BOSTON UNIVERSITY

COLLEGE OF ENGINEERING

Dissertation

PERSON RE-IDENTIFICATION USING FISHEYE CAMERAS WITH APPLICATION TO OCCUPANCY ANALYSIS

by

MERTCAN COKBAS

B.S., Sabancı University, 2018
M.S., Boston University, 2023

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2023

Approved by

First Reader

Janusz Konrad, Ph.D.
Professor of Electrical and Computer Engineering

Second Reader

Prakash Ishwar, Ph.D.
Professor of Electrical and Computer Engineering
Professor of Systems Engineering

Third Reader

Brian Kulis, Ph.D.
Associate Professor of Electrical and Computer Engineering
Associate Professor of Systems Engineering
Associate Professor of Computer Science

Third Reader

Eshed Ohn-Bar, Ph.D.
Assistant Professor of Electrical and Computer Engineering
Assistant Professor of Computer Science

Acknowledgments

I want to start by thanking my advisors, Prof. Janusz Konrad and Prof. Prakash Ishwar. It would not have been possible to put this work together without their assistance and guidance. Whenever I needed their help, they were always there for me. They guided me through not only research but all sorts of other aspects of my professional life. Out of the all the important things that they taught me, communication skills and attention to detail are two things that I will carry with me throughout my life. I would also like to thank to my other dissertation committee members, Prof. Brian Kulis and Prof. Eshed Ohn-Bar, for their valuable feedback and time. Furthermore, I would like to thank the Advanced Research Projects Agency – Energy (ARPA-E) and the College of Engineering, Boston University for financially supporting my research.

Throughout this journey, I had the chance to collaborate with many amazing people aside from my committee members. I would like to express my gratitude to Dr. Ozan Tezcan, Josh Bone, John Bolognino, Zhangchi Lu, Yu Xiao, Unay Dorken Gallastegi, Dr. Minxu Peng and Prof. Vivek Goyal. I also want to thank the annotation team that I worked with. Many thanks to Ragib Ahsan, Christopher Alonzo and Annette Hong.

This journey had its emotionally and mentally challenging moments that would have been insurmountable without the help of my friends: Dr. Ozan Tezcan, Can Ozbalkan, Dr. Uroš Kuzmanović, Dr. Novak Boškov, Ida Boškov, Dr. Mert Toslali, Burcu Ozek, Dr. Sebastian Carrasco Pro, Onur Altintas, Unay Dorken Gallastegi, Dr. Cagatay Karakan, Dr. Celalettin Yurdakul and all the members of the BM3 intramural basketball team.

I cannot thank my parents, Zeynep Cokbas and Recai Cokbas, and my brother, Sarper Cokbas, enough. Despite the distance between Boston and home, I still felt their love and support to the fullest. Finally, I would like to express my love and

gratitude to the biggest supporter that I had in this journey, my significant other, Rachel Portnoy. She was always there with me, in the best and the worst days of this chapter of my life. Her endless support and love turned Boston into a home far from home.

PERSON RE-IDENTIFICATION USING FISHEYE CAMERAS WITH APPLICATION TO OCCUPANCY ANALYSIS

MERTCAN COKBAS

Boston University, College of Engineering, 2023

Major Professors: Janusz Konrad, Ph.D.
Professor of Electrical and Computer Engineering

Prakash Ishwar, Ph.D.
Professor of Electrical and Computer Engineering
Professor of Systems Engineering

ABSTRACT

Person re-identification (PRID), the problem of matching identity of a person between images, finds applications in video surveillance, sports analytics, and, more recently, in spatial analytics (e.g., retail). PRID has been extensively studied for the case of standard surveillance cameras equipped with rectilinear lens. However, their narrow field of view (FOV) severely limits the indoor area each camera can monitor. Recently, fisheye cameras that capture 360° FOV have penetrated the video surveillance market, but little attention has been paid in the literature to PRID for such cameras. This dissertation focuses on fisheye-camera PRID and demonstrates its effectiveness in occupancy estimation in large indoor spaces.

Since no fisheye PRID datasets were publicly available, we created one using 3 ceiling-mounted fisheye cameras in a large classroom and published it on-line (63 downloads to date). Subsequently, we evaluated 6 state-of-the-art PRID methods on

our dataset and concluded that such methods, developed for rectilinear cameras, do not perform well on fisheye images due to potential dramatic body-viewpoint and body-size differences between different camera views, and fisheye-lens distortions. To address these challenges, we developed a novel approach to PRID that relies on occupant location in the room instead of appearance. This approach is possible in our scenario since overhead fisheye cameras have overlapping FOVs; knowing location of a person in one camera view, we map this location to another camera view with knowledge of the person's approximate height. The distance between a mapped and current location allows to match identities, and we develop 4 distance metrics for this purpose using a range of typical human heights. Evaluated on our dataset, the location-based approach outperforms the 6 state-of-the-art PRID methods that use appearance by at least 10% points in accuracy, but struggles when people are very close to one another. To address this challenge, we proposed combining location-based methodology with appearance features (deep-learning embedding and color histogram) by means of a Naïve Bayes method. The additional appearance features improve the location-based re-identification accuracy by at least 2% points.

To demonstrate the practical importance of fisheye PRID, we evaluated its potential for accurate people counting in a large space with high occupancy. Firstly, we assessed the performance of occupancy sensing using single fisheye camera and concluded that high counting accuracy is possible only in small-to-medium size spaces. We then proposed a two-camera system, that employs fisheye PRID to avoid over-counting, and demonstrated an up to 20%-point accuracy boost compared to single-camera approaches. To support even larger spaces, we proposed and evaluated two extensions of PRID to N cameras. Overall, our results show that the proposed fisheye PRID methods enable high-accuracy people counting in large indoor environments, and have a great potential for improving people tracking and activity analysis.

Contents

1 Introduction **1**

1.1 Occupancy Sensing 1

1.2 Methods for Occupancy Sensing 2

1.2.1 Active Approaches 2

1.2.2 Passive Approaches 3

1.3 Occupancy Sensing Using Overhead Fisheye Cameras 4

1.4 Person Re-Identification 5

1.5 Thesis Overview and Contributions 6

1.6 Thesis Organization 9

2 Related Work **11**

2.1 Traditional Rectilinear PRID 12

2.2 Traditional Fisheye PRID 15

2.3 Cross-Frame Rectilinear PRID 15

2.4 Cross-Frame Fisheye PRID 16

2.5 PRID Focus of this Dissertation 19

3 Fisheye PRID Dataset and its Evaluation **20**

3.1 FRIDA Dataset 21

3.1.1 Annotations 21

3.1.2 Scenarios 24

3.1.3 Gallery Set with Single Sample per ID 25

3.1.4 Synchronous and Overhead Capture 26

3.1.5 Fisheye Distortions 26
3.1.6 Resolution Mismatch Between Query and Gallery Sets 26
3.2 Evaluating SOTA Algorithms on FRIDA 27
3.2.1 Training-Testing Methodology 28
3.2.2 Dataset Splits 29
3.2.3 Evaluation Metrics 30
3.2.4 Results 31
3.3 Chapter Summary and Discussion 35

4 Location-Based Person Re-Identification 37

4.1 Fisheye Image Pixel-Correspondence Model 38
4.1.1 Notation 38
4.1.2 Forward Mapping 39
4.1.3 Pixel-Correspondence Mapping 41
4.2 Camera Calibration 42
4.3 Application of the Pixel-Correspondence Model to Person Re-Identification 45
4.3.1 Distance Measures 48
4.3.2 Identity Matching 49
4.4 Experimental Results 50
4.5 Chapter Summary and Discussion 54

5 Spatio-Visual Fusion-Based Person Re-Identification 55

5.1 Methodology 56
5.1.1 Deep-Learning Features and Pairwise-Similarity Scores 58
5.1.2 Color Histograms and Pairwise-Dissimilarity Scores 59
5.1.3 Location-Based Features and Pairwise-Dissimilarity Scores . . 61
5.1.4 Fusion of Features 62
5.1.5 Matching Algorithm 63

5.2 Experimental Results 64
5.2.1 Dataset Splits 64
5.2.2 Implementation Details 65
5.2.3 Results 66
5.3 Chapter Summary and Discussion 69

6 Application of PRID to Occupancy Analysis 70
6.1 Evaluation Metrics 71
6.2 People-Counting Dataset 73
6.3 Occupancy Estimation Using Single Fisheye Camera 74
6.3.1 Performance of RAPiD at Different Distances from the Camera 75
6.3.2 Performance of RAPiD in a Medium-Sized Room 77
6.3.3 Performance of RAPiD in a Large Space with High Occupancy 79
6.4 Occupancy Estimation Using Two Fisheye Cameras 83
6.5 Occupancy Estimation using N Cameras 91
6.5.1 General Method based on N-Dimensional Score Matrix 91
6.5.2 Clustering of Real-World Locations of People 94
6.5.3 Experimental Results for N-Camera Algorithms 97
6.6 Chapter Summary and Discussion 103

7 Conclusions and Future Directions 104
7.1 Future Directions 107
7.1.1 Leveraging Temporal Information 107
7.1.2 General Matching-Score Metrics for N-camera PRID 109
7.1.3 Domain Adaptation 109
7.1.4 Distance Estimation 110

A Derivations for Pixel-Correspondence Mapping 112

B Low-Resolution Overhead Thermal Tripwire for Occupancy Estimation **116**

B.1 Methodology 117

B.1.1 Background Subtraction 118

B.1.2 Event Detection 121

B.1.3 Event Classification 124

B.2 Experimental Results 125

B.2.1 Dataset 125

B.2.2 Performance Analysis 126

B.3 Discussion 134

References **135**

Curriculum Vitae **143**

List of Tables

1.1 Key differences between traditional PRID and our PRID. 6

3.1 Comparison of FRIDA with the most popular image datasets for person re-identification. (BBox = bounding box) 22

3.2 Detailed information about FRIDA (*Fisheye Re-Identification Dataset with Annotations*). 25

3.3 Performance comparison of state-of-the-art algorithms trained on **Market-1501** and tested on **FRIDA** for different camera pairs. The highest values of QMS and mAP for each camera pair and cumulatively are shown in boldface. 32

3.4 Performance comparison of state-of-the-art algorithms trained on **FRIDA** and tested on **FRIDA** for different camera pairs. The highest values for QMS and mAP for each camera pair and cumulatively are shown in boldface. 32

3.5 Performance comparison of state-of-the-art algorithms trained on **Market-1501** and tested on **FRIDA** for different segments. The highest values for QMS and mAP for each segment and cumulatively are shown in boldface. 34

3.6 Performance comparison of state-of-the-art algorithms trained on **FRIDA** and tested on **FRIDA** for different segments. The highest values for QMS and mAP for each segment and cumulatively are shown in boldface. 34

4.1 Estimated parameter values $\widehat{\boldsymbol{\omega}}$ for our experimental setup. 51

4.2 Performance comparison of the location-based PRID for different camera pairs. The highest values for QMS and mAP for each camera pair and for the cumulative are shown in boldface. 52

4.3 Performance comparison of the location-based PRID for different segments of FRIDA. The highest values for QMS and mAP for each algorithm are shown in boldface. 52

5.1 Performance comparison of PRID on FRIDA dataset for various combinations of deep-learning (DL), color-histogram (CH) and location-based (LOC) features, for both PPD and CBD distance measures. The highest values of QMS and mAP for each camera pair (e.g., "1↔2") and for the cumulative ("Cum.") metric are shown in boldface. . . . 67

6.1 People-counting performance of RAPiD for increasing camera field of view. 77

6.2 People-counting performance (MAE, MAE_{pp} and X-Accuracy) of RAPiD in 3 selected scenarios and cumulatively over all 8 scenarios in CEPDOF dataset (Duan et al., 2020). 79

6.3 People-counting performance (MAE, MAE_{pp} and X-Accuracy) of RAPiD for the 3-day test in 2,000 ft^2 room for all three cameras. The last column shows cumulative metrics computed across the 3 days. 82

6.4 Threshold τ values that minimize MAE for each feature combination. The minimization is performed through grid search. 86

6.5 People-counting performance (MAE, MAE_{pp} and X-Accuracy) of PRID algorithms in a 3-day test in large classroom using cameras 2 and 3. . 89

6.6 People-counting performance (MAE, MAE_{pp} and X-Accuracy) of the CBD approach ($N = 2$ cameras) and N-camera PRID algorithms for $N = 3$ in a 3-day test in a large classroom. 102

B.1 Details of TIDOS (Thermal Images for Door-based Occupancy Sensing) dataset. Each 32×24-pixel frame was acquired by Melexis MLX90640 sensor at 16 fps. Data was collected by 2 sensors, one over each door of a small classroom. 127

B.2 Performance comparison of the proposed algorithms on TIDOS dataset using three metrics. The lowest values for MAE and MAE_{PP} and the highest value for CCR_{WCC} for each recording are shown in boldface. 129

List of Figures

2·1 Illustration of differences between traditional rectilinear PRID and cross-frame fisheye PRID, the focus of this dissertation. Key differences are: 1) camera type, 2) FOV overlap (which determines the Re-ID timing), 3) number of gallery samples per identity 4) viewpoint (side-view versus overhead). 13

2·2 Illustration of cross-frame fisheye PRID with two cameras. The goal is to establish a correspondence between people visible in two frames captured at the same time instant. At a given time instant, we consider the people in one camera view as "query" elements, and people in the other camera view as "gallery" elements. Yellow and green bounding boxes illustrate the two sets of elements. Either set can be treated as a query or gallery set. 18

3·1 Bird's eye view of the space where FRIDA was recorded. 22

3·2 Example of three synchronously-captured fisheye images with annotations from FRIDA (top: camera 2, middle: camera 1, bottom: camera 3). 23

3·3 Illustration of bounding-box parameters available in FRIDA. 24

3·4 Bounding-box resolution mismatch for all camera pairs. 27

4·1 Projection model for two parallel fisheye cameras with $\xi_A = \xi_B = \xi$. 39

4·2 Three synchronously-captured images showing a mobile cart with colored spherical LED light. From left to right are shown images from cameras 2, 1, 3, respectively (Figure 3·1). The LED light location is found by color thresholding and shown as a red cross. 44

4·3 Illustration of the mapping given by Equation (4.12) from camera $A = 1$ (left) to camera $B = 3$ (right) produced by the calibrated geometric model. The green points are ground-truth locations and the red points are predicted locations. 46

4·4 Illustration of location predictions for 4 people (cyan, yellow, magenta, orange). (a) Query-subject locations in camera 1 (crosses). (b) Gallery-subject locations in camera 3 (crosses) and predicted query locations for single $Z = Z_{\text{avg}} = 170\text{cm}$ (4 bullets). (c) Gallery-subject locations in camera 3 (crosses) and predicted query locations for $Z \in Z_R$ (4 sequences of bullets). 47

5·1 Illustration of location and appearance ambiguity. When people are very close to each other, distinct color of clothing and/or body shape/features may help resolve location ambiguity (left column). When people have very similar appearance (e.g., light-gray T-shirts and dark pants), knowing their location may help resolve appearance ambiguity (right column). 57

5·2 Block diagram of the proposed method. In the first step, features are extracted from the contents of bounding boxes (for appearance-based methods) or their positions within the frames (for location-based method) for the query and gallery identity sets (which are denoted by Q_n and G_n, respectively, for frame number n). In the next step, features of query-gallery pairs are converted into either pairwise-similarity or pairwise-dissimilarity scores. They are then normalized via the softmax operation to obtain match-probability matrices for each feature group which are then fused together via the Hadamard product into a fused match-probability matrix. Finally, a greedy sequential matching algorithm based on the fused match-probability matrix is used to produce the query-gallery matches. 58

5·3 Illustration of the importance of color information in fisheye PRID. The person in the red bounding box in the top frame is severely shrunk, however the distinctive color of the sweatshirt still allows to distinguish this person from others. 60

6·1 Sample frames from each day of the 3-day dataset. 74

6·2 The true people count for the 3-day dataset. 75

6·3 Overhead fisheye view of the classroom where FRIDA was recorded. The superimposed green concentric circles correspond to the true physical distance of 10-35 ft from the camera in 5 ft increments. 76

6·4 Examples of detections by RAPiD in a 500 ft^2 classroom in 3 scenarios selected from CEPDOF dataset (Duan et al., 2020). 78

6·5 Sample frame with people detections by RAPiD from day 1 of the 3-day test. 80

6·6 The true number of people (red line) and an estimate by RAPiD in a 3-day test in the same classroom as FRIDA for each of the three cameras. 81

6·7 (a-b) Sample frames from two overhead fisheye cameras overlooking a large classroom; (c-d) The same frames with people detections by RAPiD (Duan et al., 2020). 84

6·8 Ground-truth people count and four people-count estimates (two single-camera estimates, sum of single-camera estimates, and an estimate obtained using PRID (DL+CH+CBD) to eliminate double counts) for the 3-day test. 87

6·9 Illustration of PRID scenario for $N = 3$ cameras. 92

6·10 MAE values for different values of λ for the N-D Score Matrix approach from Section 6.5.1 with $N = 3$. 98

6·11 MAE values for different values of ϵ for the Real-World Location Clustering approach from Section 6.5.2 with $N = 3$. 98

6·12 The true number of people (red line) and an estimate by 3-camera Real-World Location Clustering approach for the 3-day test in 2,000 ft^2 classroom and three different values of ϵ. 99

6·13 The true number of people (red line) and an estimate by 3-camera N-D Score Matrix approach for the 3-day test in 2,000 ft^2 classroom and three different values of λ. 100

A·1 Geometry of a model for single fisheye camera. 113

B·1 Configuration of our virtual-tripwire door setup: low-resolution thermal sensor mounted above a door and facing down (left); and 32×24-pixel thermal frame captured by the sensor when a person is leaving the room (right). 118

B·2 Block diagram of the proposed approach. 118

B·3 Thermal frame and results of background subtraction for a single person passing through a door. 120

B·4 Thermal frame and results of background subtraction and centroid calculation for 2 people passing through a door. 123

B·5 People counts estimated by the **baseline algorithm**. True (blue) and estimated (red) people-count plots for the proposed algorithms across all recordings in the TIDOS dataset. To distinguish between the red and blue curves in frames where their values exactly coincide, we added a positive vertical offset of 0.1 person to the blue curves. Note that since at each time instant two frames are collected (one by each door sensor), the number of frames in this plot is one-half of the total number of frames in Table B.1. 130

B·6 People counts estimated by the **multi-person algorithm**. True (blue) and estimated (red) people-count plots for the proposed algorithms across all recordings in the TIDOS dataset. To distinguish between the red and blue curves in frames where their values exactly coincide, we added a positive vertical offset of 0.1 person to the blue curves. Note that since at each time instant two frames are collected (one by each door sensor), the number of frames in this plot is one-half of the total number of frames in Table B.1. 131

List of Abbreviations

CBD	.....	Count-Based Distance
CH	.....	Color Histograms
DL	.....	Deep-Learning Features
FOV	.....	Field of View
FRIDA	.....	Fisheye Re-identification Dataset with Annotations
IR	.....	Infrared Radiation
LiDAR	.....	Light Detection and Ranging
LOC	.....	Location-based Features
mAP	.....	Mean Average Precision
PPD	.....	Point-to-Point Distance
PRID	.....	Person Re-Identification
PSMD	.....	Point-to-Set Minimum Distance
PSTD	.....	Point-to-Set Total Distance
QMS	.....	Query Matching Score
ToF	.....	Time of Flight

Chapter 1

Introduction

This thesis is motivated by the problem of occupancy sensing in large indoor environments. We begin by describing the problem, its applications, and various approaches proposed in the literature to tackle it. We then focus on the family of methods based on overhead fisheye cameras which is the focus of this thesis. We discuss the need for person re-identification (PRID) and summarize the key contributions of this thesis. We conclude with an outline of the remainder of the thesis.

1.1 Occupancy Sensing

Occupancy sensing is a key technology for smart buildings of the future (Sruthi, 2019; Nguyen and Aiello, 2013). Knowledge of count and locations of people in a building enables, among others, smart HVAC control to save energy, space management to reduce rental costs and enhanced security, (e.g., fire, flooding, active shooter) (Hashimoto et al., 1997).

Energy savings: In 2018, energy consumption in commercial buildings in the United States amounted to 6,787 quad BTUs (quadrillion British thermal units), with 52% expended on heating, cooling and ventilation (HVAC) (CBECS, 2018). A significant fraction of the HVAC energy is wasted due to over-ventilation, in particular when rooms are only partially occupied relative to their maximum capacity. Other than saving energy, occupancy information also plays an increasingly important role in space management and security/safety.

Space management: The COVID-19 pandemic has dramatically impacted the office-building market, leading to new office-usage patterns. Many companies returning to offices opt for the so-called "flexible workspace" where desks are not assigned to employees but can be reserved whenever employees return in-person for work, meetings, etc. Real-time accurate detection of workspace occupancy is essential for an effective implementation of this concept. A similar knowledge of where people are is essential for the retail industry, e.g., how many people are waiting at a check-out counter.

Security and safety: Occupancy information is also important for security and safety in a building. For example, it can be used to ensure everyone is accounted for in an emergency situation (e.g., fire). Moreover, knowing how people are located in a space can be useful in the context of public health, such as managing a pandemic (e.g., social distancing).

1.2 Methods for Occupancy Sensing

In the last decade, a wide range of approaches have been proposed for occupancy sensing in commercial buildings. We broadly divide these approaches into *active* and *passive* categories.

1.2.1 Active Approaches

In *active* approaches, all occupants are expected to carry a device that allows them to be detected. The most common device is a magnetic-swipe or RFID-proximity card that allows an authorized person to enter a restricted space. Counting people from card swipes is straightforward but is prone to drift errors when two or more people enter using a single swipe. Such errors are typical in any "tripwire"-type system (i.e., a system that detects when a boundary is crossed by a person) when a *change* in people count rather than the count itself is being estimated; a missed detection can

be corrected only by an opposite error (false detection). Another type of device that can reveal presence is one producing radio frequencies, such as a cell phone, tablet or laptop; its presence can be detected at the network level based on WiFi, Bluetooth, or cell-band communication. Since the number of devices is being counted, rather than a change in this number, this approach is not prone to drift errors. However, it relies on each occupant carrying one device only, an unlikely scenario these days. On the other hand, one benefit of active approaches is that they can handle large spaces with high occupancy.

1.2.2 Passive Approaches

In terms of passive approaches, *passive-indirect* methods rely on harvesting information from either system sensors (e.g., reheat valve or damper position) or environmental sensors (e.g., temperature, CO_2, humidity). Inferring occupancy level from environmental data is cost-effective (sensors are often installed), but in case of CO_2 sensors, the need for frequent re-calibration due to drift errors, measurement lag due to gas mixing delays and sensitivity to installation location and room size make such an approach cumbersome and imprecise. Additionally, while using an HVAC system's control data (e.g., damper position) is simple and reliable, this approach only allows one to learn occupancy patterns with the goal of *predicting* future occupancy patterns to aid a building-management system (BMS). This approach does not explicitly estimate true occupancy in real time but rather "guesses" a likely occupancy level based on past system data (Ardakanian et al., 2018).

The most promising methods for occupancy sensing in large commercial spaces are *passive-direct* approaches that use occupant-related features such as appearance, movement, weight, body heat, sound, etc. Sensors commonly deployed in such scenarios are RGB cameras, LiDAR/ToF sensors (Lu et al., 2021), structured-light sensors (Diraco et al., 2015), microphones (Huang et al., 2016), IR/thermal sensors (Piechocki

et al., 2022), etc.

In our first attempt at occupancy estimation (Cokbas et al., 2020), we designed a system using low-resolution thermal sensors to detect body heat (see Appendix B). We opted for thermal sensors to preserve occupant privacy, and we chose low-resolution sensors over high-resolution ones for cost-efficiency. Due to the low resolution and narrow field of view of these sensors, we chose to design a tripwire-type system in which we solely monitored the entry/exit points to an indoor space. Unfortunately, given the nature of tripwire-type systems, our system suffered from drift errors. Thus, we decided not to pursue this method of people counting any further.

1.3 Occupancy Sensing Using Overhead Fisheye Cameras

Compared to other visual sensors, RGB cameras offer higher resolution, lower cost and ability for fine-grained estimations. To date, many methods have been proposed for RGB cameras (Ryan et al., 2011; Liu et al., 2013; Erickson et al., 2013; Choi et al., 2021; Wei et al., 2022). These methods primarily use cameras with rectilinear lens (e.g., surveillance cameras) which project straight lines in a room onto approximately straight lines on sensor surface. While such methods demonstrate promising results in many scenarios, their most significant deficiency is the relatively narrow field-of-view (FOV), typically around 90°. As a room gets larger, more cameras are needed which increases system cost. To cover more space, such cameras are typically mounted at an angle (i.e., side-mounted rather than ceiling-mounted), which increases the likelihood of occlusions and may lead to undercounting.

In this dissertation, to overcome such issues, we focus on occupancy estimation with overhead fisheye-lens cameras. Fisheye-lens cameras have a 360° FOV allowing overhead mounting (i.e., monitoring directly from above) which, in turn, minimizes occlusions. Some people detection algorithms for overhead fisheye cameras have been

already proposed (Tamura et al., 2019; Li et al., 2019; Duan et al., 2020). These algorithms have been shown to perform well in small-to-medium sized rooms. However, in large rooms (i.e., larger than 1,000 square feet), due to severe geometric distortions, these algorithms struggle to detect people at FOV periphery. In such scenarios, multiple cameras are needed to cover the whole room.

1.4 Person Re-Identification

Unfortunately, using multiple overhead fisheye cameras in a room for occupancy estimation is problematic. Due to the 360° FOV, multiple fisheye cameras in the same room will have a large FOV overlap, which is likely to cause overcounting. To resolve this, the same person must be identified in the FOVs of all cameras, a problem known as person re-identification (PRID). From the broadest perspective, in PRID, the goal is to re-identify a person captured by a camera in the view of another camera.

PRID with rectilinear lens cameras: PRID has been well explored for cameras equipped with a rectilinear lens. There exist many traditional, model-based PRID methods (Yang et al., 2014; Xiong et al., 2014). Some methods focus on localized features extracted from images of people (Yang et al., 2014). Other methods focus on the development of distance metrics to maximize the distance between different identities and minimize the distance between the same identities (Xiong et al., 2014). However, deep-learning methods are currently the state-of-the-art (a good review can be found in this paper (Ye et al., 2022)). All these methods have been developed for re-identifying people across *side-view* rectilinear-lens cameras that have no FOV overlap. The popular benchmark PRID datasets (Market-1501 (Zheng et al., 2015), Duke MTMC (Ristani et al., 2016), CUHK03 (Li et al., 2014)) have all been collected and set up for this PRID setting.

PRID with fisheye lens cameras: There have been very few attempts to per-

form PRID using fisheye cameras, which we will review in Chapter 2, and no datasets have been made public. Furthermore, our PRID setup is different from the traditional PRID setup – our fisheye cameras are mounted on the ceiling and their FOVs fully overlap. Since we need to match identities at the same time instant, each query identity (from one camera) may have at most one match among the gallery identities (from another camera); there is no match in the case of occlusion. However, in the traditional PRID setting with no FOV overlap, for a single query there may be numerous correct matches in the gallery set since it is captured from several cameras at many time instants. Table1.1 highlights some of the key differences between the traditional PRID scenario and our scenario. A more detailed and structured categorization of different PRID scenarios is presented in Chapter 2, where a visualization of key differences outlined in Table 1.1 is provided (see Figure 2·1).

	FOV	**Re-ID timing**	**Gallery samples per identity**
Traditional PRID	Non-Overlapping	Asynchronous	Multiple
Our PRID (this thesis)	Overlapping	Synchronous	Single/None

Table 1.1: Key differences between traditional PRID and our PRID.

1.5 Thesis Overview and Contributions

The core research problem of this dissertation is to re-identify people captured with time-synchronized, overhead, fisheye-lens cameras that have fully overlapping FOVs.

Prior to this work, there was no suitable dataset available in the literature to study this problem. Thus, we collected a first-of-its-kind dataset (described in Chapter 3) named "Fisheye Re-IDentification Dataset with Annotations" (FRIDA), that was captured by three overhead fisheye cameras in a large space (2,000+ square feet)

and includes over 240,000 bounding-box annotations of people. Also, by evaluating 6 state-of-the-art (SOTA) PRID methods on FRIDA, we demonstrated that methods developed for images captured by rectilinear-lens cameras do not perform well on images from overhead fisheye cameras.

The 6 algorithms we evaluated on FRIDA rely solely on the appearance of a person. However, appearance gets distorted by a fisheye lens, especially if a person is far away from the camera where geometric distorsions are more severe. Thus, we aimed to develop a PRID algorithm that does not rely on the appearance of people at all. It led us to a design that depends only on the location of people. The principal inspiration for this algorithm comes from the observation that a person occupies a single 3D-world coordinate at a given time instant. In other words, given a pair of fisheye images, a person may appear at different pixel coordinates in different camera views; however, his/her real-world (3D) coordinate is unique.

Our new location-based PRID algorithm is described in detail in Chapter 4 together with an automated method for calibrating it and certain PRID distance metrics needed to tackle occupant height-variability which affects location-based occupant estimation. We evaluate these PRID approaches on FRIDA. Results show that location-based methods outperform the 6 state-of-the-art appearance-based methods.

Although location-based approaches perform well, there is still room for improvement. When people stand close to each other, location-based approaches struggle. However, if such people have distinct physical appearances, then appearance-based methods might perform better (see Figure 5·1). Inspired by this observation, in Chapter 5, we introduce a framework that combines appearance-based and location-based features to perform PRID. As appearance features, we use traditional color histograms and features extracted by a state-of-the-art deep-learning model, while as the location feature we use one of the PRID distance metrics proposed in Chapter 4.

To perform identity matching, we introduce a probabilistic feature-fusion approach.

All the PRID methods that are proposed in this dissertation thus far have been tailored for two cameras. However, two fisheye cameras are not enough to perform reliable people detection and re-identification in very large indoor spaces (e.g., convention halls, supermarkets). Thus, in Chapter 6, we introduce two approaches extending the proposed two-camera methods to N cameras.

Finally, we evaluate the occupancy estimation performance of the proposed PRID methods on a 72-hour video recorded in a 2,000+ square-feet room where the occupancy reaches 87 people at times. The results prove that with the proposed methods, it is possible to estimate the number of people in the room accurately.

The main contributions of this thesis can be summarized as follows:

1. **Dataset for fisheye PRID (Chapter 3)**: There exist public PRID datasets captured by side-view rectilinear-lens cameras, but not by fisheye cameras. We introduce a first-of-its-kind PRID dataset for indoor person re-identification using time-synchronized overhead fisheye cameras. We benchmark six state-of-the-art PRID algorithms on this new dataset.

2. **Location-based PRID (Chapter 4)**: Existing PRID methods have relied on appearance-based features. However, such methods do not perform well on fisheye images. We propose a novel PRID method that relies solely on the location of occupants instead of their appearance. We also describe a novel calibration method to estimate the intrinsic and extrinsic parameters of pairs of fisheye cameras.

3. **Spatio-visual PRID (Chapter 5)**: We introduce a framework for multi-feature PRID that combines location-based and appearance-based features. As part of this framework, we propose a feature-fusion method for identity match-

ing. Through an ablation study, we demonstrate that combining these features yields a better PRID performance than using single-feature approaches.

4. **Scaling from two cameras to N cameras (Chapter 6)**: The majority of PRID methods in the literature focus on PRID between two cameras (two sets: query and gallery). In this thesis, we introduce two approaches on how to scale PRID from two cameras to N cameras. We analyze the performance of these methods in terms of occupancy-estimation performance.

5. **Application to occupancy analysis (Chapter 6)**: We demonstrate the potential of the proposed fisheye PRID methods for occupancy estimation. We provide a comparison of occupancy-estimation performance of the proposed fisheye PRID methods using two cameras against a state-of-the-art single fisheye-camera method.

1.6 Thesis Organization

The rest of this thesis is organized as follows. In Chapter 2, we discuss related work on occupancy estimation and person re-identification, and we formally state the main problem that this thesis addresses. In Chapter 3, we introduce a first-of-its-kind dataset that is captured by time-synchronized overhead fisheye cameras that have fully overlapping FOVs. We demonstrate that state-of-the-art methods do not perform as well on this dataset as they do on other benchmark PRID datasets. In Chapter 4, we introduce a location-based person re-identification method and demonstrate its superior performance over state-of-the-art algorithms. In Chapter 5, we introduce a framework to combine location-based and appearance-based features for PRID and show results demonstrating the effectiveness of this framework. In Chapter 6, we discuss occupancy estimation using overhead fisheye cameras and demonstrate performance improvements afforded by the proposed PRID methods vis-à-vis single-camera

methods. Finally, in Chapter 7, we provide a discussion of the contributions of this dissertation and discuss some potential ideas for future work.

Chapter 2

Related Work

Traditional PRID methods developed to date have considered matching identities captured at *different time instants* (asynchronously) by multiple cameras with *no FOV overlap*. In this case, a query (i.e., a person-image that we are interested in re-identifying) from one camera is allowed to have multiple matches in the gallery set (i.e., set of person-images that the query elements are allowed to match) that is extracted from other cameras (see Figure 2·1a). In principle, asynchronous PRID from cameras *with overlapping FOVs* would fall into the traditional PRID category but we are not aware of any such work.

An alternative is a different type of PRID, that we call *cross-frame PRID* and study in this dissertation. In cross-frame PRID, matching is performed between identities captured at the *same time instant* (synchronously) by two cameras with *overlapping FOVs*. In this case, person-images from one camera's video frame are considered query elements while those from the other camera's synchronous video frame are considered to be the gallery set. Consequently, a query ID is allowed to have only *one correct match* (or none, in case of occlusion) in the gallery set. Note, that synchronous PRID from cameras with no FOV overlap is not possible since a person would appear only in a single camera view (see Figure 2·1b). Table 1.1 summarizes key differences between traditional and cross-frame PRID.

In this chapter, we further subdivide traditional and cross-frame PRID based on camera type used as follows:

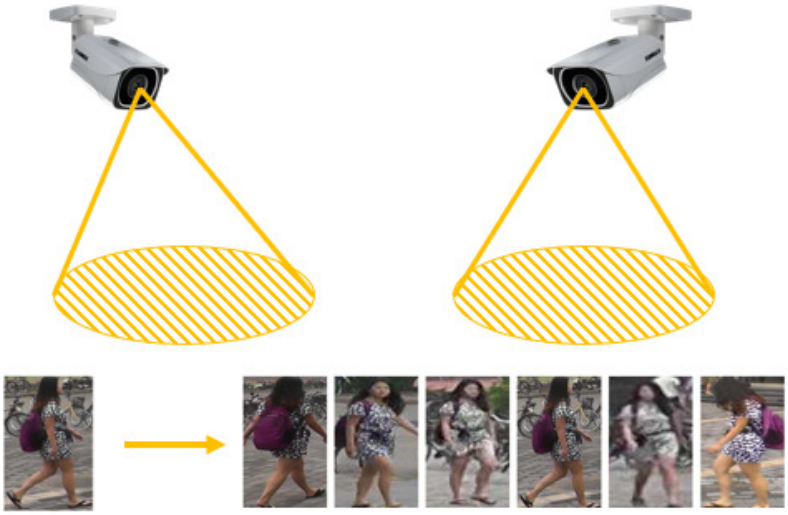
- traditional rectilinear PRID,
- traditional fisheye PRID,
- cross-frame rectilinear PRID,
- cross-frame fisheye PRID.

and perform literature review for each category separately.

