PERSON RE-IDENTIFICATION IN DISTRIBUTED WIDE-AREA SURVEILLANCE

A Dissertation

Presented to

the Faculty of the Department of Computer Science

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

By

Apurva Bedagkar-Gala

May 2014

PERSON RE-IDENTIFICATION IN DISTRIBUTED WIDE-AREA SURVEILLANCE

Apurva Bedagkar-Gala

APPROVED:

Shishir Shah, Chairman
Dept. of Computer Science

Edgar Gabriel
Dept. of Computer Science

Christoph Eick
Dept. of Computer Science

Weidong Shi
Dept. of Computer Science

Jayan Eledath
SRI International

Dean, College of Natural Sciences and Mathematics

Acknowledgements

I would like to thank to my advisor, Dr. Shishir Shah, for being an excellent teacher, mentor, and support during my time at University of Houston. It's been a privilege working with him, I consider myself lucky for having been his Ph.D. student and grateful for his patience. Thanks to members of Quantitative Imaging for many enlightening chats. I also thank all my friends, near and far, for their support and care. Finally, I want to thank my whole family, my parents: Sunita and Vinay Bedagkar, my brother: Ashish Bedagkar and most importantly my husband: Deepak M. Gala, for their support, patience and understanding. And last but not least my son, Adeet Gala for bringing me all the cheer, joy and unconditional love. This Ph.D. degree would not have been possible without their love and encouragement.

PERSON RE-IDENTIFICATION IN DISTRIBUTED WIDE-AREA SURVEILLANCE

An Abstract of a Dissertation
Presented to
the Faculty of the Department of Computer Science
University of Houston

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

By
Apurva Bedagkar-Gala
May 2014

Abstract

Person re-identification (Re-ID) is a fundamental task in automated video surveillance and has been an area of intense research in the past few years. Given an image or video of a person taken from one camera, re-identification is the process of identifying the person from images or videos taken from a different camera. Re-ID is indispensable in establishing consistent labeling across multiple cameras or even within the same camera to re-establish disconnected or lost tracks. Apart from surveillance it has applications in robotics, multimedia, and forensics. Person re-identification is a difficult problem because of the visual ambiguity and spatio-temporal uncertainty in a person's appearance across different cameras. However, the problem has received significant attention from the computer-vision-research community due to its wide applicability and utility. In this work, we explore the problem of person re-identification for multi-camera tracking, to understand the nature of Re-ID, constraints and conditions under which it is to be addressed and possible solutions to each aspect. We show that Re-ID for multi-camera tracking is inherently an open set Re-ID problem with dynamically evolving gallery and open probe set. We propose multi-feature person models for both single and multi-shot Re-ID with a focus on incorporating unique features suitable for short as well as long period Re-ID. Finally, we adapt a novelty detection technique to address the problem of open set Re-ID. In conclusion we identify the open issues in Re-ID like, long-period Re-ID and scalability along with a discussion on potential directions for further research.

Contents

1 Introduction 1

1.1 Motivation 1

1.2 Person Re-ID: Problem Definition 4

1.3 Research Contributions 6

1.4 Dissertation outline 8

2 Person Re-identification: Challenges and Scenarios 10

2.1 Task and its challenges 10

2.1.1 System-level Challenges 12

2.1.2 Component-level Challenges: Descriptor Issues 13

2.1.3 Component-level Challenges: Correspondence Issues 14

2.2 Person Re-ID Scenarios 15

2.2.1 Open-set Re-ID 17

2.2.2 Closed-set Re-ID 19

3 Related Work 20

3.1 Contextual Methods 20

3.1.1 Camera Geometry as Context 21

3.1.2 Camera Calibration as Context 22

3.2 Non-Contextual Methods 24

3.2.1 Passive Methods 25

3.3 Active Methods 29

3.3.1 Color Calibration 30

3.3.2 Descriptor Learning 31

3.3.3 Distance Metric Learning 36

3.4 Public Datasets and Evaluation Metrics 40

3.4.1 ViPER 41

3.4.2 ETHZ 41

3.4.3 i-LIDS for Re-ID 42

3.4.4 CAVIAR4REID 43

3.4.5 i-LIDS MA and AA 44

3.4.6 V-47 47

3.4.7 QMUL underGround Re-IDentification (GRID) Dataset . . . 48

3.4.8 Additional Datasets 49

3.4.9 Limitations of Datasets 49

3.4.10 Evaluation Metrics 51

3.4.11 Limitations of current work and our approach 52

4 Color & Facial features-based Person-Model 53

4.1 Color Model 54

4.1.1 Body Part Extraction 55

4.1.2 Part Color Feature Extraction 56

4.1.3 Spatio-temporal Color Model Generation 60

4.2 Facial features based Model 62

4.2.1 Face Image Selection 65

4.2.2 Facial Model Generation 66

4.3 Re-identification by Matching 67

4.4 Experimental Results 69
4.4.1 MCID Dataset 69
4.4.2 Multiple Person Re-ID Experiments 73

5 Color & Gait features-based Person-Model Generation 83

5.1 Gait Features 86
5.2 Cross-view Gait Recognition 87
5.3 Gait Features Model 88
5.3.1 Gait Energy Image (GEI) 89
5.3.2 Frame Difference Energy Image (FDEI) 91
5.4 Feature matching for Re-ID 93
5.4.1 Gait Recognition by Sparsified Representation 94
5.5 Experimental Results 95
5.5.1 SAIVT SoftBio dataset 95
5.5.2 MCID dataset 98

6 Open-Set Person Re-ID 103

6.1 Why is Re-ID an open set recognition problem? 103
6.2 Open Set Re-ID: Proposed Approach 107
6.2.1 Binary SVM for novelty detection 109
6.2.2 One-vs-Set SVM for novelty detection 111
6.3 Experimental Results 113

7 Conclusion and Future Work 122

7.1 Summary of Current Work 122
7.2 Future Work 124
7.2.1 Open Issues in Person Re-ID 124

Bibliography **128**

List of Figures

1.1 Multi-camera surveillance network illustration of Re-ID, colored dots depict different people and numbers besides the dots are the IDs assigned to them. 5

2.1 Images of the same person taken from different cameras to illustrate the appearance changes. The top row images were captured on the same day, bottom row images were captured on different days. 11

2.2 Re-ID System Diagram. 12

2.3 Re-ID as a recognition system. 16

2.4 Multi-camera tracking scenario based on open set Re-ID, colored dots depict different people and numbers besides the dots are the IDs assigned to them. 18

3.1 Method-based taxonomy of Re-ID approaches. 21

3.2 Example images from the VIPeR dataset. 42

3.3 Example images from the ETHZ for Re-ID dataset, the first, second and third rows of images come from sequences 1, 2 and 3 respectively. 43

3.4 Example images from the i-LIDS for Re-ID dataset. 44

3.5 Example images from the CAVIAR for Re-ID dataset. 45

3.6 Example images from the automatically annotated i-LIDS dataset for Re-ID. 46

3.7 Example images from the manually annotated i-LIDS dataset for Re-ID. 46

3.8 Example images from the v-47 for Re-ID dataset. 47

3.9 Example images from the GRID dataset for Re-ID. 48

4.1 Spatio-temporal Color Model Generation. 55

4.2 Representative Colors Descriptor Generation. 59

4.3 Representative Meta Colors Extraction. 62

4.4 Combined Color and Facial feature based Model Generation Pipeline. 64

4.5 Facial feature based Model Generation. 66

4.6 Outdoor placement of cameras: (a) Placement of cameras overlayed on building floor plan for perspective. (b) Example images from each outdoor cameras FOV. 71

4.7 Indoor placement of cameras: (a) Placement of cameras overlayed on building floor plan for perspective. (b) Example images from each indoor cameras FOV. 72

4.8 Re-identification Rate in different scenarios for comparative study of the Color Invariant Histograms. 74

4.9 Closed Set Re-identification Performance: (a) and (b) show the comparative Re-ID rate for N = 2 and 10, respectively. Each graph shows the results obtained by our color model and SDALF [19]. 75

4.10 Closed Set Re-identification Performance: the bar graph shows the results obtained by our color model, color and face model and SDALF [19]. 77

4.11 Open Set Results: Accuracy vs. FAR curves obtained using (a) Optimal Assignment, and (b) Suboptimal Assignment. The top, middle and bottom row are results obtained on the Outdoor-Outdoor, Indoor-Indoor and Outdoor-Indoor scenarios respectively. 80

5.1 Images from SAIVT SoftBio dataset and extracted silhouettes of the same person from different cameras, (a) images with visible face regions, (b) images without visible face regions. 84

5.2 Example images from SAIVT SoftBio dataset and erroneous extracted silhouettes. 85

5.3 Gait Period Estimation: (a) Sample image sequence. (b) Sample silhouette sequence. (c) Bounding box aspect ratio after average filtering. (d) Auto-correlation signal of aspect ratio signal. (d) First order derivative of auto-correlation. (f) Detected peak positions. 89

5.4 GEI features extracted from sample sequences from the 2 datasets, (a) Lower body GEI from MCID dataset and (b) GEI from the SAIVT Soft-Bio dataset. 90

5.5 GEI features extracted from sample sequences from the 2 datasets, (a) Lower body FDEI from MCID dataset and (b) FDEI from the SAIVT Soft-Bio dataset. 92

5.6 CMC Curves on the SAIVT SoftBio dataset, combining color and gait features with varying weights for Re-ID. 100

5.7 CMC Curves on the SAIVT SoftBio dataset, using only color, only GEI and only FDEI features for Re-ID. 101

5.8 Examples of Re-ID gallery ranking, showing the first 5 ranked on SAIVT SoftBio dataset using: (a) only color, (b) color+GEI and (c) color+FDEI, for two different IDs. The images highlighted by the red border denote the true match. 101

5.9 Closed set Re-ID performance: (a)Bar graph shows the results obtained by our color, color+GEI and color+FDEI on the entire MCID dataset and (b) Bar graph shows the results obtained by our color, color+GEI, color+FDEI, only GEI and only FDEI on subset of MCID dataset. 102

5.10 Examples of Re-ID gallery ranking, showing the first 5 ranked on MCID dataset using: (a) only color, (b) color+GEI and (c) color+FDEI for two different IDs. The images highlighted by the red border denote the true match. 102

6.1 Re-identification setup and gallery generation in a multi-camera scenario. 104

6.2 Open Set Re-ID: Novelty Detection followed by Gallery Ranking. . . 107

6.3 Binary SVM: separating hyperplane. 110

6.4 Positive open space: all the space on the right side of the positive plane is the positive open space, PO and the space denoted by the red circle is the positively labeled space, PT (defined by the positive labeled training data). 112

6.5 Dataset split to train and test the proposed novelty detection SVMs. 114

6.6 Novelty Detection Stage Performance: Top row shows the changes in classification accuracy with the probe set openness and the bottow row images shows the F-measure. 118

6.7 Novelty Detection Stage Performance: Top and bottom row show evolution of TNs and FPs, respectively, with the probe set openness. . . 119

6.8 Novelty Detection Output: Comparison of the classification performance of both the oneSet and binary SVMs for probe openess = 7% case, i.e. probe set consists of all 15 gallery IDs plus 5 unknown IDs. 120

6.9 Open-Set Re-ID: Top row shows the changes in Re-ID rate (accuracy) with the probe set openness and the bottom row images shows the Re-ID false acceptance rate. 121

List of Tables

3.1 Summary of recent non-contextual person Re-ID approaches, ✓ and × indicate whether the method incorporates the indicated function or not, respectively. 25

3.2 Summary of public person Re-ID datasets. 41

3.3 Summary of performance of state-of-the-art approaches on popular Re-ID datasets. The results correspond to the single-shot case, for LMNN-R, Impostor Learning, PRDC and RankSVM methods; the accuracy on i-LIDS corresponds to gallery size of 30 out of total 119 (not the complete gallery) and on VIPeR corresponds to gallery size 316 out of total 632. 50

4.1 Comparison of computation times (in seconds) taken by each stage of Re-ID by our method and the SDALF method. 76

4.2 Open Set (Discrimination) results for all the models. The top portion of the table is generated using the Munkers assignment algorithm, whereas the lower portion comes from the suboptimal assignment algorithm. 79

5.1 Rank 1, 5 and 20 matching accuracy and nAUC measures for color, color+GEI and color+FDEI with varying weights. 96

5.2 Rank 1, 5 and 20 matching accuracy and nAUC measures for color, GEI and FDEI features. 97

Chapter 1

Introduction

1.1 Motivation

In the past two decades, increased terrorist activity has sparked heightened focus on safety and security in and around public and private facilities worldwide. This has led to significant changes in security techniques and technologies adopted by law enforcement agencies to discourage future attacks and avoid undesirable events. Governments and businesses increasingly employ surveillance cameras. In Britain, for example, there are up to 4.2 million surveillance cameras - about one for every 14 people [2]. The US homeland Security Agency supports millions of dollars in funding for local, state and federal agencies to install modern video surveillance systems [121]. Law enforcement agencies have come to rely heavily on high-tech surveillance of wide areas like airports, train stations and office buildings using network of large number of cameras [59]. These cameras are usually connected to a recording device and video is monitored by a human operator.

Cameras and recording equipment used to be relatively expensive to install but infrastructure advances in terms of cameras, network capabilities and storage space provide the ability to capture, store and distribute huge amounts of video data at affordable costs. Traditional surveillance systems are limited in their effectiveness not by technological capabilities but by the attention span and vigilance of the person monitoring the surveillance videos. Human monitoring of large amounts of surveillance video is extremely inefficient. It is not just labor-intensive, expensive and time consuming but also error prone due to lapses of attention. A US National Institute of Justice study found that even a dedicated and well-intentioned person's attention will degrade below acceptable levels within just 20 minutes of monitoring surveillance video [58]. Most of the video collected is archived and stored for forensic or data mining applications leaving the huge potential of camera networks unrealized. Traditional surveillance systems are essentially passive monitoring systems and hence there is a need for continuous active warning capability that can alert security officials during or even before a crime occurs.

Automated video surveillance (AVS) provides an effective means of detecting and preventing incidents, making the system proactive. An extensive study [102] conducted in Europe in 2007 demonstrated that AVS could effectively perform a variety of functions like detecting unusual events to help reduce operator overload. It reduces the manpower burden for monitoring and searching through surveillance videos [51] by automatic detection of events and situations that require the attention of security personnel [125]. The Defense Advanced Research Projects Agency

(DARPA) was among the first to realize the potential of AVS and facilitated the initial research and development of automated surveillance systems through their Visual Surveillance And Monitoring (VSAM) program [1].

AVS can provide the level of information that can change the security paradigm from *investigative* to *preemptive* [61] and has several other advantages over traditional surveillance systems like:

- Ability to prevent incidents by setting off alarms in response to abnormal behavior.
- Enhanced forensic capabilities through content based video retrieval.
- Situation Awareness through knowledge about location, identity and activity of objects in the monitored space.

One of the fundamental enablers to realizing the benefits of AVS is the ability to track people across a scene under surveillance. The tasks of people detection and tracking are extremely difficult to achieve even within the field-of-view (FOV) of a single camera. Although there has been considerable amount of work in the past decade to accomplish these tasks, they remain far from solved [40, 47, 48, 29, 4, 105, 108, 138, 144]. Typically, wide-area surveillance systems employ multiple cameras as it is not possible to observe the complete area of interest with a single camera due to the limited FOV of individual cameras. Thus, tracking needs to be performed across multiple cameras and with the capability to handle non-overlapping camera views. Consistent tracking over multiple cameras requires the ability to re-identify people as the leave the FOV of one camera and enter the FOV of another camera. Re-identification is a fundamental task in wide-area tracking and is challenging to accomplish due to drastic changes in a person's appearance between cameras. Hence, there is a need to develop more robust and reliable algorithms for wide-area tracking to enable automated surveillance.

1.2 Person Re-ID: Problem Definition

Understanding of a surveillance scene through computer vision requires the ability to track people across multiple cameras, perform crowd movement analysis and activity detection. Tracking people across multiple cameras is essential for wide area scene analytics and person Re-ID is a fundamental aspect of multi-camera tracking. Re-identification (Re-ID) is defined as a process of establishing correspondence between images of a person taken from different cameras. It is used to determine whether instances captured by different cameras belong to the same person, in other words, assign a stable ID to different instances of the person. Fig. 1.1 shows a typical surveillance area monitored by multiple cameras with non-overlapping FOVs. The figure shows the top view of a building floor plan and the relative placement of the cameras with respect to the building. Colored dots depict different people and numbers besides the dots are the IDs assigned to the people. The dotted arrows represent the directions in which certain people move through the camera network. As a person moves from one camera's FOV into another camera's FOV, Re-ID is used to establish correspondence between disconnected tracks to accomplish multiple

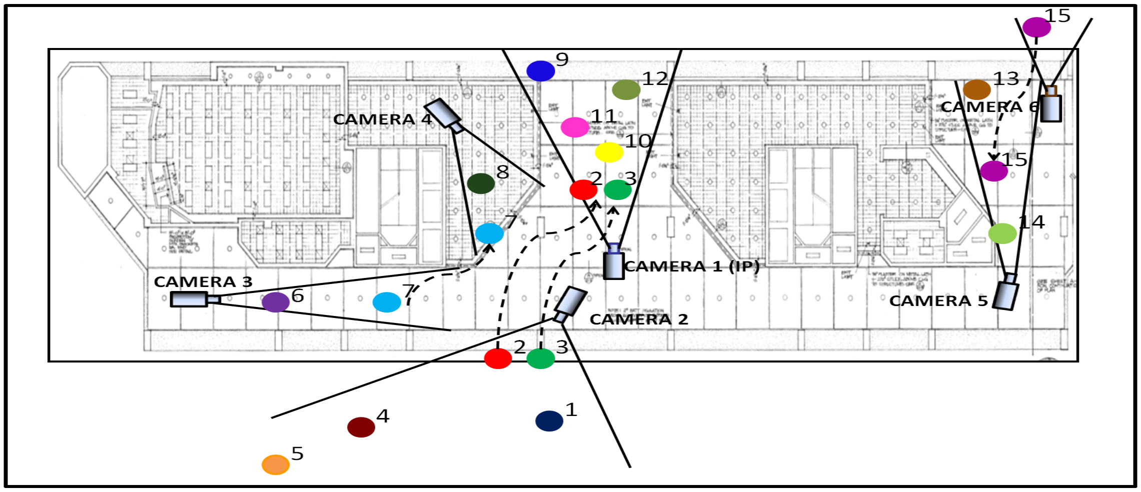


Figure 1.1: Multi-camera surveillance network illustration of Re-ID, colored dots depict different people and numbers besides the dots are the IDs assigned to them.

camera tracking. Thus, single camera tracking along with Re-ID across cameras allows for the reconstruction of the trajectory of a person across the larger scene. Person Re-ID is a non-trivial task, but is critical in improving the semantic coherence of analysis. Re-ID is relevant for surveillance applications with a single camera as well. For example, to determine if a person visits a particular location multiple times or if the same or different person picks up an unattended package/bag. Beyond surveillance it has applications in robotics, multimedia, and more popular utilities like automated photo tagging or photo browsing [124].

