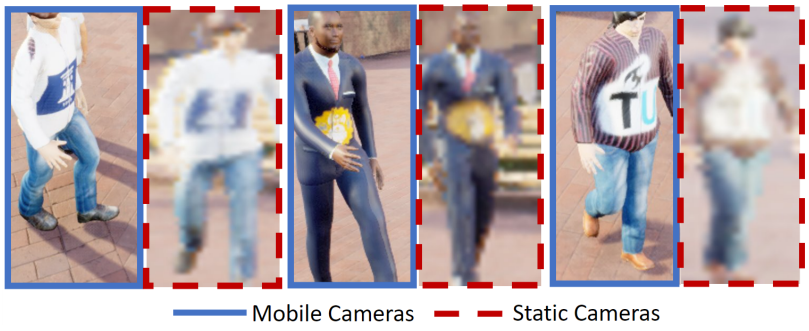


Figure 4.8: Examples of people images captured from the correct viewpoint: mobile cameras (blue box) and static cameras (red dashed box) in the *Street Busy* sequence. Every pair of columns displays images of the same person.

Final Evaluation. As a summary, we present the performance of the proposed framework in both photo-realistic environments, *Street* and *Font*. The results obtained are shown in Table 4.5, from which we can conclude that the method is consistent under various conditions, including different numbers of people, size of the space, and type of environments. Specifically, the experiments focus on evaluating sparse, medium, and busy scenarios, with 5, 10, and 15 pedestrians, respectively. Moreover, the *Font* environment is larger than the *Street* environment with static cameras located further away from the path where people walk, making it more challenging for monitoring. Finally, the mean time required by

Sequence	Multi-target Tracking			Classification	
	↑MOTA%	↑IDF1%	↑MOTP%	↑TC%	↑P%
Street Sparse	54.18	48.34	61.43	71.5	64.3
Street Medium	43.62	42.57	60.45	76.2	66.7
Street Busy	38.83	42.52	60.1	81.5	74.1
Font Sparse	38.24	41	58.1	100	75
Font Medium	47.22	55.52	59	93.3	53.35
Font Busy	40.34	47.96	63.63	72.73	63.63

Table 4.5: Results of the evaluated metrics in the *Street* and the *Font* sequences where sparse, medium, and busy environments are analyzed.

Figure 4.9: Example of images captured by the hybrid system. First row *Street Busy* sequence and second row *Font Medium* sequence. Static cameras are mainly responsible for the global understanding of the scene while mobile cameras (drones) capture pedestrian images from the desired viewpoint.

each of the modules comprising the proposed framework is measured: detection 0.0198 s, local data association per tracker 0.038 s, distributed Kalman filter per tracker 0.002 s, distributed tracker manager 0.0015 s, class information fusion per tracker 0.00004 s, viewpoint control policy 0.005 s. The complete evaluation is conducted in one computer with an Intel® Core™ i7-9700 CPU @ 3.00GHz ×8 and a Nvidia GeForce GTX 1070. Both tracking modules, with mobile and static cameras, and classification modules, with mobile cameras, work in parallel. Provided that the poses of new targets are estimated and relayed to the mobile cameras within τ_h, our framework operates in real time. This is not a strict constraint as there is allowable latency; however, it is crucial that tracked targets remain within the recommended viewpoint FOV during any such delays.

In addition to the numerical results, Figure 4.9 displays examples of images captured by the hybrid system at a specific time. The first row corresponds to images from the *Street* environment and the second row from the *Font* scenario. The overall understanding of the scene is mainly performed by the static cameras although the drones also assist in the distributed tracking, while the close-up person images are gathered from the mobile cameras. For example, in the first row, Drone2 correctly captures the viewpoint of the target with local identity 14, and in the second scenario, Drone1 accomplishes its goal with local identity 7.

4.8 Conclusions

This chapter has explored the potential of collaborating heterogeneous systems, specifically static and mobile cameras, as a powerful approach for improving functionality and information gathering in pedestrian monitoring applications. The major benefit of our hybrid system is the high-resolution visualization of specific pedestrian attributes. The ability to capture images of enhanced quality has been shown to hold great potential for tackling challenging perception tasks. The implementation of the whole framework has been performed using a photo-realistic environment in which we deploy all the essential tools to easily design pedestrian scenarios for robotic-oriented applications, i.e., a user-friendly tool for creating pedestrian trajectory, pedestrian models ready-to-use by simply dragging and dropping, and an API to automatically collect the pedestrians and cameras' metadata. The proposed collaborative framework performs multi-camera distributed tracking providing a global understanding of the scene for which the static cameras are mainly responsible. The experiments performed demonstrate that by allowing collaboration between cameras through sharing information with the closest nodes, the multi-target tracking improves up to 21 points in the IDF1 metric and up to 15 points in MOTA. Global scene awareness and the current state of drones are used by the viewpoint control policy to provide a new position and orientation for mobile cameras whose goal is capturing a desired viewpoint of the people as quickly as possible. In comparison with a static multi-camera system, mobile cameras are able to capture the required viewpoint with higher precision in most of the scenes evaluated. Overall, the integration of the hybrid system with ROS to handle communications between sensors and its implementation using a photo-realistic environment that enables the processing of visual information along with robot control, brings the presented framework closer to real-world applications.

Chapter 5

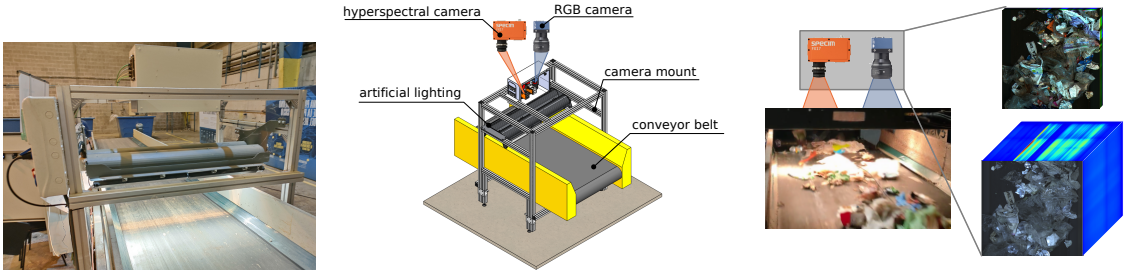
Multi-modal object identification.

In order to explore not only heterogeneous systems based on movement constraints but also alternative information sources beyond traditional RGB cameras, this thesis studies the benefits of capturing additional non-visible light intensity using hyperspectral sensors. Unlike conventional RGB cameras, hyperspectral cameras capture hundreds of narrow and contiguous bands across the spectrum. These spectral signatures facilitate the identification of materials, making hyperspectral information valuable in applications where a deeper understanding of materials within the scene is essential. Therefore, our research focuses on analyzing the combined use of multimodal data (RGB and hyperspectral) to enhance the automation of recycling facilities. Recycling facilities play a crucial role in reducing the non-biodegradable waste generated, but their automation is hindered by the complex characteristics of waste recycling lines like clutter or object deformation. Additionally, the lack of publicly available labeled data for these environments makes developing robust perception systems challenging.

The research presented in this chapter delves into the advantages of multimodal perception for object segmentation in real waste management scenarios. First, we present SpectralWaste, the first dataset collected from an operational plastic waste sorting facility that provides synchronized hyperspectral and conventional RGB images. This dataset contains labels for several categories of objects that commonly appear in sorting plants and need to be detected and separated from the main trash flow for several reasons, such as security in the management line or reuse. Moreover, we propose a pipeline employing different object segmentation architectures and evaluate the alternatives on our dataset, conducting an extensive analysis of multimodal and unimodal alternatives. Our evaluation pays special attention to efficiency and suitability for real-time processing and demonstrates how hyperspectral imaging can bring a boost to RGB-only perception in these realistic industrial settings without much computational overhead.

5.1 Introduction

The global issue of waste production intensifies as societies grow and consumption rises. The sheer volume of waste generated, particularly non-biodegradable waste such as plastics, has reached concerning proportions (Sukno & Palunko, 2022). Recycling and reuse are key strategies to lessen the environmental burden of waste. Hence, automating waste management facilities not only increases the volume of properly processed waste but also



(a) (b) (c)

Figure 5.1: The multimodal setup in the waste sorting facility contains two synchronized line-scan cameras that gather RGB and hyperspectral data. (a) Real prototype installed in the facility. (b) Diagram of the setup for data capture. Both hyperspectral and RGB cameras are housed in an industrial enclosure and completed with artificial lighting to ensure correct image capture. (c) Example of a scene as captured by the RGB camera and by the hyperspectral camera (with a false-color for visualization purposes).

safeguards worker health and comfort. To achieve automated manipulation of relevant elements in these real industrial environments, the first step is to improve and adapt existing perception systems to this domain.

Despite the great advances in automated visual recognition tasks in recent years, real industrial settings often present recurring problems that hinder real-world applicability, such as the lack of annotated data to achieve precise domain adaptation or high computational requirements (Linder et al., 2021). The most common method to identify and localize elements of interest in automated tasks is through the segmentation of RGB images (Lee et al., 2022; Lu et al., 2023). However, accurate detection based only on visual features is extremely challenging in waste management scenarios with severe clutter, high materials diversity, deformable or broken objects, and translucent elements (see sample images in Figure 5.2).

To overcome these issues, the use of more complex sensing modalities like hyperspectral imaging (HSI), commonly used for raw material classification (Henriksen et al., 2022), can provide insights beyond the visual appearance of objects. While conventional RGB cameras capture the visible spectrum, hyperspectral cameras are able to acquire light across a wide range of wavelengths. Thus, leveraging the combined information from both modalities improves the performance of complex perception tasks such as segmentation of buildings (Habili et al., 2022), terrain classification (Kodgule et al., 2019) or object tracking (Z. Liu et al., 2022). Unfortunately, the advantages derived from using hyperspectral information along with RGB images remain unexplored in object recognition for automatic waste sorting, where the task is mostly approached using RGB information (Bashkirova, Abdelfattah, et al., 2022).

This chapter demonstrates the benefits of using multimodal segmentation approaches in waste management and contributes to mitigating two key challenges that currently hinder their adoption, the scarcity of public multimodal waste datasets and the high computational demands associated with HSI data. Specifically, our main contributions are twofold: 1) We introduce SpectralWaste[1], the first multimodal dataset obtained from an operational

[1]Dataset website: https://sites.google.com/unizar.es/spectralwaste/home.

waste sorting facility, featuring in-the-wild industrial data from both hyperspectral and RGB cameras (Figure 5.1). This dataset addresses the identification of critical objects that frequently appear in real trash flows and impact sorting efficiency by either clogging machinery if not removed or holding value if recovered. 2) We present a comprehensive object segmentation analysis that underscores the boost in performance when combining both modalities and, for the first time, explores the suitability of using HSI for object segmentation in waste sorting scenarios. The proposed pipeline places particular emphasis on employing efficient architectures. Furthermore, to ensure consistency in annotated masks between modalities and reduce the labeling effort required, this research proposes a novel label transfer algorithm that automatically adapts RGB-annotated masks to HSI without any calibration needed.

5.2 Related Work

This section summarizes existing works regarding waste datasets for object identification, and image segmentation methods using both hyperspectral imaging exclusively and multimodal information.

Waste Object Identification Datasets. Numerous datasets spanning diverse domains have focused on collecting data for waste identification. Stanford TrashNet (Yang & Thung, 2016) consists of classifying single objects on an empty white background. Aiming to tackle the localization problem in addition to classification, Trash Annotations in Context (TACO) (Proença & Simões, 2020) presents a dataset in open and real environments, such as streets, lakes, or beaches where world litter is shown. Following the idea of reducing waste in natural environments, Floating Waste (FloW) (Cheng et al., 2021) is dedicated to the efficient cleaning of inland water areas with autonomous boats. Due to the demands of this task, a multimodal sub-dataset, FloW-RI, is included, providing millimeter wave radar data synchronized with the images. Similarly, other works aim to tackle challenging tasks by providing complementary information alongside RGB images. For instance, Kim et al., 2023 introduces a multimodal dataset comprising RGB-D, thermal infrared and object poses to address the issue of transparent object identification.

Closer to our work is the ZeroWaste dataset (Bashkirova, Abdelfattah, et al., 2022) and its extension ZeroWaste-v2, proposed in the VisDA2022 challenge (Bashkirova, Mishra, et al., 2022). In these papers, they provide a dataset comprised of conventional RGB images for industrial waste object segmentation that have been collected from a real sorting plant. In contrast, our dataset includes hyperspectral data synchronized with the RGB images. Thus, this spectral information combined with data from the visible spectrum holds significant potential for enhancing object identification tasks within the recycling processes.