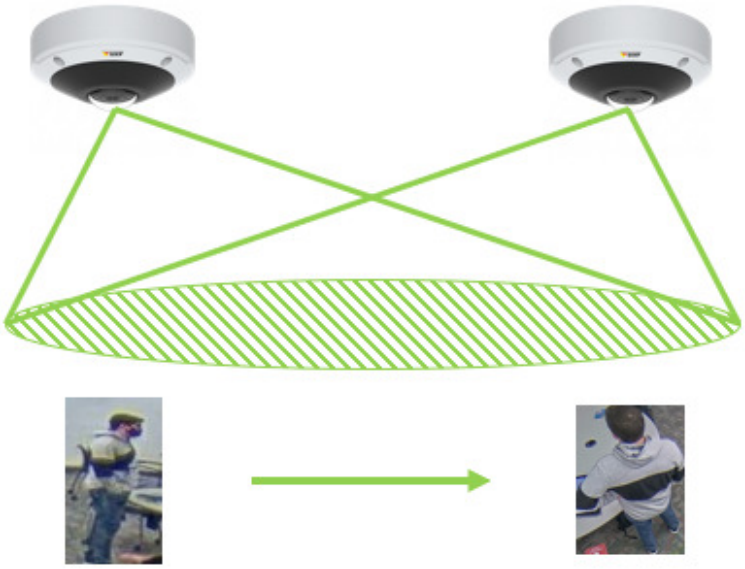
2.1 Traditional Rectilinear PRID

Among the four types of PRID, *traditional rectilinear PRID* has been explored the most. Its goal is to match the image of a person from the query set to an image from the gallery set, where images are captured by side-mounted rectilinear-lens cameras. The query and gallery sets consist of images captured by *different* cameras. Moreover, different cameras have no field-of-view overlap so query and gallery images of the same identity have been captured at *different time instants*. Thus, in most of the traditional rectilinear PRID datasets there are, typically, *multiple* gallery images with the same ID as the query image. This scenario is depicted in Figure 2·1a.

The majority of public PRID datasets are tailored towards traditional rectilinear PRID due to its popularity in the research community. The most commonly-used ones are VIPeR (Gray and Tao, 2008), iLIDS (Zheng et al., 2009), PRID 2011 (Hirzer et al., 2011), GRID (Loy et al., 2013), CUHK03 (Li et al., 2014), Market-1501 (Zheng et al., 2015), MSMT17 (Wei et al., 2018) and Airport (Karanam et al., 2019). Some of these datasets were collected in public places in a real-world setting with intense foot traffic, which makes PRID challenging. For example, GRID was collected in a busy underground station, Market-1501 was collected in front of a supermarket and Airport was collected in an airport.



(a) Traditional rectilinear PRID (images are from Market-1501 dataset (Zheng et al., 2015))



(b) Cross-frame fisheye PRID (focus of this dissertation)

Figure 2·1: Illustration of differences between traditional rectilinear PRID and cross-frame fisheye PRID, the focus of this dissertation. Key differences are: 1) camera type, 2) FOV overlap (which determines the Re-ID timing), 3) number of gallery samples per identity 4) viewpoint (side-view versus overhead).

Early approaches to traditional rectilinear PRID were mostly model-based and relied on hand-crafted features (Gray and Tao, 2008; Farenzena et al., 2010; Zheng et al., 2011; Köstinger et al., 2012; Hirzer et al., 2012; Xiong et al., 2014; Yang et al., 2014; Liao et al., 2015). Some methods focused on localized features extracted from images of people (Gray and Tao, 2008; Farenzena et al., 2010; Yang et al., 2014; Liao et al., 2015). Many methods focused on the development of distance metrics to maximize the distance between samples from different identities and minimize the distance between samples from the same identities (Zheng et al., 2011; Köstinger et al., 2012; Hirzer et al., 2012; Xiong et al., 2014).

However, the advances made by deep learning in many computer-vision tasks did not bypass PRID. The best-performing PRID algorithms in the last decade have been based on deep learning (Sun et al., 2018; Chen et al., 2019b; Chen et al., 2019a; Bryan et al., 2019; Zheng et al., 2019a; Li et al., 2018b; Yu et al., 2019; Zheng et al., 2019b; Zhou et al., 2019; Zhihui et al., 2020; Wieczorek et al., 2021). (Sun et al., 2018) proposed PCB in which feature vectors are uniformly partitioned in an intermediate layer to obtain part-informed features. This structure allows to separately focus on different parts of an image and extract local information for each part. (Zheng et al., 2019a) developed a "Pyramid" network which extends part-based image matching by simultaneously incorporating local and global features through a coarse-to-fine model. (Chen et al., 2019b) proposed an attention-based network called ABD-Net, which instead of a small portion of an image focuses on its wider aspect by means of a diverse attention map. This is accomplished by combining two separate modules: one module focuses on context-wise relevance of pixels while the other module focuses on spatial relevance of these pixels. (Zhihui et al., 2020) proposed a network called VA-reID that allows matching of people regardless of the viewpoint from which they were captured. Instead of creating a separate space for each viewpoint (i.e., front, side, back), they

create a unified hyperspace which accommodates viewpoints in-between the main viewpoints (e.g, side-front, side-back). Recently, (Wieczorek et al., 2021) proposed a CTL model (Centroid Triplet Loss model) , which extends the triplet loss. When working with triplet loss, it is typical to choose one positive sample and one negative sample for an anchor. However, in the CTL model, instead of choosing a single sample, a centroid is computed over a set of samples which significantly improves performance.

2.2 Traditional Fisheye PRID

Recently, traditional PRID methods have been extended to overhead fisheye cameras, but we are aware of only two attempts. An early approach, proposed by (Barman et al., 2018), matches images of people who appear at the same radial distance from a camera. The authors recognize difficulties arising from fisheye-lens distortions and potentially-dramatic viewpoint differences. However, limiting the search space to the same radial distance from each camera is restrictive, and leads to sub-par performance since people often appear at different distances from FOV centers in different cameras. Another algorithm proposed by (Blott et al., 2019) applies tracking to extract front-, back- and side-view images of a person. The proposed PRID algorithm uses a person's descriptor built by fusing features extracted from individual views. However, there is no guarantee that a person will appear in all 3 viewpoints during a recording, thus limiting performance. Moreover, both works report results on non-public fisheye data only.

2.3 Cross-Frame Rectilinear PRID

Cross-frame PRID finds fewer applications than traditional PRID since it requires multiple time-synchronized cameras to monitor the same space (increased cost and

complexity). It has been primarily used in people detection and tracking.

(Fleuret et al., 2014) applied cross-frame PRID to improve tracking in sporting events. Their approach is appearance-based and uses jersey color, jersey number and facial features for re-identification between concurrent rectilinear camera views. (Hu et al., 2022) also focused on improving tracking of people but in a smaller space of a hospital operating room. They first perform 3-D tracking of each person's skeleton from calibrated rectilinear cameras and then re-identify people between different camera views by clustering 3D trajectories. Each 3D-trajectory cluster is processed to produce a robust 3D trajectory for each person. This approach uses no appearance features for re-identification. Finally, (Wang et al., 2021) proposed precise localization of people indoors using up to 8 calibrated, time-synchronized, rectilinear cameras. They estimated 2-D skeletons of people using OpenPose library and projected them to 3-D space for distance-based clustering. This 3-D skeleton clustering is a form of location-based cross-frame PRID.

In these studies, cameras have a side view of the scene which can cause occlusions and no-match scenarios. To address this, one possible solution is to mount cameras overhead and point them down. However, due to a relatively-narrow FOV of rectilinear cameras, this solution would be impractical as it would require many cameras to cover a large space.

2.4 Cross-Frame Fisheye PRID

There exists a rich body of literature describing algorithms focused on detecting and tracking people in images captured by overhead fisheye cameras. To support this research a number of people-focused datasets captured by such cameras have been introduced, for example: BOMNI (Demiroz et al., 2012), MW (Ma et al., 2018), HABBOF (Li et al., 2019), CEPDOF (Duan et al., 2020), PIROPO (del Blanco

et al., 2021), WEPDTOF (Tezcan et al., 2022). These datasets were crucial for the development and validation of people detection methods, such as (Tamura et al., 2019; Li et al., 2019; Duan et al., 2020). One of the applications of these methods is in people counting and this works well in small-to-medium sized rooms. However, as the room gets larger, people appear smaller at the periphery of fisheye frames which makes it difficult even for a human labeler to reliably detect people. Clearly, for effective people counting in large spaces multiple cameras are needed. Unfortunately, this may lead to overcounting if the same person is seen in FOVs of several cameras. This is particularly serious for overhead fisheye cameras since their FOVs fully overlap. To avoid overcounting, PRID is needed so that each person is counted only once. However, to the best of our knowledge, this type of PRID has never been studied prior to this dissertation.

This dissertation introduces and develops *cross-frame fisheye PRID*. An illustration of the cross-frame fisheye PRID scenario is shown in Figure 2·1b. Note that appearance-matching is more challenging in this case on account of fisheye-lens geometric distortions as can be seen in Figure 3·2. Unlike in traditional PRID, we are interested in performing PRID for a pair of frames that are captured at the same time instant. This means there can be at most one match for a person in another camera view at this time instant. However, in traditional PRID a query element can have multple matches in the gallery set, which makes it more likely to find a correct match compared to cross-frame fisheye PRID that we study. Prior to the research work that forms this dissertation, there were no datasets that would have allowed us to study cross-frame fisheye PRID. None of the datasets listed earlier can be used as they were all collected with a single fisheye camera.

We would like to point out that cross-frame rectilinear PRID methods discussed in Section 2.3 are specialized to their respective use cases (e.g., recognizing jersey

Figure 2·2: Illustration of cross-frame fisheye PRID with two cameras. The goal is to establish a correspondence between people visible in two frames captured at the same time instant. At a given time instant, we consider the people in one camera view as "query" elements, and people in the other camera view as "gallery" elements. Yellow and green bounding boxes illustrate the two sets of elements. Either set can be treated as a query or gallery set.

numbers) or cannot be easily generalized to the overhead fisheye camera setting due to the unique lens distortion and viewpoint, e.g., skeletons and facial features cannot be recognized right below the camera.

2.5 PRID Focus of this Dissertation

In this dissertation, we focus on cross-frame fisheye PRID. The goal is to perform person re-identification between time-synchronized overhead fisheye cameras that have fully-overlapping field of views. To be more specific, we perform re-identification between sets of person-images, where each set of person-images is captured at the same time instant by different overhead fisheye cameras (where all cameras have fully-overlapping FOVs). For the majority of this thesis, we focus on the scenario where we have two sets of person-images at each time instant. We call these sets *query* and *gallery* sets. An illustration of a typical scenario for two cameras (i.e., two sets of person-images) is shown in Figure 2·2. In this dissertation we propose using cross-frame fisheye PRID to accurately estimate the number of occupants in a room using overhead fisheye cameras.

Chapter 3

Fisheye PRID Dataset and its Evaluation

As discussed in Chapter 2, the most thoroughly researched type of PRID is *traditional rectilinear PRID*, so the majority of PRID datasets comprise images that were captured by side-view, rectilinear-lens cameras with no FOV overlap (Gray and Tao, 2008; Hirzer et al., 2011; Karanam et al., 2019; Li et al., 2014; Loy et al., 2013; Wei et al., 2018; Zheng et al., 2015; Zheng et al., 2009). However, in this thesis, we explore PRID between images captured at the same time with fisheye cameras that have fully overlapping FOVs, which we call *cross-frame fisheye PRID*. The main goal in *cross-frame fisheye PRID* is to match identities between two fisheye images captured at the same time instant by cameras that have fully overlapping FOV (see Figure 2·2). Prior to the research of this dissertation, no public work and no datasets existed on cross-frame fisheye PRID.

In this chapter[1], we introduce the first-of-its-kind dataset for cross-frame fisheye PRID called *Fisheye Re-Identification Dataset with Annotations* (FRIDA). We also evaluate and compare the performance of 6 state-of-the-art traditional PRID methods on FRIDA under two training conditions: 1) when trained on a part of FRIDA and 2) when trained on the non-fisheye Market-1501 dataset (Zheng et al., 2015).

[1]This work was published at the 2022 IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS) (Cokbas et al., 2022)

3.1 FRIDA Dataset

FRIDA is the first PRID dataset captured indoors by multiple overhead fisheye cameras and is publicly available[2]. In FRIDA, the cameras have fully-overlapping FOVs (360° horizontally, i.e., parallel to the floor plane, and 185° vertically, i.e., orthogonal to the floor), unlike in typical PRID datasets, and are time-synchronized (frames are captured at the same time). FRIDA was collected in a 2,000 ft^2 room using 3 ceiling-mounted fisheye cameras (100 in above the floor). The bird's eye view of the room is shown in Figure 3·1, and an example of a time-synchronized frame triplet is shown in Figure 3·2. along with annotations. The frames were captured by three Axis M3057-PLVE cameras at 2,048×2,048-pixel resolution and 1.5 frames/sec. Annotations in FRIDA consist of 243,439 bounding boxes manually drawn around people. Table 3.1 compares FRIDA against the most widely used PRID datasets, developed since 2007, in terms of the number of bounding boxes and cameras, and frame resolution.

FRIDA can be used in a number of ways: as a still-image dataset for PRID, as a video dataset for people tracking, or as an image dataset for people detection and counting. In this thesis, to demonstrate its most unique features, we treat it as a still-image PRID dataset.

In the following subsections, we discuss the unique characteristics and challenges associated with FRIDA.

3.1.1 Annotations

At each time instant, three video frames are available with manually-drawn, human-aligned bounding boxes for all people visible in each frame. Each bounding box is represented by 6 parameters: x, y, w, h, α, ID, where (x, y) are the coordinates of its center, (w, h) are its width and height, α is its counter-clockwise rotation angle

[2] `vip.bu.edu/frida`

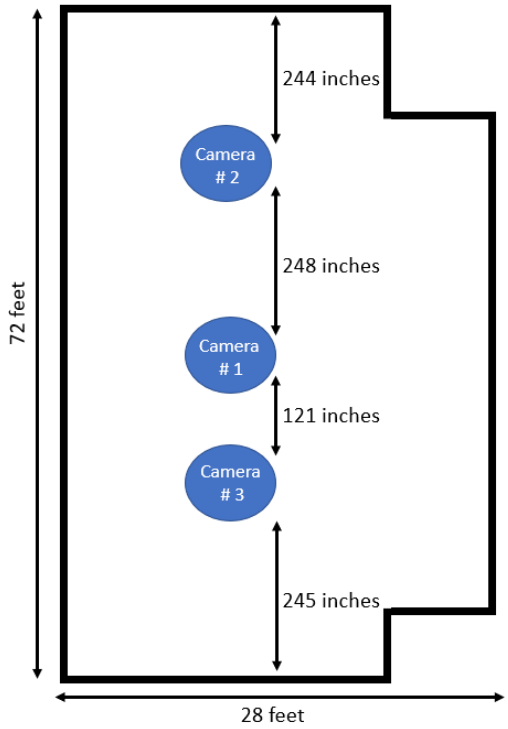


Figure 3·1: Bird's eye view of the space where FRIDA was recorded.

Table 3.1: Comparison of FRIDA with the most popular image datasets for person re-identification. (BBox = bounding box)

Dataset	Year	# BBoxes	# Cameras	Frame Resol.
VIPer (Gray and Tao, 2008)	2007	1,264	2	Fixed
iLIDS (Zheng et al., 2009)	2009	476	2	Variable
GRID (Loy et al., 2013)	2009	1,275	8	Variable
PRID 2011 (Hirzer et al., 2011)	2011	24,541	2	Fixed
CUHK03 (Li et al., 2014)	2014	13,164	2	Variable
Market-1501 (Zheng et al., 2015)	2015	32,668	6	Fixed
Airport (Karanam et al., 2019)	2017	39,902	6	Fixed
MSMT17 (Wei et al., 2018)	2018	126,441	15	Variable
FRIDA	2022	243,439	3	Fixed

Figure 3·2: Example of three synchronously-captured fisheye images with annotations from FRIDA (top: camera 2, middle: camera 1, bottom: camera 3).

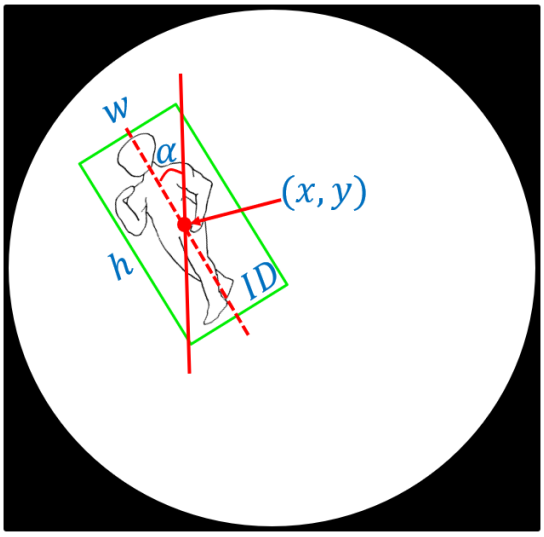


Figure 3·3: Illustration of bounding-box parameters available in FRIDA.

with respect to the vertical axis of the image, and ID is the ID number of a person (Figure 3·3). Each person in the dataset is assigned a unique ID which is consistent across all frames of the dataset. There are 20 unique ID numbers in FRIDA.

3.1.2 Scenarios

FRIDA consists of four segments where each segment captures a different type of challenge (Table 3.2). In segment #1, people enter the room, walk and sit down (people are evenly distributed in the room). This segment resembles a lecture where people remain seated for most of the time and their lower bodies are mostly occluded. Segment #2 is the most crowded and dynamic segment. People are constantly moving which occasionally causes severe occlusions, especially when people are close to each other. This segment resembles a social meeting where people are wandering around the room and talking to each other. Segment #3 is the longest one and has over 100,000 bounding boxes. Participants gather at either end or in the middle of the room, and stand close to each other leading to severe occlusions. Segment #4 is the shortest, with people leaving the room and causing occasional occlusions at the doors.

Table 3.2: Detailed information about FRIDA (*Fisheye Re-Identification Dataset with Annotations*).

Segment	Number of frames	Number of BBoxes	Number of BBoxes per frame	Scenarios/Challenges
#1	7,017	66,810	3-15	People coming in and settling down; evenly distributed around the room; mostly sitting (lower bodies mostly occluded)
#2	3,471	53,460	13-18	People walking around the room; significant occlusions
#3	6,207	103,141	13-17	Concentration of people in parts of the room; people standing and staying close to each other; people strongly occluding each other
#4	1,623	20,028	5-16	People leaving the room; occasional occlusions at entry/exit points

3.1.3 Gallery Set with Single Sample per ID

In typical PRID datasets, for a given query element there are multiple samples in the gallery set with the query ID. In FRIDA, however, frames are captured at the same time instant and the identities in one frame are treated as the query set while identities in another frame are treated as the gallery set. Therefore, for a given query element there can be at most one sample with the query identity in the gallery set. In some cases, due to occlusions, a person may not be visible in a camera's view. This may lead to a no-match scenario at certain time instants for some query elements. Note, that FRIDA can also be used for typical PRID by constructing the gallery from multiple images of the same ID captured at different times, but this is not in the focus of our work.

3.1.4 Synchronous and Overhead Capture

Due to the overhead placement of cameras and simultaneous capture, the viewpoint of a person directly under one camera may be dramatically different from the viewpoint from another camera. This is unlike in most other PRID datasets where it is common to capture a person from similar viewing angles (e.g., front, back, side, top) using different cameras. Then, if one of the gallery elements has the same viewpoint as the query, the chance of a match increases. However, in FRIDA, since the query and gallery elements are synchronously recorded by different overhead cameras, people never appear from the same viewpoint. This can be seen in Figure 3·2 where person #14 is seen from the top in camera #1 view, from the front in camera #3 view and from the back in camera #2 view. This makes the problem of PRID more challenging compared to other datasets.

3.1.5 Fisheye Distortions

Since FRIDA was recorded by fisheye cameras, images are subject to radial geometric distortions, especially close to FOV periphery. When a person is located at a different distance to each camera, the person's appearance is geometrically distorted to a different degree in each camera view. This makes the problem of PRID even more challenging compared to other datasets.

3.1.6 Resolution Mismatch Between Query and Gallery Sets

The synchronous, overhead capture and fisheye distortions often lead to very differently-sized bounding boxes for the same person (resolution mismatch). Examples can be seen in Figure 3·2, e.g., person #15 appears with very different resolutions in camera #2 and camera #3 views. In Figure 3·4, we demonstrate this resolution mismatch quantitatively. The resolution ratio R between two bounding boxes B_1 and B_2 is

defined as follows:

$$R = \frac{min(Area(B_1), Area(B_2))}{max(Area(B_1), Area(B_2))}. \tag{3.1}$$

Each data point in the plot shows the number of bounding-box pairs such that $R \leq \rho$ with $0 \leq \rho \leq 1$. Note that, the resolution mismatch is the largest (highest curve) between cameras 2 and 3 since they are farthest apart (Figure 3·1).

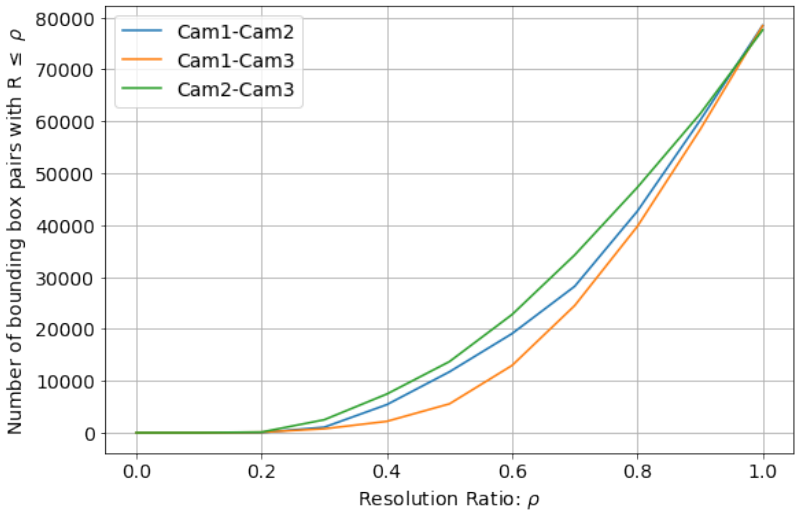


Figure 3·4: Bounding-box resolution mismatch for all camera pairs.

3.2 Evaluating SOTA Algorithms on FRIDA

In order to gauge challenges offered by FRIDA, we evaluated the performance of the following six SOTA traditional PRID algorithms on it: PCB (Sun et al., 2018), Pyramid (Zheng et al., 2019a), ABD-Net (Chen et al., 2019b), VA-reID (Zhihui et al., 2020), CTL (Wieczorek et al., 2021) and ResNet-50 (He et al., 2016). Detailed descriptions of these algorithms can be found in Section 2.1. They are all CNN-based and use person's appearance for re-identification.

When evaluating each algorithm, we used the training strategy and hyper-parameters suggested in the corresponding paper. To create query and gallery sets, we used the

ground-truth bounding boxes from FRIDA. It should be noted that in a real-world scenario (i.e., when no ground truth is available), bounding boxes are obtained by a people detection algorithm. Mis-detections produced by such an algorithm will degrade the PRID results.

3.2.1 Training-Testing Methodology

The same training and testing procedures (based on best practices commonly followed in the PRID literature) were applied to all methods. We trained each CNN as a classifier where we treated each identity as a different class. During testing, for a given pair of video frames, we treated all people from one frame as the query set and those from the other frame as the gallery set. Then, we fed the image of a person (within the person's bounding box) into the trained network and extracted a feature vector from the final convolutional layer to serve as this image's descriptor. We computed the cosine similarity between all feature vectors of the query and gallery sets, resulting in a score matrix $\mathbf{S}$ for each pair of frames[3]. We applied greedy matching to the score matrix to match the query and gallery elements as follows:

1. pick the maximum value in $\mathbf{S}$ and assume the corresponding query and gallery identities match,
2. remove the row and column of the matching identities from the matrix,
3. repeat the first 2 steps until no more matches are possible.

We used this algorithm since in FRIDA the fisheye cameras have a fully-overlapping FOV and, therefore, a person can have at most *one* match in another camera's FOV (and can be removed from the score matrix). This procedure yields a *matching* that associates query identities with gallery identities. A detailed pseudo-code for a

[3]Since cosine distance is symmetric, it is not important which frame is chosen as the query set and which as the gallery set.

possible implementation of the greedy algorithm is provided in Algorithm 1. In this code instead of deleting rows and columns of matched pairs, we assign a value of $-\infty$ to them. This is a convenient method to keep track of identities without changing the size of $\mathbf{S}$.

Algorithm 1 Greedy Algorithm

Ensure: $\mathbf{S}_{|Q_n|\times|G_n|} \in \mathbb{R}^2$

▷ $\mathbf{S}$ is the score matrix $\mathbf{S}_{|Q_n|\times|G_n|} = (s_{ij})_{1\leq i\leq|Q_n|,1\leq j\leq|G_n|}$

$0 \leq (s_{ij})_{1\leq i\leq|Q_n|,1\leq j\leq|G_n|} \leq 1$

while $\max_{1\leq i\leq|Q_n|,1\leq j\leq|G_n|}(s_{ij}) \neq -\infty$ **do**

$k, l \leftarrow \arg\max_{1\leq i\leq|Q_n|,1\leq j\leq|G_n|}(s_{ij})$

$k \sim l$ ▷ Match the k^{th} query and l^{th} gallery identity

$s_{kj}{}_{1\leq j\leq|G_n|} \leftarrow -\infty$

$s_{il}{}_{1\leq i\leq|Q_n|} \leftarrow -\infty$

end while

An alternative to this greedy approach is a minimization of the total distance between all query/gallery matches, e.g., by means of the Hungarian algorithm, but we found this method did not outperfrom the greedy approach while being significantly slower.

3.2.2 Dataset Splits

Despite more than 240,000 bounding boxes, FRIDA has only 20 different identities, Since this is insufficient for creating separate training, validation and testing sets, we evaluated the algorithms using 2-fold *identity-wise* cross-validation. We used half of the identities in training (training fold) and the other half in testing (testing fold),

and then we swapped the roles of identities and repeated the process. Specifically, we created the training set by choosing 50 random time stamps[4] for each identity from the training fold and taking 3 images of this identity (one from each camera) captured at this time (cameras are synchronized). This allowed a rich training set with many different viewpoints of the same person. The training set is specified in FRIDA. The testing set was composed of images of identities from the testing fold extracted from *all* frames containing such identities.

We also trained the networks on Market-1501 (Zheng et al., 2015) and tested them on FRIDA. Market-1501 is a commonly-used PRID dataset composed of images captured by side-view, rectilinear-lens cameras (different cameras capture a person at different times). For fairness, we used the same cross-validation sets as when both training and testing on FRIDA.

3.2.3 Evaluation Metrics

To evaluate the algorithms, we modify the commonly-used Correct Matching Score. We call our new performance metric the Query Matching Score (QMS), defined as follows:

$$QMS = \frac{\sum_{n=1}^{M} \sum_{q \in Q_n} \mathbb{1}(q = \widehat{q})}{\sum_{n=1}^{M} \left| Q_n \cap G_n \right|}$$

where M denotes the number of frames, Q_n, G_n are the sets of query and gallery identities in frame number n, respectively, and $\widehat{q}$ is the predicted identity of query q or "null" if there is no match. The important difference between QMS and CMS is that QMS accounts for situations when there is no match between a query and gallery elements ($|Q_n \cap G_n|$ in the denominator). Basically, QMS gives the ratio of the number of correct matches to the number of true matches.

[4]Since in most of the traditional PRID datasets the number of samples per person in the training set varies between 20 and 60, for a fair comparison we picked 50 time stamps from FRIDA for each person during training.

In addition to QMS, we also compute the commonly-used mean average precision (mAP) (Zheng et al., 2015). It is important to note that in our scenario there exists *at most* one matching gallery-frame identity for a given query element. Unlike in traditional PRID, we can encounter a query whose identity is absent from the gallery (due to complete occlusion). We exclude such cases from the mAP calculation.

3.2.4 Results

In Table 3.3, we report results when the algorithms are trained on Market-1501 and tested on FRIDA. In Table 3.4, we report results for the same algorithms, but here they are both trained and tested on FRIDA. These results are computed over all 4 segments of FRIDA for each camera pair. We also report the cumulative QMS value which is computed as the total number of correct matches from all camera pairs and all segments divided by the total number of possible correct matches from all camera pairs and all segments. In addition to QMS, we report mAP (the cumulative mAP is computed in a manner analogous to cumulative QMS). The common trend in both tables is that all algorithms achieve the highest QMS/mAP for cameras 1 and 3, and the lowest for cameras 2 and 3. This was to be expected since cameras 1 and 3 are the closest to each other (Figure 3·1); people are captured at a more similar resolution, viewpoint and geometric distortion compared to other camera pairs. Conversely, the distance between cameras 2 and 3 is the largest which makes PRID more challenging.

As Table 3.3 shows, when trained on Market-1501, Pyramid (Zheng et al., 2019a) performs the best among the six appearance-based methods and outperforms the second-best algorithm, CTL (Wieczorek et al., 2021), by 6.87% points in terms of cumulative QMS, and by 5.0% points in terms of cumulative mAP.

When these algorithms are trained on FRIDA (Table 3.4), CTL (Wieczorek et al., 2021) outperforms other networks by 3.87-14.11% points in cumulative QMS and by 1.75-7.57% points in cumulative mAP. For cameras 1 and 3, CTL performs above 90%

Table 3.3: Performance comparison of state-of-the-art algorithms trained on **Market-1501** and tested on **FRIDA** for different camera pairs. The highest values of QMS and mAP for each camera pair and cumulatively are shown in boldface.

	QMS [%]				mAP [%]			
	$1 \leftrightarrow 2$	$1 \leftrightarrow 3$	$2 \leftrightarrow 3$	Cum.	$1 \leftrightarrow 2$	$1 \leftrightarrow 3$	$2 \leftrightarrow 3$	Cum.
ResNet-50 (He et al., 2016)	57.63	73.99	45.33	59.04	70.03	79.41	59.33	69.63
PCB (Sun et al., 2018)	56.63	74.64	45.62	59.02	70.91	79.28	59.79	70.04
ABD (Chen et al., 2019b)	61.26	73.68	44.22	59.78	70.80	77.93	58.71	69.19
Pyramid (Zheng et al., 2019a)	**74.58**	**84.66**	**54.88**	**71.46**	**78.72**	**86.38**	**64.89**	**76.72**
VA-ReID (Zhihui et al., 2020)	60.79	74.18	44.99	60.06	71.21	78.68	59.31	69.78
CTL (Wieczorek et al., 2021)	66.92	83.68	42.88	64.59	72.57	84.99	57.44	71.72

Table 3.4: Performance comparison of state-of-the-art algorithms trained on **FRIDA** and tested on **FRIDA** for different camera pairs. The highest values for QMS and mAP for each camera pair and cumulatively are shown in boldface.

	QMS [%]				mAP [%]			
	$1 \leftrightarrow 2$	$1 \leftrightarrow 3$	$2 \leftrightarrow 3$	Cum.	$1 \leftrightarrow 2$	$1 \leftrightarrow 3$	$2 \leftrightarrow 3$	Cum.
ResNet-50 (He et al., 2016)	64.93	75.79	50.11	63.67	76.20	81.60	68.00	75.30
PCB (Sun et al., 2018)	63.30	74.79	51.77	63.33	75.79	81.23	67.91	75.01
ABD (Chen et al., 2019b)	75.31	83.18	62.05	73.57	82.43	85.16	74.81	80.83
Pyramid (Zheng et al., 2019a)	67.79	80.78	53.48	67.42	75.38	81.59	68.61	75.23
VA-ReID (Zhihui et al., 2020)	67.52	79.46	54.59	67.24	76.58	82.74	68.00	75.81
CTL (Wieczorek et al., 2021)	**77.30**	**90.11**	**64.76**	**77.44**	**82.7**	**89.79**	**75.17**	**82.58**

in terms of QMS. When trained on FRIDA, all networks achieve cumulative QMS above 63% and cumulative mAP above 75%.

Comparing the performance of algorithms trained on Market-1501 versus those trained on FRIDA, all the networks performed better when trained on FRIDA except for Pyramid. In terms of cumulative QMS, the improvement achieved by training ResNet-50, PCB, ABD, VA-ReID and CTL on FRIDA ranges from 4.31% to 13.79% points. In terms of cumulative mAP, these networks improve by 4.97% to 11.64% points by training on FRIDA. Considering the large number of bounding boxes in FRIDA, these margins correspond to thousands of correct matches between identities. It is impressive that training on Market-1501 using 750 identities and 9,928 bounding boxes is outperformed by training on FRIDA with only 10 identities and less than 1,500 bounding boxes. This suggests that for an effective PRID on overhead fisheye images, having a higher variability of the viewpoint (including overhead) for each identity is more important than having more identities with less viewpoint variability. We note, however, that Pyramid is an exception to this observation. This seems to suggest that Pyramid is able to leverage a plurality of identities more effectively than viewpoint variability.

In addition to Table 3.3 and Table 3.4, we report results for each segment separately in Table 3.5 and Table 3.6. In Table 3.5, the algorithms are trained on Market-1501 and in Table 3.6 they are trained on FRIDA.

In Table 3.5, all algorithms perform the best on Segment 4. Segment 4 is the shortest segment out of all segments. Moreover, in the second half of Segment 4, people remaining in the room have a good variability in the color of their outfits (purple sweatshirt, white sweatshirt, green tshirt etc.). In the experiments for Table 3.5, there was a large domain gap between training (rectilinear, side-view) and testing (fisheye, overhead) data which we believe caused algorithms to mostly rely

Table 3.5: Performance comparison of state-of-the-art algorithms trained on **Market-1501** and tested on **FRIDA** for different segments. The highest values for QMS and mAP for each segment and cumulatively are shown in boldface.