Person Re-ID as a task is quite simple to understand. As humans, we do it all the time without much effort. Our eyes and brains are trained to detect, localize, identify and later re-identify objects and people in the real world. Re-ID implies that a person that has been previously seen is identified in their next appearance using a unique descriptor of the person. Humans are able to extract such a descriptor based on the person's face, height and built, clothing, hair color, hair style, walking pattern, etc. A person's face is the most unique and reliable feature that humans use to identify people. Automation of person Re-ID on the other hand is quite difficult to accomplish without human intervention.

1.3 Research Contributions

Re-ID has been a topic of intense research in the past five years [33, 20, 19, 54, 133]. In almost all of the research, the problem of Re-ID has been widely treated as a retrieval or recognition problem. Given an image or multiple images of an unknown person (probe) and a gallery set that consists of a number of known people, the objective is to produce a ranked list of all the people in gallery based on their similarity with the unknown person. The expectation is that the highest ranked match in the gallery will provide an ID for the unknown person, thereby identifying the probe. Here the assumption is that the probe ID is a subset of the gallery of known individuals, i.e., closed-set Re-ID. Current state-of-the-art methods attempt to solve the closed set Re-ID problem.

Most of the current approaches rely on appearance-based similarity between images to establish correspondences. The typical features used to quantify appearance are low-level color and texture extracted from clothing. A review of appearance descriptors for Re-ID is presented in [115]. However, such appearance features are only stable over short time intervals as people dress differently on different days. Thus, appearance based models are only suited for short-period Re-ID. All of the state-of-the-art approaches attempt solutions to short period Re-ID. Earlier research on Re-ID focused on combining inter-camera relationships with the matching process, but more recent efforts have focused on developing discriminative features, learning distance models, or both, for robust matching.

The assumption of short-period Re-ID is unrealistic, hence clothing-based features need to be augmented by discriminative descriptions not based on clothing. In this work, our focus is to develop person models that combine clothing features with non-clothing based features. This not only helps boost the Re-ID accuracy for short-period Re-ID but also gives us insight into features that can be leveraged for long-period Re-ID. Typically, in a multi-camera scenario video data is available so we leverage features that are extracted over multiple frames.

Further, person Re-ID in the context of tracking across multiple cameras is an open-set matching problem where the gallery evolves over time, the probe set dynamically changes for each camera FOV, and all the probes within a set are not necessarily a subset of the gallery. Thus, effective Re-ID requires the ability to detect probe IDs that are not a part of the gallery. This is termed as *novelty detection*. Additionally, there might be several subjects that co-exist in time and need to be re-identified simultaneously. Thus, Re-ID is not a single-person but a multiple-person matching problem.

In addressing the Re-ID problem, the specific contributions of this work are following:

1. Developed non-clothing based features like face and gait that combined with clothing-based features generates a person model that can both boost short-period Re-ID as well as provide performance estimates for long-period Re-ID.
2. Developed a strategy for multiple person Re-ID problem, we achieve ID assignment under one-to-one correspondence constraint.
3. Proposed person-model-similarity-based false acceptance reduction for partially addressing open set Re-ID.
4. Proposed a principled framework for *novelty detection* applicable for open set Re-ID.

1.4 Dissertation outline

The organization of the remainder of this dissertation is as follows. We begin in Chapter 2 by a discussion of the general problem of person Re-ID and its broader challenges. However, person Re-ID can be constrained by the context in which it is applied. We will also explore the context specific nature of the problem. Chapter 3, will provide an overview of the current research work in the field. We adopt a methodology based taxonomy to classify the methods and better understand the current trends. Further, we discuss the evaluation techniques and present the datasets currently used to conduct Re-ID experiments. Chapter 4 describes the proposed color and facial features based model and presents results of the proposed multiple person Re-ID technique. Chapter 5 details the proposed color-and-gait-features-based model and presents results under a closed-set scenario with applicability of gait to long-period Re-ID. Chapter 6 introduces the proposed novelty detection technique designed for open-set Re-ID. Finally, the last chapter highlights the deficiencies of current Re-ID models, and more importantly, points out the un-addressed issues in person Re-ID. It further summarizes the dissertation, and its contributions with the future perspective of this work.

Chapter 2

Person Re-identification: Challenges and Scenarios

2.1 Task and its challenges

In general, person Re-ID is difficult to automate for a number of reasons, which we will discuss later in this section, but the main challenge to Re-ID comes from the variation in a person's appearance across different cameras. Figure 2.1 shows images of a person taken by different cameras on the same and different days, highlighting the variations in appearance. The top row illustrates the changes in appearance of a person across different cameras. It is also interesting to note that the appearance changes significantly within the same camera view as well.

A typical Re-ID system has two basic components: capturing a unique person

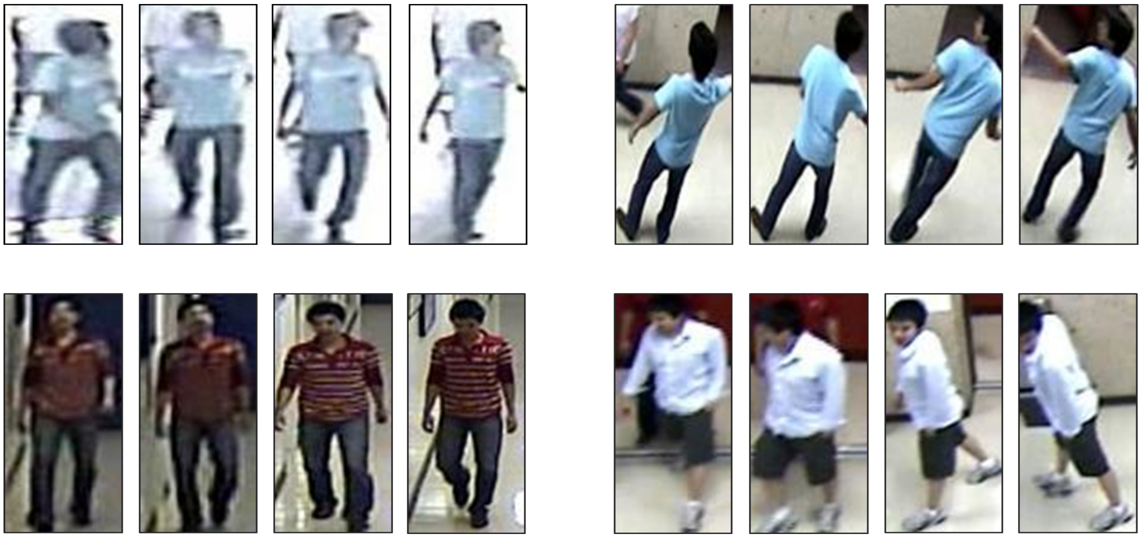


Figure 2.1: Images of the same person taken from different cameras to illustrate the appearance changes. The top row images were captured on the same day, bottom row images were captured on different days.

descriptor or model and then comparing two models to infer either a match or a non-match. In order to learn a unique person descriptor, the ability to automatically detect and track people in images or videos is required. Figure 2.2 shows a schematic representation of a Re-ID system, its two components and the subcomponents within each component. To automate each component, a series of tasks need to be accomplished, which present their own challenges and contribute to the complexity of Re-ID.

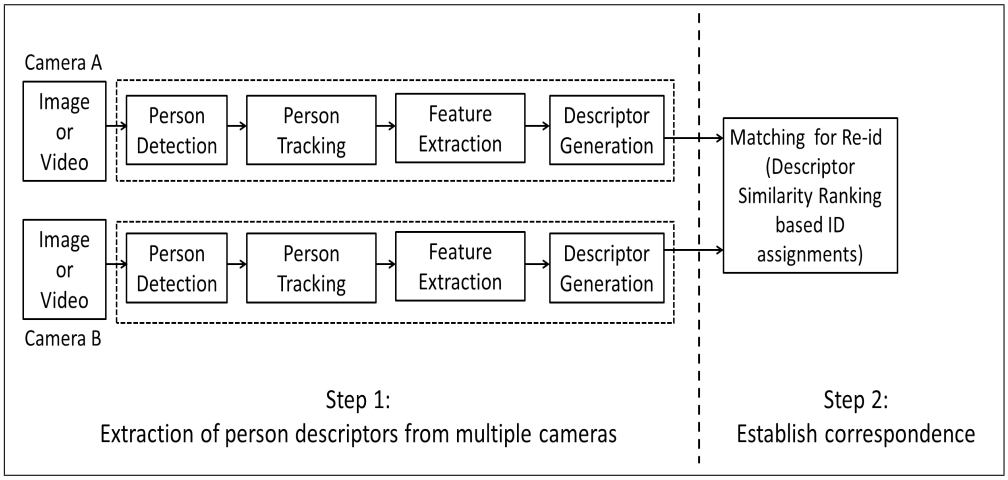


Figure 2.2: Re-ID System Diagram.

2.1.1 System-level Challenges

A typical Re-ID system may have an image (single-shot) or a video (multi-shot) as input for feature extraction and descriptor generation. For an image input the person must be reliably detected and localized for accurate feature extraction. If multiple images are available, in order to ensure that the features extracted belong to the person of interest, we need the ability to establish correspondence between detected subjects across frames. This process is also called tracking and it provides a consistent label to each subject in multiple frames. Thus, multiple instances of a person can be used for feature extraction and subsequent descriptor generation to be used for Re-ID.

Person detection and multiple-person tracking are difficult problems with their own hurdles. Significant amount of work has gone into the problem of person detection over the years [40, 47, 48, 29]. Multiple Object Tracking (MOT) within a single camera's FOV has also been widely researched and many algorithms have been proposed [4, 105, 108, 138, 144] over the past two decades, but sustained tracking under varying observation environments remains an open problem.

2.1.2 Component-level Challenges: Descriptor Issues

Assuming that accurate person detection and single camera tracking is possible, the first step in Re-ID is to learn a person's visual descriptor. Robust and discriminative visual descriptors need to be extracted from data that is captured in unconstrained environments where people may not be co-operative and the environments are uncontrolled. Besides, people can be partially or completely occluded due to crowds or clutter. It is difficult to ensure high quality of visual data as factors like resolution, frame rate, imaging conditions and imaging angles vary widely and cannot always be controlled. Thus, extracting a reliable descriptor is dependent upon availability of good quality observations. Incorrect detections and faulty trajectory estimation introduce errors in the descriptor extraction and generation process that directly affect the quality of Re-ID.

The simplest and most obvious descriptor of a person that can be easily obtained from video data is *appearance*, characterized by features like color and texture. Shape is another descriptor that is extractable. However, these descriptors are hardly unique and prone to variations. Color/texture descriptors are not sufficiently descriptive and vary drastically due to cross view illumination variations, pose variations or view angle or scale changes inherent in a multi-camera setting. Articulated nature of human body leads to deformable shapes of silhouettes and different camera geometries make shape descriptors less discriminative.

Since the person descriptors come from different cameras, the nature of separation between the cameras dictates the difficulty in Re-ID. For example, if the two images are taken only a few minutes or hours apart then appearance based descriptors could prove reasonable to use in Re-ID. The assumption being that people will most probably be in the same clothes, as clothing is a major contributor to appearance. This does not mean that clothing is the best descriptor in this scenario but is a reasonable one. We will refer to this type of Re-ID scenario as ***short-period Re-ID***. Whereas, if the images/video are taken days or months apart, the Re-ID is called ***long-period Re-ID***. In figure 2.1, the images shown in the bottom row are of the same person captured from different camera on different days. This figure perfectly illustrates the fragile nature of appearance based descriptors. The temporal separation between the images is a factor in the complexity of Re-ID. Thus, person Re-ID requires robust, yet unique descriptors, which are extremely difficult to extract automatically.

2.1.3 Component-level Challenges: Correspondence Issues

Comparing person descriptors is challenging due to the uncertainty attributed to the possible lack of prior known spatio-temporal relationships between cameras. Appearance of the same person can change dramatically due to other objects like bags, unzipped jackets across front and back views, etc. At the same time appearance of different people might be rather similar. This implies that within-class variations can be large where as inter-class variations may be relatively smaller. Moreover, even if the person's descriptors can be captured effectively, matching them across cameras in the presence of large number of people observed is non-trivial. Comparing person descriptors across large number of potential candidates is a hard task as the descriptors are captured in different locations, time instants, and over different durations. Complexity of the matching process further increases, as increase in the number of candidates leads to loss of descriptor specificity, increasing the possibility of matching errors. It is also a compute- and memory-intensive process.

2.2 Person Re-ID Scenarios

In the previous section, we presented the general definition of person Re-ID and discussed the implementation pipeline and associated challenges. However, the Re-ID problem can be split into two scenarios: ***open set Re-ID*** and ***closed set Re-ID***. A Re-ID system is similar to a recognition system, which comprises of a gallery set (set of known people) and the probe (unknown person) on which the recognition has to be performed. Figure 2.3 depicts the Re-ID system setup as a recognition system. Let the gallery set be represented as $G = (g_1, g_2, ..., g_N)$. Thus, the set of known IDs are given by $id(G) = (id(g_1), id(g_2), ..., id(g_N))$, where the function $id(.)$ specifies the ID assigned to its argument. Let $P = (p_1, p_2, ..., p_M)$ represent the probe set, which means that the set of unknown IDs is given by $id(P) = (id(p_1), id(p_2), ..., id(p_M))$. Typically in a recognition framework, when the probe is presented to the system,

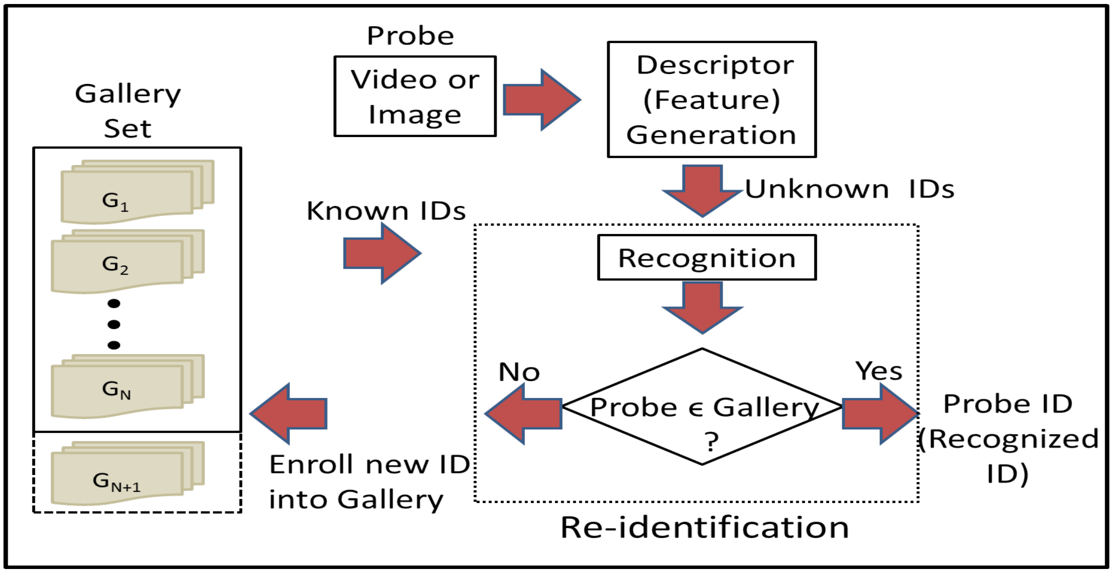


Figure 2.3: Re-ID as a recognition system.

it is compared to each gallery and similarity measures are computed. The gallery is ranked using the similarity in order to determine the probe ID. The same setup applies to the problem of Re-ID. Closed set Re-ID is the scenario where the probe is a subset of the gallery, i.e. the probe ID exists in the gallery and the objective is to determine the true ID of the probe. Thus, given that $id(P) \subseteq id(G)$, the true probe ID for a given probe p_j is $id(p_j) = id(g_{i*})$, such that,

$$i* = \underset{i \in 1,..,N}{\operatorname{argmax}}\; p(g_i|p_j) \tag{2.1}$$

where, $p(g_i|p_j)$ is the likelihood that $id(p_j) = id(g_i)$ and is most often represented by a similarity measure. This implies that the top-ranked gallery ID is assigned to the probe. In open-set Re-ID on the other hand, the probe may or may not be a subset of the gallery. This implies the open-set Re-ID objective is to first establish if the probe ID is a part of the gallery, and if so, determine the true probe ID. Thus, in order to find the true ID, in addition to ranking the gallery elements and determining $i*$ using equation 2.1, the following condition also needs to be satisfied,

$$p(g_{i*}|p_j) > \tau. \tag{2.2}$$

In equation 2.2, τ is the acceptable level of certainty above which we can be reasonably assured that $id(p_j) \subseteq id(G)$. If this condition is not satisfied, then it is determined that the probe is not a part of the gallery. If so, the probe ID is then to be enrolled into the gallery. The process of determining a previously unknown ID is called *novelty detection*. Similar to identification tasks, the closed set Re-ID is a constrained form of open set Re-ID. The Re-ID application dictates the matching scenario. For instance, to achieve consistent tracking over multiple cameras for global trajectory of a person over a camera network requires open set Re-ID. On the other hand, identity based retrieval (for forensic applications), i.e., the ability to identify multiple observations of a particular person is a closed set Re-ID problem.