Identification with Hyperspectral Data. A wide variety of techniques have been explored to leverage hyperspectral sensors for raw material identification, including spectral angle mapper (Kruse et al., 1993), multi-layer perceptrons (Hanson et al., 2023) and CNNs (Seidlitz et al., 2022). However, the limited amount of data in the existing HSI datasets poses a challenge for training data-based models. Common methods involve pixel-

wise training and testing on a reduced set of images (Wendel & Underwood, 2016), leading to information leakage and suboptimal generalization capabilities (Nalepa et al., 2019).

Final applications with HSI typically focus on fields where the information captured by RGB cameras lacks sufficient detail, e.g., environmental monitoring, agriculture, medical imaging, or remote sensing (Grewal et al., 2023). More specifically, the use of hyperspectral sensors for automatic plastic sorting in recycling facilities is a widespread technique. For example, the structure of plastic material is analyzed by sparse pixels (Henriksen et al., 2022; Shiddiq et al., 2023), or hyperspectral imaging is used to densely label images based on per-pixels classifications (Karaca et al., 2013). Unlike these works focused on raw material identification, we study the use of HSI for object segmentation, which can potentially overcome the limitations of RGB data in waste sorting tasks by leveraging spectral properties for material differentiation.

Regarding multimodal sensors for segmentation tasks, multiple works have addressed this problem through sensor fusion (Z. Chen et al., 2020; Valada et al., 2017). In particular, the combination of HSI with RGB information has been used in several fields. For instance, combining HSI with RGB among other modalities is exploited in environmental monitoring with UAV's (Qin et al., 2022) and autonomous terrain classification (Kodgule et al., 2019). Moreover, perception approaches combine both modalities for different tasks like classification of building materials (Habili et al., 2022) or rise seeds inspection (Fabiyi et al., 2020). However, the exploration of multimodal segmentation in real-world industrial waste sorting scenarios remains an open area of research.

5.3 SpectralWaste Dataset

This section describes the novel multimodal SpectralWaste dataset, explaining the data acquisition followed and the annotation process performed.

5.3.1 Data Acquisition

The dataset was collected in a real waste sorting industry specializing in plastics, cartons, and cans, with a true-to-life prototype of the conveyor belt installed on the waste separation line (see Figure 5.1). This prototype closely mimics the real installation, ensuring that the captured waste streams accurately mirror those arriving at the facility for separation.

The setup involved two synchronized cameras for multimodal data capture: a line-scan RGB camera (Teledyne DALSA Linea) and line-scan hyperspectral sensor (Specim FX17) that captures 224 contiguous spectral bands in a range from 900 to 1700 nm. Both cameras were housed in an industrial enclosure situated at a height of 1.7m. This installation was supplemented with a set of LED illuminators and infrared halogen lighting to ensure a suitable image capture in the spectral domain. RGB images were stored with a resolution of 1200×1184 pixels and 8-bit color depth, while HSI images were stored as data cubes of size $600 \times 640 \times 224$ with 16-bit precision. Due to the physical distance between the cameras and the differences in their internal settings, the image captured by each camera slightly differs.

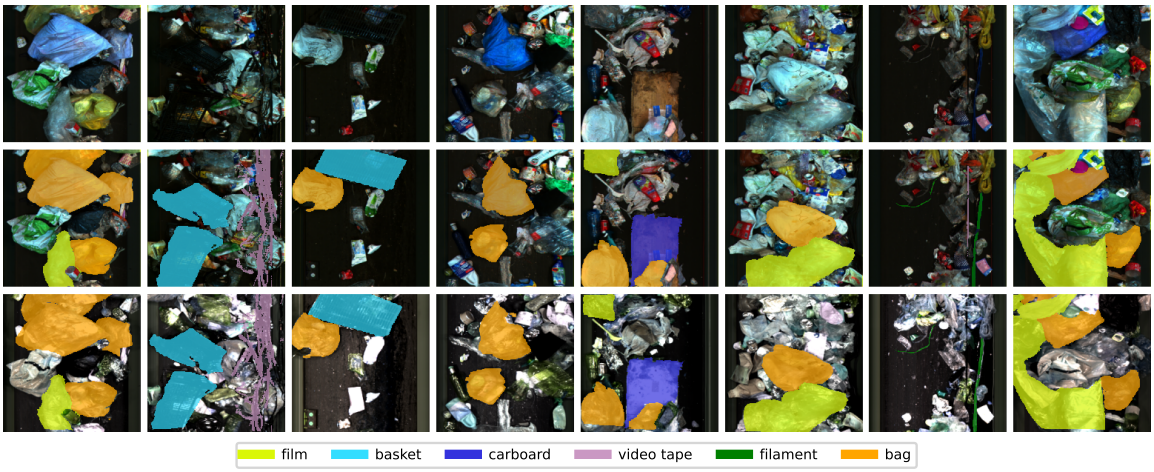


Figure 5.2: Examples of images included in the dataset. The first row shows images captured by the RGB camera, the second row shows the ground-truth RGB annotations and the third row shows the same scenes captured by the hyperspectral camera, annotated with labels obtained using the proposed label-transfer algorithm. For visualization purposes, we manually select three out of all the hyperspectral bands.

5.3.2 Data Annotation

The classes chosen for annotation in the presented dataset were selected according to the requirements of the facility. Labeled objects represent elements that commonly cause operational problems in recycling lines, impacting the efficiency of the sorting process. Among these problems, machinery jams pose a significant issue, causing a complete stoppage of the waste separation until the obstructing object is removed. Thus, the selected objects for automatic identification include *film* and *basket*, large objects that can clog the conveyor belts as they are not easily breakable; *video tape* and *filament*, representing long objects prone to entangle in waste separation zones and requiring manual intervention; *trash bag*, which encompasses closed bags containing waste that need to be mechanically opened for further processing; and *cardboard*, the paper material received at the facility whose recovery adds value sent to another recycling process.

To streamline the time-consuming process of labeling and facilitate this tedious task, we developed an interactive segmentation tool leveraging the point-prompt feature from Segment Anything Model (SAM) (Kirillov et al., 2023). In essence, the user can select points on the image belonging to an object, and with each selection, a new mask is generated and displayed over the image for further refinements or saving. In the presented dataset, we used our tool to manually generate the ground truth masks of the defined objects in the RGB image set.

5.3.3 Dataset Content

The result of the entire data acquisition and annotation process is encompassed in the SpectralWaste dataset. The dataset provides annotations for six object classes: *film, basket, video tape, filaments, trash bags*, and *cardboard*, totaling 2059 annotated instances across a set of 852 non-overlapping images. Table 5.1 presents the overview of the annotations, while Figure 5.2 and 5.6 illustrate sample images from the different classes.

In addition to the labeled set, SpectralWaste contains 6803 unlabeled multimodal images (RGB-HSI). We believe that releasing these unlabeled images is valuable for the community, enabling the researchers to explore the advantage of hyperspectral or multimodal object identification in a real industrial waste facility through further study of different techniques. These techniques may include refining labeled segmentation using semi-supervised or self-supervised methods, as well as exploring unsupervised segmentation approaches.

	Instances per Class					
Total	*Film*	*Basket*	*Cardboard*	*Video Tape*	*Filaments*	*Trash Bag*
2059	339	300	68	287	111	954

Table 5.1: Summary of the instances annotated in SpectralWaste.

5.4 Waste Segmentation

This section describes the proposed pipeline for waste object segmentation using RGB and HSI images. Figure 5.3 summarizes the key steps of the process where we consider different configurations that can take one or both modalities. A detailed description of the steps involved in the baselines and the adapted architectures is provided in the following.

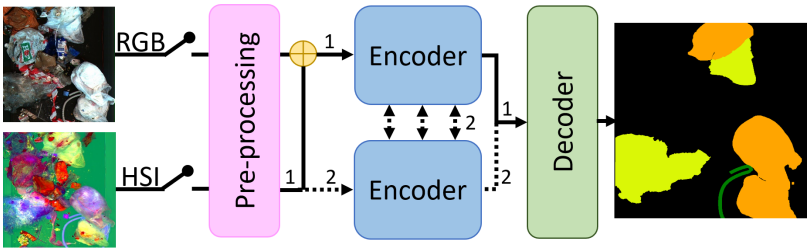


Figure 5.3: Pipeline of the segmentation process. Path 1 represents the data flow for the initial architectures designed for unimodal RGB segmentation (MiniNet-v2 and SegFormer). The dashed path 2 represents the HSI data flow in the multimodal CMX architecture.

5.4.1 Data Preprocessing

General Preprocessing. The pipeline includes a series of common preprocessing steps. We first crop the mismatched areas captured by each sensor to align the space shown in each image, resize the RGB and HSI images to 256×256 pixels and scale them to 32-bit floating-point values in the range $[0, 1]$.

HSI Channel Reduction. Hyperspectral imaging imposes significant computational demands due to the large amount of information contained in each pixel. Aiming to explore efficient solutions within our pipeline, a dimensionality reduction across the spectral channels is implemented using principal component analysis (PCA) for selecting the three main components (HYPER3). This PCA analysis is conducted on the pixel values from the training set of the dataset and then applied to the validation and testing sets to obtain their reduced version. The reduction process maintains 99.7 % of the explained variance

over the training set. By reducing the HSI channels to three components, we not only obtain a compressed representation of the information but also provide a balanced input when combining HSI with RGB images in the multimodal configurations.

5.4.2 Segmentation Architectures

Considering the application addressed of segmenting objects in an industrial setting, we compare three different architectures within our pipeline paying special attention to alternatives suitable for real-time inference.

The first architecture is *MiniNet-v2* (Alonso et al., 2020), a lightweight convolutional neural network designed for segmentation tasks. This model uses multi-dilation depth-wise separable convolutions and two convolutional branches instead of skip connections to achieve a favorable trade-off between accuracy and computation. The second architecture is *SegFormer* (Xie et al., 2021), a well-known transformer-based segmentation network. Specifically, we opted for the version with the smallest encoder (SegFormer-B0) as it is reported to be suitable for real-time environments and closer to MiniNet-v2 in terms of computational requirements and number of parameters. Since both MiniNet-v2 and SegFormer were originally designed for processing only RGB images, we adapt them for multimodal segmentation by early-fusion both modalities. This early fusion involves concatenating the RGB and HSI images across the channel dimension before feeding them into the network (see path 1 in Figure 5.3). Finally, we evaluate *CMX* (J. Zhang et al., 2023), a hybrid fusion transformer based on SegFormer. This network receives RGB and HSI data separately, processing them with two encoders that share information throughout the network (see path 2 in Figure 5.3). This model also integrates feature rectification techniques to mitigate the effects of noisy measurements from different modalities, which is crucial in our case. It is noteworthy that, while CMX has been previously assessed with multiple modalities (depth, polarization, event and LiDAR), including multispectral ones (thermal and infrared) (K. Chen et al., 2023; Z. Wang et al., 2023; Yan et al., 2023), it has not been evaluated on HSI data before.

In summary, our pipeline offers flexibility in handling different data types. MiniNet-v2 and SegFormer can be used for processing individual RGB or HSI images (unimodal inputs). Additionally, all three architectures, i.e., MiniNet-v2, SegFormer, and CMX, can be employed for multimodal segmentation combining RGB and HSI data.

5.4.3 Label Transfer

To explore the unimodal HSI configurations of our pipeline, we address the challenge of data misalignment, i.e., the inaccurate alignment of manually RGB-annotated masks to the corresponding objects in HSI. This effect is due to the physical distance between the cameras and the heterogeneous internal settings of each sensor, which causes the perspective captured by each camera to differ. Since the employed cameras are non-conventional (line-scan cameras) and we only have two views with no additional spatial information such as depth, the well-established calibration methods for pinhole cameras are not applicable (Behmann et al., 2015). Therefore, to ensure consistency in annotated masks between modalities while improving labeling efficiency, a novel label transfer algorithm is proposed in this chapter. Our approach automatically adapts annotated segmentation masks from

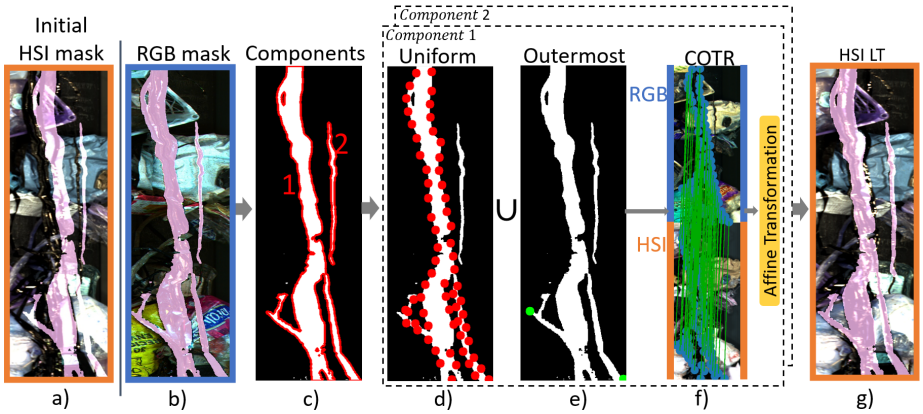


Figure 5.4: Visualization of the main steps of the proposed Label Transfer: a) initial mask superposed to HSI b) original manually annotated mask (in pink) superposed to RGB image; c) contour components extracted; d) uniformly-sampled points (red); e) outermost points (green); f) COTR matching for selected points in RGB (top) to the HSI image (bottom) used to compute an affine transformation for the mask; g) resulting mask after projection into the HSI image.

one camera (RGB) to another (HSI), relying exclusively on both images. The goal is to find affine transformations per mask that adapt the existing segmentation to the size and perspective of the corresponding object in the target image.