	QMS [%]					mAP [%]				
	Seg.1	Seg.2	Seg.3	Seg.4	Cum.	Seg.1	Seg.2	Seg.3	Seg.4	Cum.
ResNet-50	52.77	57.41	62.63	66.18	59.04	66.44	70.87	70.43	73.50	69.63
PCB	54.86	59.16	60.39	67.18	59.02	67.31	71.75	70.25	74.78	70.04
ABD	52.20	58.04	64.39	66.73	59.78	65.32	69.95	70.46	74.71	69.19
Pyramid	**70.12**	**70.50**	**72.75**	**76.56**	**71.46**	**78.05**	**77.32**	**75.40**	**80.20**	**76.72**
VA-ReID	53.20	58.32	64.05	68.77	60.06	65.87	71.34	70.81	74.88	69.78
CTL	57.23	67.81	67.02	68.35	64.59	67.43	74.82	72.23	75.28	71.72

Table 3.6: Performance comparison of state-of-the-art algorithms trained on **FRIDA** and tested on **FRIDA** for different segments. The highest values for QMS and mAP for each segment and cumulatively are shown in boldface.

	QMS [%]					mAP [%]				
	Seg.1	Seg.2	Seg.3	Seg.4	Cum.	Seg.1	Seg.2	Seg.3	Seg.4	Cum.
ResNet-50	62.87	62.67	63.74	68.49	63.67	75.19	75.42	74.75	79.17	75.30
PCB	63.69	63.61	63.79	62.19	63.33	76.23	76.13	74.03	75.24	75.01
ABD	**76.27**	77.80	70.40	73.84	73.57	**84.14**	83.09	77.94	81.66	80.83
Pyramid	64.36	68.33	68.18	74.39	67.42	74.21	76.99	74.55	79.99	75.23
VA-ReID	65.72	68.75	67.86	67.88	67.24	76.26	77.97	74.63	77.08	75.81
CTL	74.60	**81.66**	**77.42**	**79.66**	**77.44**	81.68	**86.00**	**81.33**	**84.82**	**82.58**

on color information since "color" does not change much between the two domains. Similarly to Table 3.3, the best performing algorithm on all segments, when trained on Market-1501, is Pyramid (Zheng et al., 2019a), its performance exceeding 80% points in terms of mAP in Segment 4.

In Table 3.6, almost all algorithms perform better compared to Table 3.5. This is due to the fact that both training and testing sets consist of fisheye frames. In Table 3.6, there is no single segment on which all algorithms perform the best, unlike in Table 3.5. Pyramid (Zheng et al., 2019a) performs the best on Segment 1, VA-ReID (Zhihui et al., 2020) and CTL (Wieczorek et al., 2021) perform the best on Segment 2, and the remaining algorithms perform the best on Segment 4. We believe this is due to the fact that the domain-gap between training and testing sets is small (fisheye frames in training and testing sets), which leads each network to focus on different parts of the appearance due to different architectures. In Table 3.6, the best performing algorithm is CTL and its performance is improved by 10%-17% points for different segments compared to Table 3.5.

3.3 Chapter Summary and Discussion

We introduced FRIDA, the first image dataset for person re-identification from overhead fisheye cameras. The dataset is unique not only for the camera type used but also for their overlapping fields of views that is often encountered when counting people in large spaces. This leads to a new type of PRID - matching of people "seen" by different cameras at the same time.

We evaluated the performance of 6 state-of-the-art PRID algorithms on FRIDA. These algorithms were all CNN-based and used appearance of people for re-identifying them. We demonstrated that training these algorithms on fisheye images improves performance when testing on fisheye images, which is not surprising. However, even

when trained on FRIDA, the best one achieved below 83% in cumulative mAP. This suggests there is much space for improvement.

Chapter 4

Location-Based Person Re-Identification

In the previous chapter, we demonstrated the performance of 6 state-of-the-art *traditional rectilinear* PRID algorithms on FRIDA. All these algorithms used appearance of people to extract meaningful features for PRID. However, these algorithms did not perform as well on FRIDA as they do on rectilinear PRID datasets (even when they were trained on fisheye images). This performance drop is largely due to differences in people's appearance between images from different cameras that are caused by fisheye distortions and significant viewpoint change. However, in *cross-frame* PRID, when groups of images are captured at the same time, an *unoccluded* person in the scene appears at specific pixel locations in fisheye images from *all* cameras. The pixel location is unique in each image since only one person can occupy a 3-D location in the scene. We leverage this observation and propose a *location-based* approach to *cross-frame fisheye* PRID, which uses location instead of appearance to match identities.

In this chapter [1], we are introducing a cross-frame fisheye PRID algorithm that solely depends on the location of people. Compared to appearance-based methods, this location-based approach is less sensitive to body-viewpoint variability, body-size distortions and occlusions. However, it relies on image correspondence between two fisheye cameras. We develop a mathematical relationship for this correspondence using the *unified spherical model* (Geyer and Danilidis, 2001; Courbon et al., 2012) for our cameras and propose a new calibration procedure to jointly estimate their

[1] This work was published in the 2021 IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS) (Bone et al., 2021)

intrinsic and extrinsic model parameters including a novel calibration-data collection method. The image-to-image correspondence depends on the height of a person which is unknown. The challenge is in developing a methodology to handle these unknown and varying heights of people. We propose 4 PRID algorithms that use different metrics to quantify location-based match of identities for a range of typical human heights. We evaluate the 4 algorithms on FRIDA using 2-fold cross-validation with the same procedure for dataset splitting as proposed in Chapter 3. We demonstrate that all 4 algorithms perform significantly better than the state-of-the-art appearance-based methods that were evaluated in the previous chapter.

4.1 Fisheye Image Pixel-Correspondence Model

In this section, we develop a mathematical relationship that maps a point in an image captured by one fisheye camera to the corresponding point in an image captured by another fisheye camera. Our development uses the *unified spherical model* (USM) proposed by Geyer and Danilidis (Geyer and Danilidis, 2001) and validated for fisheye cameras by Courbon et al. (Courbon et al., 2012). The USM can be described as a function $F : \mathbb{R}^3 \mapsto \mathbb{Z}^2$, where the domain consists of 3D-world coordinates and the range – 2D pixel locations in an image.

4.1.1 Notation

Consider a system composed of two overhead fisheye cameras (mounted on the ceiling of a room) with overlapping fields of view. We assume the cameras have parallel optical axes, orthgonal to the ceiling and floor, and are installed at the same height. This is a common configuration in most indoor surveillance scenarios. Figure 4·1 depicts this configuration, where both cameras (A and B) are represented by USM. Let the projection centers of cameras A and B be located at $\boldsymbol{O}_A = [0, 0, 0]^T$ and $\boldsymbol{O}_B = [d_x, d_y, 0]^T$, respectively. Clearly, the 3D-world coordinate system (X, Y, Z)

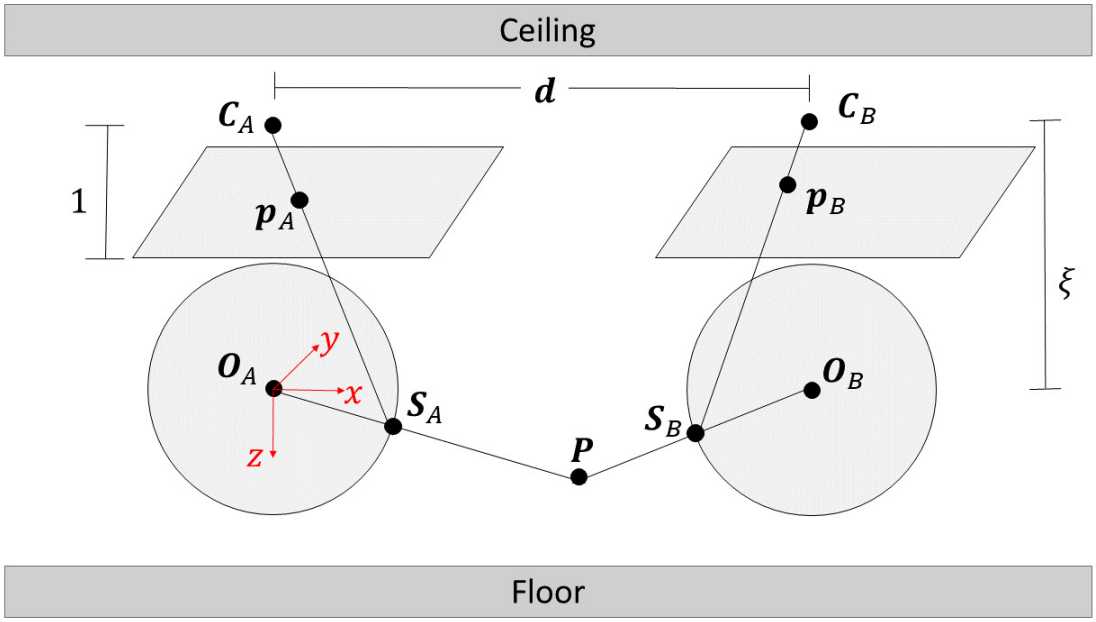


Figure 4·1: Projection model for two parallel fisheye cameras with $\xi_A = \xi_B = \xi$.

is centered at $\boldsymbol{O}_A$, i.e., associated with camera A. Let the optical centers of both cameras be located at $\boldsymbol{C}_A = [0, 0, -\xi_A]^T$ and $\boldsymbol{C}_B = [d_x, d_y, -\xi_B]^T$, respectively where $\xi_A, \xi_B > 0$. Finally, let each camera's normalized image plane be orthogonal to the Z axis and at the distance of 1 from the respective optical centers.

4.1.2 Forward Mapping

In this section, we describe the function which maps a 3D-world point at location $\boldsymbol{P} = [P_x, P_y, P_z]^T$ (Figure 4·1) to the pixel coordinates in each camera view. Given the displacement $\boldsymbol{d} = [d_x, d_y, 0]^T$ between cameras A and B, the orthogonal projection of $\boldsymbol{P}$ onto the unit spheres centered at $\boldsymbol{O}_A$ and $\boldsymbol{O}_B$ is, respectively, given by:

$$\boldsymbol{S}_A = \frac{\boldsymbol{P}}{\|\boldsymbol{P}\|}, \quad \boldsymbol{S}_B = \frac{\boldsymbol{P} - \boldsymbol{d}}{\|\boldsymbol{P} - \boldsymbol{d}\|} + \boldsymbol{d}.$$

A perspective projection of points at $\boldsymbol{S}_A$ and $\boldsymbol{S}_B$ onto the normalized (homogeneous) image planes of each camera (with $\boldsymbol{C}_A$ and $\boldsymbol{C}_B$ used as the respective projection centers), results in the following homogeneous coordinates $\boldsymbol{p}_A$ and $\boldsymbol{p}_B$ (relative to the origin at $\boldsymbol{C}_A$):

$$\boldsymbol{p}_A := [p_{A,x}, p_{A,y}]^T = \left[\frac{P_x}{P_z + \xi_A \|\boldsymbol{P}\|}, \frac{P_y}{P_z + \xi_A \|\boldsymbol{P}\|}, 1 \right]^T \tag{4.1}$$

$$\boldsymbol{p}_B := [p_{B,x}, p_{B,y}]^T = \left[\frac{P_x - d_x}{P_z + \xi_B \|\boldsymbol{P} - \boldsymbol{d}\|} + d_x, \frac{P_y - d_y}{P_z + \xi_B \|\boldsymbol{P} - \boldsymbol{d}\|} + d_y, 1 \right]^T \tag{4.2}$$

In order to obtain pixel coordinates $\boldsymbol{x}_A, \boldsymbol{x}_B$, we transform the normalized image-plane coordinates $\boldsymbol{p}_A, \boldsymbol{p}_B$ using intrinsic-parameter matrices $\boldsymbol{K}_A$ and $\boldsymbol{K}_B$ of cameras A and B, respectively. Since in practice it is difficult to keep the x and y axes of the two cameras perfectly aligned, we allow a rotation of camera B by angle θ about the Z axis. These two operations are expressed as matrix multiplications:

$$\boldsymbol{x}_A = \boldsymbol{K}_A \boldsymbol{p}_A \tag{4.3}$$

$$\boldsymbol{x}_B = \boldsymbol{R}_\theta \boldsymbol{K}_B \boldsymbol{p}_B \tag{4.4}$$

The intrinsic parameters of camera $i \in \{A, B\}$ are given by:

$$\boldsymbol{K}_i = \begin{bmatrix} k_{i,x} & 0 & \gamma_{i,x} \\ 0 & k_{i,y} & \gamma_{i,y} \\ 0 & 0 & 1 \end{bmatrix}$$

where $k_{i,x}$ and $k_{i,y}$ are scaling factors in the horizontal and vertical directions, respectively, and $(\gamma_{i,x}, \gamma_{i,y})$ is the projection of a camera's optical center onto the normalized image plane (Courbon et al., 2012). The rotation by angle θ about the Z axis is

expressed through the matrix:

$$\boldsymbol{R}_\theta = \begin{bmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The composition of mappings described in equations (4.1) and (4.3), along with the final rounding of $\boldsymbol{x}_A$ to integer pixel locations, describes function F_A that maps $\boldsymbol{P}$ to $\boldsymbol{x}_A$:

$$\boldsymbol{x}_A = F_A(\boldsymbol{P}; \boldsymbol{\omega}_A), \tag{4.5}$$

where $\boldsymbol{\omega}_A := \{\xi_A, \boldsymbol{K}_A\}$ are parameters of the function. Similarly, the composition of mappings described in equations (4.2) and (4.4), along with the rounding of $\boldsymbol{x}_B$, describes function F_B that maps $(\boldsymbol{P}, \boldsymbol{d})$ to $\boldsymbol{x}_B$:

$$\boldsymbol{x}_B = F_B(\boldsymbol{P}, \boldsymbol{d}; \boldsymbol{\omega}_B), \tag{4.6}$$

where $\boldsymbol{\omega}_B := \{\xi_B, \boldsymbol{K}_B, \boldsymbol{R}_\theta\}$ are parameters of the function.

4.1.3 Pixel-Correspondence Mapping

In order to establish a mapping from pixel coordinates $\boldsymbol{x}_A$ in camera A to pixel coordinates $\boldsymbol{x}_B$ in camera B, we need to invert F_A (4.5). Since all 3D-world points located on a line through camera's projection center are mapped to the same pixel coordinates, F_A is not a 1-to-1 mapping and, therefore, not invertible. However, if the z-coordinate P_z of a 3D-world point $\boldsymbol{P}$ projected to a pixel at location $\boldsymbol{x}_A$ is available, we can recover the x and y coordinates of $\boldsymbol{P}$ as follows. First, we compute

$$\boldsymbol{p}_A = \boldsymbol{K}_A^{-1} \boldsymbol{x}_A. \tag{4.7}$$

Given $\boldsymbol{p}_A$ and P_z, the x and y coordinates of $\boldsymbol{P}$ are:

$$[P_x, P_y]^T = P_z \cdot [u \cdot p_{A,x}, u \cdot p_{A,y}]^T \tag{4.8}$$

where:

$$u = \frac{1 + \xi_A \sqrt{1 + (1 - \xi_A^2)(p_{A,x}^2 + p_{A,y}^2)}}{1 - \xi_A^2 (p_{A,x}^2 + p_{A,y}^2)} \tag{4.9}$$

Equation (4.9) is derived in the supplementary material (Appendix A) where we also show that:

$$0 \leq \xi_A \leq \frac{1}{\sqrt{p_{A,x}^2 + p_{A,y}^2}}. \tag{4.10}$$

The composition of mappings described in equations (4.7) and (4.8) defines a mapping from the tuple $(\boldsymbol{x}_A, P_z)$ to the 3D-world point $\boldsymbol{P}$, which we denote as G_A:

$$\boldsymbol{P} = G_A(\boldsymbol{x}_A, P_z; \boldsymbol{\omega}_A). \tag{4.11}$$

Composing the mappings described in equations (4.11) and (4.6) we get:

$$\begin{aligned} \boldsymbol{x}_B &= F_B(G_A(\boldsymbol{x}_A, P_z; \boldsymbol{\omega}_A), \boldsymbol{d}; \boldsymbol{\omega}_B). \\ &=: G_{AB}(\boldsymbol{x}_A, P_z, \boldsymbol{d}; \boldsymbol{\omega}) \end{aligned} \tag{4.12}$$

where $\boldsymbol{\omega} := \{\boldsymbol{\omega}_A, \boldsymbol{\omega}_B\}$. Equation (4.12) defines the pixel-correspondence map from $\boldsymbol{x}_A$ to $\boldsymbol{x}_B$ with knowledge of the two additional inputs P_z and $\boldsymbol{d}$, and parameters $\boldsymbol{\omega}$ which will be learned from training data during camera calibration as described in the next section.

4.2 Camera Calibration

The geometric model described in Section 4.1 is characterized by

$$\boldsymbol{\omega} = \{\xi_A, \xi_B, \boldsymbol{K}_A, \boldsymbol{K}_B, \boldsymbol{R}_\theta\},$$

containing a total of 11 scalar parameters which must be estimated for each camera type and relative placement of the cameras. Clearly, if both cameras are identical,

then $\xi_A = \xi_B = \xi$ and $\boldsymbol{K}_A = \boldsymbol{K}_B = \boldsymbol{K}$ and then only 6 instead of 11 scalar parameters need to be estimated. We estimate $\boldsymbol{\omega}$ by minimizing the sum of squared Euclidean distances between N given matched pairs of pixel locations from two identical cameras $(\boldsymbol{x}_A^j, \boldsymbol{x}_B^j), j = 1, \ldots, N$, with each pair corresponding to the same 3D-world point:

$$\widehat{\boldsymbol{\omega}} = \arg\min_{\boldsymbol{\omega}} \sum_{j=1}^{M} \|\boldsymbol{x}_B^j - G_{AB}(\boldsymbol{x}_A^j, P_z, \boldsymbol{d}; \boldsymbol{\omega})\|^2. \tag{4.13}$$

The displacement $\boldsymbol{d}$ is assumed to be known and all 3D-world points used for camera calibration have, by design (as described below), the same (known) z-coordinate P_z.

In order to solve the minimization in equation (4.13), a calibration dataset of matched pixel pairs $(\boldsymbol{x}_A^j, \boldsymbol{x}_B^j)$ is needed. We collected and annotated a calibration dataset, with 3 overhead fisheye cameras arranged in the linear configuration depicted in Figure 3·1.

In the interest of minimizing the human effort required for calibration, we applied a semi-automated process to collect matched $(\boldsymbol{x}_A^j, \boldsymbol{x}_B^j)$ pairs. We rolled a cart around the room equipped with a colored spherical LED light mounted at a known height (known P_z) and synchronously captured images by all three cameras (Figure 4·2). We applied color thresholding to each image to automatically find the LED location in pixel coordinates. In total, we collected M=1,173 matched point triplets (pixel coordinates from all three cameras corresponding to the same 3D-world point) for calibration.

We used stochastic gradient descent to minimize the objective in equation (4.13). During each update step, we applied constraint (4.10) if ξ was to violate this constraint after updating. Because each camera can be treated as either camera A or camera B, there are six ordered pairs of cameras in our experimental setup (Figure 3·1). We train on data from all 6 camera pairs in a fixed order, and repeat this cycle 5,000 times with a learning rate of 10^{-5}. Figure 4·3 shows an example of the accuracy of

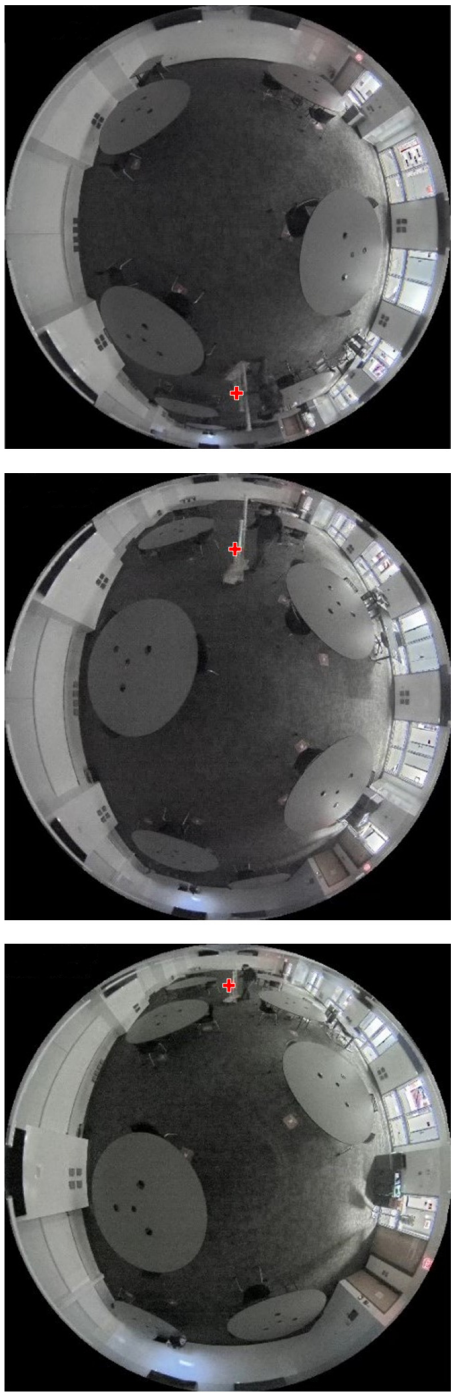


Figure 4·2: Three synchronously-captured images showing a mobile cart with colored spherical LED light. From left to right are shown images from cameras 2, 1, 3, respectively (Figure 3·1). The LED light location is found by color thresholding and shown as a red cross.

image-to-image mapping obtained with the estimated parameters $\widehat{\boldsymbol{\omega}}$.

4.3 Application of the Pixel-Correspondence Model to Person Re-Identification

In this section, we develop a re-identification method based on geometric, rather than appearance, constraints. Let $\{\boldsymbol{x}_A^i\}_{i=1}^K$ be the set of known pixel locations of K people (query subjects) in a fisheye image captured by camera A, and $\{\boldsymbol{x}_B^j\}_{j=1}^L$ the pixel locations of L people (gallery subjects) in an image simultaneously captured by camera B. Due to occlusions and detection errors, it is sometimes the case that $K \neq L$. Our goal is to find which query subjects in camera A correspond to gallery subjects in camera B based on their respective locations.

Using equation (4.12), for the i-th query person's location in camera A, namely $\boldsymbol{x}_A^i$, we can estimate this person's location in camera B as follows:

$$\widehat{\boldsymbol{x}}_B^i(Z) = G_{AB}(\boldsymbol{x}_A^i, Z, \boldsymbol{d}; \widehat{\boldsymbol{\omega}}), \tag{4.14}$$

if we know Z, which is related to the height of the person. We assume that the person's pixel location in an image is the center of a close-fitting rectangular box bounding the person's body in the image. Taking the average person's height as 168 cm and knowing the camera installation height of 254 cm, we set the expected value of the center of a person's body Z_{avg} to 170 cm (254 cm minus one half of person's average height). In order to accommodate a range of people's heights and also different body poses (e.g., sitting, bending over a laptop, leaning on a table) we also consider a range of Z values from 150 cm to 190 cm in steps of 2 cm: $Z \in Z_R := \{150, 152, 154, \ldots, 190\}$ cm, which corresponds to heights from 128 cm to 208 cm.

Figure 4·4(a) shows query locations $\boldsymbol{x}_A^i$ (crosses) of 4 people in camera $A = 2$, each associated with one color. Figure 4·4(b) shows their true locations $\boldsymbol{x}_B^i$ (gallery)

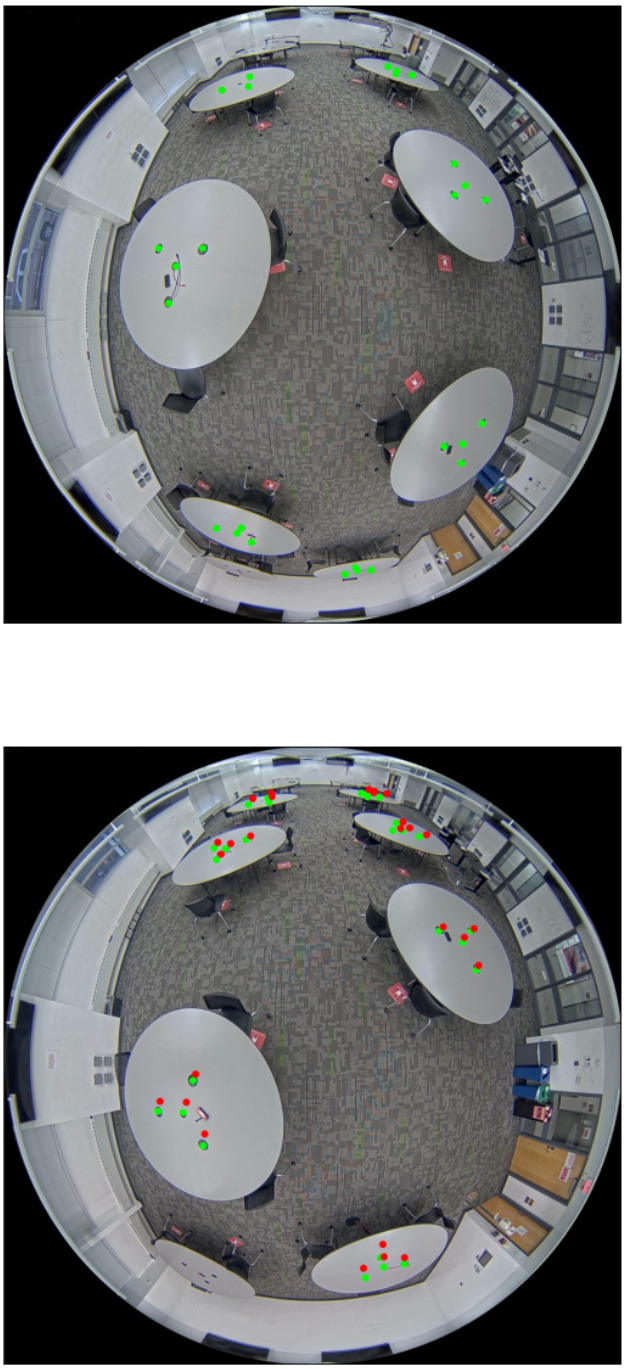
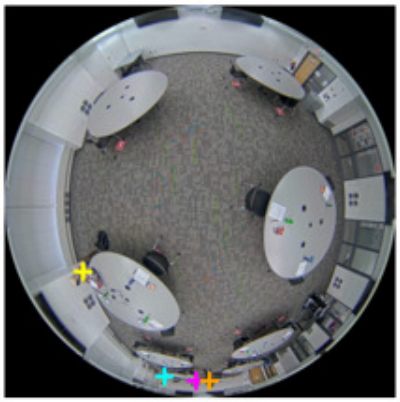
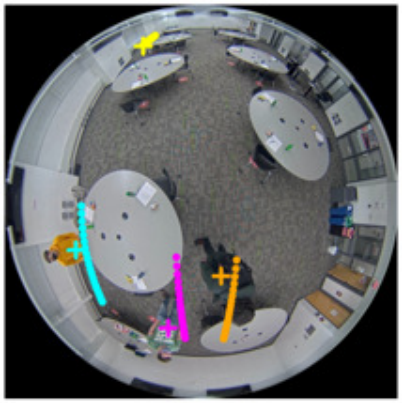


Figure 4·3: Illustration of the mapping given by Equation (4.12) from camera $A = 1$ (left) to camera $B = 3$ (right) produced by the calibrated geometric model. The green points are ground-truth locations and the red points are predicted locations.



(a) Camera 2

(b) Camera 3, $Z = 170$cm



(c) Camera 3, $Z \in Z_R$

Figure 4·4: Illustration of location predictions for 4 people (cyan, yellow, magenta, orange). (a) Query-subject locations in camera 1 (crosses). (b) Gallery-subject locations in camera 3 (crosses) and predicted query locations for single $Z = Z_{\text{avg}} = 170$cm (4 bullets). (c) Gallery-subject locations in camera 3 (crosses) and predicted query locations for $Z \in Z_R$ (4 sequences of bullets).

in camera $B = 3$, also denoted by crosses, as well as their predicted locations $\widehat{\boldsymbol{x}}_B^i(Z)$ for $Z = Z_{\text{avg}}$, denoted by bullets. Note that the predicted locations are quite accurate except for the person who is far away from camera 1 but almost directly under camera 3 (orange). This is the worst-case scenario since the same distance in 3-D space is severely contracted at FOV periphery and expanded in FOV center. A mapping in the opposite direction would result in a much smaller error. The distances between predicted and known locations in camera B can be used to match identities between the two cameras. Similarly, Figure 4·4(c) shows true locations $\boldsymbol{x}_B^i$ in camera B and *sequences* of predicted locations $\widehat{\boldsymbol{x}}_B^i(Z)$ for $Z \in Z_R$. Each predicted location in a sequence is associated with a different assumed height of this person and can be used to match identities between the two cameras.

4.3.1 Distance Measures

We consider four methods, described below, to measure the distance $\boldsymbol{D}_{ij}$ between the pixel location of the j^{th} gallery subject in camera B and the predicted pixel location (or a set of locations) in camera B of the i^{th} query subject:

- **Point-to-Point Distance (PPD):** Taking $Z_{\text{avg}} = 170$ cm, we measure the distance between the i^{th} query subject's predicted location in camera B and the j^{th} gallery subject's location as:

$$\boldsymbol{D}_{i,j} = \|\widehat{\boldsymbol{x}}_B^i(Z_{\text{avg}}) - \boldsymbol{x}_B^j\|.$$

  This method invokes the geometric mapping only once for each query location and is therefore the fastest among all methods that we propose.

- **Point-to-Set Minimum Distance (PSMD):** We accommodate different people's heights and body poses by producing a set of query subject's predicted locations $\{\widehat{\boldsymbol{x}}_B^i(Z)\}_{Z \in Z_R}$ in camera B. Then, we measure the minimum distance

between the set of query subject's predicted locations and each gallery subject's location:

$$\boldsymbol{D}_{i,j} = \min_{Z \in Z_R} \|\widehat{\boldsymbol{x}}_B^i(Z) - \boldsymbol{x}_B^j\|.$$

- **Point-to-Set Total Distance (PSTD):** Again, we use the same range of Z values as in PSMD, but we measure the sum of distances between the query subject's predicted locations and each gallery subject's location:

$$\boldsymbol{D}_{i,j} = \sum_{Z \in Z_R} \|\widehat{\boldsymbol{x}}_B^i(Z) - \boldsymbol{x}_B^j\|.$$

- **Count-Based Distance (CBD):** We use the same range of Z values as in PSMD and PSTD, but instead of computing a distance we conduct a "vote". The vote of the j^{th} gallery subject's location is the number of Z values for which the j^{th} gallery location is closest to the predicted query location among all other members in the gallery. We define:

$$\boldsymbol{D}_{i,j} = |Z_R| - \sum_{Z \in Z_R} \mathbb{1}\left[\left(\arg\min_k \ \|\widehat{\boldsymbol{x}}_B^i(Z) - \boldsymbol{x}_B^k\|\right) = j\right]$$

where $\mathbb{1}$ is the indicator function and ties in $\arg\min$ are broken randomly. We note that unlike other distance measures, $\boldsymbol{D}_{ij}$ in CBD depends on not just $\boldsymbol{x}_B^j$, but on all the gallery subject's locations $\boldsymbol{x}_B^1, \ldots, \boldsymbol{x}_B^L$.

4.3.2 Identity Matching

Given two sets of identities in camera A and camera B, the goal is to find a match between the sets: $\{\boldsymbol{x}_A^i\}_{i=1}^K$ and $\{\boldsymbol{x}_B^j\}_{j=1}^L$. In the previous section, we showed how to come up with the distance matrix $\boldsymbol{D}_{i,j}$ which consists of distances between the predicted location of the i-th query subject and the location of the j-th gallery subject both in camera B. However, this can be reversed, and we can map the query locations

from camera B to camera A. In other words, we can assign $\boldsymbol{x}_B^j$ to be a query subject's location in camera B and similarly assign the set $\{\boldsymbol{x}_A^i\}_{i=1}^K$ to be the gallery subject locations in camera A.

Similarly to mapping (4.14), for the j-th person's location in camera B, namely $\boldsymbol{x}_B^j$, we can estimate this person's location in camera A as follows:

$$\widehat{\boldsymbol{x}}_A^j(Z) = G_{BA}(\boldsymbol{x}_B^j, Z, \boldsymbol{d}; \widehat{\boldsymbol{\omega}}), \tag{4.15}$$

and compute distances $D_{j,i}$ by replacing $\|\widehat{\boldsymbol{x}}_B^i(Z) - \boldsymbol{x}_B^j\|$ with $\|\widehat{\boldsymbol{x}}_A^j(Z) - \boldsymbol{x}_A^i\|$. Since it is unclear which matching direction would produce more accurate results, we propose to use a combination of two distance matrices: $\boldsymbol{D} = \boldsymbol{D}_{AB} + \boldsymbol{D}_{BA}^T$, where $\boldsymbol{D}_{AB}$ denotes the distance matrix when a query is in camera A and the gallery is in camera B. Then, we apply greedy algorithm to $\boldsymbol{D}$ to find matches. We perform the greedy algorithm in a similar way to that described in Section 3.2.1, however this time we perform *minimization* instead of maximization.

Note, that bidirectional matching was not needed in Section 3.2.1 since the similarity between two identities was computed through cosine similarity which is symmetric. Thus, regardless of which camera identities were chosen as query or gallery, the result would have been the same.

4.4 Experimental Results

Our experimental setup includes three cameras, so we need to learn two θ values: angle θ_1 between cameras 1 and 2, and angle θ_2 between cameras 1 and 3. Since we use 3 identical cameras (Axis M3057-PLVE), there are a total of 7 scalar parameters to be learned, that is ξ, four entries of the $\boldsymbol{K}$ matrix, and angles θ_1 and θ_2. Estimates of these parameters produced by the camera calibration method described in Section 4.2 are shown in Table 4.1.

ξ	θ_1	θ_2	k_x	k_y	γ_x	γ_y
2.0129341	0.003265	-0.0086144	2.0210802	1.9735336	0.0059010	0.0111225

Table 4.1: Estimated parameter values $\widehat{\boldsymbol{\omega}}$ for our experimental setup.

We evaluated the PRID performance of our algorithms on FRIDA (Chapter 3). For a fair comparison, we used the same data splits as proposed in Section 3.2.2. Note, that this time instead of using RGB values from each bounding box, we used the pixel location of the center of each bounding box. In each pair, one camera serves as a source of query locations and the other serves as the source of gallery locations. We used the same bounding boxes as in Section 3.2, however, this time instead of using the content of the bounding boxes, we only used coordinates of their centers. People detection was not in the scope of this study.

Similarly to Section 3, we use QMS and mAP as evaluation metrics. Table 4.2 shows both metrics for location-based PRID for different camera pairs. Over all camera pairs CBD is a consistent winner. The main reason CBD outperforms PSMD and PSTD is that it does not depend on the distance but instead depends on counts. During a height sweep, some predicted query locations may be far away from the correct gallery location thus biasing PSMD and PSTD by such distance outliers. On the other hand, in CBD all query locations contribute the same value (count of 1) to $\boldsymbol{D}$ thus improving robustness to outliers.