2.2.1 Open-set Re-ID

Person Re-ID in the context of tracking across multiple cameras is an open set matching problem where the gallery evolves over time, the probe set dynamically changes for each camera FOV, and all the probes within a set are not necessarily a subset of the gallery. Additionally, there might be several subjects that co-exist in time and need to be re-identified simultaneously. Thus, it is not a single person but a multiple person Re-ID problem. We will explore the open set Re-ID problem by

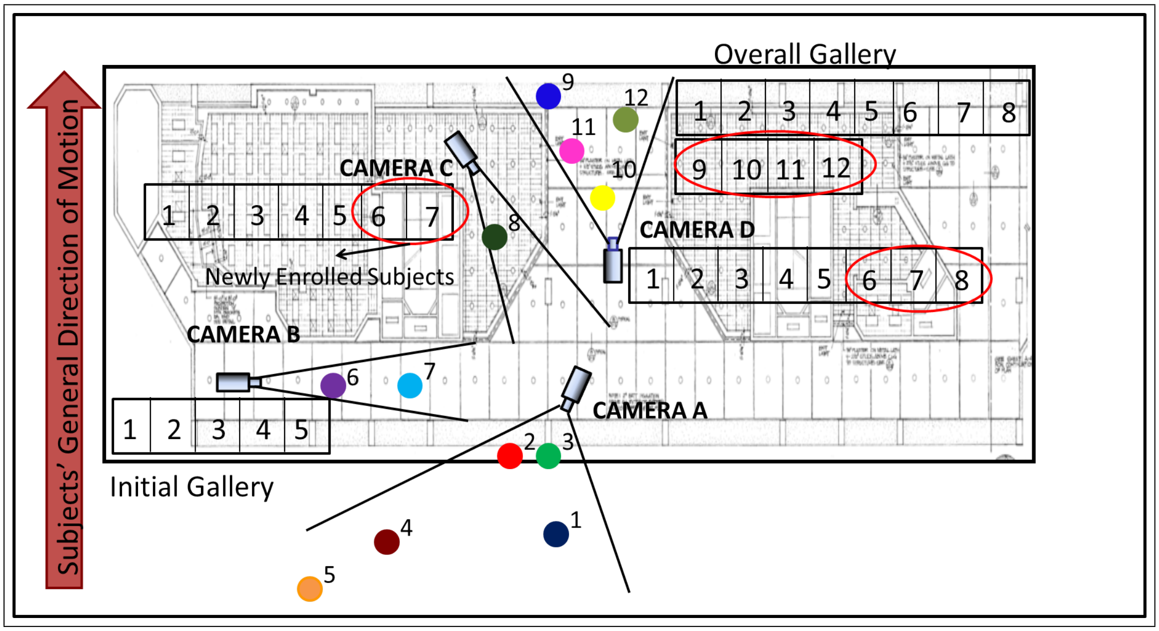


Figure 2.4: Multi-camera tracking scenario based on open set Re-ID, colored dots depict different people and numbers besides the dots are the IDs assigned to them.

illustrating Re-ID in multiple camera tracking. Figure 2.4 shows a schematic of a camera network (with 4 cameras) and the evolution of the gallery, where the Re-ID is done across each camera pair. For ease of illustration, let us assume that all subjects in the figure are moving in the direction depicted by the red arrow and the tracking across the network starts at $t = 0$ from camera A. In other words, there is no prior gallery set and tracking (Re-ID) progresses from camera A through D. Additionally, the subjects appear in the FOV of camera C before they appear in the FOV of camera D. The first time a person is seen in camera A, his/her appearance model is learned, and the subject is enrolled in the gallery set. Thus, all people observed in the camera B form the probe set. After Re-ID, all the people observed in the camera B who were previously unseen are enrolled into the gallery. As the Re-ID moves to the next camera pair (camera B and C), the gallery set is extended. The open set Re-ID can be summarized as a many-to-many matching problem. In tracking scenario, Re-ID provides a means of connecting subjects' tracks that were disconnected due to the subject entering an area not in the FOV of the camera network.

2.2.2 Closed-set Re-ID

Person Re-ID in the context of identity retrieval is closer to the classic closed set matching problem, where a single probe is presented and the gallery size is fixed. In a typical multi-camera identity retrieval scenario, the person whose multiple observations throughout the network are to be detected is the probe subject and his/her appearance model is assumed to be available. The gallery set is a set of people IDs seen in selected or all the cameras over a specified period of time. The time interval specified can be different for different cameras. In other words, the gallery is comprised of subjects seen in many different cameras constrained by time and space. Additionally, the probe can simultaneously match to gallery subjects coming from different cameras, i.e., multiple instances of the probe can be detected within the gallery. After Re-ID, multiple observations of the probe subject across the gallery cameras are detected. As the probe subject to be re-identified changes, the cameras and time intervals to be searched change and so does the gallery. However, for Re-ID of a particular probe the gallery remains fixed. Thus, the closed set Re-ID is a one-to-many matching problem.

Chapter 3

Related Work

In general, recent approaches have focused on two aspects of the solution: 1) design of discriminative, descriptive and robust visual descriptors to characterize a person's appearance; and 2) learning suitable distance metrics that maximize the chance of a correct correspondence. Overall the methods for Re-ID can be broadly classified into Contextual and Non-contextual methods. Figure 3.1 provides a methodology based taxonomy that summarizes the state-of-the-art research in person Re-ID.

3.1 Contextual Methods

These methods rely on external contextual information either for pruning correspondences or extracting features for Re-ID. They can be further classified as those that utilize camera geometry information or those that incorporate camera calibration as the context.

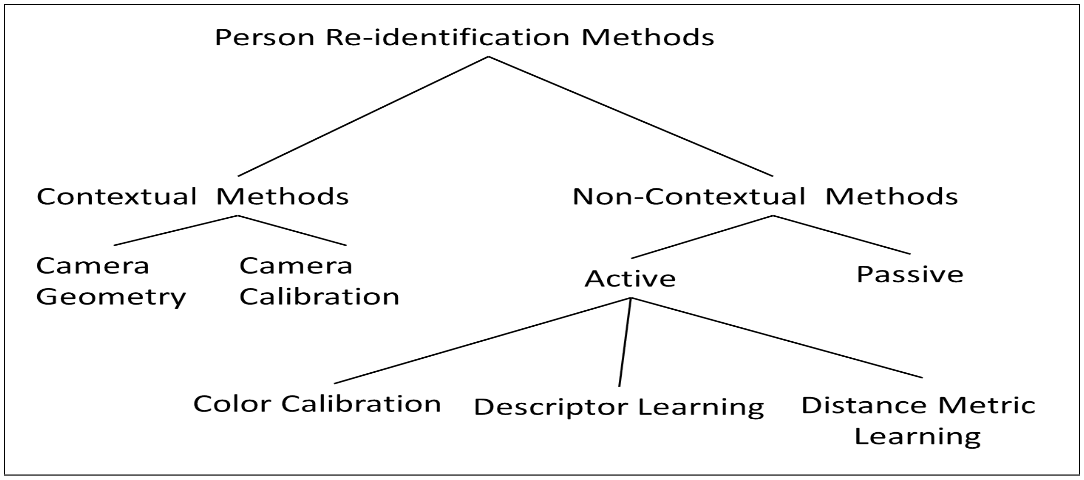


Figure 3.1: Method-based taxonomy of Re-ID approaches.

3.1.1 Camera Geometry as Context

The early work in person Re-ID focused on leveraging spatial and temporal relationships between cameras to reduce the Re-ID errors by limiting the size of the gallery set. Space-time cues are exploited in [71] to learn inter-camera relationships that are in turn used to constrain correspondences across cameras. These relationships are modeled as a probability density function of space-time parameters like entry and exit locations, velocities, and transition times between cameras. Entry-exit points of each camera and transition times between cameras are learned in [96], in order to calibrate all the cameras in the network. The calibrated cameras are used to learn the topology of the camera network as a bipartite graph. The topology is further augmented with temporal information to achieve a tempo-topographical model of the camera network. A similar approach is used to calibrate the camera network and estimate trajectory of targets in the network using MAP estimation in [113].

Propagation of people trajectories are used in [99] to identify areas of interest in the unobserved regions within cameras. These areas are further used to choose potential paths people might take, limiting the reappearance areas in the subsequent camera's FOV to constrain Re-ID.

The topology of cameras is determined by correlating activities across cameras with disjoint FOVs in [90, 91] and hence do not rely on tracking information. The FOVs of the cameras are segmented into regions within which activity patterns are similar. Spatial and temporally causal relationships across these regions in different cameras are modeled using canonical correlation analysis [68]. Affinity matrices are used to infer camera spatio-temporal camera topologies to aid Re-ID. A similar idea is built on in [92]. Here the relationships between activities are learned using MAP estimate that is continually updated at each time instant. A comprehensive review of camera topology estimation methods is presented in [132] and a study of scalability of topology estimation is performed in [42].

3.1.2 Camera Calibration as Context

In these methods, camera calibration or homography is exploited to extract unique and discriminative features to augment the visual descriptors used for Re-ID. The height of a person is determined using homography based 3D position estimation in [81]. The human silhouette is divided into three parts from top to bottom at specified proportions to the blob height. Each region is then represented with dominant color and edge energy texture descriptors. Integrated region matching is used for human silhouette similarity computations to achieve Re-ID. Similar height extraction method is used in [107]. The height along with clothing color and body build are used as features to establish a match.

A panoramic appearance map (PAM) proposed in [53] extracts and combines information from all the cameras that view the object to generate a single object signature. Multiple camera triangulation is used to determine the position of the object and a cylindrical surface model is placed at its location. A parametric surface grid is projected onto all cameras where the object is visible and corresponding image patches are extracted. The features or pixel colors from these extracted patches are integrated to form the PAM, which is used for Re-ID. Maps from different tracks are compared using weighted sum of a squared distance metric. However, to generate the appearance signature, the object needs to be visible in at least 3 cameras with overlapping views simultaneously. Camera calibration and accurate 3D positioning is needed to accomplish Re-ID. The principal axis of each person, i.e. the axis of symmetry of the human body is detected in [70] to match people across camera views. Landmarks on the ground plane shared by two cameras are used to estimate homography. The intersection of the principal axis of a person in one view and transformed principal axis of a person in another view using homography is used to compute a degree of match between people from different cameras. The degree of match is used to compute correspondence likelihood to establish Re-ID. However, the accuracy of the detected principal axis depends on the accurate segmentation of the human silhouette from the foreground and hence is prone to errors in crowded scenes and cluttered backgrounds.

A 3D point process model is used for detection and representation for person matching in [14]. The placement and orientation of the 3D model is determined using camera calibration and tracking information. Each model vertex is represented by a number of appearance features, namely, HSV histogram, mean color, direction of normal to vertex, optical reliability of vertex and vertex uniqueness. Re-ID score is a product of distances between HSV histograms weighted by vertex reliabilities and vertex saliency distances. As is evident, the main drawback of these methods is their reliance on camera calibration. With large camera networks, calibration of all cameras is not feasible.

3.2 Non-Contextual Methods

Several approaches have been proposed that rely entirely on the analysis of visual descriptors and no external contextual information is incorporated to assist the correspondence process. These methods can be further classified as active and passive methods. Most of the recent research is focused on non-contextual methods. A popular classification within this class is based on whether single image (*single-shot*) or multiple images (*multi-shot*) are used to generate and compare the appearance descriptors. There are many different non-contextual techniques for Re-ID and in order to provide an overview of some of the more prominent approaches, a tabular summary is presented in table 3.1. The approaches are distinguished based on the type of features used, single/multi-shot incorporation and the incorporation of false

Approach Type	Approaches	Structural Information	Images used for Descriptor	Features	False match Rejection
Passive	Spatiotemporal Model [54]	✓	Multiple	Color,Edges	×
	SDALF [19]	×	Single/Multiple	Color,Texture	×
	SCR [10]	✓	Single	Position,Color,Gradients	×
	Multi-feature Model [22]	✓	Multiple	Color,Face	✓
	BiCov [94]	✓	Multiple	Color,Texture	×
	CPS [36]	✓	Single/Multiple	Color	×
Descriptor Learning	ELF [57]	✓	Single	Position,Color,Gradients	×
	PLS [122]	×	Single	Color,Texture,HOG	×
	Shape Context [133]	✓	Single	Shape,Color,Texture,HOG	×
	Group Context [149]	×	Single	Color,Texture,HOG	×
	Boosted Re-ID [11]	✓	Multiple	Position,Color,Gradient	×
	Re-ID with Attributes [83]	✓	Single	Color,Texture	×
	Correlation Space Re-ID [8]	✓	Multiple	Position,Color,Gradient	×
	Re-ID by Saliency [147]	✓	Single	Position,Color,Texture	×
Metric Learning	LMNN-R [43]	×	Single	Color	✓
	PRDC [151]	×	Single	Color,Texture	×
	RankSVM [112]	×	Single	Color,Texture	×
	Impostor Learning [66]	✓	Single	Color,Texture	✓
	Fisher Vector [95]	✓	Multiple	Position,Color,Texture	×

Table 3.1: Summary of recent non-contextual person Re-ID approaches, ✓ and × indicate whether the method incorporates the indicated function or not, respectively.

match rejection (novelty detection) in the matching framework.

3.2.1 Passive Methods

These methods deal with design of descriptive visual descriptors to characterize the person's appearance and compare these by computing similarity measures to achieve Re-ID. These methods are termed as passive as they do not reply on learning techniques, supervised or unsupervised, for descriptor extraction and matching.

A color and shape features-based appearance model from the detected blob is proposed in [77]. The blob is segmented into multiple polar bins and the color Gaussian model and edge pixels counts from each bin form the descriptor. A match is established using three similarity measures and the optimal match is the one that maximises all the similarity measures. A spatiotemporal segmentation algorithm based on watershed segmentation and graph partitioning is used in [54] to detect stable spatiotemporal edges called edgels. The appearance of a person is a combination of color (hue and saturation) and edgel histograms and the intersection histogram is used to establish a match between observations. A non-surveillance application of person Re-ID was explored in [124], where the objective was to find all occurrences of person in a sequence of photographs taken over a short period of time. A two step approach is adopted, where the first step, identifies different people that exists in the photographs by clustering frontal face detections. The clustering is based on 16-bin RGB histograms extracted from clothing. In the second step, color features based pictorial structures are used to find each person identified in the previous step, even in photographs where their frontal faces cannot be seen. Each part identified by the pictorial structure is represented by 5 component Gaussian mixture model. This approach assumes that each person is facing the camera in at least one photograph in the sequence and that people are distinguishable by their clothing color.

The human silhouette is represented by two complementary appearance features in [20]. The first feature is an HSV histogram that encodes the global appearance while the local appearance is encoded using a set of recurrent local patches using epitomic analysis. The appearance matching is based on a weighted sum of feature similarities. The features are extracted over multiple images of a person and is termed as Histogram Plus Epitome (HPE). The human silhouette is divided into head, torso and legs regions by detecting 2 horizontal axes of asymmetry and one vertical axis of symmetry in [19]. Each part is then described using 3 features, weighted HSV histogram, maximally stable color regions (MSCR) [50] and recurrent highly textured local patches. Again, the appearance matching is based on a weighted sum of feature similarities. The features extracted are combined over multiple images of a person to form an appearance model called SDALF. The Re-ID performance of the features proposed in [20] and augmented by applying it as a human part descriptor adopting the asymmetry driven part detection proposed in [19], thus defining a structure feature named Asymmetry-based HPE [21]. Pictorial structures model [5] was employed in [36] for part based human detection and each part is used to extract HSV histograms and MSCRs. The part-based representation is then used for Re-ID. They proposed a slight modification of pictorial structures to better localize body parts using multiple images of a person to guide the MAP estimates of the body configuration. This approach is known as Custom Pictorial Structures (CPS). This aids the extraction of more reliable visual features to improve Re-ID.

Spatial covariance regions (SCR) are extracted from human body parts in [10] and spatial pyramid matching is used to design a dissimilarity measure. HOG based body part detector is used to detect 4 parts: torso, left arm, right arm and legs. Each detected body part is characterized by a covariance descriptors based on region colors, gradient magnitudes and orientations. These descriptors encode variances in region features, their covariances, and spatial layout. Covariance matrix distance is used to compute dissimilarity between descriptors. Covariance matrices are also adopted in [94] but the underlying features are Gabor filter response magnitude images extracted from different spatial scales (BiCov). Neighboring scale responses are grouped to form a single band and magnitude images are computed using the MAX operator within each band. The appearance model is not the covariance matrices but differences between matrices between consecutive bands.

A multiple component matching approach inspired by multiple component learning concept in object recognition is proposed in [118]. Multiple frames of a person are treated as multiple instances of the person. Thus, the descriptor is basically a collection of features extracted from multiple frames. Each image feature set is treated as an instance and gallery and probe are considered a match if at least a few pairs of instances match. The body is represented by randomly selected rectangular patches whose appearance is captured by HSV histograms. This framework is extended in [116] where person descriptors are formed by a vector of dissimilarity values to a set of predefined visual prototypes.

Interest-point-based descriptors collected over a number of images of a person are utilized in [60] to characterize the person's appearance. Hessian based interest points are detected using efficient integral image implementation. A histogram of Haar wavelet responses in a 4x4 region centered around the interest points are used as the descriptors. The descriptors are matched using sum of absolute differences metric and Re-ID is established using a best bin first (BBF) search on a KD-tree containing all gallery models. The Re-ID model is generated from tracking data in [75]. The model is generated by encoding SIFT features extracted during tracking by Implicit Shape Model [84] codebook learned offline. The spatial distance between the SIFT points also contributes towards the model. Matching high dimensional models is computationally very expensive and the major drawback of this approach. A comparative study of local features for the task of Re-ID was reported in [18] and concluded that GLOH [100] and SIFT features outperform other local features.

FisherFaces [23] based facial features and dense sampling of colors in luminance-chrominance space (LCC) along with horizontal and vertical edges from clothing are combined for Re-ID in [52]. Person recognition is based on a nearest neighbor classifier. Color position histogram is constructed in [37] by splitting the silhouette into fixed number of horizontal bands and characterizing each band with its mean color. Sparsified representation [141] is utilized for Re-ID. A person is represented as a graph in [32] with color features representing the nodes and region proximity dictating the graph edges. Graph edit distances based graph kernel is used for classification and hence Re-ID. Covariance descriptors based on color, Gabor and LBP features is used to characterize appearance in [145]. Fuzzy color quantization in the Lab color space is used to extract probabilistic histograms in [41]. Re-ID is based on a k-nearest neighbor classifier. Color-position histogram is used to characterize silhouettes in [38] and are subsequently subjected to non-linear dimensionality reduction to form the descriptor vector.

3.3 Active Methods

These methods are termed as active as they employ supervised or unsupervised learning techniques for descriptor extraction or matching. Such learning based methods can be further classified into color calibration methods, descriptor learning and distance metric learning methods. The last two sub-categories depend on whether learning is employed to learn optimal appearance features or to learn optimal distance metrics for Re-ID.

3.3.1 Color Calibration

These methods attempt to model the color relationships between a given camera pair using color calibration techniques and they need a learning stage to develop the calibration model that needs to be updated frequently to capture all desired relationships. In order to model the changes in appearance of objects between two cameras a brightness transfer function (BTF) between each pair of cameras is learned from training data in [72]. The BTF is used as a cue in establishing appearance correspondence. Learning the brightness transfer function between a pair of cameras was first proposed in [109]. Once such a mapping between cameras is learned, the Re-ID problem is reduced to one of matching transformed appearance models. However, such a mapping is not unique and it changes from frame to frame depending on varying factors like illumination, scene geometry, focal length, exposure time and aperture size of each camera. Thus, a single BTF cannot be used for matching models consistently. Javed *et al.* [71] show that all the BTFs between a given camera pair lie in a low dimensional subspace even in the presence of large number of unknown parameters. They propose a method to learn the low dimensional subspace from training data and use this information to determine the probability that observations taken from two different cameras belong to the same person. Prosser *et al.* [111]

propose a cumulative BTF computation method that requires only a spare color training set and the BTF is computed by relying on the mean operation taken over multiple learned BTFs. A novel bi-directional matching criterion is also proposed for comparing individuals in order to reduce false matches.

3.3.2 Descriptor Learning

This class of methods either attempt to learn the most discriminative features or a discriminative weighting scheme for multiple features to achieve Re-ID or employ a learning stage to generate descriptive dictionaries of features that better represent a person's appearance using a bag-of-features approach.