The different steps of the algorithm are visualized in Figure 5.4. First, we extract the contours of the segmentation mask and process each connected component independently. For example, in Figure 5.4c) two different local affine transformations are computed, one per component. Then, we sample each contour to obtain a sparse representation of the shape, generated by uniformly-sampled points (Figure 5.4d)), always including the outermost points for better representativeness (Figure 5.4e)). The next step involves identifying the corresponding points in the target image. To accomplish this, we leverage the feature matching system COTR (Jiang et al., 2021), which uses a transformer-based model for finding correspondences of a query point belonging to one image in another (Figure 5.4f)). Finally, the affine transformation is obtained from both sets of points. Figure 5.4g) illustrates the resulting mask transferred with the proposed algorithm, which compared to the initial one (Figure 5.4a)), shows a significant improvement in fitting the object.

5.5 Experiments

This section runs several experiments to evaluate the waste segmentation pipeline (Section 5.4) in the SpectralWaste dataset (Section 5.3).

5.5.1 Experimental Settings

Training Configuration. The training of our pipeline runs with the following configuration based on the recommendations from the original works (Alonso et al., 2020; Z. Wang et al., 2023; Xie et al., 2021): MiniNet-v2 uses the Adam optimizer with 1×10^{-3} initial learning rate, 1×10^{-4} weight decay and a polynomial schedule with the power set to 0.9. SegFormer models are trained using AdamW with 1×10^{-3} initial learning rate and a polynomial scheduler with 0.1 power. The selected CMX architecture is based on SegFormer-B0

	IoU ↑						
Alignment Method	*Film*	*Basket*	*Card*	*Tape*	*Filam.*	*Bag*	**mIoU ↑**
Manual Alignment (MA)	69.0	61.1	81.2	27.4	25.1	74.0	56.3
Label Transfer (LT)	78.7	78.3	93.5	57.1	81.2	88.8	79.6

Table 5.2: Evaluation of the annotations alignment between modalities of manual alignment (MA) and automatic label transfer (LA). Results are computed with the IoU of the obtained masks with a set of 81 instances manually labeled in a subset of 20 hyperspectral images.

			IoU ↑						
Backbone	**Modality**	**Labels**	*Film*	*Basket*	*Card.*	*Tape*	*Filam.*	*Bag*	**mIoU ↑**
MiniNet-v2	HYPER	MA	59.2	57.2	76.2	17.2	19.2	49.2	46.3
MiniNet-v2	HYPER	LT	61.2	61.0	78.8	28.8	30.5	56.3	52.8
MiniNet-v2	HYPER3	MA	56.8	52.8	62.9	19.2	6.9	45.9	40.7
MiniNet-v2	HYPER3	LT	58.8	61.9	69.4	30.2	23.0	50.5	49.0
SegFormer-B0	HYPER	MA	61.8	59.1	85.0	18.6	26.3	52.0	50.5
SegFormer-B0	HYPER	LT	65.4	63.2	85.2	21.9	33.1	57.2	54.3
SegFormer-B0	HYPER3	MA	56.6	55.4	84.7	14.4	27.5	46.7	47.5
SegFormer-B0	HYPER3	LT	60.4	58.4	86.6	22.6	43.0	49.9	53.5

Table 5.3: Hyperspectral image segmentation with different supervision: labels resulting from manual alignment (MA) and the proposed automatic label transfer (LT).

and is trained using AdamW with 1×10^{-3} initial learning rate and a polynomial schedule with power 0.9. In all cases, the loss is calculated using the softmax cross entropy function and the models are trained during 200 epochs. The classes are weighted in the loss calculation using pixel-level median frequency balancing. Regarding data augmentation, the training set is augmented with random rotations of ± 30 degrees and vertical and horizontal flips.

Metrics. To evaluate the implemented architectures, preprocessing steps, and fusion methods, the intersection over union (IoU) per class is computed. As the final metric, we present the mean of the classes (mIoU).

5.5.2 Label Transfer Evaluation

To obtain a reliable evaluation of the proposed algorithm for transferring labels, we manually annotate 20 hyperspectral images with 81 object instances, ensuring that all the classes appear in the set.

Table 5.2 presents the results of the proposed label transfer approach (LT) in comparison to the base manual alignment (MA) performed in the general preprocessing step. The manual alignment process involves cropping the excess image captured by each camera to align the space shown and resizing them to the same shape. The evaluation of both methods is based on the intersection over union (IoU) of the resulting masks with the set of 81 instances manually labeled in the hyperspectral images. In the case of big objects, i.e., film, basket, cardboard, and trash bag, label transfer demonstrates an improvement on every class ranging between 9.7 and 17.2% compared to the manual alignment, mainly

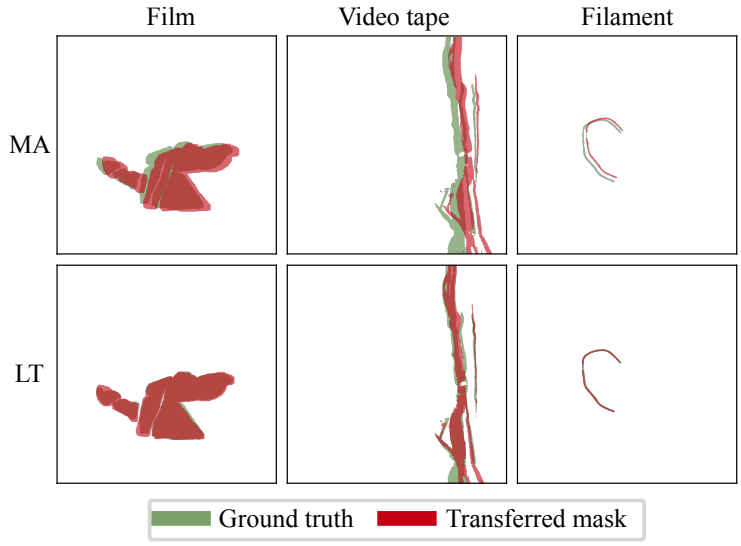


Figure 5.5: Qualitative results of the label transfer evaluation. The manually annotated masks are shown in green and the resulting mask from each method is in red. First row visualizes the proposed label transfer (LT) and second row the manual alignment (MA).

due to the high volume of the segmented objects. On the more complex classes with thin objects, i.e., video tape and filaments, gains increase to 29.7 and 56.1%, respectively. In fact, in these thin objects, the baseline does not even reach 28% of IoU with the ground truth. In summary, the proposed label transfer improves the average mIoU by 23.3% with respect to the baseline and achieves a mIoU of 79.6%.

Figure 5.5 shows qualitative results from both algorithms for the challenging thin classes (video tape and filaments) and the larger class in the dataset (film). The ground truth is shown in green and the resulting masks in red. The first row corresponds to MA and the second row to LT alignment. The visualization shows how the mask transferred with the proposed label transfer algorithm matches significantly better the ground truth than the manual alignment process.

In addition, we also evaluate the impact of training the unimodal models, MiniNet-v2 and SegFormer, with the annotations resulting from each of the alternative methods, MA and LT, as labels. Table 5.3 summarizes this study. The results demonstrate a higher mIoU using LT than MA in every configuration analyzed. Thus, we validate that our label transfer approach generates masks that better fit the objects in the target images, hyperspectral in our case, by reducing the noise of the segmentation annotations.

5.5.3 Segmentation Architectures Evaluation

Unimodal Segmentation Analysis. First we analyze the potential of using hyperspectral data for the object segmentation task. The same architectures are evaluated using hyperspectral or RGB data. The obtained results are contained in Table 5.4 and demonstrate the high value of the hyperspectral imaging information (HYPER) for object segmentation, achieving a higher mean intersection over union (mIoU) than conventional RGB images in both architectures, i.e., MiniNet-v2 and SegFormer.

Backbone	Modality	Fusion	IoU ↑						mIoU↑	I/s↑	P(M)↓	GFLOPs↓
			Film	*Basket*	*Card.*	*Tape*	*Filam.*	*Bag*				
MiniNet-v2	RGB	-	63.1	58.9	55.4	30.6	10.0	49.2	44.5	126.7	0.522	1.343
MiniNet-v2	HYPER	-	61.2	61.0	78.8	28.8	30.5	56.3	52.8	125.8	0.585	3.429
MiniNet-v2	HYPER3	-	58.8	61.9	69.4	30.2	23.0	50.5	49.0	125.9	0.522	1.431
MiniNet-v2	RGB-HYPER	early	67.3	59.1	82.6	24.1	6.1	55.4	49.1	124.6	0.586	3.457
MiniNet-v2	RGB-HYPER3	early	57.9	53.3	69.6	13.5	6.5	49.6	41.7	125.0	0.523	1.459
SegFormer-B0	RGB	-	66.9	71.3	48.9	33.6	15.2	54.6	48.4	156.2	3.716	3.508
SegFormer-B0	HYPER	-	65.4	63.2	85.2	21.9	33.1	57.2	54.3	152.2	4.062	6.347
SegFormer-B0	HYPER3	-	60.4	58.4	86.6	22.6	43.0	49.9	53.5	152.9	3.717	3.596
SegFormer-B0	RGB-HYPER	early	71.3	62.9	87.5	21.2	22.0	56.9	53.6	155.9	4.067	6.385
SegFormer-B0	RGB-HYPER3	early	57.7	59.2	80.9	10.2	34.4	48.6	48.5	157.1	3.721	3.634
CMX-B0	RGB-HYPER	hybrid	77.7	74.9	80.2	31.1	20.7	64.5	58.2	54.7	11.539	8.365
CMX-B0	RGB-HYPER3	hybrid	71.7	71.6	71.7	27.8	37.7	59.4	56.6	55.1	11.193	5.615

Table 5.4: Object segmentation evaluation on SpectralWaste dataset with different architectures (MiniNet-v2, SegFormer and CMX) and different modalities (RGB, HYPER, HYPER3, RGB-HYPER and RGB-HYPER3).

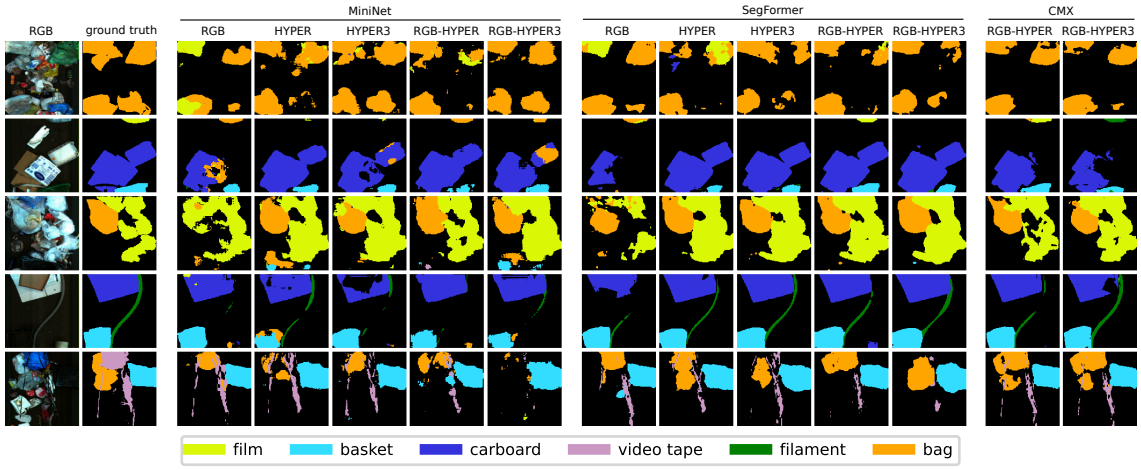


Figure 5.6: Qualitative results of the implemented architectures using the different modalities.

Another relevant aspect to note arises when comparing the results using all hyperspectral bands directly as input (HYPER) with the PCA 3-channels reduction (HYPER3). The good results achieved with the reduced input confirm the high variance covered with just 3 components in the dimensionality reduction performed with PCA.