We expected that PPD should have the lowest computational complexity among all 4 algorithms since it uses a single Z whereas other algorithms sweep a range of Z's. Running the algorithms for all camera pairs over all frames on Intel(R) Core(TM) i7-4790K CPU@4.00GHz resulted in average time per frame-pair of 4.63 ms for PPD and between 70.49 ms and 71.28 ms for the other algorithms. This about 15-fold slowdown is consistent with the 21-fold increase in the number of Z values tested. However, the

Table 4.2: Performance comparison of the location-based PRID for different camera pairs. The highest values for QMS and mAP for each camera pair and for the cumulative are shown in boldface.

	QMS [%]				mAP [%]			
	1↔2	1↔3	2↔3	Cum.	1↔2	1↔3	2↔3	Cum.
PPD	85.39	90.73	87.92	88.01	93.35	95.67	92.74	93.93
PSMD	92.52	90.44	89.25	90.75	95.37	95.92	93.87	95.06
PSTD	86.81	90.81	88.66	88.75	93.71	95.67	92.77	94.06
CBD	**94.63**	**92.62**	**92.07**	**93.11**	**97.13**	**97.22**	**96.55**	**96.97**

Table 4.3: Performance comparison of the location-based PRID for different segments of FRIDA. The highest values for QMS and mAP for each algorithm are shown in boldface.

	QMS [%]					mAP [%]				
	Seg.1	Seg.2	Seg.3	Seg.4	Cum.	Seg.1	Seg.2	Seg.3	Seg.4	Cum.
PPD	99.51	87.39	80.60	90.77	88.02	99.49	94.47	90.13	93.95	93.93
PSMD	99.58	91.21	85.69	87.99	90.75	**99.83**	95.62	91.98	94.16	95.06
PSTD	**99.69**	88.08	81.79	91.03	88.76	99.60	94.69	90.22	94.19	94.06
CBD	99.17	**92.18**	**89.81**	**93.69**	**93.11**	99.69	**96.82**	**95.41**	**96.51**	**96.97**

PPD approach is sensitive to the selection of Z value (in these experiments, we set the person's height to 168 cm, the US average).

Table 4.3 shows the performance of geometry-based PRID algorithms for each segment of FRIDA. Clearly, all 4 algorithms did extremely well on segment 1 in which people are spread out fairly uniformly in the room and are never very close to each other. On the other hand, in segment 3 people stand very close to each other posing difficulties for location-based matching, resulting in the lowest performance among all segments. As expected, the algorithm based on the PPD distance metric (single query location mapping using an average person's height) achieves the lowest

performance among the four algorithms. Algorithms that use a range of heights of people perform better with the CBD-based algorithm achieving the best performance in terms of the cumulative QMS (93.11%) and cumulative mAP (96.97%).

It is interesting to compare the performance of appearance-based algorithms from Table 3.4 and the performance of location-based methods from Table 4.2. It is clear that the location-based algorithms perform significantly better than the appearance-based methods. In fact, the worst performing location-based algorithm PPD achieves cumulative QMS of 88.01% and cumulative mAP of 93.93%, which is significantly better than the best-performing appearance-based algorithm which achieves cumulative QMS of 77.44% and cumulative mAP of 82.58%.

Another interesting point is that, both types of algorithms struggle on different segments of FRIDA. When we compare Table 3.6 and Table 4.3, we can see that location-based methods did extremely well (all of them achieve QMS above 99%) on segment 1 in which people are spread out fairly uniformly in the room and are never very close to each other. On the other hand, appearance-based methods achieve at most 76.27% QMS because people are coming into the room and spending some time in their coats which look similar to each other. However, on segment 3 all location-based algorithms perform worse than on the other segments, because in this segment people stand very close to each other. On the other hand, some appearance-based methods perform better on segment 3 than they performed on segment 1. The maximum performance difference of appearance-based methods between different segments is 7.4%-points in QMS and 6.2 %-points in mAP. A similar difference for location-based methods is larger: 18.91%-points in QMS and 9.36%-points in mAP.

4.5 Chapter Summary and Discussion

We have proposed a supervised training methodology that leverages the unified spherical model to jointly estimate the extrinsic and intrinsic parameters of a network of overhead fisheye cameras. We have shown that this model can be used to accurately predict pixel locations across camera views, and that this can be used for person re-identification between time-synchronized overhead fisheye cameras with fully overlapping FOVs (i.e., in cross-frame fisheye PRID). In this setting, we have shown that our location-based approaches can vastly outperform state-of-the-art appearance-based methods. The best-performing location-based method reaches almost 97% in cumulative mAP computed across all FRIDA segments and camera pairs. This is close to a perfect re-identification. Only in high-density scenarios (people close to each other causing severe occlusions), does its performance drop to about 95%. However, location-based algorithms require calibration of each camera type used and additional measurements for each camera layout.

Chapter 5

Spatio-Visual Fusion-Based Person Re-Identification

In Chapter 3, we evaluated six state-of-the-art appearance-based PRID algorithms. As discussed in Section 3.2.4, these algorithms did not perform as well as they do in traditional rectilinear PRID. Among the key reasons for the performance drop are fisheye-lens distortions, different viewpoints of the same person and different resolutions of the corresponding person-images, especially pronounced when the person is located at very different distances from both cameras. In Chapter 4, we by-passed these issues by introducing a cross-frame fisheye PRID approach that solely relies on the location of people. Thanks to its resistance to fisheye distortions and resolution mismatch, location-based approach outperformed the appearance-based PRID methods with a significant margin. However, as discussed in Section 4.4, the performance of location-based PRID degrades when people are standing close to each other.

In this chapter we propose a framework for cross-frame fisheye PRID that combines appearance and location information to improve performance. The methods and results of this chapter were published in (Cokbas et al., 2023) When people are very close to each other, matching identities using location information is often perilous due to location-estimation errors, as discussed in Section 4.4. In this case, visual characteristics of a person may help disambiguate any confusion as, for example, in the left column of Figure 5·1. On the other hand, when people look very similar (e.g., due to color of clothing or body shape), appearance-based methods frequently fail to

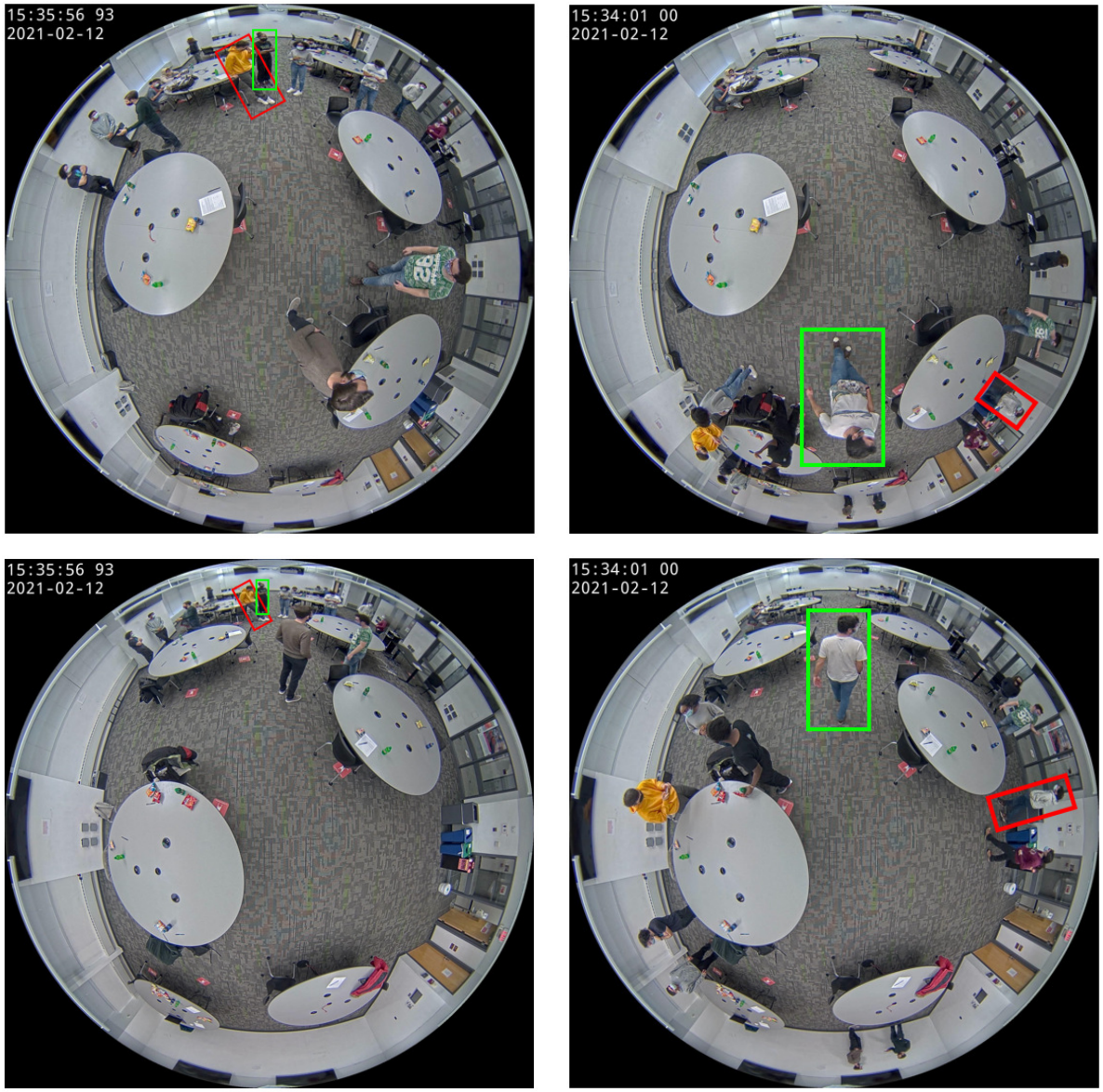
differentiate between them. However, a 3-D location in a room can be occupied by at most one person, so location information should help with correct matching as in the example shown in the right column of Figure 5·1. Only when two similar-looking people are next to each other would this approach would fail, but then a human eye would likely fail as well.

For appearance information, we propose to combine hand-crafted and deep learning-based features. As the former, we use a color histogram computed from a bounding box around each person. As the latter, we use an embedding from a state-of-the-art PRID deep-learning method (Wieczorek et al., 2021) developed for rectilinear cameras and fine-tuned on fisheye data (domain transfer). More detailed description of this method can be found in Chapter 3. In order to obtain person-location information, we use the method developed in Chapter 4. We use Naïve Bayes fusion to combine the color, embedding and location data by converting each to a similarity score (between two identities) and multiplying the scores after *suitable* normalization. We perform identity matching by a greedy algorithm.

We evaluate the proposed framework on FRIDA (Chapter 3) and include an ablation study where we demonstrate that a combination of all three feature types performs up to 33% points better in QMS and up to 27% points better in mAP than when using a feature individually.

5.1 Methodology

In order to perform cross-frame PRID between two fisheye frames, as we have been doing in Chapter 3 and 4, we designate one of the frames as a *query* frame and the other frame as a *gallery* frame. We denote the sets of identities of people in the query and gallery frames at time t_n as Q_n and G_n, respectively. We compute a $|Q_n| \times |G_n|$ score matrix where, $|\cdot|$ is the cardinality operator. The score in the i^{th} row and j^{th}



Location ambiguity Appearance ambiguity

Figure 5·1: Illustration of location and appearance ambiguity. When people are very close to each other, distinct color of clothing and/or body shape/features may help resolve location ambiguity (left column). When people have very similar appearance (e.g., light-gray T-shirts and dark pants), knowing their location may help resolve appearance ambiguity (right column).

column of this matrix represents the similarity between the i^{th} query element and the j^{th} gallery element. Each score is a combination of 3 scores, each derived from a different type of feature: (1) neural-network embedding; (2) color histogram; and (3) location, as shown in Figure 5·2. Below, we discuss how each feature is computed, how it is converted to a pairwise-similarity or -dissimilarity score, how these scores from different features are converted into match-probabilities and fused together and, finally, how the matching between query and gallery elements is performed based on the fused match-probabilities.

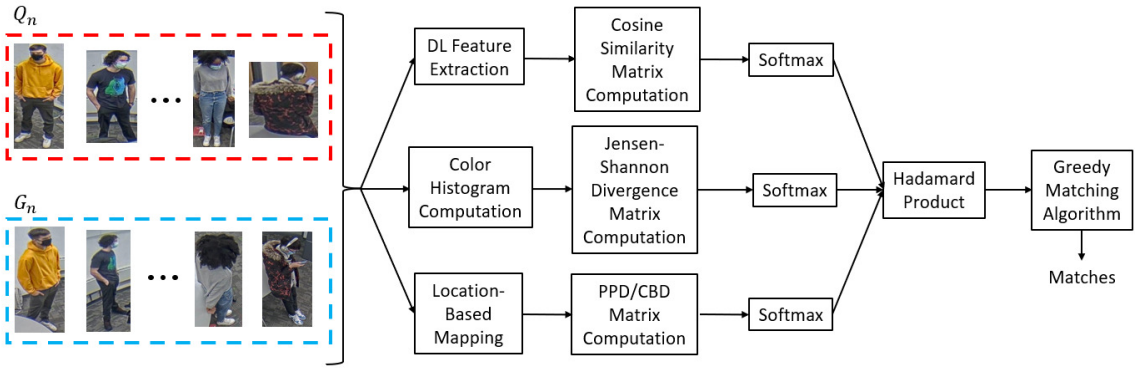


Figure 5·2: Block diagram of the proposed method. In the first step, features are extracted from the contents of bounding boxes (for appearance-based methods) or their positions within the frames (for location-based method) for the query and gallery identity sets (which are denoted by Q_n and G_n, respectively, for frame number n). In the next step, features of query-gallery pairs are converted into either pairwise-similarity or pairwise-dissimilarity scores. They are then normalized via the softmax operation to obtain match-probability matrices for each feature group which are then fused together via the Hadamard product into a fused match-probability matrix. Finally, a greedy sequential matching algorithm based on the fused match-probability matrix is used to produce the query-gallery matches.

5.1.1 Deep-Learning Features and Pairwise-Similarity Scores

Deep-learning methods perform exceedingly well in many visual inference tasks, including rectilinear-camera PRID. It is only natural to consider features extracted by

such methods for fisheye PRID as well.

Features: To extract deep-learning features, we pass the content of each bounding box (assumed given) through the CTL deep neural network (Wieczorek et al., 2021). We opt for CTL since, as we demonstrated in Section 3.2.4, it is the top-performing model among 6 state-of-the-art PRID models evaluated on fisheye data. We use the output of the last convolutional layer in the CTL model as the feature vector.

Pairwise-Similarity Scores: We compute cosine similarity between the features of the query and gallery elements to obtain a $|Q_n| \times |G_n|$ pairwise-similarity matrix at time t_n. Since cosine similarity is symmetric, its values are unaffected by whether a given camera view is designated as query or gallery.

5.1.2 Color Histograms and Pairwise-Dissimilarity Scores

In Chapter 3, we showed that CTL's performance improved when trained on fisheye images instead of rectilinear images. However, even when trained on fisheye images, its features are still affected by the challenges mentioned in Section 3.1 (e.g., fisheye distortions, resolution mismatch, viewpoint mismatch). One feature that is not impacted by these challenges is *color*; an example can be seen in Figure 5·3. We leverage this observation, by using color histogram as a hand-crafted identifier of each person.

Features: Rather than using a 3-D RGB histogram, we convert RGB values to HSV space and then compute a 1-D histogram of hue H and normalize it to obtain a probability distribution over discretized hue values. In this way, we reduce the impact of illumination variations (change in value V) and color variations under different lighting (change in saturation S).

Pairwise-Dissimilarity Scores: If $\mathbf{q}$ and $\mathbf{g}$ are query and gallery hue probability distributions, respectively, we measure their dissimilarity by calculating the Jensen-

Figure 5·3: Illustration of the importance of color information in fisheye PRID. The person in the red bounding box in the top frame is severely shrunk, however the distinctive color of the sweatshirt still allows to distinguish this person from others.

Shannon (JS) divergence D_{JS} between them as follows:

$$D_{JS}(\mathbf{q}||\mathbf{g}) \;=\; \frac{1}{2}D_{KL}(\mathbf{q}||\mathbf{p}) + \frac{1}{2}D_{KL}(\mathbf{g}||\mathbf{p})$$

where $\mathbf{p} = (\mathbf{q}+\mathbf{g})/2$ and $D_{KL}(\cdot||\cdot)$ is the Kullback-Leibler (KL) (Kullback and Leibler, 1951) divergence between two probability distributions.

Although both KL divergence and JS divergence are commonly used as measures of distance between two density/distribution functions, we opt for JS divergence because it is symmetric (unlike KL divergence) and it is always finite (the KL divergence $D_{KL}(\mathbf{a}||\mathbf{b})$ is infinity if there is a component i where $b_i = 0$ but $a_i > 0$). Note that the final result is a $|Q_n| \times |G_n|$ divergence matrix (larger values denote larger dissimilarity) unlike in the case of pairwise-similarity matrix in Section 5.1.1.

5.1.3 Location-Based Features and Pairwise-Dissimilarity Scores

The deep-learning features and color histograms capture visual appearance of people. However, as mentioned earlier, due to the overhead viewpoint and fisheye-lens distortions the appearance of people may change dramatically between cameras. This difficulty was observed and addressed in Chapter 4 where we proposed to use the location of people rather than their appearance for re-identification.

In this method, as described in Section 4.1.3, we first *inverse-map* the 2D-pixel coordinates (i.e., centers of bounding boxes) from a camera's FOV to their 3D-world coordinates. Then, we map these 3D-world coordinates to another camera's FOV through *forward-map* described in Section 4.1.2. Finally, we match identities using a distance measure (Section 4.3.1) between locations of the mapped query elements and the gallery elements. One caveat to this approach is that the mapping model requires knowledge of a person's height. In Section 4.3.1, we addressed this issue by either assuming that every person has the same height (PPD) or by sweeping a range

of typical heights for each person (PSMD, PSTD, CBD).

Features: We use two distances proposed in Section 4.3.1, specifically the Point-to-Point Distance (PPD), which is computationally efficient, and the Count-Based Distance (CBD), which is best-performing but significantly more complex.

Pairwise-Similarity Scores: Since neither PPD nor CBD metric is symmetric, we compute a symmetric version of pairwise-dissimilarity matrix as detailed in Section 4.3.2. More specifically, we swap the query and gallery designations of cameras, and compute another distance matrix. Then, we transpose this matrix and add it to the unswapped distance matrix to arrive at the final pairwise-dissimilarity matrix.

5.1.4 Fusion of Features

While each feature type can be used individually to perform PRID, our motivation is to leverage appearance features (deep-learning and color histogram) to disambiguate location uncertainty while relying on location to differentiate individuals with similar appearance. Ideally, one would select suitable features for each potential match, but this is a more difficult problem that should be considered in future research.

In this work, we combine features for all potential matches in the same way by means of a probabilistic information fusion mechanism. Recall that for each feature type and query/gallery sets we obtain a $|Q_n| \times |G_n|$ pairwise-similarity or pairwise-dissimilarity matrix. We normalize each row of either matrix by applying the *softmax* operator with positive sign in the exponent to pairwise-similarity matrices and with negative sign to pairwise-dissimilarity matrices as follows:

$$(\mathbf{S}_{N-DL})_{ij} = \sigma[(\mathbf{S}_{DL})_{ij}] = \frac{e^{\frac{(\mathbf{S}_{DL})_{ij}}{T}}}{\sum_{k=1}^{K} e^{\frac{(\mathbf{S}_{DL})_{ik}}{T}}} \tag{5.1}$$

$$(\mathbf{S}_{N-CH})_{ij} = \sigma[(D_{CH})_{ij}] = \frac{e^{-\frac{(D_{CH})_{ij}}{T}}}{\sum_{k=1}^{K} e^{-\frac{(D_{CH})_{ik}}{T}}} \tag{5.2}$$

$$(\mathbf{S}_{N-LOC})_{ij} = \sigma[(D_{LOC})_{ij}] = \frac{e^{-\frac{(D_{LOC})_{ij}}{T}}}{\sum_{k=1}^{K} e^{-\frac{(D_{LOC})_{ik}}{T}}} \tag{5.3}$$

where σ denotes the softmax operator, T is a temperature parameter, and K is the gallery size for a given video frame pair. The similarity matrices for deep-learning features before and after normalization are denoted by $\mathbf{S}_{DL}$ and $\mathbf{S}_{N-DL}$, respectively. Similarly, D_{CH} and D_{LOC} denote the dissimilarity matrices for color-histogram features and location-based features before normalization and $\mathbf{S}_{N-CH}$ and $\mathbf{S}_{N-LOC}$ denote their normalized versions. Note that in Equations (5.2) and (5.3), the exponents of softmax are negative to convert the dissimilarity matrices into similarity matrices. This converts both types of matrices into row-stochastic matrices where each row represents the conditional probabilities of gallery elements (columns) matching a given query (row).

Finally, we use the Naïve Bayes methodology to fuse the conditional probabilities of different features by taking the Hadamard product (element-wise multiplication) of the conditional probability matrices of different features to obtain the final match-probability matrix (Figure 5·2):

$$(\mathbf{S}_{DL+CH+LOC}) = \mathbf{S}_{N-DL} \odot \mathbf{S}_{N-CH} \odot \mathbf{S}_{N-LOC} \tag{5.4}$$

where $\odot$ is the Hadamard-product operator.

5.1.5 Matching Algorithm

Regardless of whether each feature type is used separately or is combined with one or two other feature types, the final match-probability matrix contains elements that describe the normalized degree of similarity between a query identity and a gallery identity. In order to match query and gallery identities, one could maximize the sum of logarithms of match-probabilities (or equivalently the product of match-probabilities)

for all possible matches via the Hungarian algorithm, but this is computationally expensive. Instead, as discussed in Section 3.2.1, we apply greedy matching to matrix **S** of match-probabilities between query and gallery elements. The *probability of the matching* is taken to be the product of the match-probabilities for all the query-gallery matches in the matching.

Due to the row-wise normalization described in Section 5.1.4, swapping query and gallery sets would likely produce different results. Therefore, we consider both cases by applying row-wise normalization to the original matrix (either pairwise-similarity or pairwise-dissimilarity) and to its transposed version. Then, we apply the greedy algorithm outlined above to both normalized matrices and compute the *probability of the matching* (for the best greedy sequential matching) for each matrix. As the final identity match, we use the pairings provided by the matrix with the higher probability of matching. We apply this approach to individual features and to all combinations thereof.

5.2 Experimental Results

Just as in Sections 3.2 and 4.4, in this section we focus on the *re-identification* problem, independent of the people *detection* problem. Therefore, we assume that people detections (i.e., bounding boxes) are provided by an annotated dataset. In practice, application of PRID will also be affected by people mis-detections, which will be further discussed in Chapter 6.

5.2.1 Dataset Splits

In order to validate the proposed algorithms, we used FRIDA (Cokbas et al., 2022), as we have been doing in Chapters 3 and 4.

Among the three feature types we proposed, only deep-learning features need training (CTL). To train and evaluate CTL, we adopt the 2-fold cross-validation

methodology that was described in Section 3.2.2 – half of the identities are in one fold and the remaining half are in the other fold. Although color-histogram and location features do not require training, in order to ensure fair comparison, we adopt the same testing approach that was used for CTL to evaluate the performance of all our methods. We note that in this testing methodology, for a given pair of time-synchronized video frames, all people from one frame (whose identities belong to a fold) are treated as the query set and those from the other frame (again with identities belonging to the same fold) are treated as the gallery set.[1]

5.2.2 Implementation Details

We extracted deep-learning features from each bounding box using CTL with a ResNet backbone (Wieczorek et al., 2021). We trained CTL on NVIDIA Tesla V100 GPUs using Adam optimizer with a learning rate of 3.5e-4, weight decay of 5e-4 and momentum of 0.937 over 300 iterations.

In order to match the bounding box size to the one that CTL accepts, we applied zero padding to maintain the aspect ratio and then resized the image. When computing color histograms, we used 256 bins and normalized each histogram to sum up to 1. When computing location-based features, we used an average person's height of 168 cm in the PPD metric and 21 heights ranging from 128 cm to 208 cm in the CBD metric. In the softmax normalization of features, we set the temperature parameter to $T = 10$ [2].

[1]Note that in real-life scenarios, there might be cases where an identity in the query frame may not appear in the gallery frame due to occlusion or mis-detection by a person-detection algorithm. These scenarios will be discussed in Chapter 6.

[2]The value of T was determined heuristically via a loose grid search using values that are an order of magnitude apart.

5.2.3 Results

In this section, we report the PRID performance for deep-learning features (DL), color histograms (CH) and location-based features (LOC) separately, and also for three combinations of two feature types (DL+CH, CH+LOC, DL+LOC) and for all three feature types together (DL+CH+LOC) for two distance metrics: PPD and CBD. All the results are summarized in Table 5.1. To evaluate the performance of the proposed algorithms, we use the same evaluation metrics (i.e., QMS and mAP) that have been used in Chapters 3 and 4.

We note that since we are using an annotated dataset, each bounding box corresponds to a person truly visible in a camera's FOV. This allows us to demonstrate the re-identification performance of each algorithm without the confounding influence of errors in people detection.

In addition to reporting camera-pairwise QMS and mAP values (e.g., "1↔2"), we also report cumulative values ("Cum.") computed by using the total number of correct matches and the total number of *possible* correct matches over all camera pairs rather than for a single pair.

It can be observed, that among individual features the location-based one significantly outperforms DL and CH. Furthermore, the color histogram performs at least 12% points below DL in cumulative QMS and mAP. Clearly, color is hardly enough to distinguish people, but DL features which capture richer properties of objects, including color, perform much better. However, a person's location captured by time-synchronized cameras is a much better indicator of who is who.

Combining the appearance-based features (DL+CH), improves DL's performance by about 4% points in cumulative QMS and about 2% points in cumulative mAP. However, the DL+CH combination still performs well below location-based PRID. Interestingly, for Camera 1-Camera 3 pair the QMS performance of DL+CH combina-

Table 5.1: Performance comparison of PRID on FRIDA dataset for various combinations of deep-learning (DL), color-histogram (CH) and location-based (LOC) features, for both PPD and CBD distance measures. The highest values of QMS and mAP for each camera pair (e.g., "1↔2") and for the cumulative ("Cum.") metric are shown in boldface.

				QMS [%]				mAP [%]			
DL	CH	LOC (PPD)	LOC (CBD)	1↔2	1↔3	2↔3	Cum.	1↔2	1↔3	2↔3	Cum.
✓				80.34	91.12	66.89	79.50	83.31	90.78	76.00	83.40
	✓			60.41	85.84	44.91	63.80	67.68	85.60	58.46	70.63
✓	✓			83.58	94.79	70.65	83.06	84.82	93.58	76.55	85.02
		✓		94.76	95.83	92.51	94.37	94.84	96.78	93.97	95.20
	✓	✓		95.41	96.33	93.66	95.14	95.13	97.12	94.24	95.50
✓		✓		97.09	98.25	95.15	96.84	95.78	97.73	94.83	96.12
✓	✓	✓		97.31	**98.42**	95.45	97.07	95.92	97.88	94.93	96.25
			✓	96.63	95.01	93.28	94.98	98.18	97.92	96.92	97.68
	✓		✓	97.23	96.37	94.67	96.10	98.48	98.27	97.22	98.00
✓			✓	**98.07**	97.66	96.05	97.27	**98.83**	99.24	**97.63**	**98.57**
✓	✓		✓	98.03	97.75	**96.22**	**97.34**	98.66	**99.26**	97.18	98.37

tion is close to the one for the location-based algorithms (LOC/PPD and LOC/CBD). This is so because these two cameras are closer to each other in the physical world compared to the other two camera pairs (see Figure 3·1 for camera layout). Thus, the resolution difference for a person positioned between these two cameras is the smallest. This makes appearance-based PRID more accurate compared to the other camera pairs. Indeed, the DL+CH combination performs worst for Camera 2-Camera 3 pair which are farthest apart. The location-based approach (single feature) also performs better when the cameras are closer to each other, but the performance difference between the best and the worst cases is no more than 4% points in terms of QMS or mAP (compared to over 24% points QMS and over 17% points mAP for DL+CH).

The most significant performance boost, when using two feature types, comes from the combination DL+LOC as these are the two best performing approaches individually. Combining LOC/PPD with DL boosts its performance by 2.47% points in terms of cumulative QMS. Similarly, combining LOC/CBD with DL delivers performance boost of 2.29% points in cumulative QMS. The performance improves further, although very slightly, when color histogram (CH) is combined with DL+LOC features. When we use PPD distance metric, performance reaches 97.07% in cumulative QMS and 96.25% in cumulative mAP, whereas when we use CBD distance metric it achieves 97.34% and 98.37%, respectively. However, one should note that the inclusion of CH in the DL+LOC/CBD combination can have a slightly detrimental effect.

Our results also support conclusions reached in Chapter 4, namely that algorithms that involve the CBD location metric perform better than the ones that involve the PPD metric. Indeed, LOC/CBD outperforms LOC/PPD by 0.61% points of cumulative QMS. However, when we compare DL+CH+LOC/PPD and DL+CH+LOC/CBD in Table 5.1 the performance gap between the two decreases to 0.27% points of

cumulative QMS. Considering that PPD is around 17 times faster than CBD (see Section 4.4 for a computational complexity analysis), it seems a better option for real-time system implementation.

5.3 Chapter Summary and Discussion

We proposed a multi-feature PRID framework for time-synchronized fisheye cameras with overlapping fields of view. To the best of our knowledge, this is the first work that explores combining appearance- and location-based features for PRID. A key technical contribution of our work is a novel probabilistic feature-fusion methodology for identity matching.

Our experiments show that methods which utilize location information have a high identity-matching accuracy. However, this requires knowledge of camera parameters (both intrinsic and extrinsic). In some scenarios, this information may not be available. Appearance-based methods, on the other hand, do not make use of such information and can be applied to any camera type and any camera-layout topology. However, such methods lag performance-wise behind those that use location-based features. Still, appearance-based features are valuable and do provide a boost to the identity-matching performance when combined with location-based features. Clearly, there is still much room for improvement in appearance-based PRID using overhead fisheye cameras.

Chapter 6

Application of PRID to Occupancy Analysis

In Chapters 3, 4 and 5, we reported PRID performance on FRIDA, where *human-annotated ground-truth* bounding boxes were available for each person in all frames. However, in real-world scenarios ground-truth bounding boxes are not available. In this chapter, we focus on the application of PRID to *people counting* in a real-life, large-space scenario where multiple fisheye cameras are needed to cover the whole area.

We start by introducing evaluation metrics for assessing the people-counting performance of algorithms. Then, using 3 key metrics, we evaluate performance of a state-of-the-art people-detection algorithm using a single fisheye camera. These results show that while a single camera is sufficient for accurately counting people in small-to-medium sized rooms, it is insufficient in large spaces.

Therefore, we propose to use two fisheye cameras for counting people in large-space scenarios. However, with two cameras in a room, a person may be visible in views from both cameras, potentially leading to over-counting. Applying PRID is essential for avoiding such errors. We demonstrate the people counting performance of several PRID approaches proposed in the previous chapters. In this real-life scenario, people-detection algorithms will occasionally miss a detection or produce a false one. Therefore, people-counting results presented in this chapter are impacted by both people detection and people re-identification errors.

The largest space we conducted our people-counting experiments in has a 2,000 ft^2 area, and two fisheye cameras are sufficient to obtain high accuracy. However, there exist larger spaces, such as large lecture halls, convention centers, supermarkets, airports, bus/train stations, etc. In such scenarios, more than two fisheye cameras may be needed for reliable occupancy estimation. To address this problem, we present solutions to scale the two-camera PRID algorithms proposed in Chapter 4 to N cameras ($N > 2$). Finally, we evaluate the occupancy estimation performance of these N-camera cross-frame fisheye PRID algorithms for $N = 3$.

6.1 Evaluation Metrics

Our goal is to perform fine-grained occupancy estimation (people counting, a regression problem) rather than occupancy detection (binary classification into empty/occupied classes). Although some prior works (Elkhoukhi et al., 2022; Szczurek et al., 2017) treat people counting as a classification problem, in practice it is very difficult to train a classifier for 100 classes (occupancy up to 100) since many diverse examples for each class in various scenarios are needed.

Let η_i be the true people count in fisheye frame number i, and let $\widehat{\eta}_i$ be the corresponding people-count estimate. Let M be the total number of fisheye frames from which occupancy is being estimated. The two most often used performance metrics in regression are the Mean-Absolute Error (MAE) and the Root Mean-Squared Error (RMSE):

$$MAE := \frac{1}{M}\sum_{i=1}^{M}|\widehat{\eta}_i - \eta_i|, \quad RMSE := \sqrt{\frac{1}{M}\sum_{i=1}^{M}(\widehat{\eta}_i - \eta_i)^2}. \tag{6.1}$$

Both are frequently used in the people-counting literature, but we opt for MAE in this thesis since it is more robust to outliers than $RMSE$.

However, MAE is not useful when comparing results for different occupancy sce-

narios (e.g., 10 versus 100 people). To account for this, relative MAE (and similarly relative RMSE) has been used, for example in (Kim et al., 2019)):

$$MAE_{rel} := \frac{1}{M}\sum_{i=1}^{M}\frac{|\widehat{\eta}_i - \eta_i|}{\eta_i}.$$

However, it is undefined for frames with true occupancy equal to zero (empty room) which is a major deficiency. Therefore, normalization by the dynamic range of occupancy has been proposed (Choi et al., 2021), which for MAE is defined as follows:

$$NMAE := \frac{\frac{1}{M}\sum_{i=1}^{M}|\widehat{\eta}_i - \eta_i|}{\eta_{max} - \eta_{min}},$$

where η_{min} and η_{max} are the minimum and maximum of the true people count for frames $i = 1, ..., M$, respectively. Still, $NMAE$ is undefined when the occupancy is constant for all M frames, which is not an unlikely scenario.

To address this, in (Cokbas et al., 2020) (also in Appendix B), we proposed a new metric. Rather than scaling MAE (or $RMSE$) by the dynamic range of true occupancy, we proposed to scale it by the *average* of true occupancy as follows:

$$MAE_{pp} := \frac{\frac{1}{M}\sum_{i=1}^{M}|\widehat{\eta}_i - \eta_i|}{\frac{1}{M}\sum_{i=1}^{M}\eta_i}. \tag{6.2}$$

The only scenario when MAE_{pp} is undefined happens when a space remains completely empty throughout the whole experiment, which is a very special case that can be handled separately. MAE_{pp} expresses MAE value as a fraction of average occupancy so different occupancy scenarios (e.g., 10 versus 100 people) can be fairly compared. Moreover, MAE_{pp} can be thought of as a percentage value which may be easier to interpret (e.g., MAE_{pp} of 0.1 can be thought of as a 10% error).