Shape and appearance context models were used in [133] where co-occurrences between *a priori* learned shape and appearance words form the person descriptor. The human silhouette is split into parts using a modified shape context algorithm that builds on shape dictionary learned *a priori*. The bag-of-features based approach is used to learn code words to characterize appearance based on HOG [40] features computed in the log-RGB space. The human silhouette is first represented as a collection of shape labels and then the appearance descriptor is constructed using the spatial occurrence of the appearance words with respect to each shape label. Since the model is based on appearance words, learned on a training dataset, the applicability of the model is limited. Similar appearance words and visual context using local and global spatial relationships (group context) are used to describe an individual's appearance in [149]. The appearance words are based on SIFT [89] and average RGB color as features. Groups of people are represented by two descriptors. The first one aims to describe the ratio information of appearance words within and across rectangular ring regions centered on the group. The second descriptor captures more local spatial information between the labels. The obtained group descriptors are utilized as a contextual cue for person Re-ID by combining person descriptor matching cost and group matching cost. Group information is used to reduce ambiguity in person Re-ID if a person would appear in the same group. Cai and Pietikinen [33] utilize spatial distributions of self similarities with respect to learned appearance words and combine them to form the appearance descriptor. The appearance words are based on a weighted hue histogram and then for each label, its occurrence frequency in each bin of a log-polar grid centered on the image center is computed. It is used as a global color context representation to model self similarities of image patterns. Re-ID is established between person images using a nearest neighbor classifier.

Adaboost learning is employed in [57] to simultaneously learn discriminant features and ensemble of weak classifiers (ELF) for pedestrian recognition. Weak classifiers are learnt on a training set to determine the features that impart maximum discriminative ability. The color features used are histograms of RGB, HSV and YCbCr channels and texture features are histograms of Schmid and Gabor filter responses. The most discriminative characteristics of these features, such as location of features, most discriminative bin, and likelihood ratios determined by boosting, are used. The Adaboost classifier assigns a positive label to pair of images from the same person and a negative label to pair of images from different people. This study concluded that the Hue, Saturation, R, G and B channels are most discriminative in that order. Partial least squares (PLS) technique is employed to not only find discriminative weighting for color, texture and edge features but also as a means to reduce descriptor dimensionality in [122]. The observation that appearance variations across multiple cameras are multi-modal in nature is utilized in [85] by learning multiple classifiers in the joint appearance feature space across multiple cameras. Re-ID is achieved by ranking combined scores generated by all the learned classifiers.

A two step process is applied to learn discriminative features in [65]. In the first step, covariance descriptors are used to rank the gallery images as per similarity to a probe image. The first 50 images are shown to a human operator who decides whether the true match exits in this set for the probe. If not, as a second step, boosting is performed over a set of covariance descriptors based on RGB color and Haar features to select a fixed number of discriminative features that are then used to establish a match.

Haar-like features and dominant color descriptors are used as features for Re-ID and to guide detection of upper and lower body of a person in [9]. Adaboost classifier is used to find the most appropriate appearance model to use for matching images of people. An extension of the covariance descriptors used in [10] is proposed in [11], where instead of using predefined body parts to extract low level features, a spatio-temporal grid over multiple images is used to extract the features. Each region of an image is used to extract the covariance matrices and these are combined over multiple frames using Riemannian mean of the covariances (MRC). The spatio-temporal MRCs that contribute to the final descriptor are selected by a boosting algorithm and a matching scheme based on Riemannian manifolds is used for Re-ID.

A binary SVM classifier is trained in [7] to learn camera-pair specific variations in the feature space. The features used to train the SVM are formed by concatenating the appearance descriptors of people across a given camera pair. If both the descriptors belong to the same person, then these are considered positive samples otherwise they are treated as negative samples. In other words, it solves the camera-specific Re-ID problem. HSV histograms extracted from 5 horizontal regions of the silhouette is used as the appearance feature. This same idea is extended in [31], given classifiers trained on camera pairs A-B and B-C, they attempt to infer the classifier for camera pair A-C. The inference is based on the notion of statistical marginalization which is approximated by summation over descriptors coming from camera B.

A view that different regions of the subject should be matched using different matching strategies and features is explored in [8]. The location of different regions of the body are represented by their distance from the body center along with color and texture features. Covariance matrices are used as a feature and human body specific matching criteria are learned using correlation based feature selection. The distance model for matching exists in the covariance feature space.

3.3.2.1 Attribute based person re-identification

The idea that certain features can be more important than others is explored in [86]. In the context of Re-ID, they attempt to determine features that are unique that distinguish a given subject from another even if their overall appearances are very similar. To determine such features they link low level features to attributes. Attributes are defined a *midlevel* features or visual concepts that have semantics attached to them, namely: stripped, furry, tall, short and so on [106]. An unsupervised approach for learning adaptive weights of different features based on their unique and inherent appearance attributes is proposed. First, a clustering stage is applied to a set of training images to discover representative prototypes of attributes. The assumption being that each prototype represents certain attributes specific to that subject. The feature's weights are then computed based on its ability to discriminate between different prototypes. An incoming probe image is then assigned to one of the prototypes to generate an attribute driven representation. A combination of all appearance features and attribute weighted features is used to rank gallery images and establish Re-ID. The underlying features are the color and texture features proposed in [57]. A similar idea is explored in [147, 146]. Here distinctive or salient regions are defined as regions that discriminate an individuals appearance and is general enough to identify the person across different views. For instance, these regions could be a distinctive textured backpack or bright jacket. Person images are represented by appearance characterized patches and patch matching under spatial adjacency constraints is used to generate an appearance descriptor in an unsupervised fashion. Re-ID is achieved by combining the salient patches with global appearance (SDALF features). We term this approach as Re-ID by saliency.

Attributes have been successfully applied over the past few years to various problems like face recognition [79]and object recognition [46, 128, 134]. In order to augment the insufficient discriminative ability of low level features, the concept of an attribute for Re-ID is refined in [82, 83]. For instance, a human observer can distinguish between two people wearing very similar clothing based on distinct shoes or hair styles. Thus, attributes represent features that can be interpreted distinctly based on their perceptual semantics. 15 binary attributes: type of clothing (skirts, jeans, etc), shoes, hair, gender and accessories are defined based on human operators and their detection using SVM based on color and texture features [57] has also been proposed. A combination of these attributes along with global appearance descriptors (SDALF features) defined in [19] is used to compute weighted similarity between gallery and probe images to establish Re-ID. We term this approach as Re-ID with attributes.

3.3.3 Distance Metric Learning

These methods shift the focus from feature selection based efforts to improve Re-ID to learning appropriate distance metrics that can maximize the matching accuracy regardless of the choice of appearance representation. Distance-metric learning methods [143] are extensively explored in the recognition and image retrieval problems, and they attempt to learn a metric in the space defined by image features that keep features coming from same class closer, while, the features from different classes are farther apart. In the context of Re-ID, the image features are appearance descriptors across camera views and the aim is to learn a distance metric in the appearance space that maximizes the distance between descriptors of different people and minimizes the distance for descriptors of the same person. Metric learning is done in a supervised fashion under pairwise constraints. The training features are paired appearance descriptors and the training labels are either positive or negative depending on whether the appearance descriptors belong to the same person or different, respectively.

Let the training appearance descriptor pairs be denoted by $x_1, x_1, ..., x_n$, where, n denotes the number of training samples. Let the dimensionality of each sample be denoted by m. Metric learning aims to learn a distance metric, denoted by matrix $D \epsilon R^{mxm}$, such that, distance between two appearance pairs x_i and x_j is defined as:

$$d(x_i, x_j) = (x_i - x_j)^T D(x_i - x_j) \tag{3.1}$$

$d(x_i, x_j)$ is a true metric as long as matrix D is symmetric positive-semidefinite. This problem is solved using convex programming as shown below:

$$\min_D \sum_{(x_i,x_j)\epsilon Pos} \|(x_i - x_j)\|_D^2 s.t.D \succnsim 0, and \sum_{(x_i,x_j)\epsilon Neg} \|(x_i - x_j)\|_D^2 \geqslant 1 \tag{3.2}$$

where, Pos and Neg denote positive and negative label training sample sets denoting appearance pairs the belong to the same person and different ones, respectively.

A large margin nearest neighbor (LMNN-R) distance metric is learnt in [43] such that it minimizes the distance between true matches and maximizes false match distances. The cost was computed using 8-bin RGB and HSV histograms after subjecting them to principal component analysis for dimensionality reduction. The metric learned has the capacity to reject matches based on a universal learnt threshold on the matching cost. It was shown in [131] that by using a slight modification in the feature vector extraction using overlapping regions of the human blob, the LMNN-R metric can provide greater robustness to Re-ID under occlusion and scale changes. A similar idea is explored in [151] that solves metric learning in a probabilistic manner termed as probabilistic relative distance learning (PRDC). They focus on maximizing the probability that a true match pair has a smaller distance than a false matched pair. Re-ID is cast as a tracklet matching problem in [123] and dynamic time warping distance is used as a metric to match tracklets. Dynamic time warping distance based large margin nearest neighbor metric learning is adopted.

The person Re-ID problem is treated as a relative-ranking problem in [112], the idea being not comparing direct distance scores between correct and incorrect matches, instead to learn a relative ranking of these scores that captures the relevance of each likely match to the probe image. A set of weak SVM based rankers are learned using color and texture features [57] on small training datasets and combined to build a stronger ranker using ensemble learning. In other words, they attempt to learn a subspace where true matches are ranked highest. This method is called RankSVM [74]. A Re-ID by verification approach based on transfer learning is proposed in [150], i.e., the learning process aims at extraction of transferable discriminative information using a set of non-target people (unknown IDs). In the verification scenario, a probe's identity is verified against a small set of target people (known IDs). Re-ID is performed by learning not only the separation between target and non-target IDs, but also the separation between different targets IDs.

The distance model is learned using the PRDC and the RankSVM frameworks. An iterative refinement of the ranking is proposed in [93] where gallery is modeled by a graph in the visual appearance space. The graph weights and structure are modified to accommodate the probe image and a ranking function that optimizes the graph Laplacian is computed and used for gallery ranking.

A set based discriminative ranking (SBDR) model is adopted in [140], where distance between a sequence of images of a gallery person and probe is computed using geometric distance of their approximated convex hulls. A maximum margin based ranking scheme is employed that makes distance between a true-match pair smaller than the false-matching pair. The metric-learning process is an iterative one and hence compute intensive. The color and texture features defined in [57] are used as the underlying features. Image intensity, color, position and gradient based 7-d features are extracted and represented by Gaussian mixture models in [95]. They are combined with weighted HSV histograms and stable color regions. The sparse pairwise constraints based distance metric learning suited for high dimensional data is used for Re-ID.

Mahanolobis metric distance learning is adopted in [67] by posing correspondence detection as a two class classification problem. The learning occurs in the distance space with only two labels for each point. In other words, distances between same person's different views have the same label. By relaxing the positivity constraint on the learned matrix the above optimization problem is simplified without the need for multiple iterations. The person image is split into overlapping rectangular regions and HSV, Lab colors, and LBP [104] are extracted from each region to form the person descriptor used to learn the metric. A similar pairwise metric is learnt in [66] using the large margin nearest neighbor framework. During the learning process, the samples in the training data that are difficult to separate from the matching samples are given more priority. These samples are called impostors as they invade the perimeter of a matching pair in the distance space (Impostor learning). The features used are the same as in [67].

3.4 Public Datasets and Evaluation Metrics

The visual characteristics of a person vary drastically across cameras, introducing variability in illumination, poses, view angles, scales and camera resolutions. Factors like occlusions, cluttered background and articulated bodies also contribute. Thus, in order to develop robust Re-ID techniques it is important to acquire data that captures these factors effectively. Along with high quality data emulating real world conditions, there is also a need to compare and contrast Re-ID approaches being developed and identify improvements to techniques and the datasets. There are several public datasets that have been used to test Re-ID models. ViPER [56], i-LIDS for Re-ID [149] and ETHZ [44] are currently, most commonly used for Re-ID evaluations. Table 3.2 provides a summary of the widely used Re-ID datasets.

Dataset	Multiple Images	Multiple Camera	Illumination Variations	Pose Variations	Occlusions	Scale Variations
ViPER		✓	✓	✓	✓	
ETHZ	✓		✓		✓	✓
i-LIDS	✓	✓	✓	✓	✓	
CAVIAR4REID	✓	✓	✓	✓	✓	✓
i-LIDS MA and AA	✓	✓	✓	✓	✓	✓
V-47	✓	✓	✓	✓		
GRID		✓	✓	✓	✓	✓

Table 3.2: Summary of public person Re-ID datasets.

3.4.1 ViPER

The ViPER dataset [56] consists of images of people from two different camera views, but it has only one image of each person per camera. The dataset was collected in order to test viewpoint invariant pedestrian recognition and hence the aim was to capture as many same viewpoint pairs as possible. The view angles were roughly quantized into 45 degree angles and hence there are 8 same viewpoint angle pairs or 28 different viewpoint angle pairs. The dataset contains 632 pedestrians image pairs taken by two different cameras. The cameras have different viewpoints and illumination variations exist between them. The images are cropped and scaled to be 128x48 pixels. This is one of the most challenging datasets yet for automated person Re-ID. Figure 3.2 shows some example images from this dataset.

3.4.2 ETHZ

ETHZ dataset consists of images of people taken by a moving camera [44] and this camera setup provides a range of variations in person appearances. The images of

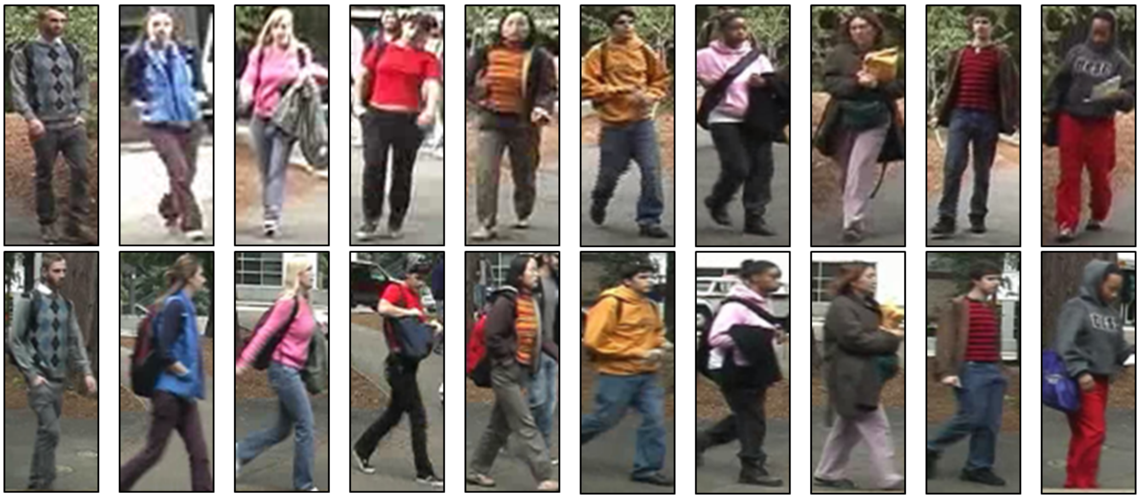


Figure 3.2: Example images from the VIPeR dataset.

pedestrians do not come from different cameras but multiple images of the person taken from a moving camera are present. The dataset has three sequences and multiples images of a person from each sequence are provided. Sequence 1, 2 and 3 have 83, 35, and 28 pedestrians respectively. The dataset consists of considerable illumination changes, scale variations and occlusions. The images are of different sizes. Figure 3.3 shows some example images from this dataset.

3.4.3 i-LIDS for Re-ID

This dataset was extracted from the i-LIDS multi- camera tracking scenario [103] or i-LIDS MCTS dataset which is widely used for tracking evaluation purposes and was acquired in crowded public spaces. The dataset contains a total of 476 images of 119 pedestrians taken from two non-overlapping cameras. On an average there

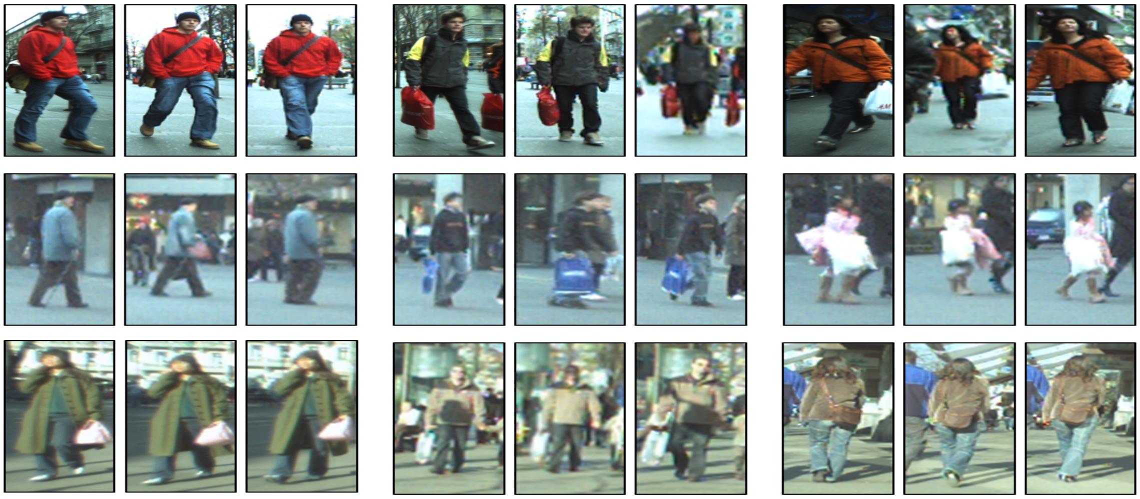


Figure 3.3: Example images from the ETHZ for Re-ID dataset, the first, second and third rows of images come from sequences 1, 2 and 3 respectively.

are 4 images of each pedestrian and a minimum of 2 images. The dataset has considerable illumination variations and occlusions across the two cameras. All images are normalized to 128x64 pixels. Figure 3.4 shows some example images from this dataset.