Figure 5.6 shows some resulting masks for a qualitative comparison between the different architectures studied. For instance, note how the filament (fourth row and green mask) is barely identified by SegFormer using RGB data. However, the resulting mask using hyperspectral information is highly accurate for this challenging class.

Multimodal Segmentation Analysis. Regarding the assessment of the multimodal results summarized in Table 5.4, the CMX architecture with hybrid fusion outperforms the early fusion baselines. Considering that the images from both modalities are not perfectly aligned, it sounds reasonable that hybrid feature fusion is able to smooth out the noise and leverage the combination of information better than the early fusion methods.

Efficiency Analysis. This evaluation is conducted on a computer equipped with an AMD Ryzen 9 5950X CPU and a NVIDIA GeForce RTX 4090 GPU. Table 5.4 examines the computational load (GFLOPs) of each configuration when processing one image at a time. Measurements also include memory, i.e., number of parameters (P(M)), and inference throughput, i.e., images per second (I/s). Note that batch inference can improve the processing time per image. When HYPER3 is used, the overhead of the additional dimensionality reduction is considered. The throughput results showcase the real-time running capability of the implemented architectures while a significant increase in GFLOPs occurs when using the 224-channel hyperspectral images (HYPER). Employing HYPER data as input brings slight improvements in accuracy with respect to the reduced version HYPER3 in the analyzed configurations. However, the efficiency analysis clearly suggests that the reduced version of hyperspectral imaging is a better choice, offering the best trade-off between accuracy and computational load. Each of the architectures evaluated presents distinct trade-offs in performance and efficiency. MiniNet-v2 is designed to run efficiently on CPU, prioritizing low computational and memory demands (GFLOPs and number of parameters respectively). This design, optimized for CPU execution, results in slower processing speeds (images/s) on GPUs compared to SegFormer. Conversely, CMX requires significantly higher computational demands but surpasses both MiniNet-v2 and SegFormer in terms of mIoU, achieving the highest segmentation accuracy.

5.6 Conclusions

This chapter explores the advantages of using heterogeneous systems that combine data from multiple sensor modalities. Specifically, we analyze the benefits of leveraging both RGB and hyperspectral images for object segmentation in recycling facilities. Our research aids in mitigating two key challenges currently hindering the automation of sorting facilities: the scarcity of datasets and the high computational demands of HSI. Thus, this thesis introduces SpectralWaste, the first multimodal dataset collected in a real waste sorting facility comprising RGB and hyperspectral images. The installed cameras acquire the data synchronously, but due to their non-conventional features, there is no calibration between the gathered images. To fit the individual RGB-annotated mask onto their corresponding object in the hyperspectral images, the proposed label transfer algorithm has demonstrated its effectiveness improving up to 50.53% of mIoU with respect to the baseline in the most challenging classes. The presented dataset is a valuable resource for the research community, facilitating new developments to further streamline waste management processes, increasing the efficiency of sorting plants and ultimately benefiting both the environment and society. The low performance in segmenting some of the classes with current state-of-the-art architectures, remarks the open challenges and opportunities that the presented dataset poses. In the context of waste object segmentation, we have also proposed a pipeline that pays special attention to employing efficient architectures and exploiting the synergies between multiple sensing modalities. Our comprehensive study conducted with SpectralWaste highlights the potential of both multimodal and single hyperspectral imaging for object segmentation in waste facilities. The simple but effective reduction of hyperspectral data prior to model learning results is essential to maintaining a good efficiency-accuracy trade-off.

Chapter 6

Conclusions

Scene understanding is a broad and complex problem that can leverage the use of multi-camera systems for gathering richer and more comprehensive information. These systems facilitate the development of solutions for challenging tasks such as detection, tracking, or activity recognition, making them especially valuable in the setup of various applications including surveillance, autonomous vehicles, and monitoring. Despite their widespread use in real-world applications, multi-camera systems still face many open challenges in reaching their full potential. This thesis presents contributions aimed at improving the capabilities of existing systems:

Scalability. As the number of cameras on a system grows, bottlenecks arise because of resource constraints like processing power and storage capacity, leading to reconfiguration and infrastructure upgrades whenever a new device is added to the network.

Adaptability. Continuous environment monitoring demands algorithms robust to temporal variations. These algorithms must be able to dynamically update their internal knowledge while maintaining a balance between available resources and the volume of information to process.

Collaboration. Ensuring seamless collaboration among cameras to provide valuable information and an accurate understanding of the environment, requires efficient data fusion methods and sophisticated coordination algorithms.

The presented thesis addressed the aforementioned challenges by designing solutions that bring multi-camera scene understanding closer to real-world applications, including flexible algorithms for growing systems and novel knowledge collection. We specifically focus on three crucial tasks: distributed multi-target tracking, open-world person re-identification, and heterogeneous sensor collaboration. In the following, we summarize our contributions to each of these topics and discuss the conclusions drawn from the conducted studies.

6.1 Contributions

Distributed multi-target tracking

The limited flexibility of multi-agent systems to scale is a problem extensively tackled by developing distributed solutions. However, the majority of distributed multi-target tracking works mainly focus on theoretical concerns, like improving consensus on the target's

state (Shorinwa & Schwager, 2023), while assuming resolved the challenges associated with processing visual information.

This thesis, more concretely the research presented in Chapter 2, contributes to narrowing the gap between multi-target tracking applications and distributed setups. Our research includes the distributed visual understanding of the scene into the problem formulation, leveraging this knowledge for decision-making, such as determining when to communicate information (Casao et al., 2022), and providing a fully distributed approach for multi-target tracking (Casao et al., 2021). The presented approach processes all received data locally, offering a flexible solution that requires minimal effort to include new nodes in the system. Moreover, information is shared among cameras only when needed, maintaining the trade-off between efficient use of bandwidth to avoid communication bottlenecks and multi-target accuracy. Our analysis of results from multiple studies disclosed a key difference from theoretical works that assume perfect perception information. In real-world systems, the centralized setup, while theoretically considered the optimal value, may not always achieve the most accurate results. The presence of outliers, measurement errors, and noise in the data can hinder accurate fusion. Incorporating information from all nodes at every iteration may be less effective than reaching a consensus based exclusively on data from reliable nodes. Therefore, distributed setups can achieve comparable performance or even better to centralized systems under similar conditions (Y. Xu et al., 2016), although with extra effort for ensuring inter-camera consensus for both low- and high-level information.

Open-world person re-identification

Chapter 3 of this thesis focuses on the challenge of adapting the system's knowledge of the environment online. Specifically, we address the problem of person re-identification, where researchers commonly centered on obtaining the most representative features to correctly match a query person-of-interest in a gallery of known people (Hou et al., 2021; H. Luo et al., 2019). Other works extend this problem to scenarios in which the query may be unknown to the system, thus relaxing the assumption of knowing every person presented in the scene (Huang et al., 2020; Martini et al., 2020). However, no prior work has addressed the crucial issue of automatically building this gallery in a real-world monitoring system, where new individuals continuously appear and resources are limited.

In this thesis, we introduce a novel unsupervised construction method for creating the gallery in an open-world person re-identification setup (Casao, Azagra, et al., 2023). Our method is able to identify new people and adapt the appearance model of each individual that composes the gallery over time. Furthermore, the limited memory resource of the system is considered by constraining the size of the appearance models and selecting the most representative data seeking a balance between uncertainty and diversity of the samples saved. Through the performed experimentation, we realize that both metrics are critical in the creation of the gallery. Hence, galleries with poorly diversified models but high confidence in the samples belonging to the same person, tend to create new classes for already existing identities. In contrast, highly diverse models create uncertain galleries, leading to failures in the identification of new samples. Therefore, among the information selection methods studied, ours maintains the best balance between providing a correct identification of the samples and having a correct gallery structure, i.e., identifying all the people while reducing the number of redundant classes assigned to the same person.

Heterogeneous sensor collaboration

Finally, the last chapters of the presented thesis (Chapter 4 and Chapter 5) explore the benefits of introducing heterogeneous sensors into the camera network. Under the premise that only by allowing the collection of information different from that of the sensors already in use, i.e., RBG static cameras, can the system gain a deeper understanding of the scene, we study two types of heterogeneity: *static-mobile* cameras and *multi-modal* sensors.

Our collaborative framework, detailed in Chapter 4, introduces a pedestrian monitoring system and enables high-resolution capture of certain people's attributes through the cooperation of static and mobile cameras (Casao et al., 2024). While static cameras focus on tracking targets in the scene, mobile cameras (drones) leverage this information to obtain the best viewpoint for target attribute classification. Analysis conducted in a photo-realistic environment (Casao, Otero, et al., 2023) confirms that the quality of images captured up close by drones far surpasses those collected by distance static cameras. Our results demonstrate that the observed disparity in image quality leads directly to higher precision in attribute classification, especially in complex scenarios, as evidenced by the designed proof of concept. This enhancement in image quality has the potential to be crucial for other challenging tasks such as search and re-identification.

Chapter 5 of this thesis explores the advantages of combining RGB with hyperspectral sensors, which capture hundreds of narrow and contiguous spectral bands beyond the visible range. Our research focuses on the task of segmenting specific objects in a waste sorting facility, selected for their potential to either clog machinery or offer value through recovery (Casao, Peña, et al., 2023). Using the collected dataset as a benchmark to evaluate several baselines, reveals that hyperspectral imaging (HSI) provides precious information for object segmentation, yielding better results in unimodal than RGB images alone. Further analysis of both accuracy and inference times concludes with recommending the reduction of channel dimensionality in HSI data. While dimensionality reduction results in a slight decrease in accuracy, the significant drop in efficiency makes this strategy advantageous. Lastly, the best results are achieved by fusing RGB and HSI in a refined fashion that smooths out the noise between sensors, indicating that the acquired information presents complementary data that boosts the final performance.

6.2 Limitations and future work

Despite significant progress made in addressing the challenges of using multiple cameras for scene understanding, there is still considerable room for improvement in developing robust real-world applications. The methods proposed in this thesis for different tasks introduce solutions more flexible, adaptable to changes, and closer to practical application. However, some limitations remain, alongside opportunities to further extend our work.

For instance, our **distributed multi-target tracking approach** from Chapter 2 locally combines the received predicted state of targets from neighboring cameras, considering their covariance in the measure, but not incorporating the covariance of the prediction itself in the computed consensus. To enhance state estimation, we propose including more sophisticated consensus algorithms that account for both covariances in the information fusion (Sebastián et al., 2021), thereby improving tracking accuracy. Additionally, our proposed solution tracks each person in the scene individually based on their historical state.

However, the behavior and trajectories of people in the real world are often interconnected. While this topic has traditionally been studied in autonomous navigation for predicting future trajectories (Yuan et al., 2021), integrating this information into the tracking system could improve anticipation of sudden trajectory changes influenced by other agents, scenarios in which tracking systems currently exhibit a lack of robustness.

In the case of the proposed **self-adaptive gallery for open-world person re-identification** presented in Chapter 3, the main limitations arise from inherent challenges in re-identification systems, particularly in long-term person re-id when individuals undergo clothing changes or significant appearance alterations. In such scenarios, our system might mistakenly initiate new identities in the scene. To mitigate this issue, we could enhance the proposed method by incorporating features from neural networks designed to strengthen long-term person re-identification robustness (F. Liu et al., 2023; P. Xu & Zhu, 2023). These networks aim to disentangle identity from non-identity components, such as pose or texture, and learn discriminative features from unique elements like 3D body shape. Besides, privacy preservation is a growing concern affecting multiple applications, including surveillance systems. Therefore, our re-identification system could introduce an adaptation of people's appearance features to the acquired knowledge using federated incremental learning methods (Dong et al., 2022), thus improving feature description while preserving people's privacy.

Regarding our study of **heterogeneous sensor systems**, a crucial step for future development in the collaborative static and mobile camera network from Chapter 4, involves removing the only assumption made in the entire framework, i.e., known drones' position. Real-time drone position estimation could be achieved through the fusion of multiple sensor information sources, such as cameras (Isaac-Medina et al., 2021), LiDAR (Caballero & Merino, 2021), and GPS. Consequently, the proposed framework would address the primary challenges in real-world deployment and facilitate the transition from simulation to practical applications. Finally, our exploration of combining RGB and HSI for object segmentation, detailed in Chapter 5, highlights the potential of multimodal approaches in recycling processes. Future research could focus on exploiting the unlabeled set of our collected data using novel semi-supervised (Maheshwari et al., 2024) or unsupervised (Vobecky et al., 2022) multimodal methodologies. Moreover, the compression of hyperspectral imaging information, highly rich but complex, has been extensively explored for material identification, yet presents open challenges in object classification. Studying advanced techniques beyond Principal Component Analysis for compression and processing of the data (Dua et al., 2021) could enhance both the accuracy and efficiency of the final model.