Works that consider people counting as a classification problem also report accuracy defined as the number of frames when $\widehat{\eta}_i = \eta_i$ expressed as a fraction of M in percent. Since this requires an exact match between the true and estimated accuracy,

in practice this fraction is not very high. We propose a modified version of accuracy, that we call X-Accuracy and define as follows[1]:

$$Acc_X = \frac{1}{M}\sum_{i=1}^{M} \mathbf{1}_X(\widehat{\eta}_i - \eta_i), \qquad \mathbf{1}_X(z) := \begin{cases} 1 & \text{if } |z| \leq X \\ 0 & \text{if } |z| > X \end{cases} \tag{6.3}$$

For $X = 0$ this definition reverts to the original definition of accuracy, but for larger values of X it tolerates the departure of $\widehat{\eta}_i$ from η_i by up to X.

6.2 People-Counting Dataset

In order to evaluate the people-counting performance of PRID methods proposed in this thesis, we collected video frames over 3 days using the same camera setup in the same 2,000 ft^2 room where FRIDA was captured (Figure 3·1). We captured time-synchronized frame triplets about every 5 seconds. Over the three days, the occupancy ranged from 0 (empty room, mostly at night) to a crowded lecture with almost maximum allowed occupancy of the classroom. In particular, on the first day there were 11 high-occupancy periods (lectures) with up to 87 students (Figure 6·1(a)). On the second day there were 4 high-occupancy periods with up to 65 students (Figure 6·1(b)), while on the third day the classroom was mostly empty with only one period when up to 9 students were present (Figure 6·1(c)). At each time instant, we counted the number of people present in the classroom by visually inspecting the captured frames. This provides ground-truth people count, shown in Figure 6·2, for performance evaluation.

[1] A similar measure was introduced in (Tezcan et al., 2018) for the special case of $X = 1$.

(a) Day 1

(b) Day 2

(c) Day 3

Figure 6·1: Sample frames from each day of the 3-day dataset.

6.3 Occupancy Estimation Using Single Fisheye Camera

In Chapter 1, we mentioned that there exist people-detection algorithms developed for fisheye images. Out of these algorithms, the best-performing one to-date is the 'Rotation-Aware People Detection" (RAPiD) algorithm (Duan et al., 2020). In this chapter, we demonstrate RAPiD's performance for occupancy estimation while mon-

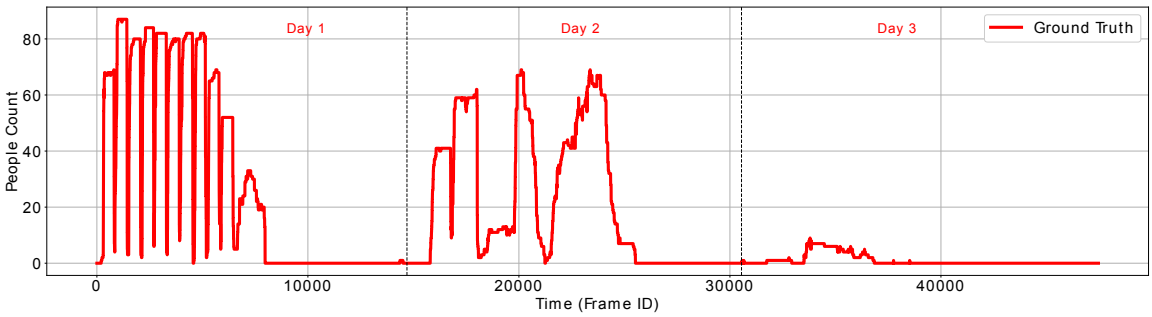


Figure 6·2: The true people count for the 3-day dataset.

itoring a room with a single fisheye camera. Firstly, we quantify changes in RAPiD's performance with increasing distance of occupants to the camera. Then, we demonstrate the impact of various occupancy scenarios on the performance of RAPiD. These results indicate that while RAPiD performs well in small-to-medium sized rooms, in large rooms it struggles to detect people at the FOV periphery, thus motivating the need for additional cameras.

Since RAPiD is a CNN, it requires training. In all experiments reported in this chapter, we used RAPiD with parameters obtained by initial training on the Microsoft COCO 2017 dataset (Lin et al., 2014) and fine-tuning on the MW-R and HABBOF datasets (for training details see (Duan et al., 2020)). RAPiD produces bounding boxes and confidence values that tell how likely a bounding box is to contain a person (from 0 for "impossible" to 1 for "certain"). By selecting a confidence threshold γ, the algorithm can be tuned to specific scenarios. Unless otherwise stated, we used $\gamma = 0.05$ in all experiments [2].

6.3.1 Performance of RAPiD at Different Distances from the Camera

In order to understand limitations of RAPiD, we first quantify its people-counting performance as a function of the monitored area. Figure 6·3 shows a video frame from

[2]The value of γ was determined heuristically via a loose grid search using values that are between 0.01 and 0.2.

camera 1 (Figure 3·1) with concentric green circles superimposed, each circle corresponding to a true physical distance of 10-35 ft from the camera in 5 ft increments. It is clear that the circles are not equally spaced out in the image, confirming radial non-linearity of the lens.

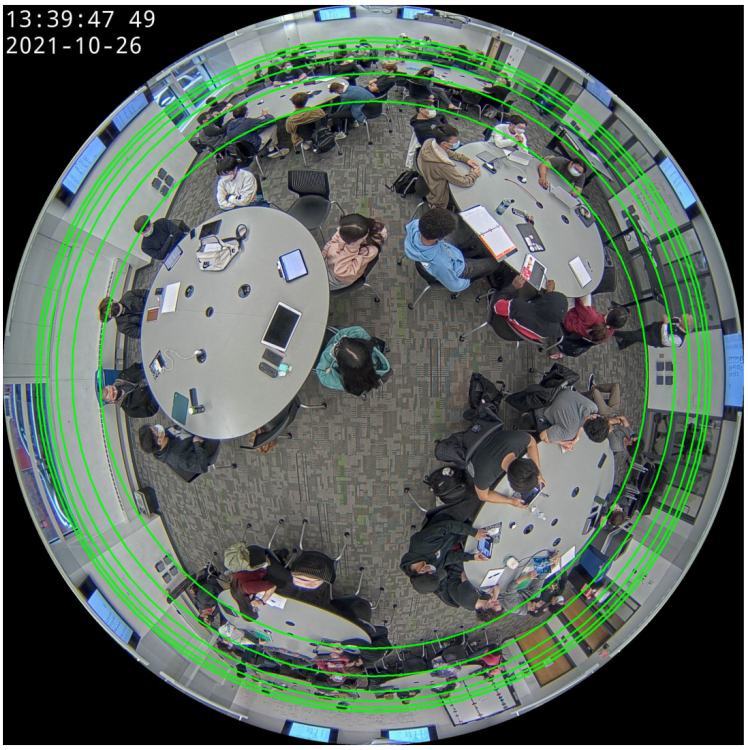


Figure 6·3: Overhead fisheye view of the classroom where FRIDA was recorded. The superimposed green concentric circles correspond to the true physical distance of 10-35 ft from the camera in 5 ft increments.

We evaluated the people-counting performance of RAPiD over a video segment from the dataset described in Section 6.2 that contains only modest movement of occupants (the number of people inside of each green circle remains constant) to ease annotation. The segment consists of 50 frames with 62 occupants each, that is 3,100 person-instances. For each circle radius, Table 6.1 gives the area inside this circle (FOV area) and the area of a square-shaped room that could be fully covered by this FOV, that is the area of a square inscribed in the circle (square room coverage).

As the FOV radius increases from 10 ft (with 23 people visible) to 35 ft (62 people visible), MAE increases from 0.64 to 16.66 while MAE_{pp} increases from 0.028 to 0.269 (or from 2.8% to 26.9%).

This increase in the error was to be expected and is due to the reduction of projected body size and increasing likelihood of occlusions with distance. Up to about 20 ft, the algorithm can count people with less than 0.07 of MAE_{pp} (7% error); it accurately counts the vast majority of the 44 people located inside this circle, despite severe occlusions by other people, tables, chairs, etc. The 20 ft radius corresponds to effective coverage of a square room with 800 ft^2 area. For larger FOVs, the error rapidly increases, so to achieve more accurate counting additional fisheye cameras are needed.

Table 6.1: People-counting performance of RAPiD for increasing camera field of view.

FOV radius	FOV area	Number of people	MAE	MAE_{pp}	Square room coverage
10 ft	314 ft^2	23	0.64	0.028	200 ft^2
15 ft	707 ft^2	34	2.36	0.069	450 ft^2
20 ft	1,257 ft^2	44	2.88	0.065	800 ft^2
25 ft	1,963 ft^2	52	6.92	0.133	1,250 ft^2
30 ft	2,827 ft^2	60	14.78	0.246	1,800 ft^2
35 ft	3,848 ft^2	62	16.66	0.269	2,450 ft^2

6.3.2 Performance of RAPiD in a Medium-Sized Room

To assess RAPiD's performance in a medium-sized space (400–800 ft^2) but in more challenging conditions, we used CEPDOF (Duan et al., 2020), an annotated dataset with overhead fisheye images captured in 8 scenarios differing in human poses, movement, occlusions, illumination, etc. with up to 13 occupants. Sample images and RAPiD's detections in 3 scenarios are shown in Figure 6·4. We selected "Edge cases"

(a) Edge cases (b) Lunch meeting 2 (c) IRfilter

Figure 6·4: Examples of detections by RAPiD in a 500 ft^2 classroom in 3 scenarios selected from CEPDOF dataset (Duan et al., 2020).

since it includes significant movement of people and unusual poses (e.g., crouching, stretching) to challenge the detector. "Lunch meeting 2" has the highest occupancy among all scenarios, but people mostly sit or stand, like in a typical meeting. In "IRfilter", lights are turned off challenging the detector by low contrast and little image detail. While in the well-lit images all people are correctly detected, in "IRfilter" (Figure 6·4(c)) there are two misses and one false positive which is not surprising since RAPiD was not trained on low-light images.

Table 6.2 summarizes RAPiD's performance on CEPDOF. While MAE is relatively low for well-lit scenarios, unsurprisingly it is quadrupled for the low-light scene. However, the cumulative MAE over all 8 scenarios (25,504 frames) is 0.827, i.e., count error of less than 1 under occupancy of up to 13 people. MAE per person (MAE_{pp}) is quite small for "Edge cases" (0.076 or 7.6%) and for "Lunch meeting 2" (0.04 or 4%) suggesting that RAPiD can handle unusual poses, movement, occlusions well. However, for "IRfilter" it is much higher at 0.236 (or 23.6%) suggesting that improvements are needed in low light. Cumulatively over all scenarios, the error of 0.122 (or 12.2%) is quite high since 35% of CEPDOF images have been captured in low light, for which RAPiD was not trained. As for X-Accuracy, we show only values

for $X = 0, 1, 2$ since occupancy is quite low (maximum of 13). While perfect counting ($X = 0$) is accomplished in about 60% of well-lit frames, in low light it is only 19%. A slack of 1 or 2 increases X-Accuracy to well over 90% in well-lit scenes suggesting that RAPiD is accurate but not so much in low light (53% and 79%). Cumulatively, X-Accuracy is high for $X = 1, 2$ but only 44% for $X = 0$ due to large fraction of low-light images in CEPDOF.

Table 6.2: People-counting performance (MAE, MAE_{pp} and X-Accuracy) of RAPiD in 3 selected scenarios and cumulatively over all 8 scenarios in CEPDOF dataset (Duan et al., 2020).

Metric	Edge cases	Lunch meeting 2	IRfilter	Cumulative
MAE	0.420	0.436	1.582	0.827
MAE_{pp}	0.076	0.040	0.236	0.122
Acc_X [%] X=0/1/2	64/95/99	59/98/100	19/53/79	44/81/94

6.3.3 Performance of RAPiD in a Large Space with High Occupancy

In the previous two sections, RAPiD was shown to perform well in small-to-mid-sized spaces in various scenarios (except for low light), but only in short-term tests (75 min in total). In order to assess RAPiD's performance over longer time span in high and dynamic occupancy, we used the full dataset described in Section 6.2 that consists of a 3-day video recording.

It is important to note that the classroom was empty during more than half of the test time (Figure 6·2) with lights turned off at night (resulting in low-contrast images), a difficult scenario for RAPiD. To adapt to these adverse conditions, we changed RAPiD's threshold γ to 0.6 in low light, while keeping the original value of 0.05 in normal light. The switching between two thresholds is automatic, based on average luminance in each frame. The higher threshold in low-light conditions helps reduce false positives. An alternative would be to train two different versions of

RAPiD for the two cases, but this would require a lot of annotated data in low-light conditions (although CEPDOF includes such frames, many more would be needed for reliable training).

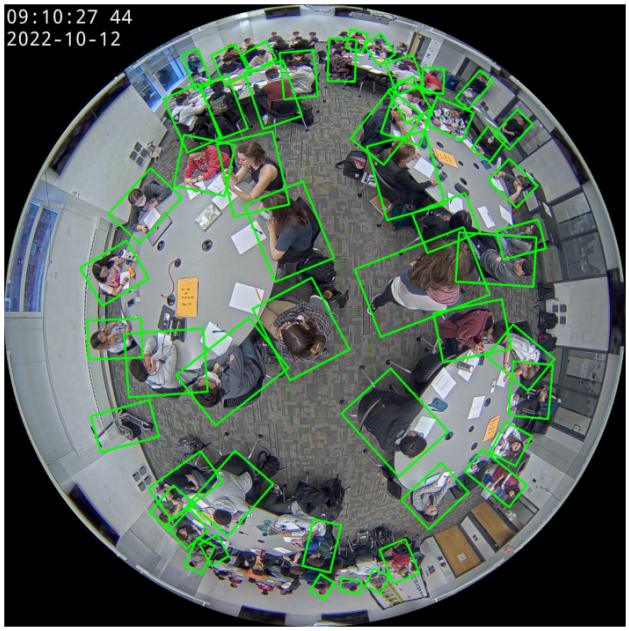
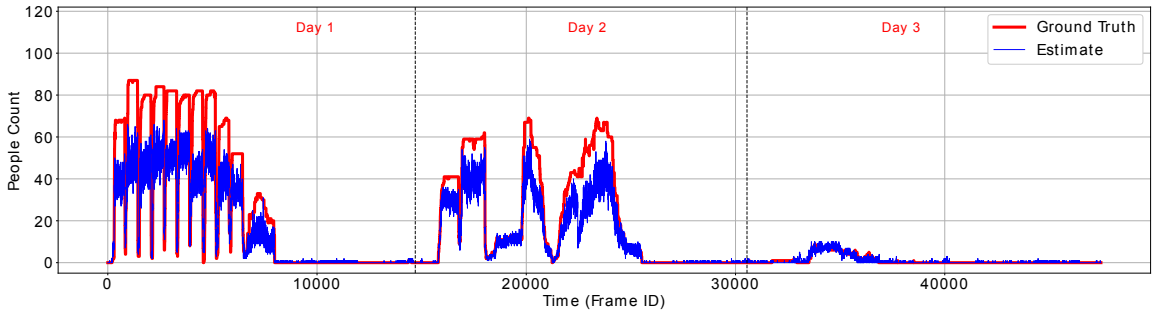


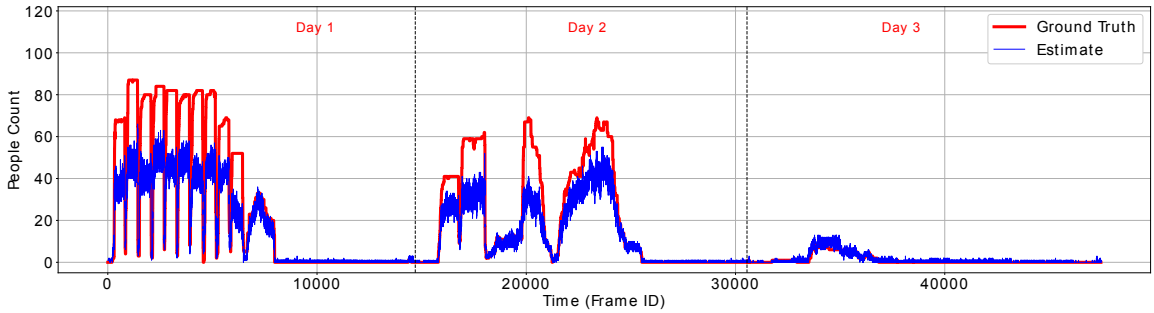
Figure 6·5: Sample frame with people detections by RAPiD from day 1 of the 3-day test.

Figure 6·5 shows a sample frame from the first day with 87 people in the classroom. Clearly, not all of them are detected by RAPiD, especially at the top and bottom of the frame, since they are too small and/or occluded. Figure 6·6 shows people-count estimates produced by RAPiD for each of the cameras. All three plots look very similar, but upon closer inspection one can see that the estimate for camera #1 is a bit closer to the ground truth compared the other two cameras. While all three occupancy estimates are highly correlated with the ground truth, they are not accurate. The true count is underestimated on days 1 and 2 by as much as 20-40 occupants. This is due to the size of the room and people present at the FOV periphery where RAPiD is unreliable. On day 3 the occupancy estimate tracks the ground truth more accurately (slight underestimation and overestimation making the

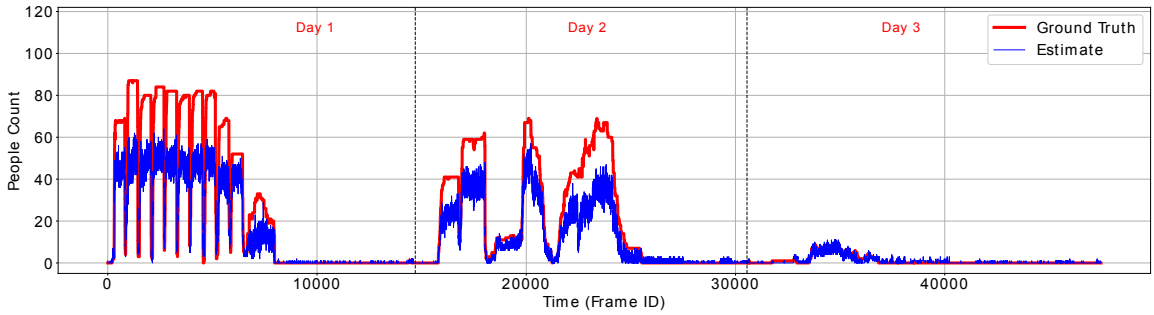
ground-truth red line not very visible).



(a) Camera #1



(b) Camera #2



(c) Camera #3

Figure 6·6: The true number of people (red line) and an estimate by RAPiD in a 3-day test in the same classroom as FRIDA for each of the three cameras.

Table 6.3 shows quantitative performance of RAPiD for the 3-day test. For all 3 cameras, MAE is largest (at about 12) on day 1 when the occupancy is very high during daytime. It drops to about 7-8 on day 2 when the classroom is occupied in daytime but with fewer people. It is the smallest (at about 0.4-0.6) on day 3 when

the room is mostly empty and the highest occupancy is only 9 during a short period. Clearly, MAE values approximately scale with average occupancy on each day. This is confirmed by MAE_{pp} values that are relatively constant around 0.4 except for camera #3 on day 3. The X-Accuracy at $X = 0$ is low on days 1 and 2 (29-47%) but quite a bit higher on day 3 (55-67%). At $X = 5$ (the count estimate may be up to 5 off), the X-Accuracy increases to 53-64% on days 1 and 2, and to 100% on day 3. Again, this is consistent with day 3 having mostly zero occupancy. At $X = 10$, the X-Accuracy is in 59-70% range for days 1 and 2, and 100% for day 3. The consistently better performance on day 3 is not surprising since it is easier to count few people spaced out than in a crowded scenario (more opportunities for mis-detection, e.g., due to occlusions).

Table 6.3: People-counting performance (MAE, MAE_{pp} and X-Accuracy) of RAPiD for the 3-day test in 2,000 ft^2 room for all three cameras. The last column shows cumulative metrics computed across the 3 days.

Metric	Camera	Day 1	Day 2	Day 3	Cumulative
MAE	#1	11.82	6.92	0.38	6.11
	#2	11.60	8.55	0.48	6.61
	#3	12.32	7.36	0.62	6.49
MAE_{pp}	#1	0.400	0.336	0.384	0.373
	#2	0.393	0.415	0.480	0.404
	#3	0.417	0.358	0.623	0.397
Acc_X [%] X=0/5/10	#1	45/53/59	43/64/70	74/100/100	55/73/77
	#2	47/54/60	29/60/65	67/100/100	48/72/76
	#3	38/59/64	42/64/70	55/100/100	46/75/79

Cumulatively across the 3 days, MAE is around 6 and MAE_{pp} is around 0.4 for all three cameras. Also, the X-Accuracy is fairly consistent between the cameras. Note,

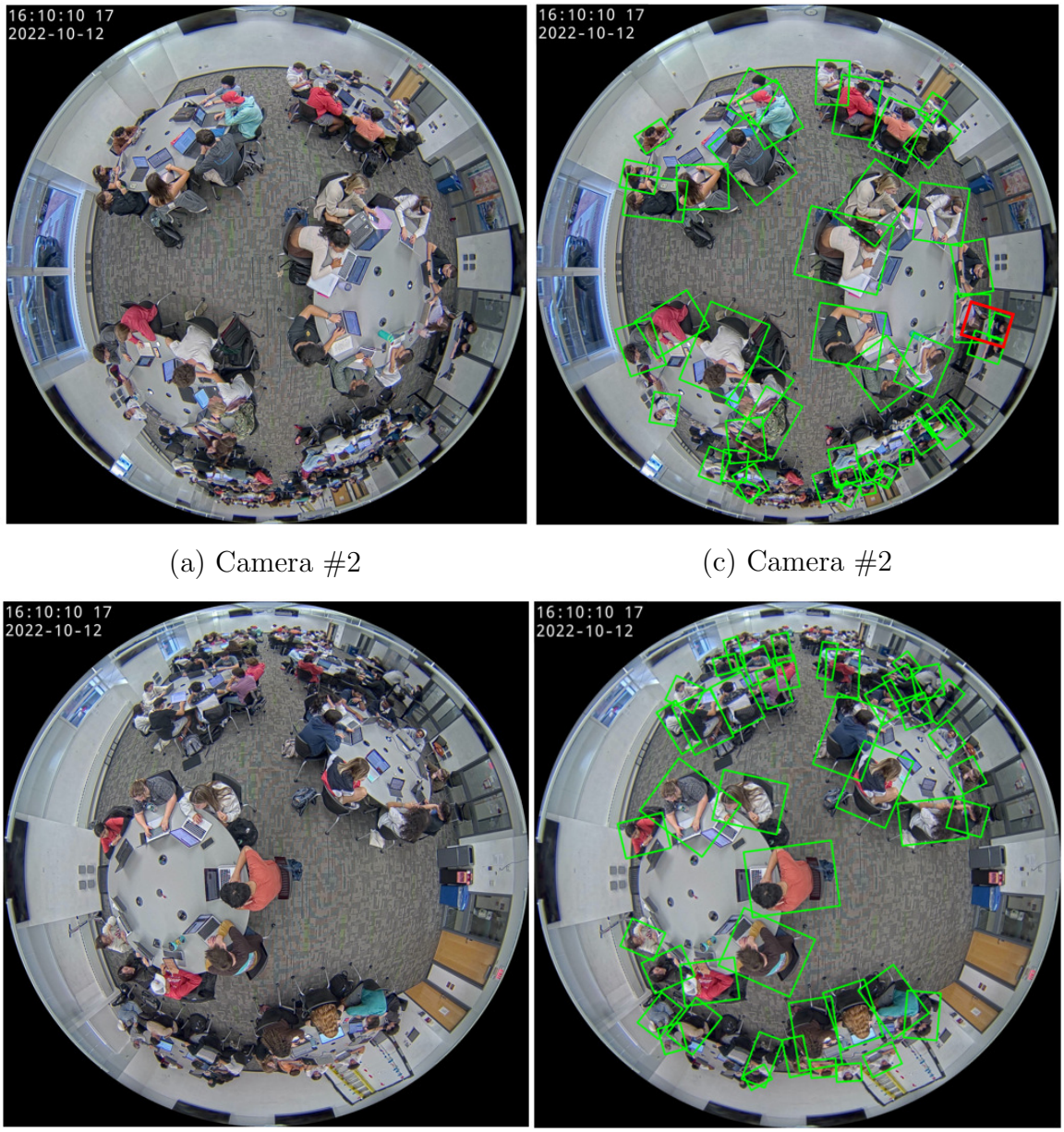
that MAE_{pp} of 0.4 is higher than 0.269 reported in Table 6.1 for 35 ft FOV. This is due to higher occupancy (87 instead of 62) and significant movement of people, which might cause occlusions, especially at the start and end of each class. MAE_{pp} for camera #1 (0.373) is slightly lower than that for the other two cameras, which is not surprising since it is mounted close to classroom's center where students tend to congregate.

These results indicate that while monitoring a space with a single fisheye camera is suitable for small-to-medium size spaces (up to about 800 ft^2), where MAE_{pp} does not exceed 0.07 (or 7%), as reported in Table 6.1, and is in 0.04-0.076 range (4-7.6%) for well-lit scenes, as reported in Table 6.2, such setup underperforms in large spaces with MAE_{pp} reaching 0.4 (or 40%) in Table 6.3 and must be redesigned.

6.4 Occupancy Estimation Using Two Fisheye Cameras

In order to improve the single-camera occupancy estimation performance, we need to understand its failure modes. Figure 6·7(c-d) shows people detections produced by RAPiD for the frame pair from Figure 6·7(a-b). One can see that some people are not detected at FOV periphery (e.g., top of the view in Figure 6·7(d)). Occasionally, the algorithm produces false detections (e.g., red bounding box in Figure 6·7(c) contains two people who already have their own green bounding boxes).

To count people, one could compute the average of counts from the two cameras, but this could result in undercounting if there are people visible in FOV of one camera but not of the other one (e.g., due to occlusions). On the other hand, adding the two counts could result in overcounting due to the double-counting of people who are visible in both cameras. These effects could be further compounded by people-detection errors. A principled approach is to *re-identify* people between different camera views and count each person only once.



(a) Camera #2 (c) Camera #2

(b) Camera #3 (d) Camera #3

Figure 6·7: (a-b) Sample frames from two overhead fisheye cameras overlooking a large classroom; (c-d) The same frames with people detections by RAPiD (Duan et al., 2020).

Motivated by the above considerations, we apply a slightly modified version of our PRID algorithms (as described below) to bounding boxes detected by RAPiD in cameras #2 and #3, and estimate the people count as follows:

$$\widehat{\eta}_i = \widehat{\eta}_i^2 + \widehat{\eta}_i^3 - \widehat{\eta}_i^{23}. \tag{6.4}$$

where $\widehat{\eta}_i^2$ and $\widehat{\eta}_i^3$ are the estimated people counts by RAPiD in frame number i from cameras #2 and #3, respectively, and $\widehat{\eta}_i^{23}$ is the number of people successfully re-identified between these two frames. Thus, the estimated people count at time i equals the sum of counts obtained from two frames reduced by the number of matched identity pairs in them. We selected cameras #2 and #3 since the FOV of each covers about one-half of the room. On the other hand, camera #1 mostly covers the room's center and is more suitable for single-camera occupancy estimation.

In order to account for people-detection errors, which can result in quite different numbers of the detected bounding boxes in each camera's view, we slightly modify the matching part of our PRID algorithm. In the matching algorithm discussed in Section 3.2.1, the greedy algorithm was exhaustive, i.e., it was applied until no matches were possible in the score matrix (step (3) in Section 3.2.1). Instead, now we stop the greedy algorithm when the remaining score-matrix elements (conditional probabilities) are below some threshold τ and treat them as corresponding to unlikely identity matches.

The threshold τ controls the trade-off between the number of matched and unmatched bounding boxes between the two views. In an ideal scenario, when all occupants in a space are detected in both query- and gallery-camera views, $\tau = 0$ will force a match of every person in the query set to a person in the gallery set. However, as we have discussed, in practice some occupants may not be detected in one of the views or there may be false detections. In this case, some query or gallery elements

may not have a match and $\tau > 0$ is needed to stop the matching process. Thus, τ serves as a match-probability threshold below which a match is unlikely.

In our people-counting experiments, we treat τ as a tuning parameter and find its best value for each method (in terms of MAE (6.1)) by searching among a finite set of uniformly-spaced choices over the interval $[0, 1]$. We found that the best values for different methods range from 1.25×10^{-4} to 0.3 (Table 6.4).

Table 6.4: Threshold τ values that minimize MAE for each feature combination. The minimization is performed through grid search.

DL	CH	LOC/PPD	LOC/CBD	Best τ
✓				0.04
	✓			0.04
✓	✓			1.6×10^{-3}
		✓		0.3
	✓	✓		6.4×10^{-3}
✓		✓		6.4×10^{-3}
✓	✓	✓		1.25×10^{-4}
			✓	0.2
	✓		✓	4.9×10^{-3}
✓			✓	4.9×10^{-3}
✓	✓		✓	1.25×10^{-4}

Figure 6·8 shows the time plot of the ground-truth people count and 4 people-count estimates[3] obtained from our 3-day test set (Section 6.1). Three estimates were

[3]The people-count estimates in Figure 6·8 are noisy since they are computed independently for each time instant and no temporal smoothing is applied. While smoothing (e.g., temporal median filtering) could be applied to the estimates, in any real-time application it would have to be causal thus creating a time delay. This smoothness/delay tradeoff has to be carefully adjusted for each practical application.

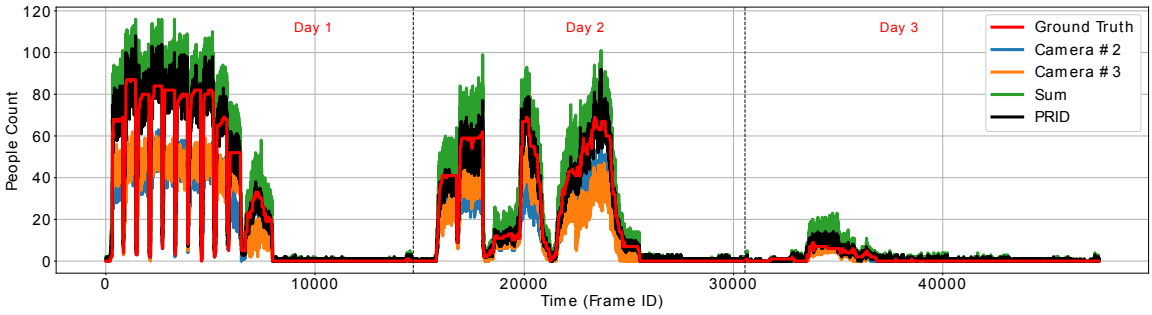


Figure 6·8: Ground-truth people count and four people-count estimates (two single-camera estimates, sum of single-camera estimates, and an estimate obtained using PRID (DL+CH+CBD) to eliminate double counts) for the 3-day test.

obtained by counting bounding boxes produced by RAPiD, either from camera #2 frames, or from camera #3 frames or by adding the two counts. It should be noted that while camera #1 (mounted in the middle of the room) slightly outperforms the other two cameras (Table 6.3), when using two cameras a better strategy is to use cameras that can more effectively capture both ends of the room.

Clearly, the counts from both camera #2 and camera #3 (blue and orange lines, respectively) severely underestimate the true count (red line) due to missed-detections at FOV periphery (the room is too large for a single camera). Averaging the two counts or taking the maximum would still result in severe undercounting. The sum of the counts from both cameras (green line) significantly overestimates the true count since many people are counted twice. The fourth estimate (black line) was obtained by first applying RAPiD to detect people in same-time frames from cameras #2 and #3, and then performing PRID on these detections using our best-performing algorithm from Table 5.1 (DL+CH+CBD). Obviously, the PRID-based people-count estimate quite accurately tracks the true people count.

While the PRID algorithm combining all three features works very well in people counting, it would be interesting to understand the impact of other feature combinations on people-counting performance. Rather than showing plots like the one in

Figure 6·8 for different feature combinations (because plots for different algorithms have significant overlap hurting discernibility), we perform this ablation study by reporting MAE (6.1), MAE_{pp} (6.2) and Acc_X (6.3) values in Table 6.5. This table shows the results for thresholds τ that minimize the cumulative MAE for each feature combination.

Similarly to re-identification experiments, algorithms involving a location-based feature (either PPD or CBD) significantly outperform algorithms based on appearance (color-histogram or deep-learning features or combination thereof), with MAE almost halved. Overall, the lowest cumulative MAE and MAE_{pp} values are obtained by CBD with cumulative $MAE = 1.59$ and $MAE_{pp} = 0.097$. However, the performance of other PRID algorithms that involve location-based features are not significantly worse, with the difference between the best- and worst-performing ones of 0.15 in cumulative MAE and 0.01 in cumulative MAE_{pp}.

In terms of cumulative $Acc_{X=0}$, all algorithms perform between 41 to 43%. However, when we check the cumulative $Acc_{X=5}$, the appearance-based algorithms all achieve 78% which is 12 to 14%-points lower than the performance of feature combinations that involve location-based features. The cumulative $Acc_{X=10}$ tells us that, the feature combinations that involve location-based features estimate the people count 98% of the time with an absolute error less than 10. For the appearance-based feature combinations, the cumulative $Acc_{X=10}$ is 86%.

Day 1 and day 2 results have similar trends to the cumulative results. An interesting observation is that, in terms of MAE the results for the first two days are worse than the cumulative MAE. However, in terms of MAE_{pp}, the results for the first two days are better than the corresponding cumulative results. The reason for this is that the average ground-truth occupancies for day 1 and day 2 are higher compared to the cumulative average occupancy. As can be seen in Figure 6·8, day 3 has low occupancy

Table 6.5: People-counting performance (MAE, MAE_{pp} and X-Accuracy) of PRID algorithms in a 3-day test in large classroom using cameras 2 and 3.