3.4.4 CAVIAR4REID

This is extracted from another multi-camera tracking dataset [34] captured at an indoor shopping mall with two cameras with overlapping views. The dataset [36] contains multiple images of 72 pedestrians, out of which only 50 appear in both cameras, where as 22 come from the same camera. The images for each pedestrian were selected with the aim of maximizing appearance variations due to resolution changes, light conditions, occlusions, and pose changes. The minimum and maximum

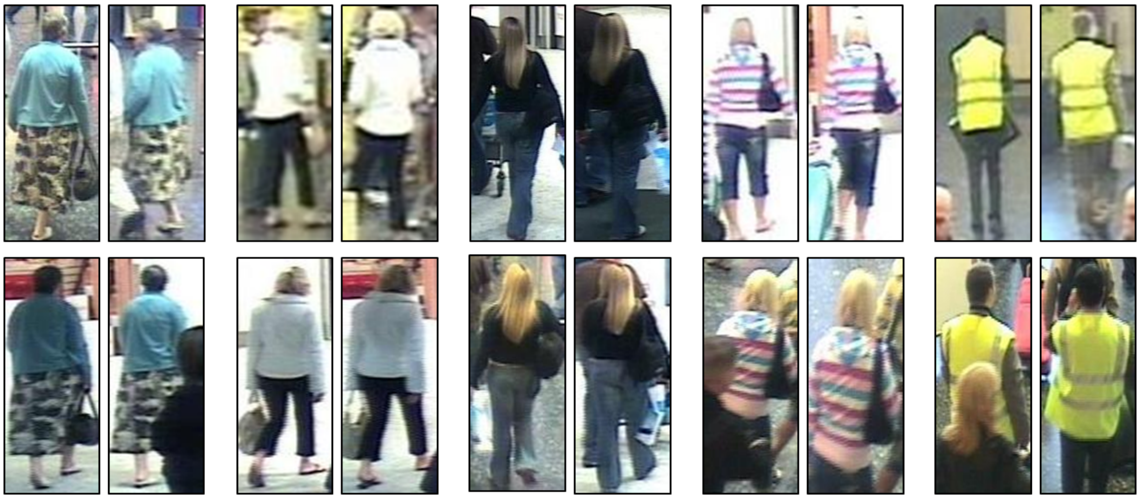


Figure 3.4: Example images from the i-LIDS for Re-ID dataset.

size of the images is 17x39 and 72x144, respectively. Figure 3.5 shows some example images from this dataset.

3.4.5 i-LIDS MA and AA

Almost all of the above datasets contain either a single image or multiple images coming from one or in some cases two cameras. None of the datasets have a significant number of multiple images of a person coming from two separate non-overlapping cameras. In order to address this deficiency, two new datasets [11] were extracted from two non-overlapping cameras from the i-LIDS MCTS dataset. Each of the two datasets contain multiple tracked frames of a number pedestrains from two different camera views. This dataset most closely resembles a multi-camera tracking scenario and captures its characteristics better than any of the above mentioned datasets.

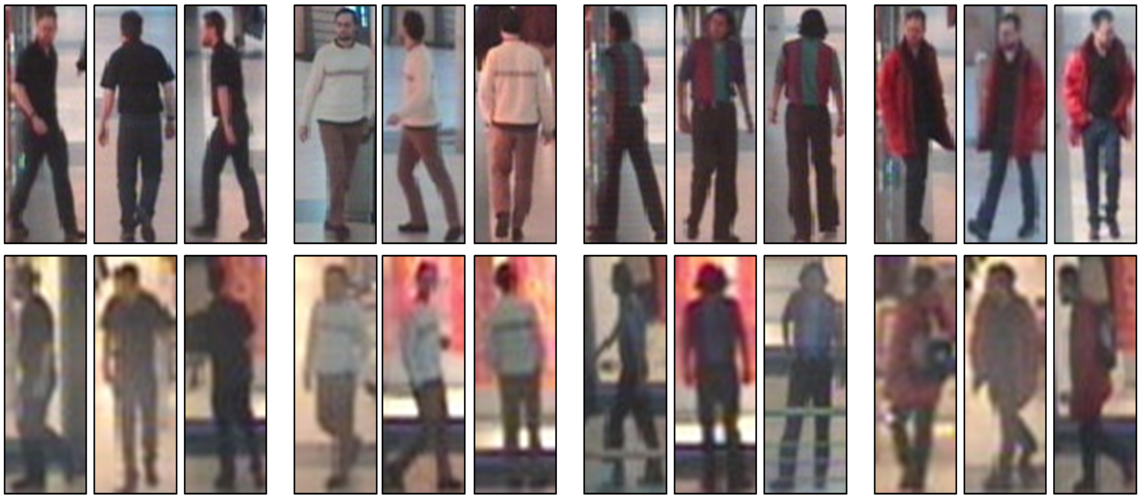


Figure 3.5: Example images from the CAVIAR for Re-ID dataset.

Figures 3.6 and 3.7 shows some example images from this dataset.

- iLIDS-MA: This dataset consists of images of 40 pedestrians taken from two different cameras. 46 manually annotated images of each pedestrain are extracted from each camera. Thus, this dataset contains a total of 3680 images of slightly different sizes.

- iLIDS-AA: This dataset consists of images of 100 pedestrians taken from two different cameras. Different number of auotmatically detected and tracked images of each pedestrain are extracted from each camera. This dataset contains a total of 10,754 images of slightly different sizes. This data presents more challenges due to the possibility of errors coming from automated detection and tracking.

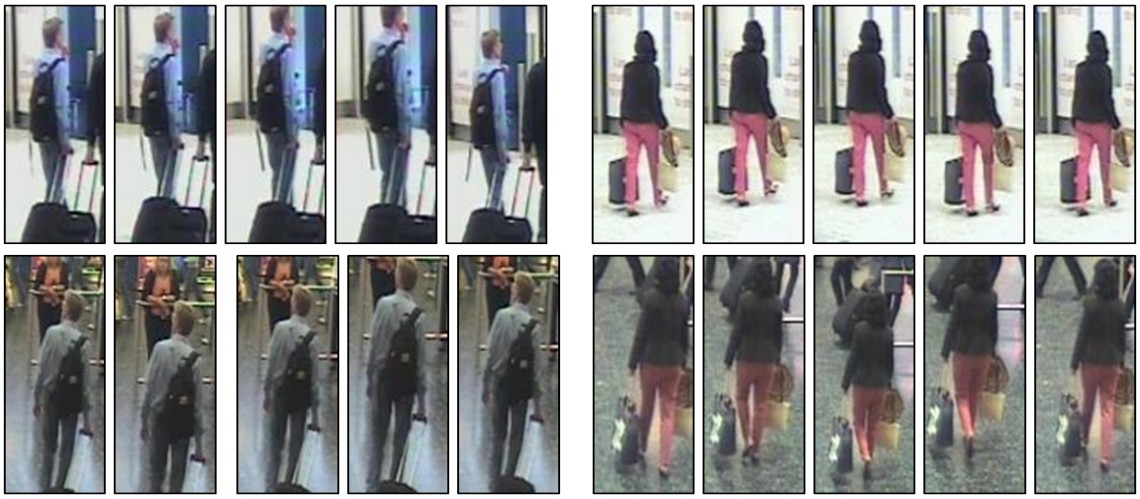


Figure 3.6: Example images from the automatically annotated i-LIDS dataset for Re-ID.

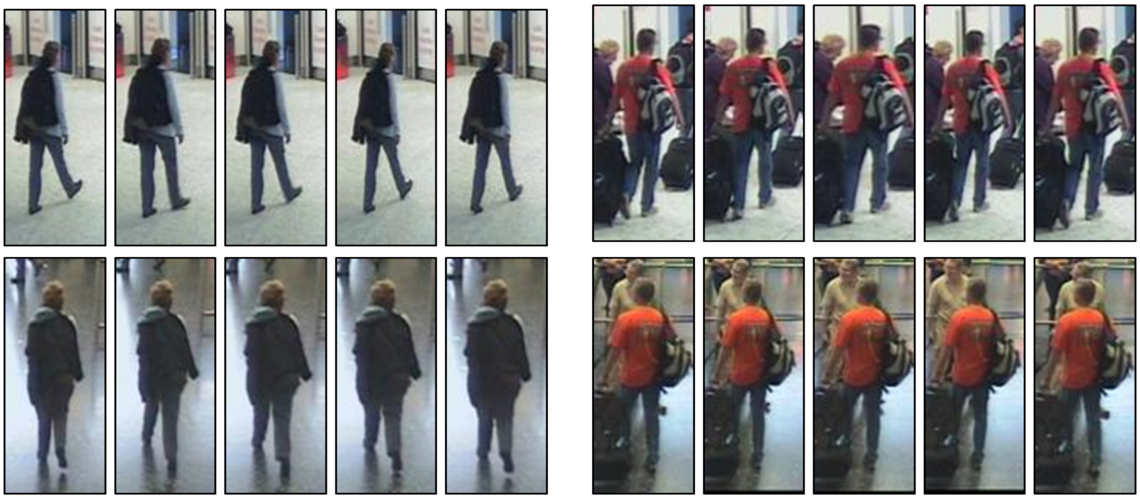


Figure 3.7: Example images from the manually annotated i-LIDS dataset for Re-ID.

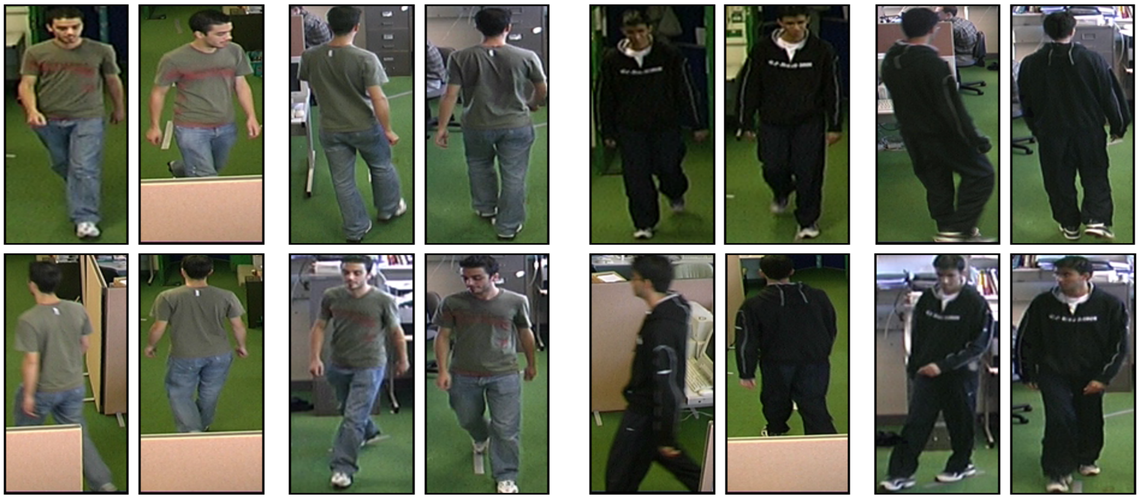


Figure 3.8: Example images from the v-47 for Re-ID dataset.

3.4.6 V-47

This dataset contains videos of 47 pedestrians captured using two cameras in an indoor setting [131]. For each camera view, two different views of each person (person walking in two different directions) are captured. The foreground masks are provided for every few frames of each video. The dataset has some illuminations variations but they are not drastic. There are few occlusions and the scene is not crowded and has very few scale variations. The dataset is important as it provides significantly large amounts of video of each pedestrian but is not sufficiently representative of typical Re-ID scenarios. Figure 3.8 shows some example images from this dataset.

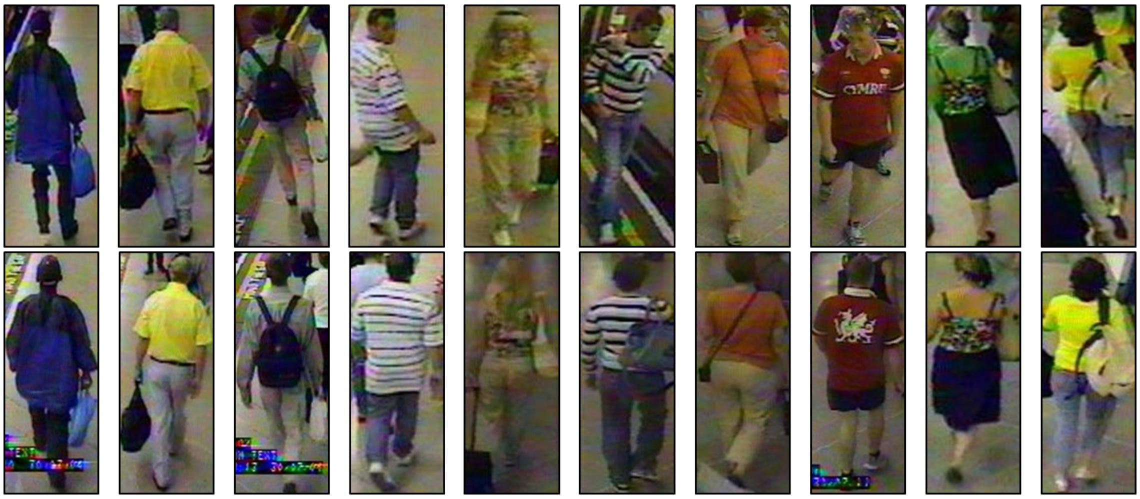


Figure 3.9: Example images from the GRID dataset for Re-ID.

3.4.7 QMUL underGround Re-IDentification (GRID) Dataset

This dataset was acquired by 8 cameras with non-overlapping FOVs, installed in a underground train station [90]. Thus, the images are low resolution and have significant illumation variations. The dataset has 250 pairs of images, i.e., 250 pedestrian images that appear in two different camera views and an additional 775 images of people in a single view. Even though acquisition is using 8 cameras, for a given pedestrian only 2 different views are available. Figure 3.9 shows some example images from this dataset.

3.4.8 Additional Datasets

A few other multi-camera datasets like Chokepoint [136], Terrascope [73], Person Re-ID dataset (PRID) [65], SAIVT-SoftBio [27] and CUHK02 [85] should be mentioned in this context as they can be very well applied to evaluation of person Re-ID methods. Sarc3D dataset [13] contains 200 images of 50 pedestrains taken from 4 predefined viewpoints captured with calibrated cameras to facilitate a 3D body model generation. The 3DPes dataset [12] further extends the Sarc3D dataset by including 600 videos of 200 people taken from 8 static and calibrated cameras. These datasets are geared towards evaluation of 3D human body model for tracking and identification and are applicable to Re-ID evaluation as well. Table 3.3 gives a summary of the Re-ID performance of some of the state-of-the-art approaches on some of the popular databases discussed above. The Re-ID performance is represented by rank 1 accuracy of Re-ID.

3.4.9 Limitations of Datasets

The currently available Re-ID datasets are fairly reasonable in terms of encompassing multi-view variations. However, they are hardly representative of real world surveillance data. For instance, in multi-camera tracking applications, video data from large number cameras with overlapping and non-overlapping views is to be analysed for Re-ID. The cameras and hence the data differ in resolution, frame rate and sensor characteristics. Most of the above mentioned databases are lacking in this respect. In addition, they do not provide means to analyse open set Re-ID performance or other

Approach Type	Approach	i-LIDS for Re-ID	ETHZ-1	ETHZ-2	ETHZ-3	ViPER
Passive	SDALF [19]	28%	65%	64%	76%	19.84%
	BiCov [94]	-	68%	71%	84%	20.66%
Descriptor Learning	Group Context [149]	23%	-	-	-	-
	ELF [57]	-	-	-	-	12%
	PLS [122]	-	79%	74%	77%	-
Metric Learning	LMNN-R [43]	-	-	-	-	20%
	Impostor Learning [66]	-	78%	74%	91%	22%
	PRDC [151]	42.96%	-	-	-	15.66%
	RankSVM [112]	44.05%	-	-	-	16.27%

Table 3.3: Summary of performance of state-of-the-art approaches on popular Re-ID datasets. The results correspond to the single-shot case, for LMNN-R, Impostor Learning, PRDC and RankSVM methods; the accuracy on i-LIDS corresponds to gallery size of 30 out of total 119 (not the complete gallery) and on VIPeR corresponds to gallery size 316 out of total 632.

evaluate system measures like scalability or space time complexity. As unconstrained and long duration video data is not available, the impact of integration of temporal or sequential data into person descriptions on Re-ID cannot be judged. Availability of video data enable learning of the inter-camera relationships [90, 91] that can greatly boost the Re-ID performance by pruning the incorrect(false positives) matches. With most of these datasets such experiments cannot be performed.

Evaluation of long period Re-ID requires data to be collected over several days using same or different set of cameras. None of the currently available datasets offers such instances of people collected on different days. A recent RGB-D person Re-ID dataset [15] captures depth information for each pedestrian and hence can be utilized for evaluation of depth-based features for Re-ID. But it is not collected over several days and hence cannot be utilized for true long period Re-ID evaluation. Also, multi-camera tracking scenarios are by nature multiple person Re-ID problems. It implies that there exists multiple probes that have to be matched simultaneously. These datasets can be setup to test a multiple person Re-ID framework but are not truly multi probe datasets. Hence, there is a definite need for a more comprehensive and extensive Re-ID dataset.

3.4.10 Evaluation Metrics

The most widely used evaluation method for person Re-ID is the performance metric known as the cumulative matching characteristic (CMC) curve. This metric is adopted since Re-ID is intuitively posed as a ranking problem, where each element in the gallery is ranked based on its comparison to the probe. The probability that the correct match is ranked equal to or less than a particular value is plotted against the size of the gallery set [56]. In order to evaluate the performance of the simultaneously matching multiple probe images of the gallery, the Synthetic Re-ID Rate (SRR) curve is derived from the CMC curve. It gives the probability that any of the given fixed number of matches are correct. The normalised area under the CMC curve (nAUC) and Rank 1 recognition rate are also important performance metrics. The nAUC is the probability that the Re-ID system will produce a true match over a false (incorrect) match. However, these metrics are inadequate for evaluating the open set Re-ID performance, more particularly, in evaluating the ability of the system to determine if a probe ID exists in the gallery or not (novelty detection). This point is discussed in detail in the next section.

3.4.11 Limitations of current work and our approach

As is evident, most of the work on person Re-ID leverages clothing appearance based features with very little focus on modeling non-clothing based features to build the person model. We propose two model: a part-based spatio-temporal appearance model that combines color and facial features and the other model combines color and gait features.

The part based description of an individual implicitly incorporates the spatial relationships between different body parts into the model. Our model leverages the temporal nature of the data by meaningfully combining these color features over time. Thus, the proposed model meaningfully encodes the spatial and temporal variations in a person's appearance. Moreover, depending on the presence of usable face images our model can decide to include the facial features or exclude them, thus giving our model flexibility in an unconstrained environment. The color and gait model combines color histograms extracted from multiple frames with gait models generated from multiple frames. Again, depending on the presence of usable gait features, we can decide to include the gait or not. The proposed gait models are also examined for their applicability for long period Re-ID. We evaluate our models in open and closed set experiments for multiple person re-identification and single person Re-ID.

Chapter 4

Color & Facial features-based Person-Model

Our approach to Re-ID is to propose strategies to handle the matching of the person models dictated by the Re-ID requirement, either single or multiple person. In this chapter, we discuss the part-based model that combines color and facial features. The feature extraction and model generation are discussed below. This model is generated using multiple frames of a person and is tested under a multiple person Re-ID scenario. The multiple person Re-ID strategy is tested in open and closed set experiments.

The proposed model in this work is a part-based spatio-temporal appearance model that combines color and facial features. A histogram of oriented gradients (HOG) based part detector is used to extract four stable human body parts: head, left torso, right torso and upper legs. The part based description of an individual implicitly incorporates the spatial relationships between different body parts into the model. The appearance of the torso and legs is characterized by two color features: color histograms and representative color descriptors. In a multi-camera scenario typically a video or sequence of frames of a particular individual are captured. Our model leverages the temporal nature of the data by meaningfully combining these color features over time. The pixel gray level values extracted from the head region images are used as facial features. Since the head region images could be blurred or no face could be present, facial feature extraction is not always possible.