Chapter 7

Conclusiones

La comprensión del entorno es un problema amplio y complejo que puede aprovechar el uso de sistemas multi-cámara para recopilar información más variada y completa. Estos sistemas facilitan el desarrollo de soluciones para tareas como la detección, el seguimiento de elementos móviles o el reconocimiento de actividades, convirtiéndolos así en recursos valiosos para el desarrollo de diversas aplicaciones, incluidas la vigilancia, los vehículos autónomos y la monitorización. A pesar de su uso generalizado en aplicaciones del mundo real, los sistemas multi-cámaras aún enfrentan muchos desafíos abiertos hasta alcanzar su máximo potencial. Esta tesis presenta contribuciones destinadas a mejorar las capacidades de los sistemas existentes:

Escalabilidad. A medida que aumenta el número de cámaras, surgen cuellos de botella debido a limitaciones de recursos como la potencia de procesamiento y la capacidad de almacenamiento, lo que lleva a la reconfiguración y actualizacion de la infraestructura cada vez que se agrega un nuevo dispositivo a la red.

Adaptabilidad. La monitorización continua del entorno exige algoritmos robustos a las variaciones temporales. Estos algoritmos deben poder actualizar dinámicamente su conocimiento interno manteniendo un equilibrio entre los recursos disponibles y el volumen de información a procesar.

Colaboración. Asegurar una colaboración fluida entre cámaras con el objetivo de proporcionar información valiosa y una comprensión precisa del entorno, requiere métodos eficientes de fusión de datos y algoritmos de coordinación sofisticados.

La tesis presentada aborda los desafíos mencionados anteriormente mediante el diseño de soluciones que acercan la comprensión del entorno con múltiples cámaras a las aplicaciones finales del mundo real, incluyendo algoritmos flexibles para incorporar fácilmente nuevas cámaras a la red y recopilación de conocimientos novedosos. Nos enfocamos específicamente en tres tareas cruciales: seguimiento distribuido de múltiples objetivos, re-identificación de personas en entornos abiertos y colaboración de sensores heterogéneos. A continuación, resumimos nuestras contribuciones a cada uno de estos temas y detallamos las conclusiones obtenidas de los estudios realizados.

7.1 Contribuciones

Seguimiento multiobjetivo distribuido

La limitada flexibilidad de los sistemas multiagente para escalar es un problema extensamente abordado mediante el desarrollo de soluciones distribuidas. Sin embargo, la mayoría de los trabajos de seguimiento distribuido de múltiples objetivos se centran principalmente en aspectos teóricos, como mejorar el consenso sobre el estado del objetivo (Shorinwa & Schwager, 2023), mientras asumen resueltos los desafíos asociados con el procesamiento de información visual.

Esta tesis, más concretamente la investigación presentada en el Capítulo 2, contribuye a estrechar la brecha entre las aplicaciones de seguimiento de múltiples objetivos y las configuraciones distribuidas. Nuestra investigación incluye la comprensión visual distribuida de la escena en la formulación del problema, proporcionando un enfoque completamente distribuido para el seguimiento de múltiples objetivos (Casao et al., 2021) y aprovechando este conocimiento para la toma de decisiones, como determinar cuándo comunicar información (Casao et al., 2022). El enfoque presentado procesa todos los datos recibidos localmente, ofreciendo una solución flexible que requiere un esfuerzo mínimo para incluir nuevos nodos en el sistema. Además, la información se comparte entre cámaras solo cuando es necesario, manteniendo el equilibrio entre el uso eficiente del ancho de banda, para evitar cuellos de botella de comunicación, y la precisión en el seguimiento de múltiples objetivos. Nuestro análisis de los resultados obtenidos en múltiples estudios expone una diferencia clave respecto a los trabajos teóricos que asumen información de percepción perfecta. En sistemas del mundo real, la configuración centralizada, aunque teóricamente considerada como el valor óptimo, igual no siempre logra los resultados más precisos. La presencia de valores atípicos, errores de medición y ruido en los datos puede obstaculizar una fusión precisa. De esta forma, incorporar información de todos los nodos en cada iteración puede ser menos efectivo que llegar a un consenso basado exclusivamente en datos de nodos fiables. Por lo tanto, las configuraciones distribuidas pueden lograr un rendimiento comparable o incluso mejor que los sistemas centralizados bajo condiciones similares (Y. Xu et al., 2016), aunque con un esfuerzo adicional para garantizar el consenso entre cámaras en información tanto de bajo como de alto nivel.

Re-identificación de personas en mundo abierto

El Capítulo 3 de esta tesis se centra en el desafío de adaptar el conocimiento del sistema sobre el entorno en línea. Específicamente, abordamos el problema de la re-identificación de personas, donde los investigadores comúnmente se centran en obtener las características más representativas para emparejar correctamente a una persona de interés con una galería de personas conocidas (Hou et al., 2021; H. Luo et al., 2019). Otros trabajos extienden este problema a escenarios en los que la consulta puede ser desconocida para el sistema, relajando así la suposición de conocer a todas las personas presentadas en la escena (Huang et al., 2020; Martini et al., 2020). Sin embargo, ningún trabajo previo ha abordado el problema crucial de construir automáticamente esta galería de forma no supervisada en un sistema de monitorización, donde continuamente aparecen nuevas personas y los recursos son limitados.

En esta tesis, introducimos un método novedoso de construcción no supervisada para crear la galería en un entorno de re-identificación de personas de mundo abierto (Casao, Azagra, et al., 2023). Nuestro método es capaz de identificar a nuevas personas y adaptar el modelo de apariencia de cada individuo que compone la galería a lo largo del tiempo. Además, se considera el recurso limitado de memoria del sistema mediante la definición de un tamaño máximo de los modelos de apariencia y la selección los datos más representativos, buscando un equilibrio entre la incertidumbre y la diversidad de las muestras guardadas. A través de la experimentación realizada, nos damos cuenta de que ambas métricas son críticas en la creación de la galería. De esta forma, las galerías con modelos poco diversificados pero con alta confianza de que las muestras pertenecen a la misma persona, tienden a crear nuevas clases para identidades ya existentes. Por otro lado, los modelos altamente diversos crean galerías inciertas, lo que lleva a fallos en la identificación de nuevas muestras. Por lo tanto, entre los métodos de selección de información estudiados, el nuestro mantiene el mejor equilibrio entre proporcionar una identificación correcta de las muestras y tener una estructura de galería correcta, es decir, identificar a todas las personas en la escena mientras se reduce el número de clases redundantes asignadas a la misma persona.

Colaboración de sensores heterogéneos

Finalmente, los últimos capítulos de la tesis presentada (Capítulo 4 y Capítulo 5) exploran los beneficios de introducir sensores heterogéneos en la red de cámaras. Bajo la premisa de que solo permitiendo la recolección de información diferente a la de los sensores ya en uso, es decir, cámaras estáticas RGB, el sistema puede obtener una comprensión más profunda de la escena, estudiamos dos tipos de heterogeneidad: cámaras *estáticas-móviles* y sensores *multi-modales*.

Nuestro entorno colaborativo, detallado en el Capítulo 4, introduce un sistema de monitorización peatonal que permite a su vez, la captura en alta resolución de ciertos atributos de las personas a través de la cooperación entre cámaras estáticas y móviles (Casao et al., 2024). Mientras que las cámaras estáticas se centran en el seguimiento de objetivos en la escena, las cámaras móviles (drones) aprovechan esta información para obtener el mejor punto de vista para la clasificación de atributos del objetivo. El análisis realizado en un simulador fotorealista (Casao, Otero, et al., 2023) confirma que la calidad de las imágenes capturadas de cerca por los drones supera con creces aquellas recopiladas por cámaras estáticas a distancia. Nuestros resultados demuestran que la disparidad observada en la calidad de la imagen conduce directamente a una mayor precisión en la clasificación de atributos, especialmente en escenarios complejos, como lo evidencia la prueba de concepto diseñada. Esta mejora en la calidad de la imagen tiene el potencial de ser crucial para otras tareas más desafiantes, como la búsqueda y re-identificación.

El Capítulo 5 de esta tesis explora las ventajas de combinar cámaras RGB con sensores hiperespectrales, los cuales capturan cientos de bandas espectrales estrechas y contiguas más allá del rango visible. Nuestra investigación se centra en la tarea de segmentar objetos específicos en una instalación de clasificación de residuos seleccionados por su potencial para obstruir maquinaria u ofrecer valor a través de la recuperación (Casao, Peña, et al., 2023). El uso del conjunto de datos recopilado como punto de referencia para evaluar varios métodos existentes, revela que las imagenes hiperespectrales (HSI) proporcionan información valiosa para la segmentación de objetos, obteniendo mejores resultados en

unimodal que las imágenes RGB solas. Como conclusión de un análisis adicional realizado de la precisión y los tiempos de inferencia, se recomienda la reducción de la dimensionalidad del canal en los datos HSI. Si bien la reducción de la dimensionalidad resulta en una ligera disminución en la precisión, la caída significativa de requerimientos en computación y memoria hace que esta estrategia sea ventajosa. Por último, los mejores resultados se logran fusionando RGB y HSI aplicando un suavizado del ruido existente entre los sensores, lo que indica que la información adquirida presenta datos complementarios que impulsan el rendimiento final.

7.2 Limitaciones y trabajo futuro

A pesar del progreso significativo logrado en abordar los desafíos derivados de utilizar múltiples cámaras para la comprensión de escenas, todavía existe un amplio margen de mejora en el desarrollo de aplicaciones robustas para el mundo real. Los métodos propuestos en esta tesis para diferentes tareas introducen soluciones más flexibles, adaptables a los cambios y más cercanas a la aplicación práctica. Sin embargo, algunas limitaciones persisten, junto con oportunidades para extender nuestro trabajo.

Por ejemplo, nuestro enfoque de **seguimiento distribuido de múltiples objetivos** del Capítulo 2, calcula localmente el consenso mediante la combinación del estado predicho de la cámara local con el recibido de las cámaras vecinas considerando la covarianza de la medición, pero no la covarianza de la predicción en sí misma. Para mejorar la estimación del estado, proponemos incluir algoritmos de consenso más sofisticados que tengan en cuenta ambas covarianzas en la fusión de información (Sebastián et al., 2021), mejorando así la precisión del seguimiento. Además, nuestra solución propuesta realiza el seguimiento individual de cada persona en la escena basándose en el histórico de su estado. Sin embargo, el comportamiento y las trayectorias de las personas en el mundo real suelen estar interconectados. Si bien este tema tradicionalmente se ha estudiado en la navegación autónoma para predecir las trayectorias futuras (Yuan et al., 2021), integrar esta información en el sistema de seguimiento podría mejorar la anticipación de cambios repentinos de trayectoria influenciados por otros agentes, escenarios en los que los sistemas de seguimiento actuales muestran una falta de robustez.

En el caso de la **galería autoadaptativa para la re-identificación de personas en mundo abierto** presentada en el Capítulo 3, las principales limitaciones surgen de los desafíos inherentes a los sistemas de re-identificación, particularmente a la re-identificación de personas a largo plazo cuando los individuos cambian de ropa o experimentan alteraciones significativas en la apariencia. En tales escenarios, nuestro sistema podría iniciar erróneamente nuevas identidades en la escena. Para mitigar este problema, podríamos mejorar el método propuesto incorporando características de redes neuronales diseñadas para fortalecer la robustez de la re-identificación de personas a largo plazo (F. Liu et al., 2023; P. Xu & Zhu, 2023). Estas redes tienen como objetivo desentrañar la identidad de los componentes no identitarios, como la pose o la textura, y aprender características discriminativas de elementos únicos como la forma corporal en 3D. Además, la preservación de la privacidad es una preocupación creciente que afecta a múltiples aplicaciones, incluidos los sistemas de vigilancia. Por lo tanto, nuestro sistema de re-identificación podría introducir una adaptación de las características de apariencia de las personas al conocimiento adquirido

mediante métodos de aprendizaje incremental federado (Dong et al., 2022), mejorando así la descripción de características mientras se preserva la privacidad de los individuos.