Metric	Algorithm	Day 1	Day 2	Day 3	Cumulative
MAE	DL	5.99	4.28	0.66	3.52
	CH	6.07	4.40	0.66	3.59
	DL+CH	6.06	4.38	0.66	3.58
	PPD	2.42	1.74	0.84	1.63
	CH+PPD	2.44	1.79	0.79	1.63
	DL+PPD	2.44	1.79	0.79	1.63
	DL+CH+PPD	2.50	1.87	0.75	1.67
	CBD	2.43	1.71	0.75	1.59
	CH+CBD	2.47	1.92	0.72	1.66
	DL+CBD	2.46	1.92	0.72	1.66
	DL+CH+CBD	2.61	2.03	0.72	1.74
Metric	Algorithm	Day 1	Day 2	Day 3	Cumulative
MAE_{pp}	DL	0.203	0.208	0.670	0.215
	CH	0.205	0.214	0.670	0.219
	DL+CH	0.205	0.213	0.670	0.218
	PPD	0.082	0.084	0.849	0.100
	CH+PPD	0.083	0.087	0.793	0.100
	DL+PPD	0.083	0.087	0.792	0.100
	DL+CH+PPD	0.085	0.091	0.756	0.102
	CBD	0.082	0.083	0.752	0.097
	CH+CBD	0.084	0.093	0.731	0.102
	DL+CBD	0.083	0.093	0.731	0.101
	DL+CH+CBD	0.088	0.099	0.731	0.107
Metric	Algorithm	Day 1	Day 2	Day 3	Cumulative
Acc_X [%] X=0/5/10	DL	38/62/74	33/71/83	52/100/100	41/78/86
	CH	38/61/74	33/70/82	52/100/100	41/78/86
	DL+CH	38/61/74	33/70/82	52/100/100	41/78/86
	PPD	41/84/96	36/92/99	50/98/100	42/92/98
	CH+PPD	41/84/96	37/91/99	50/99/100	43/92/98
	DL+PPD	41/84/96	36/91/99	50/99/100	43/92/98
	DL+CH+PPD	41/83/95	36/90/99	51/99/100	43/91/98
	CBD	41/84/96	37/93/99	51/99/100	43/92/98
	CH+CBD	41/84/96	36/89/99	51/99/100	43/91/98
	DL+CBD	41/84/96	36/90/99	51/99/100	43/91/98
	DL+CH+CBD	41/82/95	36/88/99	51/99/100	43/90/98

throughout the day which lowers the cumulative average ground-truth occupancy. In fact, the average ground-truth occupancy for day 3 is 0.99 which makes MAE_{pp} larger than MAE (see equation (6.2)), unlike for the other two days. Day 3 has the lowest MAE values because at most it has been occupied by 9 people, thus the absolute error does not get very high. On the other hand, on day 1 and day 2 the maximum occupancy is up to 87 and 69, respectively. Moreover, the average ground-truth occupancy for day 1 and day 2 is 29.53 and 20.58, respectively, which scales down the MAE significantly to get the MAE_{pp}. These results illustrate the limitations of using MAE_{pp} in scenarios with long zero occupancy periods as we discussed in Section 6.1.

While Table 5.1 reports performance of re-identification only, Table 6.5 reports a combined performance of people detection by RAPiD and of re-identification by various PRID algorithms. If RAPiD introduces people-detection errors (misses or false detections), people counts can be incorrect even with perfect PRID. However, even if PRID is imperfect, it may still lead to a correct people count, for example if the total number of matches between cameras #2 and #3 is correct but some of the matches are permuted (e.g., person A in camera #2 is matched to person B in camera #3 and person B in camera #2 is matched to person A in camera #3). PRID affects people counting only if it produces an *incorrect number* of matches between two cameras. In conclusion, whereas PRID errors have a full impact on QMS and mAP values in Table 5.1, they have only a partial impact on the MAE values in Table 6.5. This also explains the difference in the ordering of various feature combinations between person re-identification and people counting.

One final observation is in order. Examining Tables 5.1 and 6.5, it is clear that algorithms using appearance features (CH or DL or CH+DL) are significantly outperformed by algorithms that combine them with a location-based feature. However, the performance spread between the latter algorithms in both re-identification and people

counting is quite small. Since the computational complexity of calculating location-based features is far lower than obtaining a color histogram (CH) or extracting neural features (DL), in applications sensitive to complexity (e.g., real time) a single-feature algorithm using only location may be a good choice (over 94% in cumulative QMS and over 95% in cumulative mAP). Furthermore, it was mentioned in Section 4.4 that the calculation of CBD location features is significantly more complex computationally than that of PPD features. This suggests that in complexity-critical applications for **occupancy estimation** single-feature PPD algorithm would be most appropriate (Table 6.5). However, if best performance is required for a **PRID-sensitive** application, then, location features combined with color histograms and/or deep-learning features are a better option (Table 5.1).

6.5 Occupancy Estimation using N Cameras

In the previous section, we proposed an occupancy estimation method that can accurately monitor a 2,000 ft^2 space using two fisheye cameras. However, more cameras would be needed in a larger space, such as as a large lecture hall, convention center, supermarket, airport building, bus/train station, etc. Due to the wide FOV of fisheye cameras, it is inevitable that some people will be detected by multiple cameras. To address this problem, we propose two methods to scale up the two-camera PRID approaches developed thus far to $N > 2$ cameras. First, we introduce these methods and then we report occupancy-estimation results using them.

6.5.1 General Method based on N-Dimensional Score Matrix

In Chapters 3, 4 and 5, we introduced PRID methods for the case of two fisheye cameras. Each of those methods produces a 2-D score matrix $\mathbf{S}$ with elements quantifying the similarity between a query identity and a gallery identity.

Let us first consider the 3-camera setup ($N = 3$) with which we collected our data

(Section 6.2). We illustrate this case graphically in Figure 6·9 using a Venn diagram, where C_1, C_2, C_3 denote the sets of people detections (e.g., by RAPiD) in FOV of camera #1, camera #2 and camera #3, respectively.

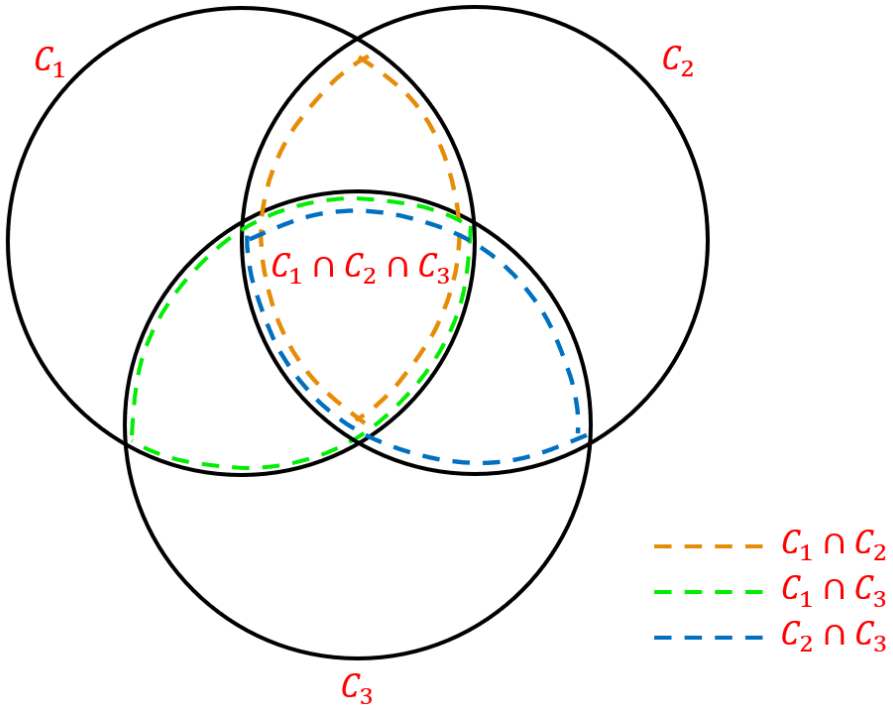


Figure 6·9: Illustration of PRID scenario for $N = 3$ cameras.

The Venn diagram allows us to compute the actual people count η as follows:

$$\begin{aligned} \eta &= |C_1| + |C_2| + |C_3| \\ &\quad -|C_1 \cap C_2| - |C_1 \cap C_3| - |C_2 \cap C_3| \\ &\quad +|C_1 \cap C_2 \cap C_3|. \end{aligned} \tag{6.5}$$

While $|C_1|, |C_2|$ and $|C_3|$ are provided by a people detection algorithm and $|C_1 \cap C_2|$, $|C_1 \cap C_3|$, $|C_2 \cap C_3|$ are identified by a two-camera PRID using 2-D score matrices, we still need to identify $|C_1 \cap C_2 \cap C_3|$. This necessitates a three-camera PRID with a 3-D score matrix, which can be thought of as a rectangular cuboid. While there are

many possible ways to compute a *joint* similarity of three bounding boxes (one from each camera), a convenient way that we adopt is to average all *pairwise* similarities as follows:

$$\mathbf{S}_{i_1,i_2,i_3} = \frac{1}{3} \times \big(f(i_1,i_2) + f(i_1,i_3) + f(i_2,i_3)\big) \tag{6.6}$$

where i_1, i_2 and i_3 are identities from camera #1, camera #2 and camera #3, respectively, $\mathbf{S}(\cdot)$ represents the similarity value between these identities and $f(\cdot,\cdot)$ is a similarity function used for two cameras, e.g., any of the two-camera PRID similarity scores from Chapters 3, 4 or 5. In Section 7.1.2, as part of future work, we will discuss alternative similarity measures that are not based on pairwise similarities.

An extension to $N > 3$ is relatively straightforward but we must carefully consider various combinations of k out of N cameras, across which identities need to be matched. For example, for $N = 4$ cameras, re-identification between 2, 3 or 4 camera views is needed in order to obtain a correct overall count. For N cameras, the total number of camera combinations to be considered is given by:

$$\sum_{k=2}^{N} \binom{N}{k} = 2^N - 1 - N \tag{6.7}$$

where the summation starts at $k = 2$ since in PRID at least two camera views are needed. For $N = 3$, this amounts to 4 camera combinations which is consistent with 4 intersections in the Venn diagram in Figure 6·9. For $N = 4$, there are 11 camera combinations (6 two-camera, 4 three-camera and 1 four-camera combinations) and for $N = 5$ there are 26 camera combinations, rapidly increasing with a growing number of cameras.

Clearly, score matrices $\mathbf{S}$ of up to N dimensions are needed for PRID. One possibility is to generalize equation (6.6) to N dimensions through the use of the well-known

inclusion-exclusion principle (van Lint and Wilson, 1992) as follows:

$$\mathbf{S}_{i_1,i_2,.....,i_{N-1},i_N} = \frac{1}{\binom{N}{2}} \sum_{k=1}^{N} \sum_{l=k+1}^{N} f(i_k, i_l). \tag{6.8}$$

For $f(\cdot,\cdot)$, we can use any of the two-camera PRID similarity scores from Chapters 3, 4 or 5.

In experiments reported in Section 6.5.3, we opt for location-based PRID discussed in Chapter 4 due to its low computational complexity compared to other methods. The computational complexity is a serious concern for larger N since the number of camera combinations that need to be considered in PRID grows exponentially with N (6.7). To compute the similarity scores $f(\cdot,\cdot)$, we chose PPD which has the lowest computational complexity of all distance metrics we proposed in Section 4.3.1. If computational complexity is not paramount, one can use any other two-camera PRID similarity for $f(\cdot,\cdot)$.

Since the score-matrix elements are computed using PPD, they are all expressed in pixels (i.e., the lower the score, the more similar the identities). To match identities, we introduce a new tuning parameter λ expressed in pixels, unlike the unit-less τ that we introduced in Section 6.4[4]. Thus, in greedy matching, instead of maximization, we will be performing minimization (score-matrix elements above λ will correspond to unmatched identities).

6.5.2 Clustering of Real-World Locations of People

The N-camera PRID method proposed in the previous section is general and can be applied to both appearance- and location-based features. The method we are proposing below applies only to location-based features.

[4]In Section 6.4, elements of the score matrix were normalized using *softmax* to facilitate feature fusion. They expressed a degree of similarity between bounding boxes; those below threshold $\tau \in [0, 1]$ corresponded to unmatched identities.

In Chapter 4, we proposed location-based PRID for two cameras. The main idea was to map pixel location of a person from query-camera image to gallery-camera image, and match identities based on distance between the mapped query locations and gallery locations. The mapping of pixel coordinates between cameras was a two-step process. First, pixel coordinates of a person were mapped from query-camera image to 3-D world coordinates. Then, these 3-D world coordinates were mapped to pixel coordinates in the gallery-camera image. In the N-camera PRID extension we proposed in the previous section, we applied the PPD distance measure to pairs of cameras and summed these distances for all pairs in each group of k cameras among the N cameras in total.

In this section, we propose an alternative idea but in 3-D world coordinates rather than in pixel coordinates. We propose to map each person's location from each of N camera images to 3-D world coordinates and cluster the mapped 3-D coordinates to match identities. Note that in this approach the number of clusters should correspond to the number of people in the room. Therefore, we cannot use a clustering algorithm such as K-means (Lloyd, 1982) since we do not know the value of K. We need a clustering method that does not require advance knowledge of the number of clusters. One such method is "Density-Based Spatial Clustering of Applications with Noise" (DBSCAN) (Ester et al., 1996). Our algorithm for clustering the mapped 3-D coordinates is inspired by DBSCAN, so it is important that we briefly review DBSCAN.

DBSCAN is a density-based clustering method that has two parameters, ϵ and *minPoints*. In DBSCAN, first one picks a random point as the point of interest and finds all points that are within an ϵ radius from this point of interest. All such points, including the point of interest, get assigned to the same cluster. Then, the process is repeated where one treats each point in the cluster as the new point of interest.

This process enlarges the cluster. One continues to spread out the cluster until there is no point within ϵ distance from any of the points in the cluster. Then, one picks another point from the dataset that has not been visited yet and repeats the process. For a group of points to be considered a cluster, there should be at least *minPoints* elements in the cluster. Also, if a certain data point has no other data points within ϵ radius, it is labelled as noise and gets discarded.

We propose a density-based 3-D coordinate clustering method that is inspired by DBSCAN. Our approach has two parameters, ϵ and *maxPoints*. The purpose and usage of ϵ are the same as in DBSCAN. The key difference is the *maxPoints* parameter. In DBSCAN, the size of a cluster has a lower bound of *minPoints* with no upper bound. In our case, we need to allow clusters with size 1 and also to limit the maximum size of a cluster (*maxPoints*). These changes are motivated by the nature of PRID and people counting that we are tackling. We want each person to have their own cluster, where each point in the cluster corresponds to a detection of the same person in a different camera view. In cross-frame PRID, some people in a room can get detected in a single camera view due to occlusions or failed detections, potentially resulting in a single point in their cluster. On the other hand, a person can be detected in at most N camera views, so a cluster may have at most N points ($maxPoints = N$). In fact, one can argue that our clustering algorithm has only one parameter, ϵ, because *maxPoints* equals N.

In experiments, we used 3-D Euclidean distance as the distance measure between data points. Note that although we are mapping pixel coordinates to 3-D world coordinates, the clustering is effectively taking place in 2-D space since the Z coordinate of a person's location is identical for all individuals (average height in the US) as described in Section 4.3.

6.5.3 Experimental Results for N-Camera Algorithms

In this section, we assess the occupancy-estimation performance of two N-camera PRID algorithms proposed in Sections 6.5.1 and 6.5.2. We evaluate these algorithms on the same 3-day dataset (Section 6.2), that we have been using throughout this chapter. While in Section 6.3 we reported occupancy-estimation results using 1 camera, in Section 6.4 we reported corresponding results using 2 cameras. Below, we report results using 3 cameras.

Recall, that the N-D Score Matrix approach has a tuning parameter λ quantifying a distance threshold expressed in pixels (identity matching is performed in image coordinates). On the other hand, the Real-World Location Clustering approach has a tuning parameter ϵ quantifying a threshold on distance in real-world (3-D) coordinates and expressed in centimeters. This is unlike parameter τ from Section 6.4, which is dimensionless, because of *softmax* normalization applied to location- and appearance-based scores before performing identity matching. However, in both N-camera approaches we do not apply any *softmax* operator so the distances are in terms of pixels or centimeters (cm). In Figures 6·10 and 6·11, we show the occupancy-estimation performance of each algorithm in terms of MAE with respect to their respective tuning parameters. The best λ value for N-D Score Matrix approach that yields the lowest cumulative MAE is $\lambda = 400$ pixels, while the best ϵ value for Real-World Location Clustering approach is $\epsilon = 250$ cm. We note that these values are large considering the fact that we are working with 2,048×2,048-pixel images in a room that has a width of 8.5 m. It is likely that some PRID matches are incorrect (as are some RAPiD detections) and yet the people count is quite accurate. However, since our dataset is labeled for people counting only (no bounding boxes or identity labels), we cannot report PRID results to support this hypothesis.

A 250 cm (2.5 m) real-world distance threshold for establishing location clusters

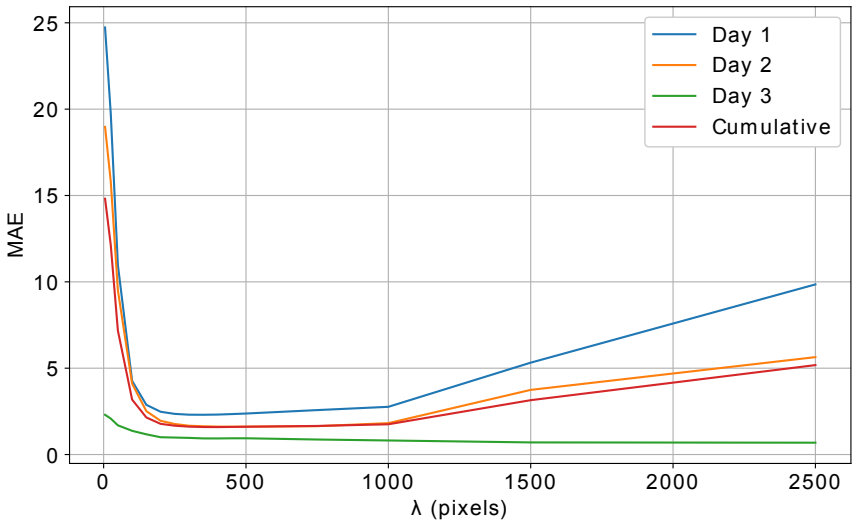


Figure 6·10: MAE values for different values of λ for the N-D Score Matrix approach from Section 6.5.1 with $N = 3$.

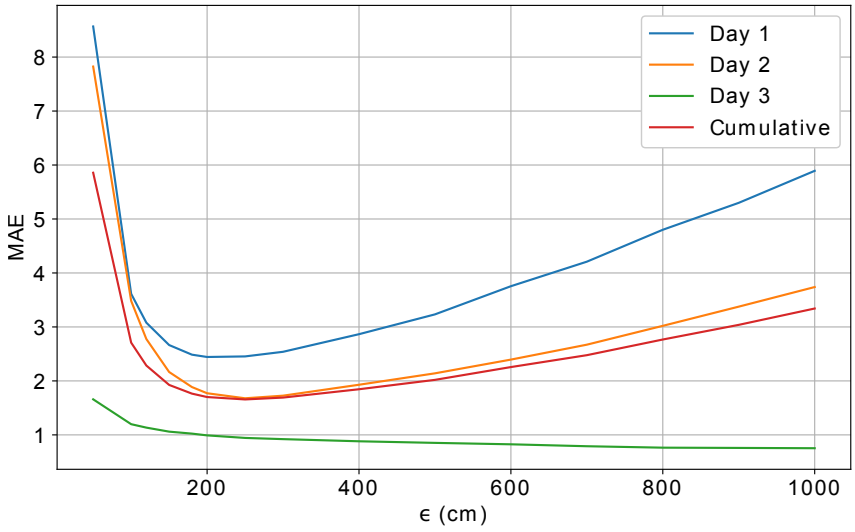
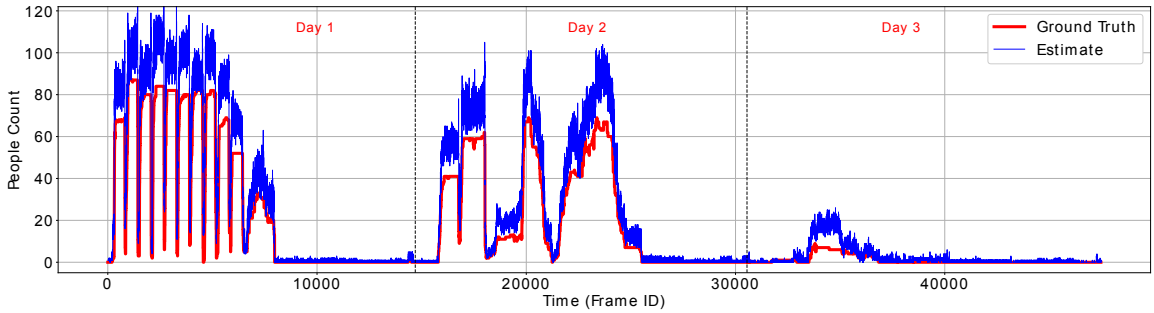
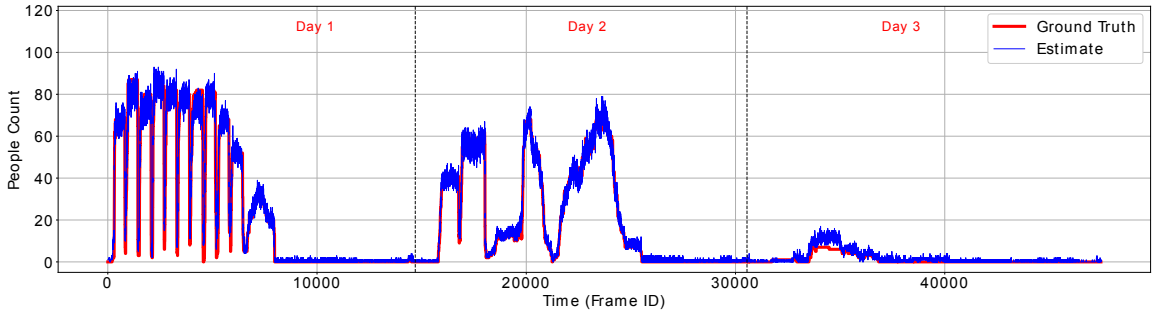


Figure 6·11: MAE values for different values of ϵ for the Real-World Location Clustering approach from Section 6.5.2 with $N = 3$.

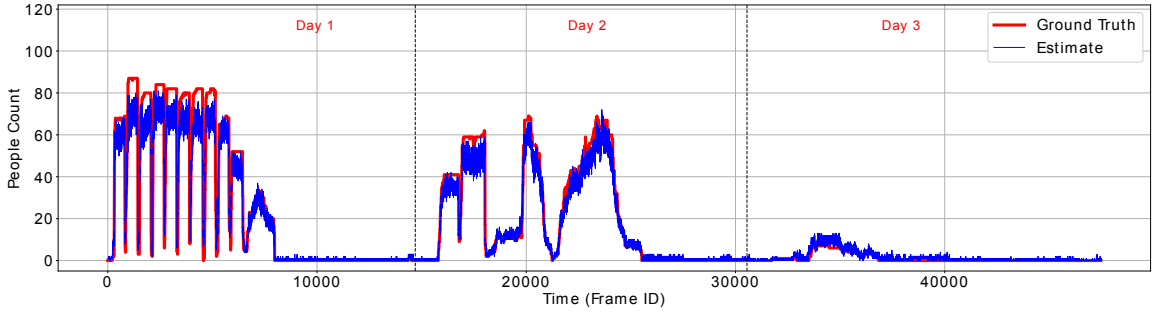
seems large for our 72×28 ft (22×8.5 m) test space. To probe deeper into this issue, in Figure 6·12 we plot the true occupancy and one estimated by Real-World Location Clustering approach across time for three values of ϵ: 50 cm, 250 cm and 800 cm. Notably, for $\epsilon = 50$ cm the algorithm significantly overcounts (Figure 6·12(a)),



(a) $\epsilon = 50$ cm



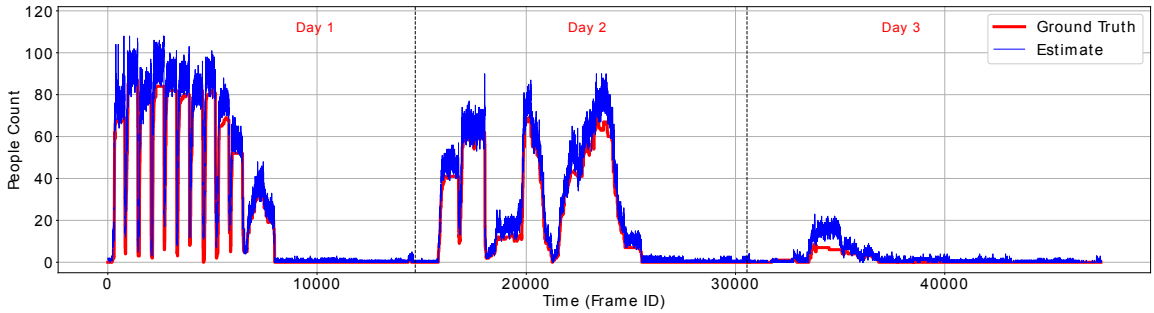
(b) $\epsilon = 250$ cm



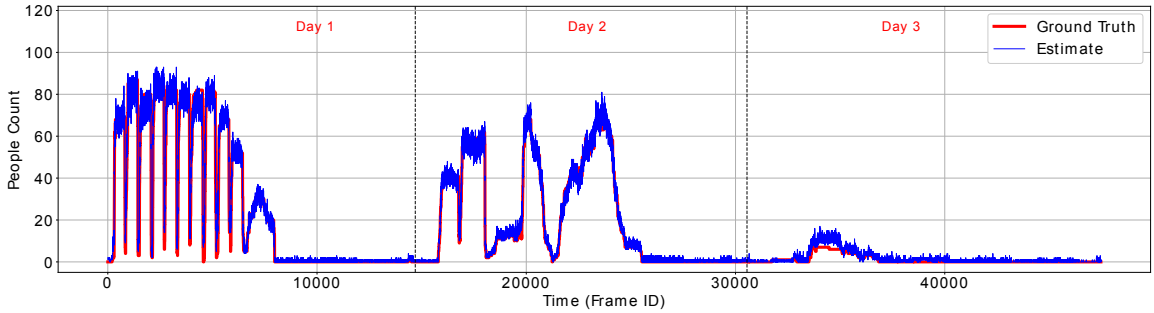
(c) $\epsilon = 800$ cm

Figure 6·12: The true number of people (red line) and an estimate by 3-camera Real-World Location Clustering approach for the 3-day test in 2,000 ft^2 classroom and three different values of ϵ.

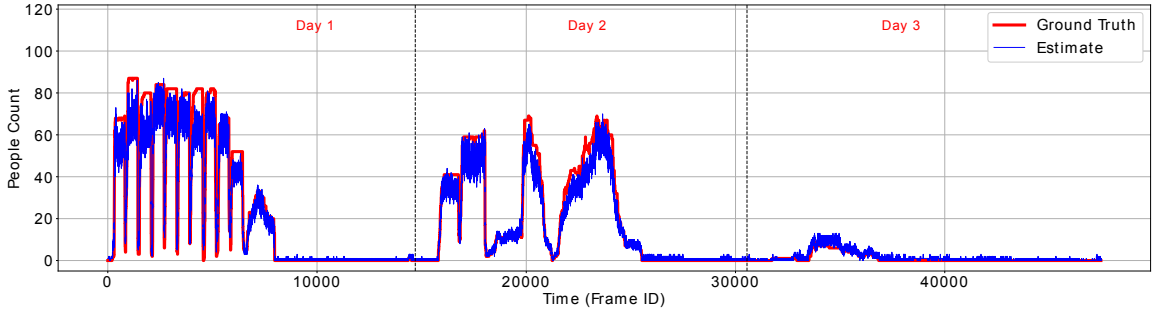
while for $\epsilon = 800$ cm it largely undercounts (Figure 6·12(c)) thus confirming large errors for extreme values of ϵ reported in Figure 6·11. This can be explained as follows. Too small values of threshold ϵ allow little room for image-to-3-D mapping errors; imprecisely mapped locations get absorbed into incorrect clusters. The more accurate the mapping algorithm, the smaller the the value of ϵ that can be used.



(a) $\lambda = 100$ pixels



(b) $\lambda = 400$ pixels



(c) $\lambda = 1500$ pixels

Figure 6·13: The true number of people (red line) and an estimate by 3-camera N-D Score Matrix approach for the 3-day test in 2,000 ft^2 classroom and three different values of λ.

In the extreme case of $\epsilon = 0$ cm, no room for mapping errors is allowed. Unless same-identity locations from all cameras are mapped to the same 3-D location, they cannot form one cluster. Since error-free mappings are very unlikely, for $\epsilon = 0$ cm very few identities would be matched resulting in significant overcounting. As ϵ increases, the degree of overcounting gets reduced. At the other extreme, if ϵ is too

large the mapped locations of different identities might fall into the same cluster thus potentially causing undercounting. Similar conclusions can be drawn for the effect of λ on N-D Score Matrix occupancy estimates (see Figure 6·13).

In Table 6.6, we list occupancy-estimation performance metrics (MAE, MAE_{pp}, Acc_X) for the tuning parameters that yielded the smallest cumulative MAE values for N-D Score Matrix ($\lambda = 400$) and Real-World Location Clustering ($\epsilon = 250$ cm). In addition to the two N-camera approaches with $N = 3$, for comparison we also provide results for the best-performing 2-camera occupancy-estimation algorithm from Section 6.4, namely location-based CBD with $\tau = 0.2$. Note, that the CBD results are identical to those reported in Table 6.5.

We observe that the 2-camera CBD and 3-camera N-D Score Matrix approaches achieve identical cumulative MAE and MAE_{pp}, but Real-World Location Clustering performs slightly worse with a cumulative MAE increase by 0.07 and MAE_{pp} by 0.04. With respect to different occupancy scenarios, the 3-camera N-D Score Matrix slightly outperforms the 2-camera CBD on days 1 and 2 in terms of both MAE and MAE_{pp} values, but fares worse on day 3. Recall, that the classroom is much more crowded on day 1 (up to 87 students, with 10 high-occupancy periods) and on day 2 (up to 65 students, with 4 high-occupancy periods) compared to day 3 (up to 9 students for a short period, and otherwise empty or almost empty). This suggests that in a large space with very many people the 3-camera N-D Score Matrix performs better than the 2-camera CBD approach. In terms of $Acc_{X=0}$, 2-camera CBD outperforms both 3-camera approaches for all 3 days. However, in terms of $Acc_{X=5}$ and $Acc_{X=10}$, N-D Score Matrix achieves the best performance compared to the other algorithms on all days except day 3.

The results in Table 6.6 may seem somewhat disappointing since a 3-camera approach only slightly outperforms a 2-camera approach and only in crowded scenarios.

Table 6.6: People-counting performance (MAE, MAE_{pp} and X-Accuracy) of the CBD approach ($N = 2$ cameras) and N-camera PRID algorithms for $N = 3$ in a 3-day test in a large classroom.

Metric	Algorithm	N	Day 1	Day 2	Day 3	Cum.
MAE	CBD	2	2.43	1.71	0.75	1.59
	N-D Score Matrix	3	2.31	1.62	0.93	1.59
	Real-World Location Clustering	3	2.45	1.68	0.94	1.66
MAE_{pp}	CBD	2	0.082	0.083	0.752	0.097
	N-D Score Matrix	3	0.078	0.079	0.941	0.097
	Real-World Location Clustering	3	0.083	0.082	0.952	0.101
Acc_X [%] X=0/5/10	CBD	2	41/84/96	37/93/99	51/99/100	43/92/98
	N-D Score Matrix	3	39/86/96	35/95/100	48/98/100	41/93/99
	Real-World Location Clustering	3	39/85/95	35/94/99	49/98/100	41/92/98

However, we should note that the 2-camera CBD approach already performs very well in this 2,000 ft^2 test space with MAE_{pp} of 0.097 (9.7% error per person) vastly outperforming the 1-camera RAPiD performance (no PRID needed) that produced a 0.373 MAE_{pp} (or 37.3% error per person) as reported in Table 6.3). Furthermore, as shown in Table 6.1, RAPiD applied to a single-camera video stream can deliver MAE_{pp} of 0.065 (6.5% error per person) up to 800 ft^2 square-room area and 0.133 (or 13.3%) up to 1,250 ft^2. Considering that the test space has 72×28 ft dimensions and that each of the cameras used by the 2-camera CBD approach (cameras #2 and #3, Figure 3·1) roughly covers one half of the classroom (about 36×28 ft space with 1,000 ft^2 area), it is clear that little improvement can be expected from additional cameras in this case. However, the N-camera approaches presented in this section are expected to be highly beneficial in larger spaces in which 2 cameras would be insufficient. Unfortunately, we were not able to access and collect data in larger spaces

due to logistical reasons (installation, networking and maintenance of the cameras). Yet, we believe that the proposed N-camera methodology would be very valuable in scaling up people-counting to much larger spaces such as convention halls, food courts, airports, train/bus stations, etc.

6.6 Chapter Summary and Discussion

In this chapter, we evaluated the *occupancy-estimation* performance of combined people detection and re-identification using 2 overhead fisheye cameras. We also introduced two N-camera occupancy-estimation approaches that can be scaled beyond $N = 2$.

Our experiments show that occupancy estimation using a single overhead fisheye cameras performs well in small-to-mid sized rooms (up to 800 ft^2). However, in larger rooms, single-camera methods struggle due to fisheye distortions, especially due to foreshortening of people's appearance at FOV periphery. To overcome this, we proposed a two-camera approach, where people are first detected in each camera FOV independently but to avoid double-counting of the same person PRID is applied. We showed that two-camera methods perform well in a 2,000 ft^2 classroom in widely-ranging occupancy. However, we acknowledge that two cameras may be insufficient for accurate people counting in larger or topologically-complex spaces. To address this, we proposed two N-camera occupancy estimation methods. Tested on a 3-camera dataset we collected ($N = 3$), these algorithms performed similarly to the two-camera approach with one having a slight edge during high occupancy periods. We believe that the proposed N-camera methods will outperform two-camera systems in rooms much larger than our 2,000 ft^2 test space. Unfortunately, we could not perform such experiments due to logistical reasons.

Chapter 7

Conclusions and Future Directions

In this dissertation, we proposed several novel person re-identification methods for overhead cameras equipped with a fisheye lens. We also introduced approaches that leverage these PRID methods for occupancy estimation in large indoor spaces.

In terms of PRID, we specifically focused on developing methods for overhead fisheye cameras that are time-synchronized and have overlapping FOVs. We named this type of PRID a *cross-frame fisheye PRID*. Unlike other types of PRID discussed in Chapter 2, cross-frame fisheye PRID has not been studied prior to this dissertation. Thus, no public datasets existed that were captured with multiple time-synchronized overhead fisheye cameras.

To study cross-frame fisheye PRID, we collected the first-of-its-kind dataset, *Fisheye Re-Identification Dataset with Annotations* (FRIDA). In Chapter 3, we described the recording setup of FRIDA and extensively discussed the unique challenges that it presents compared to other PRID datasets available in the literature. In addition to introducing FRIDA, we evaluated the performance of 6 state-of-the-art traditional rectilinear PRID methods on FRIDA. For evaluation, in addition to well-known mAP, we used a new metric, that we introduced – Query Matching Score (QMS). The unique feature of QMS compared to other common PRID performance-evaluation metrics is that it accounts for query elements which may have no match in the gallery set. We provided two sets of results. In the first set of experiments, we trained the networks on a well-known side-view rectilinear PRID dataset Market-1501 (Zheng et al.,

2015) but tested them on FRIDA. In the second set of experiments, we applied an identity-wise 2-fold cross-validation on FRIDA, where both testing and training were performed on FRIDA but on different folds. Our experimental results showed that training on FRIDA improves PRID performance on fisheye data compared to training on Market-1501. However, the matching accuracies of networks even trained on FRIDA were still well below those reported for traditional rectilinear PRID datasets, thus prompting further investigation.