Low level image cues are used to select usable face images. If usable face images are present, the persons face model is built by using the facial features. As a result of the selection process, the face images are no longer temporally adjacent, even if they come from the same video. Thus, multiple models are learnt for the same persons face to keep model inaccuracies due to feature misalignment to a minimum. Depending on the presence of usable face images our model can decide to include the facial features or exclude them, thus giving our model flexibility in an unconstrained environment.

4.1 Color Model

Matching for Re-ID has 2 steps, first is determining corresponding parts from the images to be matched and second, extracting appearance features from each corresponding part to establish a match. Different parts of the body can have different

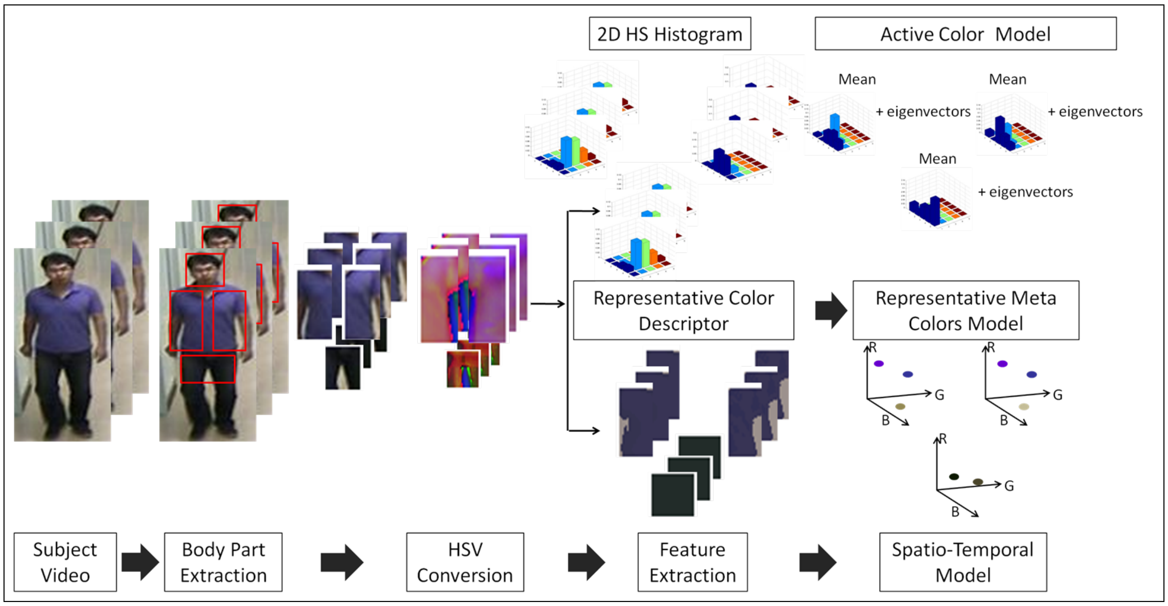


Figure 4.1: Spatio-temporal Color Model Generation.

appearance features, for example shirt could be white and pants black, this observation has prompted us to adopt segmentation of the human body into meaningful parts before matching. This not only locates corresponding parts to be matched but also greatly reduces the probability of an incorrect match. Figure 4.1 shows an overview of the clothing color feature extraction and appearance model generation pipeline.

4.1.1 Body Part Extraction

The body of a person is represented by three stable parts. This enables us to not only achieve a set of corresponding body parts to match but also imparts partial pose invariance to the model since individual body parts have fewer valid poses compared to the entire body. Further, the parts extracted are not aligned with anatomical body parts and hence pose variations do not result in drastic changes in the detected parts. These body parts were extracted using the model proposed in [48], which models the human body as a collection of parts arranged in a deformable configuration. We use the six-part person model trained on the VOC 2008 pedestrian dataset [45]. Of the six parts, we retain only three stable parts: left torso, right torso and the upper legs, since they encapsulate the area of the body that provides maximum distinguishing appearance information. The model consists of a global root template for entire body and local part filters. The global and local templates are based on Histogram of Oriented Gradients (HOG) features introduced in [40] and are learnt during a training phase using latent SVM. The local part model consists of the spatial model i.e. a set of allowable placements relative to the detection bounding box and the part filter. The local part filters are deployed at a resolution higher than the global root filter.

4.1.2 Part Color Feature Extraction

Color is the most expressive and powerful cue for object recognition and is leveraged in our model to characterize appearance. We extract two different color descriptors for each body part. The first descriptor is a color histogram that characterizes the distribution of colors within the body part and the second descriptor is a set of representative colors.

4.1.2.1 Color Histograms

The color content of each body part is characterized using color histograms. In order to provide description of the person that is invariant as much as possible to illumination variation across views and viewpoint changes, the underlying color descriptor should be photometric color invariant. We explore four different color invariant based histograms to compare and contrast each ones performance empirically deducing the best suited one for Re-ID. Typically, color histograms are extracted from color invariant transformations in the color space. The color invariant transforms impart robustness against illumination variations to the color histograms but at the cost of its discriminative ability. For Re-ID, both robustness and discriminative ability is required.

- HS Histograms: The HSV color space is a perceptually intuitive color space, defining color in terms of its hue, purity or saturation and brightness or value. Thus it is useful color space to match colors in or determine if one color is similar to another consistent with human color perception. A 2D histogram based on the H and S color channels of the HSV color space is used to characterize the chromatic content of each body part. Hue is invariant to changes and shifts in the illumination intensity while saturation is not [127]. Each channel is quantized into five bins, thus we have a 5x5 element HS histogram.
- Weighted Hue Histograms: As a result of the transformation from RGB to HSV space, the Hue value becomes unstable near the achromatic axis in the HSV space. In order to nullify this inherent instability, the hue values are

weighted by the saturation values and then the histogram is generated. Such a histogram is called the weighted Hue histogram. In our experiments we use a 16-bin weighted Hue histogram.

- Opponent Histograms: The opponent color space is given by,

$$O_1 = \frac{R-G}{\sqrt{2}}, O_2 = \frac{R+G-2B}{\sqrt{6}}, O_3 = \frac{R+G+B}{\sqrt{3}} \tag{4.1}$$

  The O_1 and O_2 channels encompass the chromatic content and O_3 channel is the intensity channel [127]. The chromatic opponent channels are invariant to shifts in illumination intensity. The opponent histogram is a combination of 1D histograms based on the O_1 and O_2 channels. Each channel is quantized into 8 bins, thus we have a 16 element opponent histogram.

- Transformed RGB Histogram: The transformed RGB values are generated by independently z-normalizing each channel. Such a transformation results in invariance to illumination color and intensity changes. The transformed RGB histogram is a 3D histogram based on the R', G' and B' channels. Each channel is quantized into 4 bins, thus we have a 4x4x4 element transformed RGB histogram. The transformed RGB color space is given by,

$$R' = \frac{R-\mu_R}{\sigma_R}, G' = \frac{G-\mu_G}{\sigma_G}, B' = \frac{B-\mu_B}{\sigma_B} \tag{4.2}$$

Experiments revealed that the HS histogram performs better than other color invariant histograms in terms of discriminative ability under illumination changes. It provides robust yet discriminative description of the color content hence is better suited to handle Re-ID challenges. The results are provided in the next chapter.

4.1.2.2 Representative Colors Descriptor

Every part is characterized using a set of representative colors. These representative colors are extracted by fitting finite mixture models to the color vector using the method proposed in [49]. This method employs a minimum-message-length type selection criteria that is part of the Expectation-Maximization algorithm. This unsupervised selection and estimation of clusters is based on Fisher information matrix. This clustering ensures that the representative colors are formed in such a way that the intra-cluster variance is minimized while the inter-cluster distance is maximized. Figure 4.2 is a pictorial representation of the clustering process that results in extraction of the representative colors. The clustering is done in the HSV space as it is perceptually intuitive, hence useful to match colors in or determine if one color

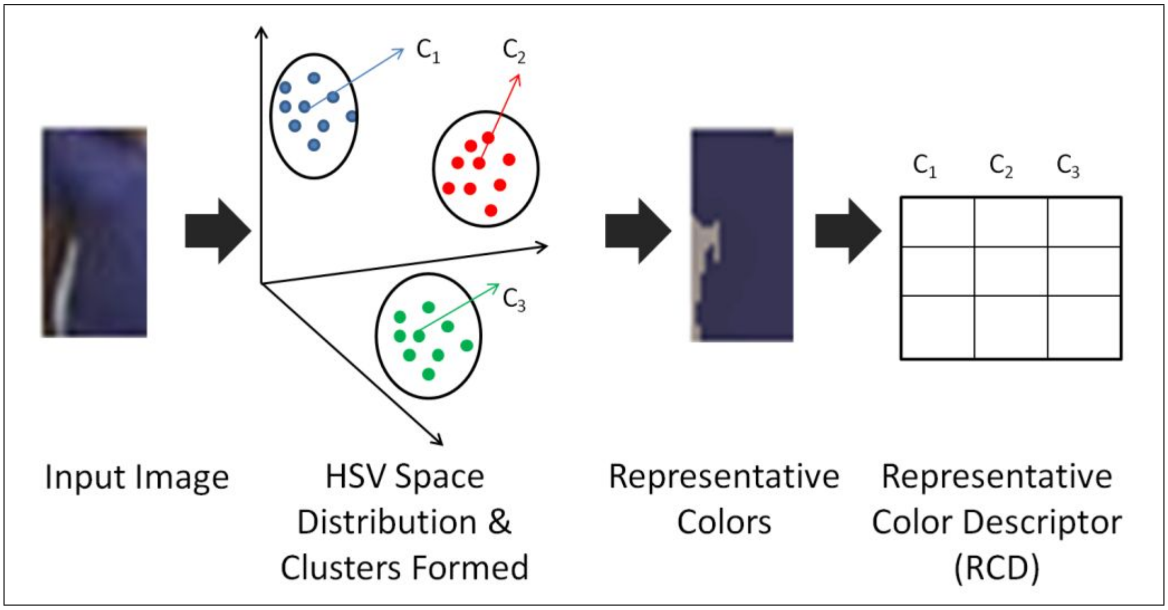

C_1
C_2
C_3
C_1
C_2
C_3
Input Image
HSV Space Distribution & Clusters Formed
Representative Colors
Representative Color Descriptor (RCD)

Figure 4.2: Representative Colors Descriptor Generation.

is similar to another consistent with human color perception. This maximizes our ability to capture the perceptually dominant color in the presence of illumination changes. The final clusters are represented by RGB triplet. Each representative color cluster is described using the cluster's average color. The representative colors descriptor (RCD) is defined as, $RCD = (\{C_i\} \mid i = 1, ..., N_c)$, where N_c is the number of color clusters. In our implementation $N_c \leq 3$.

4.1.3 Spatio-temporal Color Model Generation

Due to the varying conditions that exist between cameras in a surveillance environment the model should incorporate as much visual information as possible in order to be robust and widely applicable. Typically in a surveillance environment a sequence of frames of a particular person exists. These multiple frames can be utilized to learn not only the appearance but also the variations in the appearance as the person moves through a camera's field of view.

4.1.3.1 Active Color Model (ACM) based on Color Histograms

In order to capture the appearance variations of the chromatic content over time, we build a probabilistic model following the idea of Active Appearance Models [39] for each body part based on the underlying color histogram. The sequence of frames of a person are used to extract the 2D color histograms. Thus, the sequence of 2D histograms forms the training set used to build the ACM. The appearance model is given by $g = \overline{g} + A_g.b_g$ and it captures variations in the 2D histograms across the sequence of frames. Here, $\overline{g}$ is the mean 2D histogram and A_g is the matrix describing the modes of variations in the color histograms within the sequence of frames. Vector b_g is the parameter set of the ACM. In our experiments, we only use columns of A_g that retain 75% of the variations. This not only helps in capturing the maximum variations but also helps to eliminate redundant information and outliers.

4.1.3.2 Representative Meta Colors Model

To meaningfully combine the representative color descriptors extracted from the sequence of frames, the representative colors are matched from frame to frame. The frame to frame correspondence is established in the RGB space. Figure 4.3 illustrates the frame to frame representative colors matching and subsequent meta colors determination. The representative colors that do not match up for more than 10 consecutive frames are rejected. Then, another level of clustering is applied within the set of temporally matched colors. Once again, clustering is done in the HSV space but the meta clusters are represented by RGB triplet. Within each set of matched colors, up to 3 representative meta color clusters are built. Each representative meta color cluster or RMC is in turn described using the average color of the meta cluster. These RMCs are extracted for each body part to form the representative color based spatio-temporal model. The RMC is defined as, $RMC = (\{MC_i\} \mid i = 1, ..., N_{mc})$, where N_{mc} is the number of meta color clusters formed over the sequence.

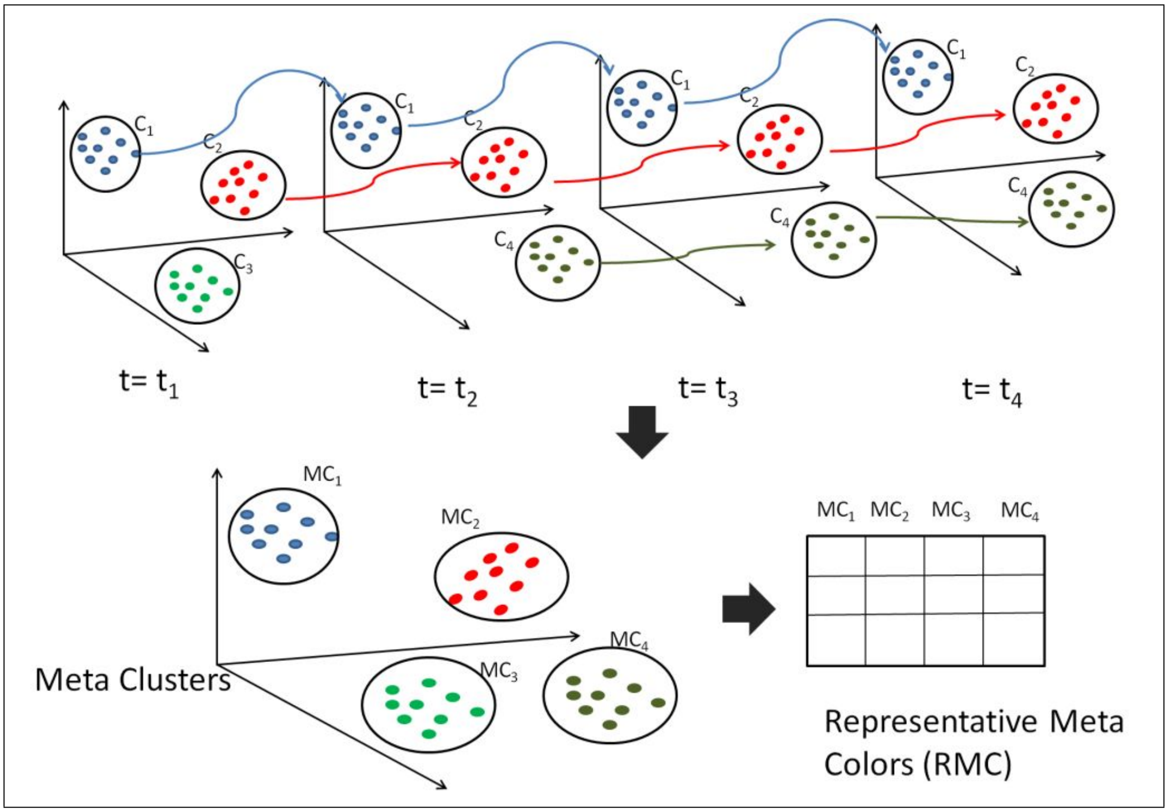


Figure 4.3: Representative Meta Colors Extraction.

4.2 Facial features based Model

Traditional face recognition systems are categorized into near distance face recognition (NRDA) which are often in cooperative applications. NRDA systems use features like, Gabor wavelet or Fourier features extracted from the face, but are not as effective in describing the facial characteristics due to the geometric shape of the face [141]. Facial features that adapt to the variations in the facial characteristics like Eigenfaces [126], Fisherfaces [23] or Laplacianfaces [64] are more relevant. Nonetheless, these features need the underlying face images to be perfectly aligned. Due to the resolution constraints of the head region, alignment between face images is difficult to achieve using standard appearance or landmark based registration methods. Typically in surveillance systems the subject under observation is non-cooperative and not necessarily facing the camera. The cameras have a wide field of view in order to visualize large areas but at the expense of image resolution. Both of these reasons cause significant degradation on performance of current face recognition systems. Further, most of these facial features used for face recognition breakdown due to changes in incident illumination, head pose, expressions and age [148]. Face recognition at a distance systems are designed to tackle most of the issues surveillance videos present and have been researched widely [6]. Zhou *et al.* [152] make use of the temporal nature of video to accomplish face recognition. A face cataloger system was developed by Hampapur *et al.* [62] which uses two calibrated cameras with overlapping FOVs for 3D multi-blob tracker and 3D head location determination. An active face tracking and recognition system is proposed in [110], where two cameras, one static and one PTZ, is used to capture face images at a distance to achieve face recognition robust to pose changes and partial occlusions. However, this enables face recognition of a single person in a video sequence. This problem is partially addressed in [24, 135] where the static camera is equipped with a person tracker and person detections coupled with tracking data are used to direct the PTZ cameras. Estimated pose is integrated into face tracking and used in [17] in order to align face images. Non-overlapping blocks are extracted from aligned faces and 2D discrete cosine transform (DCT) coefficients are concatenated to form facial appearance features that are used for person retrieval.

According to the study presented in [141], pixel gray level values of low-resolution images as features can achieve a high face recognition rate. Such facial features are better suited for Re-ID scenarios and we leverage this feature. Since the head-region images could be blurred or no face could be present, facial feature extraction is not always possible. Low-level image cues are used to select usable face images. If usable face images are present, the person's face model is built by using the facial features. As a result of the selection process, the face images are no longer temporally adjacent, even if they come from the same video. Thus, multiple models are learnt for the same person's face to keep model inaccuracies due to feature misalignment to a minimum. Depending on the presence of usable face images our model can decide to include the facial features or exclude them, this gives our model flexibility in an unconstrained environment.