En cuanto a nuestro estudio de **sistemas de sensores heterogéneos**, un paso crucial para el desarrollo futuro en la red colaborativa de cámaras estáticas y móviles del Capítulo 4, implica eliminar el único supuesto realizado en todo el entorno, es decir, la posición conocida de los drones. La estimación en tiempo real de la posición de los drones podría lograrse mediante la fusión de múltiples fuentes de información, como cámaras (Isaac-Medina et al., 2021), LiDAR (Caballero & Merino, 2021), y GPS. En consecuencia, el entorno propuesto abordaría los desafíos principales de la implementación en el mundo real y facilitaría la transición de la simulación a las aplicaciones prácticas. Finalmente, de nuestra exploración de la combinación de RGB y HSI para la segmentación de objetos, detallada en el Capítulo 5, destaca el potencial de los enfoques multimodales en los procesos de reciclaje. La investigación futura podría centrarse en explotar el conjunto no etiquetado de nuestros datos recopilados utilizando metodologías multimodales novedosas semi-supervisadas (Maheshwari et al., 2024) o no supervisadas (Vobecky et al., 2022). Además, la compresión de la información contenida en las imágenes hiperespectrales, altamente ricas pero complejas, ha sido ampliamente explorada para la identificación de materiales, pero todavía presenta desafíos abiertos para la clasificación de objetos. Estudiar técnicas avanzadas más allá del Análisis de Componentes Principales para la compresión y procesamiento de los datos (Dua et al., 2021) podría mejorar tanto la precisión como la eficiencia del modelo final.

Bibliography

Adobe Systems Incorporated. mixamo, 2022. (n.d.).

Alcántara, A., Capitán, J., Cunha, R., & Ollero, A. (2021). Optimal trajectory planning for cinematography with multiple unmanned aerial vehicles. *Robotics and Autonomous Systems*, *140*.

Aldana-López, R., Aragüés, R., & Sagüés, C. (2023). Perception-latency aware distributed target tracking. *Information Fusion*, *99*.

Alonso, I., Riazuelo, L., & Murillo, A. C. (2020). MiniNet: An efficient semantic segmentation ConvNet for real-time robotic applications. *IEEE Transactions on Robotics*, *36*(4), 1340–1347.

Ardo, H., & Nilsson, M. (2019). Multi target tracking from drones by learning from generalized graph differences. *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*.

Atanasov, N., Sankaran, B., Le Ny, J., Pappas, G. J., & Daniilidis, K. (2014). Nonmyopic view planning for active object classification and pose estimation. *IEEE Transactions on Robotics*, *30*(5), 1078–1090.

Azagra, P., Civera, J., & Murillo, A. C. (2020). Incremental learning of object models from natural human–robot interactions. *IEEE Transactions on Automation Science and Engineering*, *17*, 1883–1900.

Bang, J., Kim, H., Yoo, Y., Ha, J.-W., & Choi, J. (2021). Rainbow memory: Continual learning with a memory of diverse samples. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8218–8227.

Bashkirova, D., Abdelfattah, M., Zhu, Z., Akl, J., Alladkani, F., Hu, P., Ablavsky, V., Calli, B., Bargal, S. A., & Saenko, K. (2022). ZeroWaste dataset: Towards deformable object segmentation in cluttered scenes. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21147–21157.

Bashkirova, D., Mishra, S., Lteif, D., Teterwak, P., Kim, D., Alladkani, F., Akl, J., Calli, B., Bargal, S. A., Saenko, K., Kim, D., Seo, M., Jeon, Y., Choi, D.-G., Ettedgui, S., Giryes, R., Abu-Hussein, S., Xie, B., & Li, S. (2022). VisDA 2022 challenge: Domain adaptation for industrial waste sorting. *NeurIPS 2022 Competition Track*, 104–118.

Battistelli, G., Chisci, L., & Selvi, D. (2018). A distributed kalman filter with event-triggered communication and guaranteed stability. *Automatica*, *93*, 75–82.

Bazzani, L., Cristani, M., Perina, A., Farenzena, M., & Murino, V. (2010). Multiple-shot person re-identification by hpe signature. *International Conference on Pattern Recognition*, 1413–1416.

Behmann, J., Mahlein, A.-K., Paulus, S., Kuhlmann, H., Oerke, E.-C., & Plümer, L. (2015). Calibration of hyperspectral close-range pushbroom cameras for plant phenotyping. *ISPRS Journal of Photogrammetry and Remote Sensing*, *106*, 172–182.

Bendale, A., & Boult, T. (2015). Towards open world recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1893–1902.

Berclaz, J., Fleuret, F., Turetken, E., & Fua, P. (2011). Multiple object tracking using k-shortest paths optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(9), 1806–1819.

Bergmann, P., Meinhardt, T., & Leal-Taixe, L. (2019). Tracking without bells and whistles. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 941–951.

Bernardin, K., & Stiefelhagen, R. (2008). Evaluating multiple object tracking performance: The clear mot metrics. *EURASIP Journal on Image and Video Processing*, *2008*, 1–10.

Bisagno, N., Conci, N., & Rinner, B. (2018). Dynamic camera network reconfiguration for crowd surveillance. *International Conference on Distributed Smart Cameras*, 1–6.

Byeon, M., Yoo, H., Kim, K., Oh, S., & Choi, J. Y. (2018). Unified optimization framework for localization and tracking of multiple targets with multiple cameras. *Computer Vision and Image Understanding*, *166*, 51–65.

Caballero, F., & Merino, L. (2021). Dll: Direct lidar localization. a map-based localization approach for aerial robots. *International Conference on Intelligent Robots and Systems*, 5491–5498.

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., & Sheikh, Y. (2019). Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(1), 172–186.

Carnegie Mellon University Graphics Lab. (n.d.). Cmu graphics lab motion capture database [The data used in this project was obtained from mocap.cs.cmu.edu. The database was created with funding from NSF EIA-0196217.].

Casao, S., Azagra, P., Murillo, A. C., & Montijano, E. (2023). A self-adaptive gallery construction method for open-world person re-identification. *Sensors*, *23*(5).

Casao, S., Murillo, A. C., & Montijano, E. (2022). Data association tools for target identification in distributed multi-target tracking systems. *Iberian Robotics conference*, 15–26.

Casao, S., Naya, A., Murillo, A. C., & Montijano, E. (2021). Distributed multi-target tracking in camera networks. *International Conference on Robotics and Automation*, 1903–1909.

Casao, S., Otero, A., Serra-Gómez, Á., Murillo, A. C., Alonso-Mora, J., & Montijano, E. (2023). A framework for fast prototyping of photo-realistic environments with multiple pedestrians. *International Conference on Robotics and Automation*.

Casao, S., Peña, F., Sabater, A., Rosa, C., Suárez, D., Montijano, E., & Murillo, A. C. (2023). Spectralwaste dataset: Multimodal data for waste segmentation. *Under Review*.

Casao, S., Serra-Gómez, Á., Murillo, A. C., Böhmer, W., Alonso-Mora, J., & Montijano, E. (2024). Distributed multi-target tracking and active perception with mobile camera networks. *Computer Vision and Image Understanding*, *238*.

Castro, F. M., Marín-Jiménez, M. J., Guil, N., Schmid, C., & Alahari, K. (2018). End-to-end incremental learning. *Proceedings of the European Conference on Computer Vision*, 233–248.

Chandra, R., Bhattacharya, U., Bera, A., & Manocha, D. (2019). Densepeds: Pedestrian tracking in dense crowds using front-rvo and sparse features. *International Conference on Intelligent Robots and Systems*, 468–475.

Chang, F.-M., Lian, F.-L., & Chou, C.-C. (2015). Integration of modified inverse observation model and multiple hypothesis tracking for detecting and tracking humans. *IEEE Transactions on Automation Science and Engineering*, *13*(1), 160–170.

Chavdarova, T., Baqué, P., Bouquet, S., Maksai, A., Jose, C., Bagautdinov, T., Lettry, L., Fua, P., Van Gool, L., & Fleuret, F. (2018). Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5030–5039.

Chen, H., Wang, Y., Lagadec, B., Dantcheva, A., & Bremond, F. (2022). Learning invariance from generated variance for unsupervised person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Chen, J.-H., & Song, K.-T. (2018). Collision-free motion planning for human-robot collaborative safety under cartesian constraint. *International Conference on Robotics and Automation*.

Chen, K., Liu, J., & Zhang, H. (2023). IGT: Illumination-guided RGB-T object detection with transformers. *Knowledge-Based Systems*, *268*.

Chen, T., Ding, S., Xie, J., Yuan, Y., Chen, W., Yang, Y., Ren, Z., & Wang, Z. (2019). Abd-net: Attentive but diverse person re-identification. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8351–8361.

Chen, Z., Scott, T. R., Bearman, S., Anand, H., Keating, D., Scott, C., Arrowsmith, J. R., & Das, J. (2020). Geomorphological analysis using unpiloted aircraft systems, structure from motion, and deep learning. *International Conference on Intelligent Robots and Systems*, 1276–1283.

Cheng, Y., Zhu, J., Jiang, M., Fu, J., Pang, C., Wang, P., Sankaran, K., Onabola, O., Liu, Y., Liu, D., & Bengio, Y. (2021). FloW: A dataset and benchmark for floating waste detection in inland waters. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10933–10942.

da Costa, V. G. T., Zara, G., Rota, P., Oliveira-Santos, T., Sebe, N., Murino, V., & Ricci, E. (2022). Dual-head contrastive domain adaptation for video action recognition. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1181–1190.

De Feyter, F., Van Beeck, K., & Goedemé, T. (2019). Enhancing open-set face recognition by closing it with cluster-inferred gallery augmentation. *Asian Conference on Pattern Recognition*, 15–26.

DeCann, B., & Ross, A. (2015). Modelling errors in a biometric re-identification system. *IET Biometrics*, *4*(4), 209–219.

de Langis, K., & Sattar, J. (2020). Realtime multi-diver tracking and re-identification for underwater human-robot collaboration. *International Conference on Robotics and Automation*, 11140–11146.

Dendorfer, P., Osep, A., Milan, A., Schindler, K., Cremers, D., Reid, I., Roth, S., & Leal-Taixé, L. (2021). Motchallenge: A benchmark for single-camera multiple target tracking. *International Journal of Computer Vision*, *129*(4), 845–881.

Dong, J., Wang, L., Fang, Z., Sun, G., Xu, S., Wang, X., & Zhu, Q. (2022). Federated class-incremental learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10164–10173.

Dua, Y., Singh, R. S., Parwani, K., Lunagariya, S., & Kumar, V. (2021). Convolution neural network based lossy compression of hyperspectral images. *Signal Processing: Image Communication*, *95*.

Epic Games Incorporated. unreal engine, 2022. (n.d.).

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *kdd*, *96*, 226–231.

Fabbri, M., Lanzi, F., Calderara, S., Palazzi, A., Vezzani, R., & Cucchiara, R. (2018). Learning to detect and track visible and occluded body joints in a virtual world. *European Conference on Computer Vision*, 430–446.

Fabiyi, S. D., Vu, H., Tachtatzis, C., Murray, P., Harle, D., Dao, T. K., Andonovic, I., Ren, J., & Marshall, S. (2020). Varietal classification of rice seeds using RGB and hyperspectral images. *IEEE Access*, *8*, 22493–22505.

Feng, H., Chen, M., Hu, J., Shen, D., Liu, H., & Cai, D. (2021). Complementary pseudo labels for unsupervised domain adaptation on person re-identification. *IEEE Transactions on Image Processing*, *30*, 2898–2907.

Ferraguti, F., Landi, C. T., Costi, S., Bonfè, M., Farsoni, S., Secchi, C., & Fantuzzi, C. (2020). Safety barrier functions and multi-camera tracking for human–robot shared environment. *Robotics and Autonomous Systems*, *124*.

Fleuret, F., Berclaz, J., Lengagne, R., & Fua, P. (2007). Multicamera people tracking with a probabilistic occupancy map. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(2), 267–282.

Fontanel, D., Cermelli, F., Mancini, M., Bulo, S. R., Ricci, E., & Caputo, B. (2020). Boosting deep open world recognition by clustering. *IEEE Robotics and Automation Letters*, *5*, 5985–5992.

Ge, X., Han, Q.-L., Zhang, X.-M., Ding, L., & Yang, F. (2019). Distributed event-triggered estimation over sensor networks: A survey. *IEEE Transactions on Cybernetics*, *50*(3), 1306–1320.

Gheissari, N., Sebastian, T. B., & Hartley, R. (2006). Person re-identification using spatio-temporal appearance. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, *2*, 1528–1535.

Grewal, R., Kasana, S. S., & Kasana, G. (2023). Hyperspectral image segmentation: A comprehensive survey. *Multimedia Tools and Applications*, *82*, 20819–10872.

Guo, Y., Liu, Z., Luo, H., Pu, H., & Tan, J. (2022). Multi-person multi-camera tracking for live stream videos based on improved motion model and matching cascade. *Neurocomputing*, *492*, 561–571.

Haarnoja, T., Tang, H., Abbeel, P., & Levine, S. (2017). Reinforcement learning with deep energy-based policies. *International Conference on Machine Learning*.

Habili, N., Kwan, E., Li, W., Webers, C., Oorloff, J., Armin, M. A., & Petersson, L. (2022). A hyperspectral and RGB dataset for building façade segmentation. *European Conference on Computer Vision*, 258–267.