All methods we evaluated in Chapter 3 were appearance-based. However, in fisheye PRID the appearance of people gets distorted due to fisheye-lens geometry which hurts the performance of appearance-based methods. Therefore, in Chapter 4, we introduced a cross-frame fisheye PRID method that does not depend on appearance. This method is motivated by the fact that a person can only appear at a single 3-D world location at a given time instant. We used this observation to develop a model that allows mapping of a person's location from one fisheye image to another fisheye image. However, this mapping requires the knowledge of a person's height. We proposed either to use the average height of a person in the US or to sweep a range of reasonable human heights, and developed 4 novel distance measures to quantify the likelihood of a query/gallery identity match. Our experimental results showed that all proposed location-based methods outperform the appearance-based methods with a significant margin. However, would a combination of location-based and appearance-based approaches perform even better?

Since close proximity of people degrades performance of a location-based approach, in Chapter 5 we introduced a cross-frame fisheye PRID approach that combines appearance- and location-based features. We proposed three features to help match query and gallery elements, namely: appearance feature obtained by deep learning, color histogram, and location-based feature. We fused these features by applying nor-

malization and Naïve Bayes approach. To perform identity matching, we introduced a probabilistic approach to decide which camera serves better as a query source and which serves better as a gallery source. Through an ablation study with all possible combinations of features, we showed that a combination of appearance- and location-based features outperforms single-feature algorithms, with the best-performing combination achieving QMS of over 97% and mAP of over 98%.

While PRID has various applications, in Chapter 6 we focused on its application to occupancy estimation (i.e., people counting). The key difference compared to earlier chapters is that in Chapter 6 we used a people-detection algorithm to find occupants rather than ground-truth bounding boxes from manual annotations. In order to motivate the need for PRID in occupancy estimation, we first analyzed the occupancy-estimation performance of a state-of-the-art people-detection algorithm developed for overhead fisheye cameras - RAPiD (Duan et al., 2020). We demonstrated its good performance in a small-to-medium size spaces (up to 800 ft^2) but progressive deterioration for larger and larger spaces. This is due to fisheye-lens geometry causing people to appear very small and distorted at FOV periphery, which RAPiD often misses. To address this, we deployed 3 cameras to monitor a 2,000 ft^2 room. Although multiple cameras helped resolve the problem of missing detections, it introduced a new problem of overcounting when the same person is captured in FOVs of different cameras. To resolve this, we applied PRID between 2 cameras and showed a much improved performance on a 3-day video recording. Compared to single-camera counting which achieves people-counting error of no more than 5 in 72% of video frames, the 2-camera PRID-based methods achieves such error in 92% of video frames.

However, in larger spaces, well beyond 2,000 ft^2, two overhead fisheye cameras will not suffice. Therefore, we explored scaling the proposed two-camera PRID methods

to N cameras. We proposed one method based on the idea of scaling 2-D score matrix for two cameras to N-D score matrix for N cameras. Our second method maps pixel locations of people detected in images from N cameras to 3-D locations in room coordinates and then clusters these 3-D locations to estimate the people count. We evaluated both approaches on the 3-day video recording and showed that both N-camera people-counting methods ($N = 3$) perform comparably to the best performing two-camera PRID method. However, the N-D score-matrix approach outperformed the two-camera approach on days when the space was very crowded. While we could not demonstrate an improved performance of the 3-camera method over a 2-camera method due to limited size of the test space, we are convinced that both of the proposed N-camera approaches will perform very well in very large spaces (e.g., very large lecture halls, exhibition halls, food courts, supermarkets).

7.1 Future Directions

7.1.1 Leveraging Temporal Information

The methods we have proposed and evaluated in this dissertation fall under the umbrella of image-based PRID (Ye et al., 2022) since they do not use any temporal information. However, using temporal information is likely to improve PRID performance since people count does not change dramatically between consecutive video frames (no more than about 1 second). These types of methods would fall into category of video-based PRID, where identity matching is performed based on a group of video frames rather than on a single one. There exist video-based PRID methods developed specifically for rectilinear images (Wu et al., 2018; Li et al., 2018a; Zhou et al., 2017; McLaughlin et al., 2016). However, no such methods have been proposed to-date for overhead fisheye cameras with overlapping fields of view. We believe this would be a fruitful direction for future research, and we see two potential pathways.

In both cases, we assume that a motion track of each person is available in each camera view. These tracks can be obtained either from the ground truth or by applying a people-tracking algorithm for fisheye videos (Tezcan et al., 2022).

The first pathway we imagine can be thought of as an extension of methods we proposed in Chapter 5. The main idea is to generalize a distance measure between one query bounding box and one gallery bounding box to a measure between a sequence of query bounding boxes and a sequence of gallery bounding boxes using ideas introduced in Section 4.3.1. First, features would need to be extracted from each bounding box in the sequence, as explained in Chapter 5. Let the feature vectors of a sequence of bounding boxes from one query be $\{\boldsymbol{A}_i, i = 1, \ldots, m\}$, and those from one gallery element be $\{\boldsymbol{B}_j, j = 1, \ldots, n\}$, where m and n are the coresponding sequence lengths. One way to compute the distance between these two sets would be to follow the ideas proposed in Section 4.3.1, for example: the minimum distance $\min_{i,j} d(A_i, B_j)$ or the total distance $\sum_{i,j} d(A_i, B_j)$, where $d(\cdot, \cdot)$ can be the cosine distance or $L2$-norm.

Another pathway would be to adapt a video-based traditional rectilinear PRID method to perfrom video-based cross-frame fisheye PRID. One of the most effective architectures for handling sequential data (such as video, audio, text) are Recurrent Neural Networks (RNN). For example, in (McLaughlin et al., 2016) an RNN-based PRID method was developed for rectilinear images. A CNN was used to extract features of each bounding box separately, which were subsequently fed into to an RNN. By using temporal pooling, the features produced by the RNN were combined together and fed into a Siamese network with a contrastive loss function instead of a triplet loss. A similar architecture can be used for cross-frame fisheye PRID. However, rather than using only features extracted by a CNN, it would make more sense to combine all features that were demonstrated in Chapter 5 to be beneficial.

7.1.2 General Matching-Score Metrics for N-camera PRID

In Section 6.5.1, we introduced N-camera PRID, which requires construction of an N-dimensional score matrix for identity matching. We proposed one way to compute elements of this matrix (6.6), namely by summing camera-pairwise similarity/dissimilarity scores. However, other methods are possible. One approach would be to map query locations from $k-1$ (where $2 \leq k \leq N$) cameras to a reference camera in the group of k cameras, and develop a measure for k-tuplet of locations (each location coming from different camera) with the goal of finding k-tuplets that form "tightest" groups. Such a measure would need to capture the size of k-vertex polygons, for example:

1. the area of a k-vertex polygon,
2. the perimeter of a k-vertex polygon,
3. the average distance of vertices from the centroid of a k-vertex polygon.

Other distance measures can be considered. The principled search for the best measure would be an interesting direction to explore in future work.

7.1.3 Domain Adaptation

In Chapter 3, we demonstrated a performance gap between rectilinear PRID and fisheye PRID by training on Market-1501 (Zheng et al., 2015) (rectilinear, side-view images) and testing on FRIDA (fisheye, overhead images). Due to a significant mismatch in camera optics and viewing perspective, such an approach did not perform well. An alternative semi-supervised approach, is to use domain-adaptation methods developed for traditional rectilinear PRID (Fu et al., 2019; Deng et al., 2018; Bak et al., 2018), where only unlabeled examples from the target domain are used. However, in susch methods both source and target domains are still captured by rectilinear

cameras from a side viewpoint. Thus, our problem remains more challenging due to differences in lens geometry and acquisition viewpoint between the source domain (e.g., Market-1501 (Zheng et al., 2015), Duke MTMC (Ristani et al., 2016), CUHK03 (Li et al., 2014)) and the target domain (FRIDA). We believe that a good starting point in this case would be the approach proposed in (Fu et al., 2019), where the source dataset would be Market-1501 and the target dataset would be FRIDA.

7.1.4 Distance Estimation

In this dissertation, we focused on person re-identification and people counting using overhead fisheye cameras. As such cameras become more ubiquitous due to their wide field of view and largely-unobstructed perspective, we believe there will be increasing interest in human behavior analysis from such data, e.g., activity recognition to detect suspicious events, generation of occupancy heat-maps to determine which parts of a space are used more frequently, distance estimation between people to adhere to social distancing guidelines. These are largely unexplored topics in the context of overhead fisheye cameras.

Very recently we took a step in this direction by developing two methods for distance estimation between people using a single overhead fisheye camera (Lu. et al., 2023). We proposed and evaluated two methods: a model-based method that maps pixel location of a person to real-world (3-D) coordinates, as detailed in Chapter 4, and a data-driven one using Multi-Layer Perceptron (MLP), that requires annotated inter-person distance data. In addition to distance-estimation performance, we also reported performance of social-distancing violation detection (i.e., binary classification of distance between people, whether above or below 6 ft). To facilitate training and testing of these algorithms, we published a first-of-its-kind dataset *Distance Estimation between People from Overhead Fisheye cameras* (DEPOF) (Lu. et al., 2023). Both algorithms performed far from perfect in terms of distance estimation, with

errors ranging from 18 to 45 in. However, even with such errors the accuracy of detecting social-distance violations (below/above 6 ft) was high at 94%. We hope our work will inspire more research in this direction.

One of the difficulties we faced in this project was the use of average rather then true person's height, that contributed to errors. Thus, it would be very valuable to develop a person-height estimation algorithm (e.g., by taking bounding-box metadata and content as input) which could potentially reduce distance errors in the 3-D mapping approach. However, to improve our MLP model, that was trained on data captured for a single, known height, one would need annotated data captured for different heights. In that case, the DEPOF datset would need to be extended to several heights.

Appendix A

Derivations for Pixel-Correspondence Mapping

Below, we derive equations (4.8), (4.9) and (4.10) from Section 4.1.3. Since these derivations apply to both cameras, we omit the A, B subscripts. Figure A·1 depicts the geometric relationship between various points used in these derivations.

Let $\boldsymbol{P} = [P_x, P_y, P_z]^T$ represent a 3D-world point with $P_z \geq 0$. Then, $||\boldsymbol{P}|| = \sqrt{P_x^2 + P_y^2 + P_z^2}$. The orthogonal projection of $\boldsymbol{P}$ onto the unit sphere centered at $\boldsymbol{O}$ (Figure A·1) is given by:

$$\boldsymbol{S} = \frac{\boldsymbol{P}}{||\boldsymbol{P}||}. \tag{A.1}$$

A perspective projection of $\boldsymbol{S}$ onto the homogeneous imaging plane of the fisheye camera, with camera center $\boldsymbol{C} = [0, 0, -\xi]^T$, $\xi \geq 0$, is given by:

$$\boldsymbol{p} = \boldsymbol{C} + \frac{\boldsymbol{S} - \boldsymbol{C}}{S_z - C_z}, \tag{A.2}$$

since on the homogeneous imaging plane we must have $(p_z - C_z) = 1$.After substituting the expression for $\boldsymbol{S}$ from Eqn. (A.1) into Eqn. (A.2) and simplifying, we obtain the following mapping from $\boldsymbol{P}$ to $\boldsymbol{p}$:

$$\boldsymbol{p} = \left[\frac{P_x}{P_z + \xi||\boldsymbol{P}||}, \frac{P_y}{P_z + \xi||\boldsymbol{P}||}, 1 - \xi\right]^T \tag{A.3}$$

Now suppose that $\boldsymbol{p} = [p_x, p_y, 1 - \xi]^T$ denotes a point on the homogeneous imaging

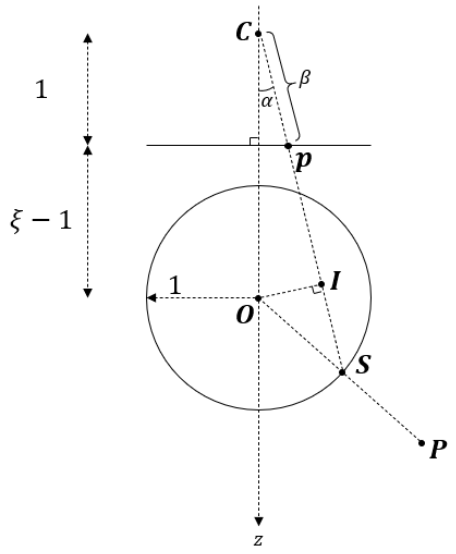


Figure A·1: Geometry of a model for single fisheye camera.

plane of the fisheye camera with coordinates expressed relative to the origin $\boldsymbol{O}$.[1] For convenience we define

$$\cos(\alpha) = \frac{1}{||\boldsymbol{p} - \boldsymbol{C}||} = \frac{1}{\beta}. \tag{A.4}$$

From Figure (A·1) we have:

$$||\boldsymbol{I} - \boldsymbol{C}|| = \xi \cos(\alpha) = \frac{\xi}{\beta} \tag{A.5}$$

$$||\boldsymbol{I}|| = \xi \sin(\alpha) = \xi \sqrt{1 - \frac{1}{\beta^2}} \tag{A.6}$$

$$\begin{aligned} ||\boldsymbol{S} - \boldsymbol{I}|| &= \sqrt{1 - ||\boldsymbol{I}||^2} \\ &= \frac{\sqrt{\beta^2(1-\xi^2) + \xi^2}}{\beta} \end{aligned} \tag{A.7}$$

$$\begin{aligned} ||\boldsymbol{S} - \boldsymbol{C}|| &= ||\boldsymbol{S} - \boldsymbol{I}|| + ||\boldsymbol{I} - \boldsymbol{C}|| \\ &= \frac{\xi + \sqrt{\beta^2(1-\xi^2) + \xi^2}}{\beta} \end{aligned} \tag{A.8}$$

[1] Coordinates of $\boldsymbol{p}$ relative to the camera center $\boldsymbol{C}$ are homogeneous and are given by $(\boldsymbol{p} - \boldsymbol{C}) = [p_x, p_y, 1]^T$.

Since, $\boldsymbol{S}$, $\boldsymbol{p}$, and $\boldsymbol{C}$ are collinear,

$$\begin{aligned}\boldsymbol{S} &= \boldsymbol{C} + (\boldsymbol{p} - \boldsymbol{C})\frac{||\boldsymbol{S} - \boldsymbol{C}||}{||\boldsymbol{p} - \boldsymbol{C}||} \\ &= \boldsymbol{C} + \tau(\boldsymbol{p} - \boldsymbol{C}),\end{aligned} \tag{A.9}$$

where from Eqns. (A.4) and (A.8),

$$\begin{aligned}\tau &:= \frac{\xi + \sqrt{\beta^2(1-\xi^2) + \xi^2}}{\beta^2} \\ &= \frac{\xi + \sqrt{1 + (1-\xi^2)(p_x^2 + p_y^2)}}{p_x^2 + p_y^2 + 1}.\end{aligned} \tag{A.10}$$

Since $\boldsymbol{P}$, $\boldsymbol{S}$, and the origin $\boldsymbol{O}$ are collinear, we should scale $\boldsymbol{S}$ by a factor to match the z-coordinate of $\boldsymbol{P}$. Thus,

$$\boldsymbol{P} = P_z \cdot \frac{\boldsymbol{S}}{S_z} \tag{A.11}$$

After substituting the expression for $\boldsymbol{S}$ from Eqn. (A.9) into Eqn. (A.11) and simplifying we get the following mapping from $\boldsymbol{p}$ to 3D-world point $\boldsymbol{P}$ whose z-coordinate equals P_z:

$$\boldsymbol{P} = P_z \left[u \cdot p_x, u \cdot p_y, 1\right]^T \tag{A.12}$$

$$\begin{aligned}u &= \frac{\tau}{\tau - \xi} \\ &= \frac{\xi + \sqrt{1 + (p_x^2 + p_y^2) \cdot (1 - \xi^2)}}{-\xi(p_x^2 + p_y^2) + \sqrt{1 + (p_x^2 + p_y^2) \cdot (1 - \xi^2)}}\end{aligned} \tag{A.13}$$

where τ is given by Eqn. (A.10). The expression for u can be simplified to

$$u = \frac{1 + \xi\sqrt{1 + (p_x^2 + p_y^2) \cdot (1 - \xi^2)}}{1 - \xi^2 \cdot (p_x^2 + p_y^2)}.$$

Constraint on ξ: We have:

$$\begin{aligned}\xi\sqrt{p_x^2+p_y^2} = \xi \cdot ||[p_x, p_y]^T|| &= \frac{\xi \cdot ||[P_x, P_y]^T||}{P_z + \xi \cdot ||\boldsymbol{P}||} \\ &\leq \frac{\xi \cdot ||\boldsymbol{P}||}{P_z + \xi \cdot ||\boldsymbol{P}||} \\ &\leq 1\end{aligned}$$

where we used equation (A.3) to obtain the second equality. Therefore,

$$0 \leq \xi \leq \frac{1}{\sqrt{p_x^2+p_y^2}}.$$

We note that the lower bound on ξ is an assumption whereas the upper bound on ξ is a consequence of the assumed geometric constraints.

Appendix B

Low-Resolution Overhead Thermal Tripwire for Occupancy Estimation

The main focus of this dissertation is on PRID and people counting using overhead fisheye cameras. However, as pointed out in Chapter 1, there are other means of estimating occupancy. In this appendix[1], we introduce a method that uses low-resolution thermal sensors for occupancy estimation. The main motivation behind choosing low-resolution sensors is to preserve privacy of occupants, i.e., the identities of people should not be recognizable. This is important especially for spaces like bathrooms, locker rooms, etc.

To date, people-counting methods using low-resolution thermal sensors have focused on assessing the state of a room's interior (Beltran et al., 2013; Tyndall et al., 2016; Amin et al., 2008). Such methods can be effective for small rooms, but in case of a large room the field of view (FOV) of a low-resolution thermal sensor might not be sufficient to capture all people in the room. In this scenario, multiple sensors are needed but this increases the cost and complexity of installation, and also requires complex processing to avoid overcounts due to FOV overlap.

In contrast, we propose to count people using a single low-resolution thermal sensor mounted above *every* entry/exit point of a room (Figure B·1) and develop a computational methodology to accomplish this. Regardless of room size, such ther-

[1]This work was published in the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (Cokbas et al., 2020)

mal tripwires can independently detect people entering/exiting a room and jointly estimate the occupancy (state) of the room. In contrast to past methods, our approach is not frame-based but event-based, that is a people count is updated only upon the completion of a door event.

The approach we propose consists of three steps: background subtraction, event detection and event classification. In the first step, we detect "warm" pixels *via* a probabilistic background-temperature model based on Running Gaussian Average (Wren et al., 1997). Since this model does not leverage spatial coherence of temperature, we combine it with a Markov Random Field (MRF) model (McHugh et al., 2009) to produce high-temperature blobs. In the second step, based on background/foreground separation, we detect door events. In a baseline version, we assume that one person passes through the door at a time and we treat all foreground pixels as associated with this person. In order to handle wider doors and multiple people, we develop an enhanced algorithm that identifies high-temperature blobs and tracks them. In the third step, we classify each event as an entry or exit based on the direction of blob movement. To validate the performance of our algorithms, we have collected and manually labeled a dataset of thermal sequences covering various scenarios, including challenging edge cases. This dataset, the first of its kind, is public and available for download. We evaluate our algorithms on this dataset and show that while both proposed algorithms work equally well in normal scenarios the enhanced algorithm outperforms the baseline algorithm on edge cases.

B.1 Methodology

In our approach, we analyze consecutive thermal frames captured by a sensor mounted above a door (Figure B·1) in three steps: (1) background subtraction to first detect the presence of one or more people in the FOV of the sensor; (2) event detection

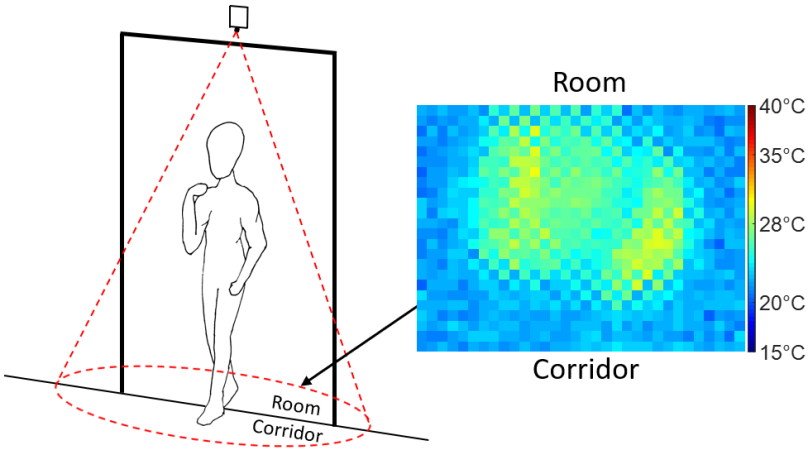


Figure B·1: Configuration of our virtual-tripwire door setup: low-resolution thermal sensor mounted above a door and facing down (left); and 32×24-pixel thermal frame captured by the sensor when a person is leaving the room (right).

to identify the beginning and end of entry or exit events spanning multiple frames; and (3) event classification as an entry or exit (Figure B·2). These three steps are discussed in detail below.

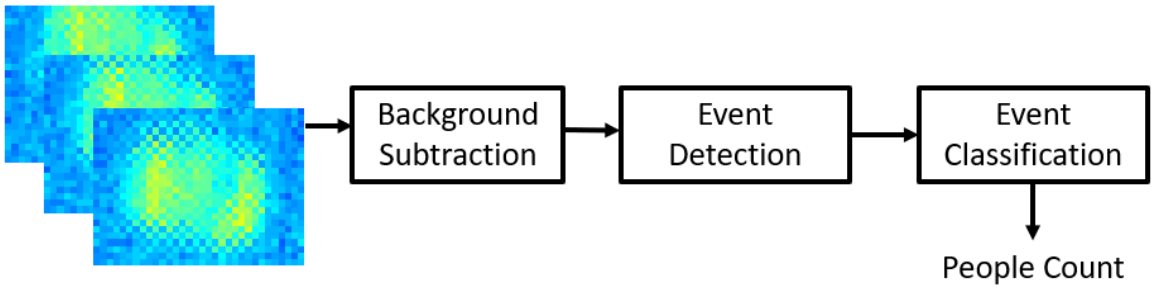


Figure B·2: Block diagram of the proposed approach.

B.1.1 Background Subtraction

In this step, our goal is to separate the pixels that correspond to a human body from those that belong to the background (floor, walls, other surroundings). Since the system is designed for indoor people counting, it is reasonable to assume that a person is warmer than the background. Despite the difference between body temperature

and room temperature, a single global threshold cannot reliably distinguish between them due to natural variations in people and indoor environments. In our approach, instead of thresholding temperature values, we model the background temperature of each pixel by a Gaussian pdf and apply a threshold to the temperature *probabilities*. Let $T_n[\boldsymbol{x}]$ denote the temperature value of a pixel at location $\boldsymbol{x}$ in frame n. We use the Running Gaussian Average (RGA) model (Piccardi, 2004), (Stauffer and Grimson, 1999) to update the mean $\mu_n[\boldsymbol{x}]$ at every background location $\boldsymbol{x}$ as follows:

$$\mu_n[\boldsymbol{x}] \quad = \quad \mathbf{1}(T_n[\boldsymbol{x}]\epsilon B)\Big[\alpha T_n[\boldsymbol{x}] \; + \; (1 \; - \; \alpha)\mu_{n-1}[\boldsymbol{x}])\Big] \; + \; \mathbf{1}(T_n[\boldsymbol{x}]\epsilon F)\mu_n[\boldsymbol{x}] \quad \text{(B.1)}$$

where the sets of background and foreground pixels are denoted by B and F, respectively, $\mathbf{1}(\cdot)$ is an indicator function, and $0 < \alpha < 1$ is a weight controlling recursive update of the mean. We model the probability that a pixel at $\boldsymbol{x}$ belongs to the background as follows:

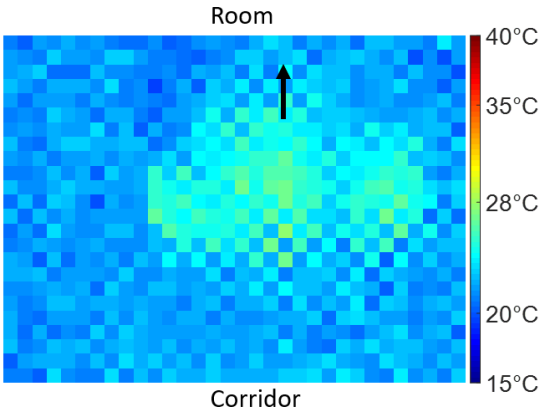
$$P_B(T_n[\boldsymbol{x}]) = \mathcal{N}(T_n[\boldsymbol{x}] - \mu_n[\boldsymbol{x}], \sigma) \tag{B.2}$$

where $\mathcal{N}(\cdot, \cdot)$ denotes the Gaussian distribution with standard deviation σ. We use the same fixed σ for all pixels and perform background subtraction by means of the following binary hypothesis test applied to $P_B(\cdot)$:

$$P_B(T_n[\boldsymbol{x}]) \underset{F}{\overset{B}{\gtrless}} \eta \tag{B.3}$$

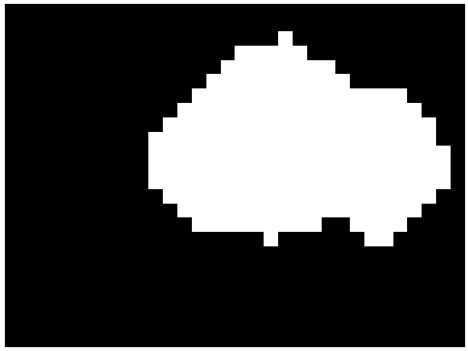
where η is a fixed threshold, identical for all pixels. We refer to this overall background subtraction model as Running Gaussian Average based Background Subtraction (RGA BS) and show a sample result in Figure B·3b.

The background subtraction model discussed so far uses temporal information to separate the foreground from the background. However, all decisions are made independently at neighboring pixels, thus leading to fragmented body-temperature areas.



(a) 32×24-pixel frame from Melexis MLX90640 sensor with person passing through a door. Rows of the frame are aligned with the door frame while columns are orthogonal to the door opening.

(b) Result of background subtraction using RGA BS algorithm.



(c) Result of background subtraction using RGA+MRF BS algorithm.

Figure B·3: Thermal frame and results of background subtraction for a single person passing through a door.

In order to address this, one needs to leverage the spatial contiguity of the human body by applying spatial constraints to foreground estimates. For this purpose, we use an approach proposed by McHugh *et al.* (McHugh et al., 2009). They used a Markov Random Field (MRF) model to ensure spatial estimate coherence within a binary hypothesis test as follows:

$$\frac{P_B(T_n[\boldsymbol{x}])}{P_F(T_n[\boldsymbol{x}])} \underset{F}{\overset{B}{\gtrless}} \theta exp\Big(\frac{Q_F[\boldsymbol{x}] - Q_B[\boldsymbol{x}]}{\gamma}\Big), \tag{B.4}$$

where $P_F(T_n[\boldsymbol{x}])$ is the probability that $T_n[\boldsymbol{x}]$ belongs to the foreground, $Q_F[\boldsymbol{x}]$ and $Q_B[\boldsymbol{x}]$ denote the number of neighboring foreground and background pixels around location $\boldsymbol{x}$, respectively, while θ and γ are parameters. Unlike $P_B(\cdot)$, we assume $P_F(\cdot)$ is a constant (uniform distribution) because we observed that the foreground (body) temperature footprint characteristics can vary significantly depending on clothing, hairstyle and height of a person. Effectively, the right-hand side of the binary hypothesis test (B.4) is a spatially-adaptive threshold. Depending on the labels of neighboring pixels, the threshold will change. If there are more foreground pixels than background pixels in the neighborhood of $\boldsymbol{x}$, the threshold will increase, and, therefore, it will be more likely that the pixel is deemed as belonging to the foreground (and vice versa). Due to the variable threshold, the MRF model increases spatial coherence of foreground estimates, which can be seen in Figure B·3c. The parameter γ can be used to adjust the degree to which the MRF model impacts the threshold.

B.1.2 Event Detection

We propose two different event detection algorithms. Our baseline algorithm assumes that no more than one person will pass under a door at a given time. Our multi-person algorithm, however, is designed to handle multiple people simultaneously passing

through the door.

(a) Baseline Algorithm

We define an event as a sequence of consecutive frames that satisfy the following conditions: (1) the frames immediately preceding and following the event are empty, i.e., have no foreground pixels, (2) each frame in the event has at least one foreground pixel, and (3) there is at least one frame in the event with at least K foreground pixels, were K is a parameter which can be adjusted to account for the height at which the sensor is mounted above the door (smaller K for greater heights).

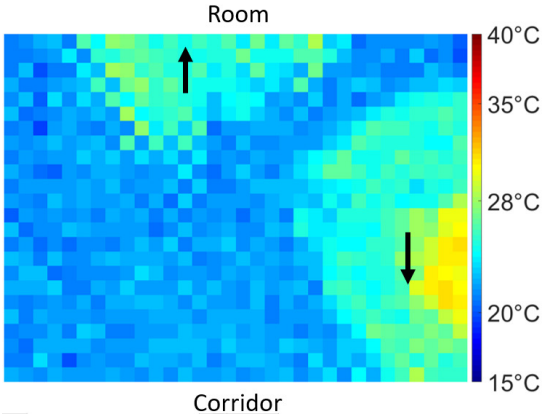
(b) Multi-Person Algorithm

In the baseline algorithm, we assumed that only one person passes under the sensor at a time. If multiple people pass through the door within the same event, the algorithm is incapable of distinguishing them (it calculates only one centroid), thus resulting in an error (Figure. B·4).

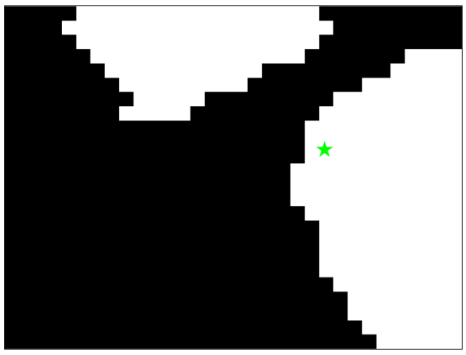
To address this, we detect blobs among foreground pixels in each frame and track their movement. A *blob* is defined as a connected component of foreground pixels of size L or more. We also define a *blob track* as a time sequence of blobs, one in each frame, that are linked between consecutive frames *via* association described below. We consider each blob track to be an event. Blob tracks start, grow and end as described below.

Blob track birth: If there are more blobs in the current frame than in the previous frame, a new blob track is created. The decision as to which blob will be associated with the new blob track is determined after data association in the growth phase.

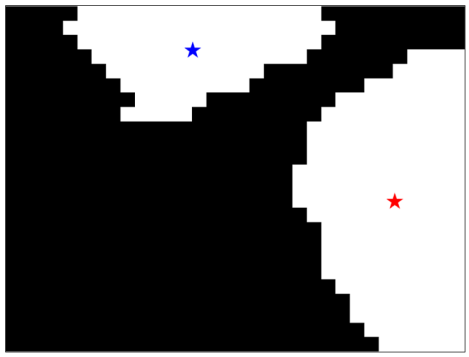
Blob track growth: If the number of blobs in the current and previous frames is the same, then a one-to-one mapping is established between blobs in those frames



(a) 32×24-pixel frame with two people passing through a door.



(b) Result of background subtraction using RGA+MRF BS algorithm with centroid (green star) computed using the baseline algorithm.



(c) Result of background subtraction using RGA+MRF BS algorithm with two centroids (red and blue stars) computed using the multi-person algorithm.

Figure B·4: Thermal frame and results of background subtraction and centroid calculation for 2 people passing through a door.

thus leading to track growth. The track to which a previous-frame blob belongs is grown by a current-frame blob with which the previous-frame blob is associated. This association is established based on the Euclidean distance between blobs' centroids. First, for each blob in the current frame the closest blob is found in the previous frame. Then, the blob pair with the smallest centroid-to-centroid distance is said to be associated with each other and removed from further consideration. The procedure is repeated for the remaining current-frame blobs. Other blob association methods could be applied as well, e.g., minimization of the sum of distances for all blob pairs. However, sophisticated methods may not work as well in our application context due to low thermal sensor resolution, short duration of events and the similarity of thermal footprints of different people.

Blob track termination: If there are fewer blobs in the current frame than in the previous frame, a blob track is terminated. The decision as to which blob is to be terminated is determined after data association in the growth phase.

B.1.3 Event Classification

Both algorithms classify each event at its completion into one of the following classes: (1) a person left the room or (2) a person entered the room. This is accomplished by analyzing the direction of movement of foreground pixels throughout the event. Let F_n be defined as follows:

- baseline algorithm: a set of all foreground pixels at time n,
- multi-person algorithm: a set of all pixels belonging to a single blob at time n (part of a blob track).

We compute the centroid $\boldsymbol{C}_n$ at time n as follows:

$$\boldsymbol{C}_n = \frac{1}{|F_n|} \sum_{\boldsymbol{x} \in F_n} \boldsymbol{x}.$$

Since columns of a thermal frame are orthogonal to the door opening (Figure B·3a), we use the vertical component C_n^v of centroid $\boldsymbol{C}_n = [C_n^h, C_n^v]$ to determine whether a person enters or leaves the room. In particular, we examine whether or not the centroid crosses the mid-line of the frame between two consecutive time instants $n-1$ and n. If C_n^v belongs to the upper part of the frame (top 32×12 pixels of the Melexis 32×24 pixel sensor) whereas C_{n-1}^v belongs to the lower part of the frame (bottom 32×12 pixels) we predict that the person is entering the room. Conversely, if C_n^v belongs to the lower part of the frame whereas C_{n-1}^v belongs to its upper part, we predict that the person is leaving the room. Based on this decision, the people count is updated.

During a hesitant entry/exit or in case of lingering, an event might involve multiple mid-line crossings. We examine the first and last crossings within an event. If the directions of these two crossings are the same, we decide as described above. If the directions differ, we consider this to be a case of lingering and do not update the people count.

B.2 Experimental Results

B.2.1 Dataset

We collected a dataset of thermal image sequences using two Melexis MLX90640 32×24-pixel sensors running at 16 Hz mounted above two doors (Figure B·1) of a small classroom. Compared to previous research (Beltran et al., 2013), (Tyndall et al., 2016) our sensor has a slightly higher spatial resolution, but still a person cannot be visually recognized from the captured data (Figs. B·3a, B·4a).