Since the facial features are simply the image pixels, noisy or non-facial pixels could misguide the facial feature extraction severely. Hence a robust face image selection process is adopted before the facial features model is generated. Figure

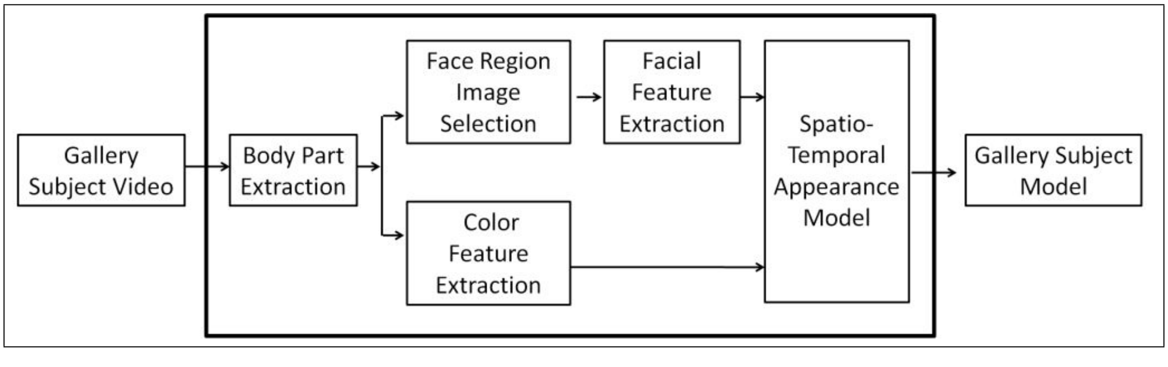


Figure 4.4: Combined Color and Facial feature based Model Generation Pipeline.

4.4 depicts the pipeline from input video to output model generation with multiple feature integration.

4.2.1 Face Image Selection

The body part detector gives the head region of a person and in order to extract facial features only the head regions with faces are retained. All the images are converted to gray scale and a 2 step selection process is used to retain usable face images. First, a threshold (τ_1) on RMS contrast is employed to reject incorrectly detected head region images as well as low contrast faces. The RMS contrast is computed by:

$$RMSContrast = \sqrt{\frac{1}{w.h}\sum_{i\in w}\sum_{j\in h}(I_{ij} - \overline{I})^2} \tag{4.3}$$

where, I_{ij} is the pixel intensity extracted from the face region image of size w by h. $\overline{I}$ is the mean intensity of the image pixels. In the next step, the retained images are then subjected to canny edge detection to detect prominent edges. The edges are summarized by fitting an ellipse. If more than half the number of pixels in the ellipse fitted region are above a threshold (τ_2) then the head image is retained and used to extract the facial features. If n_{total} is the original head region images, then the number of images retained after the selection process is n_{sel}, and these images are not all temporally adjacent. If $n_{sel} < 2$, then the facial model cannot be generated. In our experiments, $\tau_1 = 0.04$ and $\tau_2 = 0.6$.

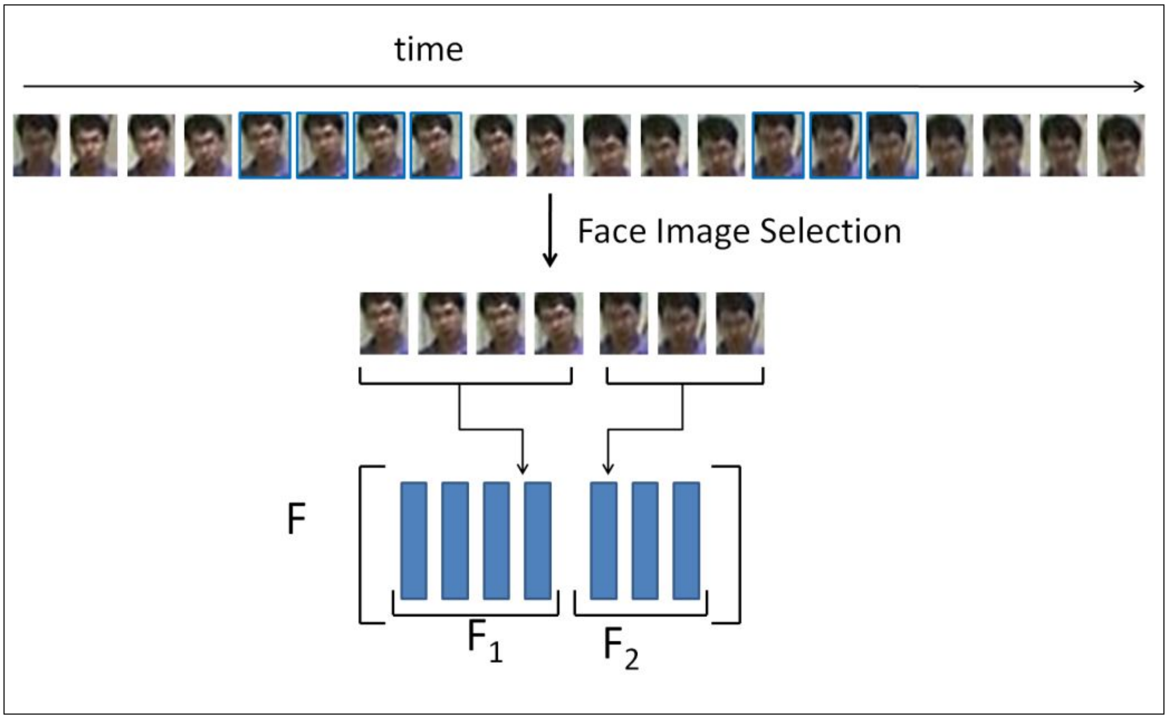


Figure 4.5: Facial feature based Model Generation.

4.2.2 Facial Model Generation

We use the selected images to extract the facial features. All the head region images are resized to a fixed dimension of 24x20. The face region images are vectorized by stacking its columns, these vectors are used as the facial features. The vector length is thus $m = w.h = 480$. Thus, the facial feature based model is simply a matrix $F = [v_1, ..., v_n]$, where v_k is the column corresponding to the kth face region. The matrix F will be of size m x n, where n is the number of selected face images.

$$F_{m,n} = \begin{pmatrix} f_{1,1} & f_{1,2} & \cdots & f_{1,n} \\ f_{2,1} & f_{2,2} & \cdots & f_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ f_{m,1} & f_{m,2} & \cdots & f_{m,n} \end{pmatrix}$$

To minimize the errors in the model the underlying features should be as aligned as possible. We make an assumption that among the selected face images temporally adjacent frames do not change drastically and are reasonably aligned. Thus, within F, F_j is a submatrix if size m x n_j, such that the columns of F_j are temporally consecutive face images. The submatrix is treated as a sub class, i.e. even if all columns from F belong to the same subject, the submatrices within F are treated as visually different instances of the same subject. F is referred to as the facial feature model (FFM). Figure 4.5 demonstrates the face model generation process.

4.3 Re-identification by Matching

The cost of matching a gallery and probe is computed by comparing the generated gallery model to selected frames of the probe sequence since multiple frame of the probe are available. The probe frames to be used for cost computation are selected at random. The matching cost of the minimum cost among all the probe frames. The color model matching cost is computed as per equation 4.4,

$$d_{color}(G,P) = w_{ACM}.d_{ACM}(G,P) + w_{RMC}.d_{RMC}(G,P) \tag{4.4}$$

where, $d_{ACM}(G, P) = \|g_m - \widehat{g}\|$ is the reconstruction error obtained by projecting the probe frame histogram $\widehat{g}$ into the ACM space of the gallery model. Gallery AAM is, $g = \overline{g} + A_g.b_g$ and the probe color histogram is given by $\widehat{g}$. Then, $g_m = \overline{g} + A_g.\widehat{b_g}$, where $\widehat{b_g} = A_g^{-1}(\widehat{g} - \overline{g})$.

$d_{RMC}(G, P)$ is the cost based on the matching of the gallery RMC model to the RCD extracted from the probe frame. The gallery RMC and the probe RCD are both treated as signatures and the matching is treated as a transportation problem. The matching cost between the gallery RMC and probe RCD is calculated using the Earth Mover's distance (EMD) [114]. d_{color} is computed for each body part and is of the form $d_{color} = \sqrt{\sum_{p \in N_{bp}} (d_{color,p})^2}$, where, N_{bp} is the total number of body parts, in our case $N_{bp} = 3$. And $w_{ACM} = w_{RMC} = 0.5$ which means that both sets of color features contribute equally to the matching cost.

The facial feature matching is established using Sparse Representation in the context of face recognition [137]. The underlying implication is that given a dictionary matrix built using labeled training images of several subjects, a test image of i-th subject is the linear combination of the training images of only the same subject from the dictionary. In our case, a given gallery FFM contains all usable images of the gallery subject but the images are arranged into subsets depending on the temporal adjacency. Given a gallery FFM F_G, if the probe ID is same as the gallery ID, the probe image will lie approximately in a linear span defined by only a subset of the images that constitute the FFM. This implies that given a probe face image f_P, it can be expressed as $f_P = F_G.\alpha$ and the intent is to find the sparest α that generated f_P in F_G. Thus among all possible solutions of α, we want the sparsest.

This amounts to solving the following ℓ_1-minimization:

$$\hat{\alpha} = \arg\min \|\alpha\|_1 s.t. f_P = F_G.\alpha \tag{4.5}$$

This optimization is solved using linear programming that leverages augmented Lagrange multipliers [142]. Thus, d_{face} is given by equation 4.6 and is an estimate of how well $\hat{\alpha}$ reproduces f_P.

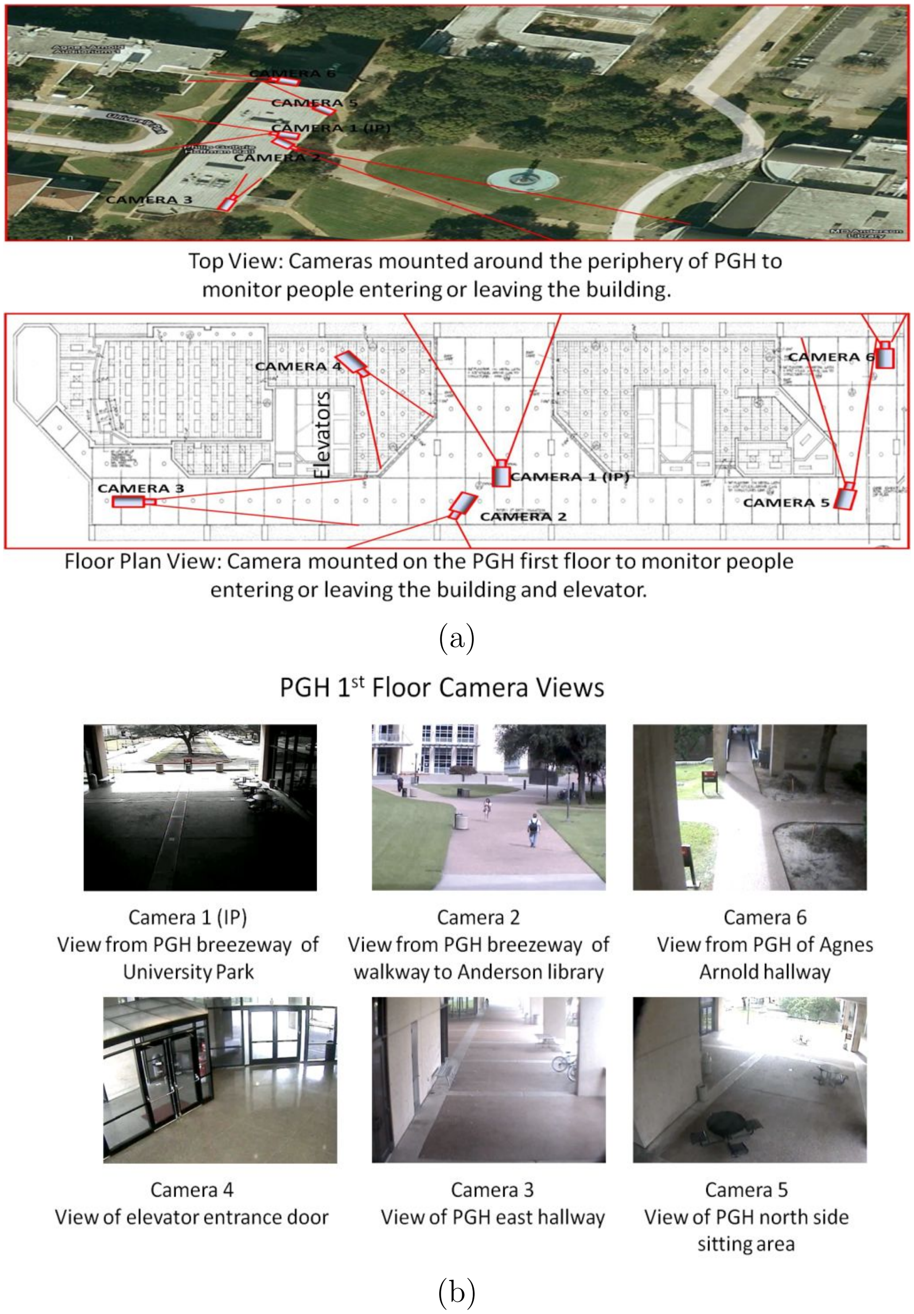
$$d_{face}(G, P) = \|f_P - F_G.\hat{\alpha}\|_2 \tag{4.6}$$

The cost matrix C is used for assignment, which is populated using $cost_{i,j} = dist(G_i, P_j)$, Here, $i = 1, .., N_P$, $j = 1, .., N_G$ and N_P is number of probe subjects and N_G number of gallery subjects. Since all the probe objects come from the same video we know their IDs are distinct so a one-to-one assignment between the probe and gallery sets is needed from multiple person Re-ID. The combinatorial optimization is solved using the Munkres algorithm [101]. Given a cost matrix C, the one-to-one assignment is such that the objective function $\sum_{i \in N_G} \sum_{j \in N_P} C(i,j) x_{ij}$ is minimized. Here, x_{ij} represents assignment of element G_i of the gallery set to element P_j of the probe set, taking value 1 if assignment is done and 0 otherwise.

4.4 Experimental Results

4.4.1 MCID Dataset

In order to obtain real world surveillance data, we setup a camera network consisting of 10 cameras in and around a university building. The camera network has cameras placed on the first floor and fifth floor of the office building. This dataset is terned as Multi-Camera Re-ID (MCID) dataset. Figure 4.6(a) shows the placement of cameras in the outdoor environment. The camera positions are overlayed on a top view of the PGH building on the University of Houston Campus as well as the floor plan of the PGH first floor. Figure 4.6(b) shows example images from each outdoor camera. Figure 4.7 illustrates the position and camera views of the indoor cameras. The Re-ID data consists of 30 videos, collected using 9 of these 10 cameras and consists of 40 subjects out of which 19 can be used to establish true Re-ID. The data is split into 3 scenarios based on the difference in environments between the camera pairs on which Re-ID is to be established. The scenarios are Outdoor-Outdoor, Indoor-Indoor and Outdoor-Indoor. The dataset is the most realistic Re-ID data with all the real world Re-ID challenges of drastic illumination variations, drastic viewpoint changes, scale and pose variations and occlusions. In the Outdoor-Outdoor scenario there exists large variations in the illumination conditions between cameras. The Indoor-Indoor scenario presents lesser illumination variations but is useful to test the model's discriminant capability. The Outdoor-Indoor scenario needs a good balance of discriminative capability and photometric invariance in the model in order to establish correct matches. Within each category the videos are placed in the order of flow of people through the camera network. Every collection of sequences is used as a single Re-ID experiment. Re-identification requires accurate single camera person detection and tracking information. The acquired data is annotated by hand labeling the images. This not only provides the tracking data needed for Re-ID but also provides ground truth for evaluation of the results. The person bounding boxes



Top View: Cameras mounted around the periphery of PGH to monitor people entering or leaving the building.

Floor Plan View: Camera mounted on the PGH first floor to monitor people entering or leaving the building and elevator.

(a)

PGH 1st Floor Camera Views

Camera 1 (IP)
View from PGH breezeway of University Park

Camera 2
View from PGH breezeway of walkway to Anderson library

Camera 6
View from PGH of Agnes Arnold hallway

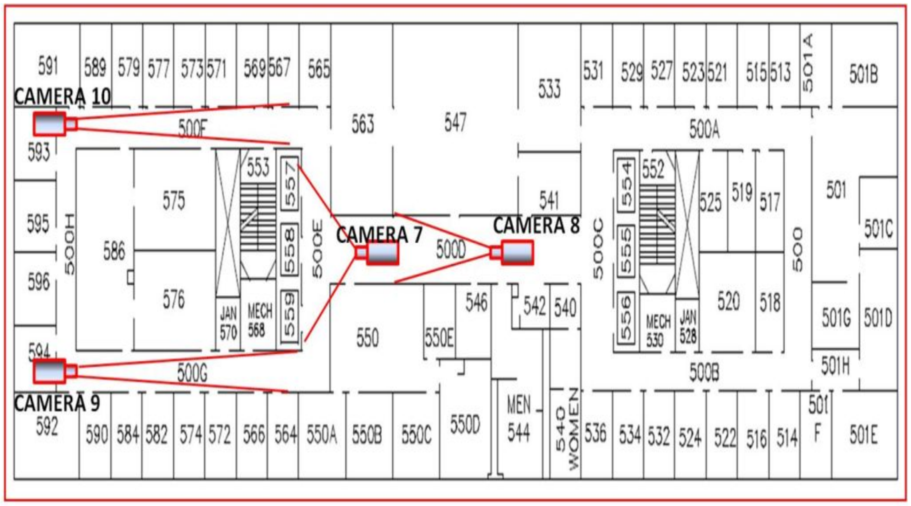
Camera 4
View of elevator entrance door

Camera 3
View of PGH east hallway

Camera 5
View of PGH north side sitting area

(b)

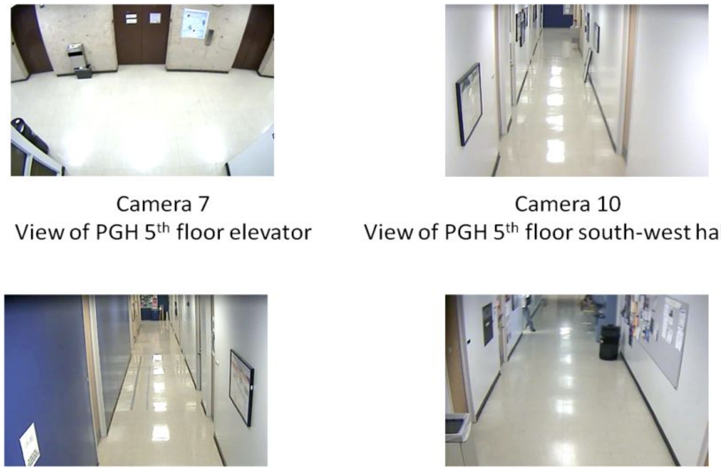
Figure 4.6: Outdoor placement of cameras: (a) Placement of cameras overlayed on building floor plan for perspective. (b) Example images from each outdoor cameras FOV.



Floor Plan View: Cameras mounted on the 5th floor of PGH to monitor people entering or leaving floor and in hallways.

(a)

PGH 5th Floor Camera Views



Camera 7
View of PGH 5th floor elevator

Camera 10
View of PGH 5th floor south-west hallway

Camera 9
View of PGH 5th floor south-east hallway

Camera 8
View of PGH 5th floor central hallway

(b)

Figure 4.7: Indoor placement of cameras: (a) Placement of cameras overlayed on building floor plan for perspective. (b) Example images from each indoor cameras FOV.

obtained from annotation are all normalized to 128x64.