Han, H., Zhou, M., Shang, X., Cao, W., & Abusorrah, A. (2020). Kiss+ for rapid and accurate pedestrian re-identification. *IEEE Transactions on Intelligent Transportation Systems*, *22*, 394–403.

Hanson, N., Lewis, W., Puthuveetil, K., Furline, D., Padmanabha, A., Padir, T., & Erickson, Z. (2023). SLURP! spectroscopy of liquids using robot pre-touch sensing. *International Conference on Robotics and Automation*, 3786–3792.

Hayes, T. L., Cahill, N. D., & Kanan, C. (2019). Memory efficient experience replay for streaming learning. *International Conference on Robotics and Automation*, 9769–9776.

He, L., Liu, G., Tian, G., Zhang, J., & Ji, Z. (2019). Efficient multi-view multi-target tracking using a distributed camera network. *IEEE Sensors Journal*.

Henriksen, M. L., Karlsen, C. B., Klarskov, P., & Hinge, M. (2022). Plastic classification via in-line hyperspectral camera analysis and unsupervised machine learning. *Vibrational Spectroscopy*, *118*.

Hill, T., & Miller, J. (2011). How to combine independent data sets for the same quantity. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *21*(3).

Hirzer, M., Beleznai, C., Roth, P. M., & Bischof, H. (2011). Person re-identification by descriptive and discriminative classification. *Scandinavian Conference on Image Analysis*, 91–102.

Hou, R., Ma, B., Chang, H., Gu, X., Shan, S., & Chen, X. (2021). Feature completion for occluded person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(9), 4894–4912.

Huang, Y., Zha, Z.-J., Fu, X., Hong, R., & Li, L. (2020). Real-world person re-identification via degradation invariance learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14084–14094.

Isaac-Medina, B. K. S., Poyser, M., Organisciak, D., Willcocks, C. G., Breckon, T. P., & Shum, H. P. H. (2021). Unmanned aerial vehicle visual detection and tracking using deep neural networks: A performance benchmark. *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 1223–1232.

Jeon, B. F., Shim, D., & Jin Kim, H. (2020). Detection-aware trajectory generation for a drone cinematographer. *International Conference on Intelligent Robots and Systems*, 1450–1457.

Jiang, W., Trulls, E., Hosang, J., Tagliasacchi, A., & Yi, K. M. (2021). COTR: Correspondence transformer for matching across images. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6207–6217.

Kamal, A. T., Bappy, J. H., Farrell, J. A., & Roy-Chowdhury, A. K. (2015). Distributed multi-target tracking and data association in vision networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(7), 1397–1410.

Kamal, A. T., Farrell, J. A., & Roy-Chowdhury, A. K. (2012). Information weighted consensus. *IEEE Conference on Decision and Control*, 2732–2737.

Karaca, A. C., Ertürk, A., Güllü, M. K., Elmas, M., & Ertürk, S. (2013). Automatic waste sorting using shortwave infrared hyperspectral imaging system. *Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*.

Kent, D., & Chernova, S. (2020). Human-centric active perception for autonomous observation. *International Conference on Robotics and Automation*, 1785–1791.

Kerim, A., Celikcan, U., Erdem, E., & Erdem, A. (2021). Using synthetic data for person tracking under adverse weather conditions. *Image and Vision Computing*, *111*.

Kim, J., Jeon, M.-H., Jung, S., Yang, W., Jung, M., Shin, J., & Kim, A. (2023). Transpose: Large-scale multispectral dataset for transparent object. *The International Journal of Robotics Research*.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. (2023). Segment anything. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Kodgule, S., Candela, A., & Wettergreen, D. (2019). Non-myopic planetary exploration combining in situ and remote measurements. *International Conference on Intelligent Robots and Systems*, 536–543.

Kruse, F. A., Lefkoff, A. B., Boardman, J. W., Heidebrecht, K. B., Shapiro, A. T., Barloon, P. J., & Goetz, A. F. H. (1993). The spectral image processing system (SIPS)—interactive visualization and analysis of imaging spectrometer data. *Remote Sensing of Environment*, *44*, 145–163.

Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval research logistics quarterly*, *2*(1-2), 83–97.

Le, Q. C., Conte, D., & Hidane, M. (2018). Online multiple view tracking: Targets association across cameras. *Workshop on Activity Monitoring by Multiple Distributed Sensing*.

Leal-Taixé, L., Milan, A., Reid, I., Roth, S., & Schindler, K. (2015). Motchallenge 2015: Towards a benchmark for multi-target tracking. *arXiv preprint arXiv:1504.01942*.

Lee, J., Hur, J., Hwang, I., & Kim, Y. M. (2022). MasKGrasp: Mask-based grasping for scenes with multiple general real-world objects. *International Conference on Intelligent Robots and Systems*, 3137–3144.

Leng, Q., Ye, M., & Tian, Q. (2019). A survey of open-world person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, *30*, 1092–1108.

Li, J., Xu, J., Zhong, F., Kong, X., Qiao, Y., & Wang, Y. (2020). Pose-assisted multi-camera collaboration for active object tracking. *AAAI Conference on Artificial Intelligence*, *34*(01), 759–766.

Li, X., Su, Y., Liu, Y., Zhai, S., & Wu, Y. (2018). Active target tracking: A simplified view aligning method for binocular camera model. *Computer Vision and Image Understanding*, *175*, 11–23.

Li, Z., Liang, Y., Xu, L., & Ma, S. (2023). Distributed extended object tracking information filter over sensor networks. *International Journal of Robust and Nonlinear Control*, *33*(2), 1122–1149.

Liang, E., et al. (2018). RLlib: Abstractions for distributed reinforcement learning. *International Conference on Machine Learning*.

Lin, Y., Dong, X., Zheng, L., Yan, Y., & Yang, Y. (2019). A bottom-up clustering approach to unsupervised person re-identification. *AAAI Conference on Artificial Intelligence*, *33*(1), 8738–8745.

Lin, Y., Xie, L., Wu, Y., Yan, C., & Tian, Q. (2020). Unsupervised person re-identification via softened similarity learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3390–3399.

Linder, T., Vaskevicius, N., Schirmer, R., & Arras, K. O. (2021). Cross-modal analysis of human detection for robotics: An industrial case study. *International Conference on Intelligent Robots and Systems*, 971–978.

Liu, F., Kim, M., Gu, Z., Jain, A., & Liu, X. (2023). Learning clothing and pose invariant 3d shape representation for long-term person re-identification. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19617–19626.

Liu, P., Liu, X., Yan, J., & Shao, J. (2018). Localization guided learning for pedestrian attribute recognition. *British Machine Vision Conference*.

Liu, Z., Wang, X., Zhong, Y., Shu, M., & Sun, C. (2022). SiamHYPER: Learning a hyperspectral object tracker from an RGB-based tracker. *IEEE Transactions on Image Processing*, *31*, 7116–7129.

Lu, Y., Fan, Y., Deng, B., Liu, F., Li, Y., & Wang, S. (2023). VL-Grasp: A 6-Dof interactive grasp policy for language-oriented objects in cluttered indoor scenes. *International Conference on Intelligent Robots and Systems*, 976–983.

Luo, H., Jiang, W., Zhang, X., Fan, X., Qian, J., & Zhang, C. (2019). Alignedreid++: Dynamically matching local information for person re-identification. *Pattern Recognition*, *94*, 53–61.

Luo, W., Sun, P., Zhong, F., Liu, W., Zhang, T., & Wang, Y. (2019). End-to-end active object tracking and its real-world deployment via reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(6), 1317–1332.

Maheshwari, H., Liu, Y.-C., & Kira, Z. (2024). Missing modality robustness in semi-supervised multi-modal semantic segmentation. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1020–1030.

Makehuman community. makehuman, 2022. (n.d.).

Mancini, M., Karaoguz, H., Ricci, E., Jensfelt, P., & Caputo, B. (2019). Knowledge is never enough: Towards web aided deep open world recognition. *International Conference on Robotics and Automation*, 9537–9543.

Martinel, N., Das, A., Micheloni, C., & Roy-Chowdhury, A. K. (2016). Temporal model adaptation for person re-identification. *European Conference on Computer Vision*, 858–877.

Martini, M., Paolanti, M., & Frontoni, E. (2020). Open-world person re-identification with rgbd camera in top-view configuration for retail applications. *IEEE Access*, *8*, 67756–67765.

Mekonnen, A. A., Lerasle, F., & Herbulot, A. (2013). Cooperative passers-by tracking with a mobile robot and external cameras. *Computer Vision and Image Understanding*, *117*(10), 1229–1244.

Milan, A., Leal-Taixé, L., Reid, I., Roth, S., & Schindler, K. (2016). Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*.

Miller, I. D., Cladera, F., Smith, T., Taylor, C. J., & Kumar, V. (2022). Stronger together: Air-ground robotic collaboration using semantics. *IEEE Robotics and Automation Letters*, *7*(4), 9643–9650.

Mohamed, S. C., Rajaratnam, S., Hong, S. T., & Nejat, G. (2019). Person finding: An autonomous robot search method for finding multiple dynamic users in human-centered environments. *IEEE Transactions on Automation Science and Engineering*, *17*(1), 433–449.

Nalepa, J., Myller, M., & Kawulok, M. (2019). Validating hyperspectral image segmentation. *IEEE Geoscience and Remote Sensing Letters*, *16*(8), 1264–1268.

Olfati-Saber, R. (2007). Distributed kalman filtering for sensor networks. *IEEE Conference on Decision and Control*, 5492–5498.

Panda, R., Bhuiyan, A., Murino, V., & Roy-Chowdhury, A. K. (2019). Adaptation of person re-identification models for on-boarding new camera(s). *Pattern Recognition*, *96*.

Patten, T., Martens, W., & Fitch, R. (2018). Monte carlo planning for active object classification. *Autonomous Robots*, *42*(02), 391–421.

Patten, T., Zillich, M., Fitch, R. C., Vincze, M., & Sukkarieh, S. (2016). Viewpoint evaluation for online 3-d active object classification. *IEEE Robotics and Automation Letters*, *1*(1), 73–81.

Popović, M., Hitz, G., Nieto, J., Sa, I., Siegwart, R., & Galceran, E. (2017). Online informative path planning for active classification using uavs. *IEEE International Conference on Robotics and Automation*, 5753–5758.

Proença, P. F., & Simões, P. (2020). TACO: Trash annotations in context for litter detection [arXiv preprint arXiv:2003.06975].

Pueyo, P., Montijano, E., Murillo, A. C., & Schwager, M. (2022). Cinempc: Controlling camera intrinsics and extrinsics for autonomous cinematography. *International Conference on Robotics and Automation*, 4058–4064.

Qin, H., Zhou, W., Yao, Y., & Wang, W. (2022). Individual tree segmentation and tree species classification in subtropical broadleaf forests using UAV-based LiDAR, hyperspectral, and ultrahigh-resolution RGB data. *Remote Sensing of Environment*, *280*.

Quach, K. G., Nguyen, P., Le, H., Truong, T.-D., Duong, C. N., Tran, M.-T., & Luu, K. (2021). Dyglip: A dynamic graph model with link prediction for accurate multi-camera multiple object tracking. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13784–13793.

Quan, R., Dong, X., Wu, Y., Zhu, L., & Yang, Y. (2019). Auto-reid: Searching for a part-aware convnet for person re-identification. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3750–3759.

Rajasegaran, J., Pavlakos, G., Kanazawa, A., & Malik, J. (2022). Tracking people by predicting 3d appearance, location and pose. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2740–2749.

Rao, D., Visin, F., Rusu, A., Pascanu, R., Teh, Y. W., & Hadsell, R. (2019). Continual unsupervised representation learning. *Advances in Neural Information Processing Systems*, *32*, 7647–7657.

Rebuffi, S.-A., Kolesnikov, A., Sperl, G., & Lampert, C. H. (2017). Icarl: Incremental classifier and representation learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2001–2010.

Ristani, E., Solera, F., Zou, R., Cucchiara, R., & Tomasi, C. (2016a). Performance measures and a data set for multi-target, multi-camera tracking. *European Conference on Computer Vision*, 17–35.

Ristani, E., Solera, F., Zou, R., Cucchiara, R., & Tomasi, C. (2016b). Performance measures and a data set for multi-target, multi-camera tracking. *European Conference on Computer Vision*, 17–35.

Ristani, E., & Tomasi, C. (2018). Features for multi-target multi-camera tracking and re-identification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6036–6046.

Robin, C., & Lacroix, S. (2016). Multi-robot target detection and tracking: Taxonomy and survey. *Autonomous Robots*, *40*(4), 729–760.

Schlichtkrull, M., Kipf, T., Bloem, P., Berg, R., Titov, I., & Welling, M. (2018). Modeling relational data with graph convolutional networks. *Extended Semantic Web Conference*, 593–607.