Our dataset, called TIDOS (Thermal Images for Door-based Occupancy Sensing), is publicly available[2] and includes several types of door activity: single person enter-

[2] `vip.bu.edu/tidos`

ing/leaving the classroom, multiple people entering/leaving through the same door, people lingering in the door, people with backpacks, in thick clothing, carrying various items, etc. Details of the dataset are provided in Table B.1. We manually annotated each frame in the dataset with a number which equals the change in the people count (if any). Such a change can only occur at the end of an event. During annotation, an event is considered to have ended when a person completely leaves the frame. We computed the ground-truth people count in the room using our annotations and the initial people count in the room (Table B.1).

B.2.2 Performance Analysis

We evaluated the performance of our algorithms on TIDOS using the following algorithm parameters: $\alpha = 0.05$ $\sigma = 0.4$ and $\eta = 0.015$ in the RGA model, $\theta P_F(T_n[\boldsymbol{x}]) = 0.015$ (a constant for all $\boldsymbol{x}$) and $\gamma = 0.2$ in the MRF-based hypothesis test, and blob-size threshold of $K = L = 100$ for both baseline and multi-person algorithms. The values of α, $\sigma, \eta, \gamma, \theta P_F$ were selected heuristically. However, the values of K and L are motivated by the typical size of a human body's projected image onto the sensor. Based on physical constraints of our setup ($55^\circ \times 35^\circ$ sensor FOV, 2.4m installation height, 1.7m average human height), we concluded that a body's projection typically occupies 200–250 pixels and this agrees with our observation of recorded data. We used 100 as our threshold to avoid misses in case of shorter people, especially children.

Since both algorithms estimate transitions in the state of a room (people-count changes), in order to estimate the state of the room (people count) an initial state of the room is needed. In our experiments, we used the true initial people count in each room reported in Table B.1.

We use three metrics to evaluate the performance of our algorithms. The first two metrics assess the raw people-count estimation performance and are based on Mean Absolute Error (MAE). Our third metric addresses the drift problem, that

Table B.1: Details of TIDOS (Thermal Images for Door-based Occupancy Sensing) dataset. Each 32×24-pixel frame was acquired by Melexis MLX90640 sensor at 16 fps. Data was collected by 2 sensors, one over each door of a small classroom.

Thermal Recording	Number of frames	Number of entries and exits	Initial people count	Challenges (*scenario*)
Lecture	7,520	2	9	Lingering in doorway (*only single-person events*)
Lunch Meeting 1	37,536	25	0	Wearing a coat; carrying various items; multiple people passing through at the same time
Lunch Meeting 2	9,344	8	12	Carrying a backpack (*only single-person events*)
Lunch Meeting 3	28,128	69	7	Lingering in doorway; wearing a hoodie or carrying a backpack; two people standing in a door and handshaking; multiple people passing through at the same time
Edge Cases	13,120	24	6	Long lingering in doorway; one or two people standing in a door while another person is passing through; multiple people passing through at the same time
High Activity	22,560	133	4	Wearing a hoodie or thick coat; carrying a backpack; pushing a chair through doorway; leaning against a closed door; one person standing in a door while another one is passing through; multiple people passing through at the same time

leads to error accumulation, and temporal misalignments between ground-truth and estimated people-count changes.

(a) Basic Metrics for Count Estimation

Our basic performance metric is the MAE between the true and estimated people counts averaged across all N frames of a thermal sequence. The value of MAE is unaffected by the initial count. However, it scales with the number of people entering/leaving a room which confounds the comparison of MAE values across different occupancy-density scenarios. Thus, we propose another evaluation metric which accounts for the number of people in a room, namely the Per-Person Mean Absolute Error MAE_{PP}, defined as follows:

$$MAE_{PP} = \frac{\sum_{n=1}^{N} |\hat{y}_n - y_n|}{\sum_{n=1}^{N} y_n}, \tag{B.5}$$

where y_n and $\hat{y}_n$ are the ground truth and estimate of the number of people in a room at time n, respectively, and N is the total number of frames in the recording. While, in principle, the denominator in (B.5) could be zero, recordings with no people entering/leaving a room are not interesting for algorithm assessment and are absent from our dataset. We show the performance of our algorithms in terms of MAE and MAE_{PP} in Table B.2 and in terms of frame-wise people count in Figures B·5 and B·6. Unlike MAE, the value of MAE_{PP} is influenced by the initial state of the room since that affects the denominator of Eq. (B.5). Moreover, for all recordings in TIDOS, the denominator of Eq. (B.5) is larger than N, the number of frames in a recording. This causes the MAE_{PP} value to be consistently smaller than the MAE value for the same algorithm applied to the same video.

Baseline algorithm: The baseline algorithm has high MAE and MAE_{PP} values for

Table B.2: Performance comparison of the proposed algorithms on TIDOS dataset using three metrics. The lowest values for MAE and MAE_{PP} and the highest value for CCR_{WCC} for each recording are shown in boldface.

	Baseline algorithm			Multi-person algorithm		
	MAE	MAE_{PP}	CCR_{WCC}	MAE	MAE_{PP}	CCR_{WCC}
Lecture	0.392	0.043	0.500	**0.003**	**0.001**	**1**
Lunch Meeting 1	0.812	0.167	0.880	**0.319**	**0.065**	**0.888**
Lunch Meeting 2	**0.009**	**0.001**	**0.777**	0.016	**0.001**	**0.777**
Lunch Meeting 3	0.973	0.137	0.826	**0.052**	**0.007**	**0.905**
Edge Cases	0.868	0.166	0.666	**0.548**	**0.105**	**0.807**
High Activity	1.431	0.239	0.651	**0.945**	**0.158**	**0.753**

"Lunch Meeting 1", "Lunch Meeting 3", "Edge Cases" and "High Activity" recordings. This is due to multiple-person events that the algorithm cannot handle. As expected, the algorithm works well for single-person events as confirmed by low error values for "Lecture" and "Lunch Meeting 2" recordings.

Multi-person algorithm: The multi-person algorithm performs very well on "Lecture" and "Lunch Meeting 2" confirming its ability to handle single-person events. It also performs well on "Lunch Meeting 1", "Lunch Meeting 3" and "Edge Cases" recordings that contain multiple-person events. Admittedly, it mishandled one of the multi-person events in "Lunch Meeting 1" (Figure B·6, around frame 18,000). The multi-person algorithm does not perform as well on "High Activity", as it is the most challenging recording in the dataset (see Table B.1). Not only does "High Activity" contain the largest number of events, its range of challenges is also widest. Overall, however, the multi-person algorithm significantly outperforms the baseline algorithm in both MAE and MAE_{PP} on all thermal recordings except for "Lunch Meeting 2" for which the error is extremely small anyway.

This performance improvement can be also seen in frame-wise people-count plots (Figures B·5 and B·6). While the baseline algorithm suffers from count drift due to mishandling multiple-person entries/exits (latter parts of "Lunch Meeting 1" and "Lunch Meeting 3"), the multi-person algorithm handles these cases correctly. Clearly, both algorithms have some difficulty with the challenging "High Activity" recording but the multi-person algorithm tracks the ground truth more accurately than the baseline algorithm, which is relfected in MAE and MAE_{PP} values.

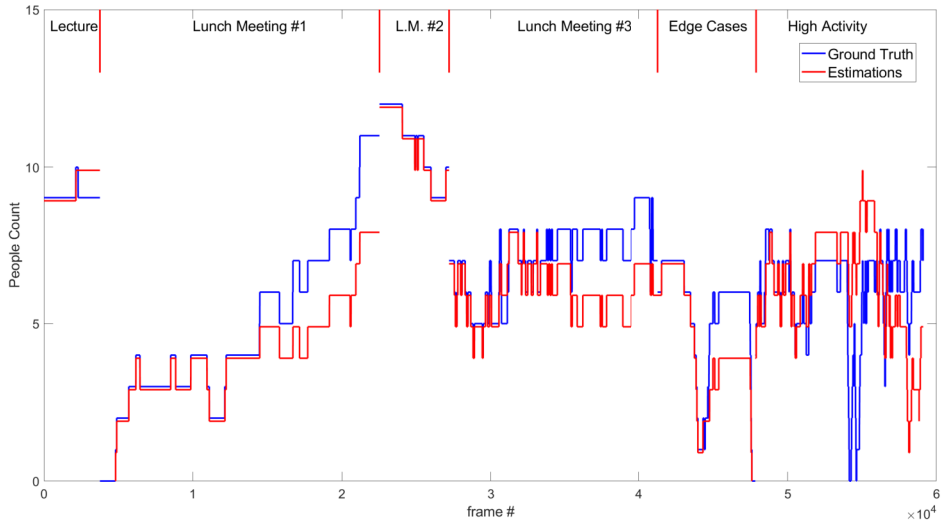


Figure B·5: People counts estimated by the **baseline algorithm**. True (blue) and estimated (red) people-count plots for the proposed algorithms across all recordings in the TIDOS dataset. To distinguish between the red and blue curves in frames where their values exactly coincide, we added a positive vertical offset of 0.1 person to the blue curves. Note that since at each time instant two frames are collected (one by each door sensor), the number of frames in this plot is one-half of the total number of frames in Table B.1.

(b) Metric Robust to Temporal Misalignments and Error Accumulation

Despite a very accurate estimate of counts by both algorithms in "Lunch Meeting 2" (Figures B·5 and B·6), their MAE and MAE_{PP} values are not zero. This is due to the fact that although all events have been correctly classified, the timings

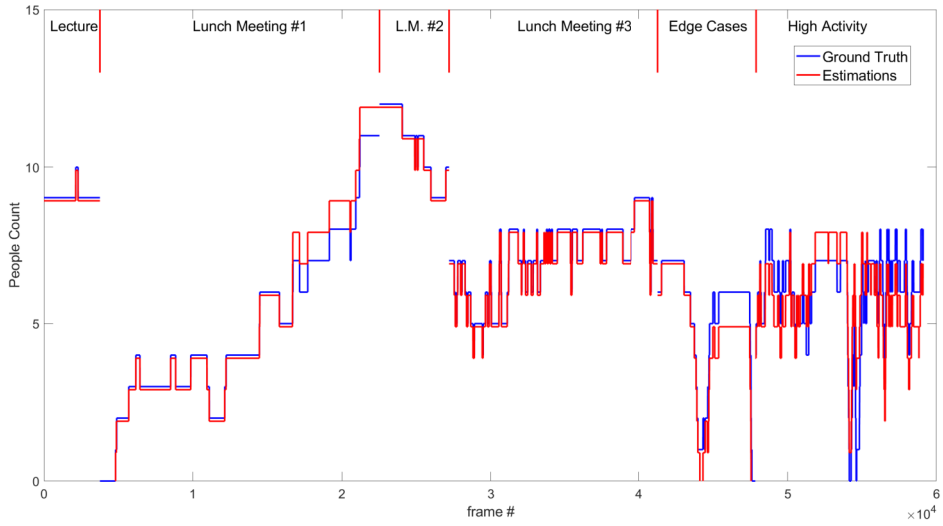


Figure B·6: People counts estimated by the **multi-person algorithm**. True (blue) and estimated (red) people-count plots for the proposed algorithms across all recordings in the TIDOS dataset. To distinguish between the red and blue curves in frames where their values exactly coincide, we added a positive vertical offset of 0.1 person to the blue curves. Note that since at each time instant two frames are collected (one by each door sensor), the number of frames in this plot is one-half of the total number of frames in Table B.1.

of a ground-truth event (marked at its completion) and of its estimate may slightly differ. For instance, in the event definition of the multi-person algorithm, a person is considered as "leaving" a frame if the associated blob has less than L pixels. However, during our manual annotation a person was considered as out of the frame if s/he left the frame completely. These slight temporal misalignments contribute non-zero values to MAE and MAE_{PP} for a few frames. We can ignore the effects of small temporal misalignments during performance assessment by examining whether the estimated count change occurs within a small temporal window w around the time that the true count change takes place.

Furthermore, MAE and MAE_{PP} apply to people counts and are sensitive to error accumulation because a single miscount could potentially contribute an MAE of 1.0

irrespective of the recording duration N. Clearly, a new evaluation metric, resistant to cumulative errors, is needed. Such a metric should focus on *changes* in people counts rather than the counts themselves.

Motivated by these dual considerations, we introduce a new metric, Windowed Count-Change (WCC) Correct Classification Rate (CCR_{WCC}), that accounts for both temporal misalignments and error accumulation, and is defined as follows:

$$\begin{aligned} e_n &= \min_{-w \le \delta \le w} |(y_{n+1} - y_n) - (\hat{y}_{n+1+\delta} - \hat{y}_{n+\delta})| \\ \delta_n &= \arg\min_{-w \le \delta \le w} |(y_{n+1} - y_n) - (\hat{y}_{n+1+\delta} - \hat{y}_{n+\delta})| \\ \widehat{\mathcal{N}} &= \bigcup_{m=1}^{N-1} \{m + \delta_m\}, \\ CCR_{WCC} &= \frac{|\{n : (y_{n+1} \neq y_n) \bigwedge (e_n = 0)\}|}{|\{n : (y_{n+1} \neq y_n) \bigvee (e_n \neq 0)\}| + M} \\ M &= |\{n \notin \widehat{\mathcal{N}} : \hat{y}_{n+1} \neq \hat{y}_n\}| \end{aligned} \tag{B.6}$$

This metric measures the fraction of frames having count changes in which the estimated count-change equals the true count-change within $\pm w$ frames. However, it ignores the frames for which both the estimated and true changes are zero (no door event) which occur very frequently and would skew the traditional definition of CCR. CCR_{WCC} is not only resistant to cumulative errors, but also to jitter: even if a prediction is delayed by $\pm w$ frames compared to ground truth, it can still be considered as correct. This metric is essential for applications where misses and false positives need to be minimized, for example monitoring of entryways to a high-security area. A more detailed explanation of CCR_{WCC} can be found on our website.[3]

However, w needs to be judiciously selected; a large w would unjustly boost CCR_{WCC}. We have considered two constraints on w, a physically-motivated one

[3] `vip.bu.edu/projects/vsns/cossy/thermal`

and a statistically-motivated one. Given our door setup (sensor's $55^\circ \times 35^\circ$ FOV and 2.4m installation height) and a typical speed of 1.2 m/sec for a person entering/exiting a room, we concluded that this person will be at least partially captured in thermal frames for about 1.3 sec. Therefore, w should be less than 1.3 sec in order to ensure that the person immediately following would not be considered as a potential match within $\pm w$. We have also computed a histogram of time differences between estimated and ground-truth entry/exit times for all events in TIDOS. Over 90% of these time differences were within 1 sec. Consequently, in all experiments we used $w = 16$ frames (1 sec).

The results of Table B.2 show that both algorithms fare equally well in terms of CCR_{WCC} on "Lunch Meeting 1" and "Lunch Meeting 2", but the multi-person algorithm clearly outperforms the baseline algorithm by a significant margin on all other recordings. It is also interesting to note that small MAE and MAE_{PP} values need not imply a higher CCR_{WCC} value. Both baseline and multi-person algorithms have lower MAE and MAE_{PP} values for "Lunch Meeting 2" than for "Lunch Meeting 1", yet their CCR_{WCC} values for "Lunch Meeting 1" are much higher than for "Lunch Meeting 2". This phenomenon may be partially attributed to the fact that in evaluation metrics such as MAE and MAE_{PP}, two errors that occur in opposite directions could cancel out each other. For example, if an algorithm misclassifies one entry event and later misclassifies one exit event, the people count errors due to these two misclassifications will "cancel" each other out resulting in zero count errors beyond the second event.

It is clear from Table B.2, that on "High Activity" the multi-person algorithm outperforms the baseline algorithm by a margin of 0.102 in terms of CCR_{WCC} value. This is a significant improvement because the "High Activity" recording has the highest number of entry and exit events and, therefore, a 0.102 fraction of events

corresponds to around 13 entries/exits. Moreover, CCR_{WCC} of 0.753 suggests that three out of four entries and exists were correctly detected and classified within 1 sec of their true occurrence. This is a very solid classification rate for a recording that is mostly composed of very challenging entry/exit scenarios (see Table B.1).

B.3 Discussion

In this study, we developed and systematically studied an overhead virtual tripwire configuration for people counting using a low-resolution thermal sensor. We believe this is the first comprehensive study of its kind encompassing sensor system design and deployment, dataset collection and annotation, algorithm development, design of new performance metrics, and performance evaluation of developed algorithms. The achieved results indicate that typically 80-90% entry and exit events are correctly classified for scenarios with a wide range of extreme challenges, while in simpler, less-active scenarios even 100% correct classification can be reached. However, since our system monitors the changes in occupancy (entry/exit), rather than its state, occasional errors in event detection cause lasting occupancy-level errors known as *drift errors*. This is not the case for occupancy estimation using fisheye cameras (Chapter 6), since the occupancy rather than its change is being estimated.

References

Amin, I., Taylor, A., Junejo, F., Al-Habaibeh, A., and Parkin, R. (2008). Automated people-counting by using low-resolution infrared and visual cameras. Measurement, 41:589–599.

Ardakanian, O., Bhattacharya, A., and Culler, D. (2018). Non-intrusive occupancy monitoring for energy conservation in commercial buildings. Energy and Buildings, 179:311–323.

Bak, S., Carr, P., and Lalonde, J.-F. (2018). Domain adaptation through synthesis for unsupervised person re-identification. In Computer Vision - European Conference on Computer Vision (ECCV) 2018, pages 193–209. Springer.

Barman, A., Wu, W., Loce, R. P., and Burry, A. M. (2018). Person re-identification using overhead view fisheye lens cameras. In 2018 IEEE International Symposium on Technologies for Homeland Security (HST), pages 1–7.

Beltran, A., Erickson, V. L., and Cerpa, A. E. (2013). Thermosense: Occupancy thermal based sensing for hvac control. In Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings, BuildSys'13, page 1–8. Association for Computing Machinery.

Blott, G., Yu, J., and Heipke, C. (2019). Multi-view person re-identification in a fisheye camera network with different viewing directions. Photogrammetrie, Fernerkundung, Geoinformation – Journal of Photogrammetry, Remote Sensing and Geoinformation Science, 87:263–274.

Bone, J., Cokbas, M., Tezcan, O., Konrad, J., and Ishwar, P. (2021). Geometry-based person reidentification in fisheye stereo. In 2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pages 1–10.

Bryan, B., Gong, Y., Zhang, Y., and Poellabauer, C. (2019). Second-order non-local attention networks for person re-identification. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pages 3759–3768.

CBECS (2018). 2018 Commercial Buildings Energy Consumption Survey (final results). `https://www.eia.gov/consumption/commercial`. Accessed: 2023-01-12.

Chen, G., Lin, C., Ren, L., Lu, J., and Zhou, J. (2019a). Self-critical attention learning for person re-identification. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pages 9636–9645.

Chen, T., Ding, S., Xie, J., Yuan, Y., Chen, W., Yang, Y., Ren, Z., and Wang, Z. (2019b). Abd-net: Attentive but diverse person re-identification. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pages 8350–8360.

Choi, H., Um, C. Y., Kang, K., Kim, H., and Kim, T. (2021). Application of vision-based occupancy counting method using deep learning and performance analysis. Energy and Buildings, 252:111389.

Cokbas, M., Bolognino, J., Konrad, J., and Ishwar, P. (2022). FRIDA: Fish-eye re-identification dataset with annotations. In 2022 18th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pages 1–8.

Cokbas, M., Ishwar, P., and Konrad, J. (2020). Low-resolution overhead thermal tripwire for occupancy estimation. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 398–406.

Cokbas, M., Ishwar, P., and Konrad, J. (2023). Spatio-visual fusion-based person re-identification for overhead fisheye images. IEEE Access, 11:46095–46106.

Courbon, J., Mezouar, Y., and Martinet, P. (2012). Evaluation of the unified model of the sphere for fisheye cameras in robotic applications. Advanced Robotics, 26(8):947–967.

del Blanco, C. R., Carballeira, P., Jaureguizar, F., and García, N. (2021). Robust people indoor localization with omnidirectional cameras using a grid of spatial-aware classifiers. Signal Processing: Image Communication, 93:116135.

Demiroz, B. E., Ari, I., Eroglu, O., Salah, A. A., and Akarun, L. (2012). Feature-based tracking on a multi-omnidirectional camera dataset. In 2012 5th International Symposium on Communications, Control and Signal Processing, pages 1–5.

Deng, W., Zheng, L., Ye, Q., Kang, G., Yang, Y., and Jiao, J. (2018). Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 994–1003.

Diraco, G., Leone, A., and Siciliano, P. (2015). People occupancy detection and profiling with 3d depth sensors for building energy management. Energy and Buildings, 92:246–266.

Duan, Z., Ozan T., M., Nakamura, H., Ishwar, P., and Konrad, J. (2020). RAPiD: Rotation-aware people detection in overhead fisheye images. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 2700–2709.

Elkhoukhi, H., Bakhouya, M., El Ouadghiri, D., and Hanifi, M. (2022). Using stream data processing for real-time occupancy detection in smart buildings. Sensors, 22(6):2371.

Erickson, V. L., Achleitner, S., and Cerpa, A. E. (2013). POEM: Power-efficient occupancy-based energy management system. In 2013 ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN), pages 203–216.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96, page 226–231. AAAI Press.

Farenzena, M., Bazzani, L., Perina, A., Murino, V., and Cristani, M. (2010). Person re-identification by symmetry-driven accumulation of local features. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 2360–2367.

Fleuret, F., Shitrit, H. B., and Fua, P. (2014). Re-identification for improved people tracking. In Person Re-Identification, pages 309–330. Springer London.

Fu, Y., Wei, Y., Wang, G., Zhou, Y., Shi, H., Uiuc, U., and Huang, T. (2019). Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pages 6111–6120.

Geyer, C. and Danilidis, K. (2001). Catadioptric projective geometry. International Journal of Computer Vision, 45:223–243.

Gray, D. and Tao, H. (2008). Viewpoint invariant pedestrian recognition with an ensemble of localized features. In Computer Vision – European Conference on Computer Vision (ECCV) 2008, page 262–275. Springer-Verlag.

Hashimoto, K., Morinaka, K., Yoshiike, N., Kawaguchi, C., and Matsueda, S. (1997). People count system using multi-sensing application. In Proceedings of International Solid State Sensors and Actuators Conference (Transducers '97), volume 2, pages 1291–1294.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778.

Hirzer, M., Beleznai, C., Roth, P. M., and Bischof, H. (2011). Person re-identification by descriptive and discriminative classification. In Proceedings of the 17th Scandinavian Conference on Image Analysis, SCIA'11, pages 91–102, Berlin, Heidelberg. Springer-Verlag.

Hirzer, M., Roth, P. M., Köstinger, M., and Bischof, H. (2012). Relaxed pairwise learned metric for person re-identification. In Computer Vision – European Conference on Computer Vision (ECCV) 2012, pages 780–793. Springer Berlin Heidelberg.

Hu, H., Hachiuma, R., Saito, H., Takatsume, Y., and Kajita, H. (2022). Multi-camera multi-person tracking and re-identification in an operating room. Journal of Imaging, 8(8):219.

Huang, Q., Ge, Z., and Lu, C. (2016). Occupancy estimation in smart buildings using audio-processing techniques. arXiv preprint arXiv:1602.08507.

Karanam, S., Gou, M., Wu, Z., Rates-Borras, A., Camps, O., and Radke, R. J. (2019). A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(3):523–536.

Kim, S., Kang, S., Ryu, K. R., and Song, G. (2019). Real-time occupancy prediction in a large exhibition hall using deep learning approach. Energy and Buildings, 199:216–222.

Köstinger, M., Hirzer, M., Wohlhart, P., Roth, P. M., and Bischof, H. (2012). Large scale metric learning from equivalence constraints. In 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2288–2295.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. Annals of Mathematical Statistics, 22(1):79–86.

Li, S., Bak, S., Carr, P., and Wang, X. (2018a). Diversity regularized spatiotemporal attention for video-based person re-identification. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 369–378.

Li, S., Tezcan, M. O., Ishwar, P., and Konrad, J. (2019). Supervised people counting using an overhead fisheye camera. In 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pages 1–8.

Li, W., Zhao, R., Xiao, T., and Wang, X. (2014). DeepReID: Deep filter pairing neural network for person re-identification. In 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 152–159.

Li, W., Zhu, X., and Gong, S. (2018b). Harmonious attention network for person re-identification. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 2285–2294.

Liao, S., Hu, Y., Zhu, X., and Li, S. Z. (2015). Person re-identification by local maximal occurrence representation and metric learning. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2197–2206.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In Computer Vision – European Conference on Computer Vision (ECCV) 2014, pages 740–755. Springer.

Liu, D., Guan, X., Du, Y., and Zhao, Q. (2013). Measuring indoor occupancy in intelligent buildings using the fusion of vision sensors. Measurement Science and Technology, 24(7):074023.

Lloyd, S. (1982). Least squares quantization in PCM. IEEE Transactions on Information Theory, 28(2):129–137.

Loy, C. C., Liu, C., and Gong, S. (2013). Person re-identification by manifold ranking. In 2013 IEEE International Conference on Image Processing, pages 3567–3571.

Lu, H., Tuzikas, A., and Radke, R. J. (2021). A zone-level occupancy counting system for commercial office spaces using low-resolution time-of-flight sensors. Energy and Buildings, 252:111390.

Lu., Z., Cokbas., M., Ishwar., P., and Konrad., J. (2023). Estimating distances between people using a single overhead fisheye camera with application to social-distancing oversight. In Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - VISAPP, volume 5, pages 528–535. INSTICC, SciTePress.

Ma, N., Knapp, R. B., Polys, N. F., Huang, J.-B., Ibrahim, A., Hurt, C., and Xiao, Y. (2018). Mirror worlds challenge. `www2.icat.vt.edu/mirrorworlds/challenge/index.html`.

McHugh, J. M., Konrad, J., Saligrama, V., and Jodoin, P. (2009). Foreground-adaptive background subtraction. IEEE Signal Processing Letters, 16(5):390–393.

McLaughlin, N., Martinez del Rincon, J., and Miller, P. (2016). Recurrent convolutional network for video-based person re-identification. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1325–1334.

Nguyen, T. A. and Aiello, M. (2013). Energy intelligent buildings based on user activity: A survey. Energy and Buildings, 56:244–257.

Piccardi, M. (2004). Background subtraction techniques: a review. In 2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No 04CH37583), volume 4, pages 3099–3104.

Piechocki, M., Kraft, M., Pajchrowski, T., Aszkowski, P., and Pieczynski, D. (2022). Efficient people counting in thermal images: The benchmark of resource-constrained hardware. IEEE Access, 10:124835–124847.

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

Ryan, D., Denman, S., Sridharan, S., and Fookes, C. (2011). Scene invariant crowd counting. In 2011 International Conference on Digital Image Computing: Techniques and Applications, pages 237–242.

Sruthi, M. S. (2019). Iot based real time people counting system for smart buildings. International Journal of Emerging Technology and Innovative Engineering, 5:83.

Stauffer, C. and Grimson, W. E. L. (1999). Adaptive background mixture models for real-time tracking. In Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149), volume 2, pages 246–252 Vol. 2.

Sun, Y., Zheng, L., Yang, Y., Tian, Q., and Wang, S. (2018). Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y., editors, Computer Vision – European Conference on Computer Vision (ECCV) 2018, pages 501–518. Springer International Publishing.

Szczurek, A., Maciejewska, M., and Pietrucha, T. (2017). Occupancy determination based on time series of CO2 concentration, temperature and relative humidity. Energy and Buildings, 147:142–154.

Tamura, M., Horiguchi, S., and Murakami, T. (2019). Omnidirectional pedestrian detection by rotation invariant training. In 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 1989–1998.

Tezcan, M. O., Duan, Z., Cokbas, M., Ishwar, P., and Konrad, J. (2022). Wepdtof: A dataset and benchmark algorithms for in-the-wild people detection and tracking from overhead fisheye cameras. In IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages 1381–1390.

Tezcan, M. O., Konrad, J., and Muroff, J. (2018). Automatic assessment of hoarding clutter from images using convolutional neural networks. In 2018 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI), pages 1–4.

Tyndall, A., Cardell-Oliver, R., and Keating, A. (2016). Occupancy estimation using a low-pixel count thermal imager. IEEE Sensors Journal, 16(10):3784–3791.

van Lint, J. H. and Wilson, R. M. (1992). The principle of inclusion and exclusion; inversion formulae. In A Course in Combinatorics, pages 89–97. Cambridge University Press.

Wang, H., Wang, G., and Li, X. (2021). Image-based occupancy positioning system using pose-estimation model for demand-oriented ventilation. Journal of Building Engineering, 39:102220.

Wei, L., Zhang, S., Gao, W., and Tian, Q. (2018). Person transfer GAN to bridge domain gap for person re-identification. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 79–88.

Wei, S., Tien, P. W., Chow, T. W., Wu, Y., and Calautit, J. K. (2022). Deep learning and computer vision based occupancy co2 level prediction for demand-controlled ventilation (dcv). Journal of Building Engineering, 56:104715.

Wieczorek, M., Rychalska, B., and Dabrowski, J. (2021). On the unreasonable effectiveness of centroids in image retrieval. In Neural Information Processing, pages 212–223. Springer International Publishing.

Wren, C. R., Azarbayejani, A., Darrell, T., and Pentland, A. P. (1997). Pfinder: real-time tracking of the human body. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(7):780–785.

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

Xiong, F., Gou, M., Camps, O., and Sznaier, M. (2014). Person re-identification using kernel-based metric learning methods. In Computer Vision – European Conference on Computer Vision (ECCV) 2014, pages 1–16. Springer International Publishing.

Yang, Y., Yang, J., Yan, J., Liao, S., Yi, D., and Li, S. Z. (2014). Salient color names for person re-identification. In Computer Vision - European Conference on Computer Vision (ECCV) 2014, pages 536–551. Springer International Publishing.

Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., and Hoi, S. C. (2022). Deep learning for person re-identification: A survey and outlook. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(06):2872–2893.

Yu, H.-X., Zheng, W.-S., Wu, A., Guo, X., Gong, S., and Lai, J.-H. (2019). Unsupervised person re-identification by soft multilabel learning. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 2143–2152.

Zheng, F., Deng, C., Sun, X., Jiang, X., Guo, X., Yu, Z., Huang, F., and Ji, R. (2019a). Pyramidal person re-identification via multi-loss dynamic training. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 8506–8514.

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., and Tian, Q. (2015). Scalable person re-identification: A benchmark. In 2015 IEEE International Conference on Computer Vision (ICCV), pages 1116–1124.

Zheng, W., Gong, S., and Xiang, T. (2009). Associating groups of people. In British Machine Vision Conference, pages 1–23.

Zheng, W.-S., Gong, S., and Xiang, T. (2011). Person re-identification by probabilistic relative distance comparison. In 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 649–656.

Zheng, Z., Yang, X., Yu, Z., Zheng, L., Yang, Y., and Kautz, J. (2019b). Joint discriminative and generative learning for person re-identification. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 2133–2142.

Zhihui, Z., Jiang, X., Zheng, F., Guo, X., Huang, F., Sun, X., and Zheng, W. (2020). Viewpoint-aware loss with angular regularization for person re-identification. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 13114–13121.

Zhou, K., Yang, Y., Cavallaro, A., and Xiang, T. (2019). Omni-scale feature learning for person re-identification. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pages 3701–3711.

Zhou, Z., Huang, Y., Wang, W., Wang, L., and Tan, T. (2017). See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 6776–6785.

CURRICULUM VITAE

Mertcan Cokbas
8 St Mary's St, Boston, MA 02215
mcokbas@bu.edu

EDUCATION

Boston University, Boston, MA *Sep 2018 – Present*
Ph.D., Electrical and Computer Engineering, GPA – 3.84

Sabancı University, Istanbul, Turkey *Sep 2014 – May 2018*
B.S., Electronics Engineering, GPA – 3.96

RESEARCH & TEACHING EXPERIENCE

Graduate Research Assistant *Sep 2018 – Present*
Visual Information Processing Lab,
Department of Electrical and Computer Engineering, Boston University

Graduate Teaching Assistant *Sep 2019 – May 2020*
Department of Electrical and Computer Engineering, Boston University

WORK EXPERIENCE

Software Engineer Intern, Machine Learning *May 2022 – Aug 2022*
Facebook, Menlo Park, CA

Software Engineer Intern *Nov 2021 – Feb 2022*
Motional, Boston, MA

PUBLICATIONS

- J. Konrad, **M. Cokbas**, P. Ishwar, Thomas D. Little, Michael Gevelber, "High-Accuracy People Counting in Large Spaces Using Overhead Fisheye Cameras," submitted to *Energy & Buildings*, 2023.

- **M. Cokbas**, P. Ishwar and J. Konrad, "Spatio-Visual Fusion-Based Person Re-Identification for Overhead Fisheye Images," in *IEEE Access*, vol. 11, pp. 46095-46106, 2023.

- Z. Lu, **M. Cokbas**, P. Ishwar, and J. Konrad, "Estimating Distances Between People using a Single Overhead Fisheye Camera with Application to Social-Distancing Oversight," *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications* (VISAPP), vol. 5, pp. 528-535 ,2023.

- **M. Cokbas**, J. Bolognino, J. Konrad, and P. Ishwar, "FRIDA: Fisheye Re-identification Dataset with Annotations," *IEEE International Conference on Advanced Video and Signal Based Surveillance* (AVSS), pp. 1-8, 2022.

- M. O. Tezcan, Z. Duan, **M. Cokbas**, P. Ishwar and J. Konrad, "WEPDTOF: A Dataset and Benchmark Algorithms for In-the-Wild People Detection and Tracking from Overhead Fisheye Cameras," *IEEE/CVF Winter Conference on Applications of Computer Vision* (WACV), pp. 1381-1390, 2021.

- L. Patino, J. Boyle, J. Ferryman, J. Auer, J. Pegoraro, R. Pflugfelder, **M. Cokbas**, J. Konrad, P. Ishwar, G. Slavic, L. Marcenaro, Y. Jiang, Y. Jin, H. Ko, G. Zhao, G. Ben-Yosef, and J. Qiu, "PETS2021: Through-foliage Detection and Tracking Challenge and Evaluation," *IEEE International Conference on Advanced Video and Signal Based Surveillance* (AVSS), pp. 1-10, 2021

- J. Bone, **M. Cokbas**, M. O. Tezcan, J. Konrad and P. Ishwar, "Geometry-based Person Re-identification in Fisheye Stereo," *IEEE International Conference on Advanced Video and Signal Based Surveillance* (AVSS), pp. 1-10, 2021.

- M. Peng, **M. Cokbas**, U. D. Gallastegi, P. Ishwar, J. Konrad, B. Kulis, and V. K. Goyal, "Convolutional Neural Network Denoising of Focused Ion Beam Micrographs," *Proceedings of IEEE International Workshop on Machine Learning for Signal Processing*, pp. 1-6, 2021.

- **M. Cokbas**, P. Ishwar and J. Konrad, "Low-Resolution Overhead Thermal Tripwire for Occupancy Estimation," *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPRW), pp. 398-406, 2020.