Schranz, M., & Andre, T. (2018). Towards resource-aware hybrid camera systems. *International Conference on Distributed Smart Cameras*, 1–7.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *ArXiv*, *abs/1707.06347*.

Sebastián, E., Montijano, E., & Sagüés, C. (2021). All-in-one: Certifiable optimal distributed kalman filter under unknown correlations. *Conference on Decision and Control*, 6578–6583.

Seidlitz, S., Sellner, J., Odenthal, J., Özdemir, B., Studier-Fischer, A., Knödler, S., Ayala, L., Adler, T. J., Kenngott, H. G., Tizabi, M., et al. (2022). Robust deep learning-based semantic organ segmentation in hyperspectral images. *Medical Image Analysis*, *80*.

Serra-Gómez, Á., Montijano, E., Böhmer, W., & Alonso-Mora, J. (2023). Active classification of moving targets with learned control policies. *IEEE Robotics and Automation Letters*, *8*(6), 3717–3724.

Shah, S., Dey, D., Lovett, C., & Kapoor, A. (2018). Airsim: High-fidelity visual and physical simulation for autonomous vehicles. *Field and service robotics*, 621–635.

Sharma, S., Ansari, J. A., Murthy, J. K., & Krishna, K. M. (2018). Beyond pixels: Leveraging geometry and shape cues for online multi-object tracking. *International Conference on Robotics and Automation*, 3508–3515.

Shiddiq, M., Arief, D. S., Zulfansyah, Fatimah, K., Wahyudi, D., Mahmudah, D. A., Putri, D. K. E., Husein, I. R., & Ningsih, S. A. (2023). Plastic and organic waste identification using multispectral imaging. *Materials Today: Proceedings*, *87*, 338–344.

Shorinwa, O., & Schwager, M. (2023). Distributed target tracking in multi-agent networks via sequential quadratic alternating direction method of multipliers. *American Control Conference*, 341–348.

Shorinwa, O., Yu, J., Halsted, T., Koufos, A., & Schwager, M. (2020). Distributed multi-target tracking for autonomous vehicle fleets. *International Conference on Robotics and Automation*, 3495–3501.

Shree, V., Chao, W.-L., & Campbell, M. (2020). Interactive natural language-based person search. *IEEE Robotics and Automation Letters*, *5*(2), 1851–1858.

Sock, J., Garcia-Hernando, G., & Kim, T.-K. (2020). Active 6d multi-object pose estimation in cluttered scenarios with deep reinforcement learning. *International Conference on Intelligent Robots and Systems*, 10564–10571.

Soto, C., Song, B., & Roy-Chowdhury, A. K. (2009). Distributed multi-target tracking in a self-configuring camera network. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1486–1493.

Sridhar Raj S, M. V. P., & Balakrishnan, R. (2022). Spatio-temporal association rule based deep annotation-free clustering (star-dac) for unsupervised person re-identification. *Pattern Recognition*, *122*.

Sukno, M., & Palunko, I. (2022). Hand-crafted features for floating plastic detection. *International Conference on Intelligent Robots and Systems*, 3378–3383.

Tallamraju, R., Price, E., Ludwig, R., Karlapalem, K., Bülthoff, H. H., Black, M. J., & Ahmad, A. (2019). Active perception based formation control for multiple aerial vehicles. *IEEE Robotics and Automation Letters*, *4*(4), 4491–4498.

Tesfaye, Y. T., Zemene, E., Prati, A., Pelillo, M., & Shah, M. (2019). Multi-target tracking in multiple non-overlapping cameras using fast-constrained dominant sets. *International Journal of Computer Vision*, *127*(9), 1303–1320.

Trujillo, J.-C., Munguía, R., Ruiz-Velázquez, E., & Castillo-Toledo, B. (2019). A cooperative aerial robotic approach for tracking and estimating the 3d position of a moving object by using pseudo-stereo vision. *Journal of Intelligent & Robotic Systems*, *96*, 297–313.

Truong, X.-T., & Ngo, T. D. (2017). Toward socially aware robot navigation in dynamic and crowded environments: A proactive social motion model. *IEEE Transactions on Automation Science and Engineering*, *14*(4), 1743–1760.

Tsai, R. Y.-C., Ke, H. T.-Y., Lin, K. C.-J., & Tseng, Y.-C. (2019). Enabling identity-aware tracking via fusion of visual and inertial features. *International Conference on Robotics and Automation*, 2260–2266.

Valada, A., Vertens, J., Dhall, A., & Burgard, W. (2017). AdapNet: Adaptive semantic segmentation in adverse environmental conditions. *International Conference on Robotics and Automation*, 4644–4651.

Valipour, S., Perez, C., & Jagersand, M. (2017). Incremental learning for robot perception through hri. *International Conference on Intelligent Robots and Systems*, 2772–2777.

Vaswani, A., et al. (2017). Attention is all you need. *Advances in Neural Information Processing System*, *30*.

Virgona, A., Alempijevic, A., & Vidal-Calleja, T. (2018). Socially constrained tracking in crowded environments using shoulder pose estimates. *International Conference on Robotics and Automation*.

Vobecky, A., Hurych, D., Siméoni, O., Gidaris, S., Bursuc, A., Pérez, P., & Sivic, J. (2022). Drive&segment: Unsupervised semantic segmentation of urban scenes via cross-modal distillation. *European Conference on Computer Vision*, 478–495.

Vorbach, C., Hasani, R., Amini, A., Lechner, M., & Rus, D. (2021). Causal navigation by continuous-time neural networks. *Advances in Neural Information Processing Systems*, *34*, 12425–12440.

Wang, D., & Zhang, S. (2020). Unsupervised person re-identification via multi-label classification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10981–10990.

Wang, G., Lai, J.-H., Liang, W., & Wang, G. (2020). Smoothing adversarial domain attack and p-memory reconsolidation for cross-domain person re-identification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10568–10577.

Wang, H., Shen, J., Liu, Y., Gao, Y., & Gavves, E. (2022). Nformer: Robust person re-identification with neighbor transformer. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7297–7307.

Wang, L., Wang, Z., Han, Q.-L., & Wei, G. (2017). Event-based variance-constrained H_∞ filtering for stochastic parameter systems over sensor networks with successive missing measurements. *Transactions on Cybernetics*, *48*(3), 1007–1017.

Wang, Z., Colonnier, F., Zheng, J., Acharya, J., Jiang, W., & Huang, K. (2023). TIRDet: Mono-modality thermal infrared object detection based on prior thermal-to-visible translation. *ACM International Conference on Multimedia*, 2663–2672.

Wang, Z., Jiang, J., Yu, Y., & Satoh, S. (2019). Incremental re-identification by cross-direction and cross-ranking adaption. *IEEE Transactions on Multimedia*, *21*, 2376–2386.

Wei, L., Zhang, S., Gao, W., & Tian, Q. (2018). Person transfer gan to bridge domain gap for person re-identification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 79–88.

Wen, L., Lei, Z., Chang, M.-C., Qi, H., & Lyu, S. (2017). Multi-camera multi-target tracking with space-time-view hyper-graph. *International Journal of Computer Vision*, *122*(2), 313–333.

Wen, Z., Sun, M., Li, Y., Ying, S., & Peng, Y. (2019). Asymmetric local metric learning with psd constraint for person re-identification. *International Conference on Robotics and Automation*, 4862–4868.

Wendel, A., & Underwood, J. (2016). Self-supervised weed detection in vegetable crops using ground based hyperspectral imaging. *International Conference on Robotics and Automation*, 5128–5135.

Wojke, N., Bewley, A., & Paulus, D. (2017). Simple online and realtime tracking with a deep association metric. *International Conference on Image Processing*, 3645–3649.

Wu, D., Zheng, S.-J., Zhang, X.-P., Yuan, C.-A., Cheng, F., Zhao, Y., Lin, Y.-J., Zhao, Z.-Q., Jiang, Y.-L., & Huang, D.-S. (2019). Deep learning-based methods for person re-identification: A comprehensive review. *Neurocomputing*, *337*, 354–371.

Wu, Y., Lin, Y., Dong, X., Yan, Y., Ouyang, W., & Yang, Y. (2018). Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5177–5186.

Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., & Girshick, R. (2019). Detectron2.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, *34*, 12077–12090.

Xompero, A., & Cavallaro, A. (2022). Cross-camera view-overlap recognition. *European Conference on Computer Vision*, 253–269.

Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, *2*, 165–193.

Xu, P., & Zhu, X. (2023). Deepchange: A long-term person re-identification benchmark with clothes change. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11196–11205.

Xu, Q., et al. (2021). Towards efficient multiview object detection with adaptive action prediction. *International Conference on Robotics and Automation*, 13423–13429.

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, *16*(3), 645–678.

Xu, Y., Liu, X., Liu, Y., & Zhu, S.-C. (2016). Multi-view people tracking via hierarchical trajectory composition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4256–4265.

Yan, C., Zhang, H., Li, X., Yang, Y., & Yuan, D. (2023). Cross-modality complementary information fusion for multispectral pedestrian detection. *Neural Computing and Applications*, *35*(14), 10361–10386.

Yang, M., & Thung, G. (2016). Classification of trash for recyclability status. *CS229 project report*.

Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., & Hoi, S. C. (2021). Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yi, D., Lei, Z., Liao, S., & Li, S. Z. (2014). Deep metric learning for person re-identification. *International Conference on Pattern Recognition*, 34–39.

Yu, J., Vincent, J. A., & Schwager, M. (2022). Dinno: Distributed neural network optimization for multi-robot collaborative learning. *IEEE Robotics and Automation Letters*, *7*(2), 1896–1903.

Yuan, Y., Weng, X., Ou, Y., & Kitani, K. M. (2021). Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9813–9823.

Zajdel, W., Zivkovic, Z., & Krose, B. J. (2005). Keeping track of humans: Have i seen this person before? *International Conference on Robotics and Automation*, 2081–2086.

Zhang, C., & Jia, Y. (2017). Distributed kalman consensus filter with event-triggered communication: Formulation and stability analysis. *Journal of the Franklin Institute*, *354*(13), 5486–5502.

Zhang, J., Liu, H., Yang, K., Hu, X., Liu, R., & Stiefelhagen, R. (2023). CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers. *IEEE Transactions on Intelligent Transportation Systems*, *24*, 14679–14694.

Zhang, R., Wu, L., Yang, Y., Wu, W., Chen, Y., & Xu, M. (2020). Multi-camera multi-player tracking with deep player identification in sports video. *Pattern Recognition*, *102*, 107260.

Zhang, T., Xie, L., Wei, L., Zhuang, Z., Zhang, Y., Li, B., & Tian, Q. (2021). Unrealperson: An adaptive pipeline towards costless person re-identification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11506–11515.

Zhang, X., Li, D., Wang, Z., Wang, J., Ding, E., Shi, J. Q., Zhang, Z., & Wang, J. (2022). Implicit sample extension for unsupervised person re-identification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7369–7378.

Zhao, R., Ouyang, W., & Wang, X. (2013). Unsupervised salience learning for person re-identification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3586–3593.

Zhao, Y., Li, Y., & Wang, S. (2019). Open-world person re-identification with deep hash feature embedding. *IEEE Signal Processing Letters*, *26*, 1758–1762.

Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., & Tian, Q. (2016). Mars: A video benchmark for large-scale person re-identification. *European Conference on Computer Vision*, 868–884.

Zheng, L., Yang, Y., & Hauptmann, A. G. (2016). Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*.

Zheng, Y., Tang, S., Teng, G., Ge, Y., Liu, K., Qin, J., Qi, D., & Chen, D. (2021). Online pseudo label generation by hierarchical cluster dynamics for adaptive person re-identification. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8371–8381.

Zhong, Y., Hu, X., Luo, C., Wang, X., Zhao, J., & Zhang, L. (2020). Whu-hi: Uav-borne hyperspectral with high spatial resolution (h2) benchmark datasets and classifier for precise crop identification based on deep convolutional neural network with crf. *Remote Sensing of Environment*, *250*.

Zhou, B., & Bose, N. (1993). Multitarget tracking in clutter: Fast algorithms for data association. *IEEE Transactions on Aerospace and Electronic Systems*, *29*(2), 352–363.

Zhou, K., Yang, Y., Cavallaro, A., & Xiang, T. (2019). Omni-scale feature learning for person re-identification. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3702–3712.

Zhou, K., Yang, Y., Cavallaro, A., & Xiang, T. (2021). Learning generalisable omni-scale representations for person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhou, Y., Xiao, J., Zhou, Y., & Loianno, G. (2022). Multi-robot collaborative perception with graph neural networks. *IEEE Robotics and Automation Letters*, *7*(2), 2289–2296.

Zhu, H., & Alonso-Mora, J. (2019). Chance-constrained collision avoidance for mavs in dynamic environments. *IEEE Robotics and Automation Letters*, *4*(2), 776–783.