4.4.2 Multiple Person Re-ID Experiments

To evaluate our model under a multiple person Re-ID framework we designed 2 kinds of experiments.

- Closed-Set Experiments: Only a subset of the probe set i.e. an intersection between the probe and gallery set is used to establish the Re-ID. This implies subjects that form the closed probe set is the intersection between the probe IDs and gallery IDs. This is a closed set experiment, in the sense that it is not possible to have false positive but rather only correct matches or mismatches. False positive is defined as the model matching a probe ID incorrectly to a gallery ID when in fact the probe ID is not in the gallery set. This experiment is more geared towards testing the sensitivity of the appearance model.
- Open-Set Experiments: The entire probe sets are used for the Re-ID, this can mean that all the subjects in the probe set might not be included in the gallery set. The probe and the gallery are truly open sets. This implies that possibilities of false positives will be increased considerably. This experiment is designed to test the appearance based false acceptance reduction criteria and a secondary objective of testing the generality of the model. In order to reduce the false acceptance, a threshold is imposed on the cost matrix C and then multiple person Re-ID is established.

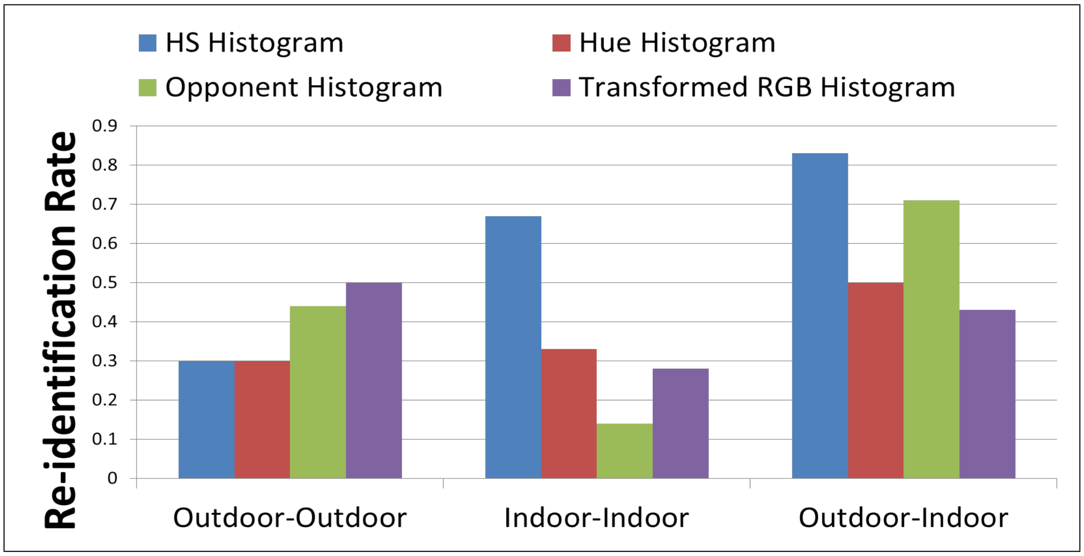
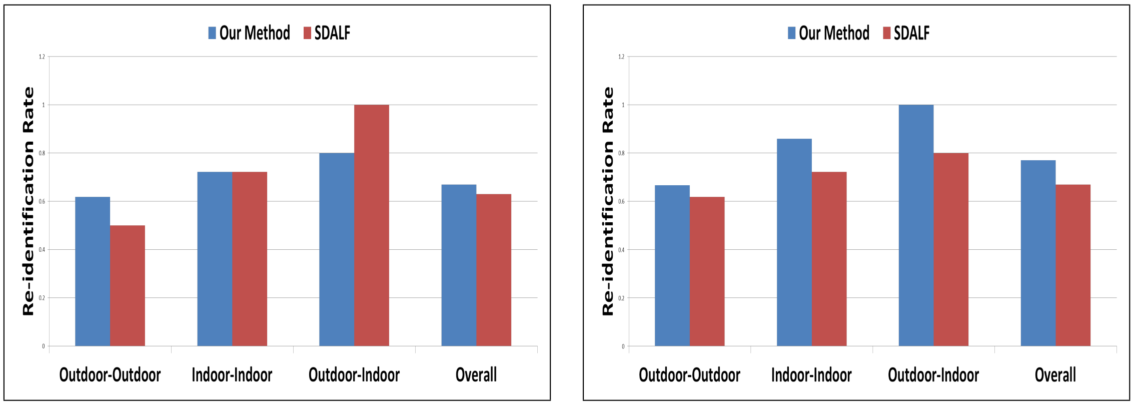


Figure 4.8: Re-identification Rate in different scenarios for comparative study of the Color Invariant Histograms.

4.4.2.1 Color Invariant Histogram Evaluation

Closed set experiments use only the gallery ACM and probe histograms for Re-ID. From figure 4.8 it is clear that not only the model's discriminative power but also its photometric invariance affects the Re-ID rate. For the rest of our experiments we use the HS histogram to build our spatio-temporal ACM. Transformed RGB histogram yeilds better Re-ID rate than the rest in the Outdoor-Outdoor scenario while HS histogram outperforms all outer in the other two scenarios. This implies that a model with greater photometric invariance is desirable in the Outdoor-Outdoor scenario, while greater discriminative ability is important in the other two scenarios.



(a) (b)

Figure 4.9: Closed Set Re-identification Performance: (a) and (b) show the comparative Re-ID rate for N = 2 and 10, respectively. Each graph shows the results obtained by our color model and SDALF [19].

4.4.2.2 Closed Set Experiment Results

Only a subset of the probe set or closed probe set; i.e. an intersection between the probe and gallery set IDs is used to establish Re-ID. This is a closed set experiment, in the sense that it is only possible to have correct matches or mismatches. This experiment is intended to test the sensitivity of the proposed model. The closed set results are evaluated using matching accuracy, i.e. number of probe subjects matched correctly. In other words, it is the percentage of rank-1 correct matches. The effectiveness of our model was compared to the SDALF model proposed in [19].First only the color features based model was compared to the SDALF in order to determine the optimal number of probe frames to use in order to ensure a correct match. Figure 5.9 shows that for all N, our method results in higher rate in the Outdoor-Outdoor scenario compared to SDALF. We believe this is so as our model

	Our Method (*sec*)	SDALF (*sec*)
Feature Extraction	0.5086	3.8486
Matching	0.0484	0.6579

Table 4.1: Comparison of computation times (in seconds) taken by each stage of Re-ID by our method and the SDALF method.

has improved photometric invariance compared to SDALF. Our model's performance on the Indoor-Indoor dataset is on par with SDALF but as N increases our model performs better. This implies that our model's discrimination is not diminished by its invariance ability. On the other hand, SDALF outperforms our model in the Outdoor-Indoor scenario for $N = 2$. This might imply that SDALF has better discriminative ability in the presence of large color variations. We surmise this is because SDALF also has a texture component in its model that is absent in ours. With an increase in N, our model outperforms across all 3 scenarios. The overall accuracy of our model is 67% compared to SDALF's 63% for $N = 2$ and 77% to SDALF's 67% for $N = 10$. SDALF's accuracy reduces in the Outdoor-Indoor scenario as N goes above 2, this might be because the possibility of picking similar frames increases which reduces discrimination during matching. We can reasonably conclude that optimal number of probe frames is $N = 10$ since for both the methods the performance over all three scenarios is better with $N = 10$. Table 4.1 shows the comparison between the computation time taken at each stage of model building and feature matching for our color model method and the SDALF method. These computation times were found using MATLAB's 'tic-toc' command on a PC with a 2.67GHz Intel Core i5

CPU and 4GB RAM. These numbers are merely used to show that the feature extraction and matching cost computation times of our method are significantly more efficient than the SDALF method. In addition, our learnt spatio-temporal model is a rather compact and hence has a lower storage footprint.

In all of our subsequent experiments we use $N = 10$ to compute the color model matching cost. Figure 4.10 shows the closed set Re-ID performance using 2 variants of our model, one based only on color features and the other using both the color and facial features. From the figure the value of incorporating facial features into the spatio-temporal model is clearly evident. In both the Outdoor-Outdoor and Outdoor-Indoor scenarios, we observe a significant improvement in the Re-ID performance. In the Indoor-Indoor scenario the performance remains unchanged by addition of

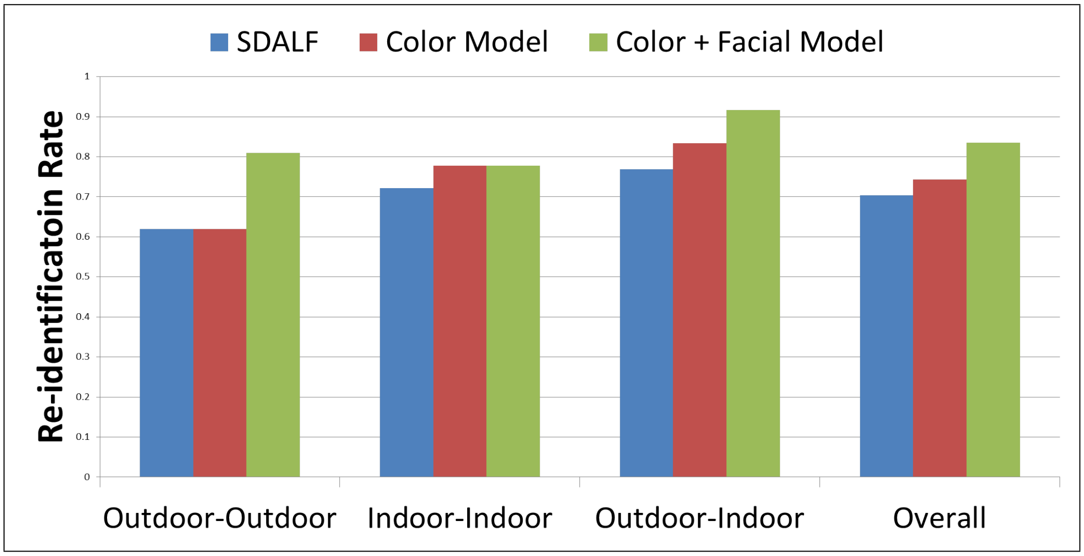


Figure 4.10: Closed Set Re-identification Performance: the bar graph shows the results obtained by our color model, color and face model and SDALF [19].

the facial features but also does not have an adverse effect on the Re-ID rate. This implies that even low-resolution face regions with varying illumination and pose can contribute towards improving the discriminative ability of our model. In the Outdoor-Outdoor scenario adding facial features causes the proportion of accurate IDs to increase from 62% to 80%, which is a considerable improvement. The overall Re-ID rate increases to 83% from 75%. Both variants of our model outperform the SDALF model in all 3 scenarios. In all the 3 scenarios our color only model gives 75% accurate IDs compared to 70% using SDALF.

4.4.2.3 Open Set Experiment Results

The entire probe set is used for Re-ID wherein all the subjects in the probe set might not be present in the gallery set. This implies that in addition to correct matches and mismatches we will also have false positives. In the case of Re-ID, true positives (TPs) are the number of probe IDs that are correctly matched. Mismatches (MMs) are the number of probe IDs that are incorrectly matched to gallery, when that probe ID does exist in the gallery. False positives (FPs) are the number of probe IDs that are matched to the gallery when the probe ID does not exist in the gallery. This experiment is designed to test the model based false acceptance reduction criteria and a secondary objective is to test the generality of the model.

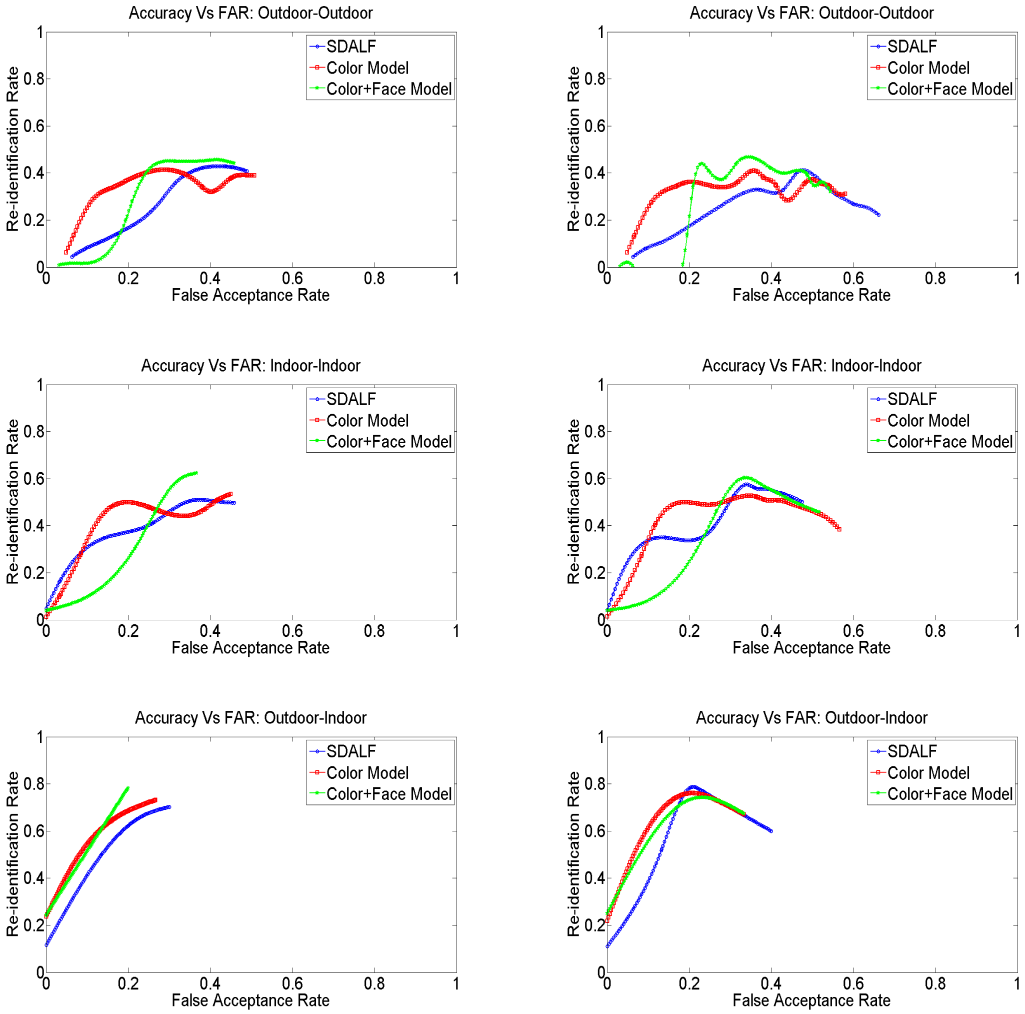
In order to reduce the false acceptance, a threshold is imposed on the cost matrix C and then multiple person Re-ID is established. The open set results are presented in terms of Accuracy vs. False Acceptance Rate (FAR) curves. The accuracy and FAR are defined as $Accuracy = \frac{TPs}{N_P}$ and $FAR = \frac{(MMs+FPs)}{N_P}$, respectively, where N_P

Assignment	Model	Outdoor -Outdoor	Indoor -Indoor	Outdoor -Indoor
Optimal	SDALF	0.38	0.32	0.35
	Color	0.34	0.32	0.35
	Color+Face	0.38	0.40	0.40
Sub-optimal	SDALF	0.34	0.3	0.28
	Color	0.33	0.26	0.29
	Color+Face	0.37	0.32	0.30

Table 4.2: Open Set (Discrimination) results for all the models. The top portion of the table is generated using the Munkers assignment algorithm, whereas the lower portion comes from the suboptimal assignment algorithm.

denotes the total number of probe subjects. The curves are obtained by varying a threshold imposed on the matching cost during the computation of the cost matrix. The threshold is varied from 0 to 0.9 in increments of 0.05. Two different assignment techniques are employed; optimal, which is the Munkers algorithm, and the sub-optimal assignment. The sub-optimal assignment technique is usually used when the cost matrices have many forbidden assignments. The assignment is suboptimal in the sense that the overall assignment cost is not the minimum possible value. In case of Re-ID, the gallery is ever increasing and most likely the intersection between gallery and probe is small compared to the size of the gallery. Thus, the possibility of incorrect assignments increases and suboptimal assignment technique could be better suited for such cases.

Figure 4.11 shows that in all 3 scenarios it is possible to find a threshold that yields the best possible trade off between accuracy and FAR. In the Outdoor-Outdoor scenario a threshold of 0.5 yields the best possible Accuracy/FAR ratio of 60%/33%



(a) (b)

Figure 4.11: Open Set Results: Accuracy vs. FAR curves obtained using (a) Optimal Assignment, and (b) Suboptimal Assignment. The top, middle and bottom row are results obtained on the Outdoor-Outdoor, Indoor-Indoor and Outdoor-Indoor scenarios respectively.

using the color and facial features model and optimal assignment technique. In the Indoor-Indoor and Outdoor-Indoor scenarios as well, color and facial features model gets the best Accuracy/FAR ratio possible. Thus, overall in both variants of our model it is possible to find a threshold that gives better Accuracy/FAR ratio than SDALF.

In order to compare the open set performance of the 3 models we use the Discrimination measure as an indicator of model's accuracy and false rejection ability. The discrimination measure is simply the area between the curve and the non-discrimination line. Table 4.2 shows that in all 3 scenarios the color and facial features model has improved discrimination over the other two models evaluated. In the Outdoor-Outdoor scenario and Indoor-Indoor scenario using optimal assignment technique, SDALF has a higher discrimination than our color only model. We surmise this is because SDALF also has a texture component in its model that is absent in ours. Overall, all the models suffer a slight drop in discrimination using the suboptimal assignment technique. Thus, we can deduce that Munkres algorithm is a better fit for Re-ID assignments.

In general, closed- and open-set experiments both suggest that using color and facial features has a distinct advantage in the Re-ID over using only color or even combined color and texture features. From the shape of these curves we can deduce that a global threshold on the matching cost is a crude yet sensible strategy for minimizing the false acceptance rate. As the threshold increases the accuracy values start increasing as well with a less drastic increase in the FAR. The most important conclusion that we can draw from these experiments is that the proposed model can be used for false acceptance reduction during Re-ID. Specifically, open set experiments can be used to select the parameters of false acceptance reduction criteria, in our case a suitable threshold.

Chapter 5

Color & Gait features-based Person-Model Generation

In this chapter we discuss the person model that combines color and gait features. The gait features are further analyzed for long period Re-ID. The combines color$gait model is tested under single person Re-ID scenario in a closed set setting. This also helps us investigate the utility of gait features alone for long period Re-ID. A closed set Re-DI experiment allows us to understand the capability of the gait using the CMC metric.

Re-ID data comes from uncontrolled environments with non-cooperative subjects where face of a person will not always be visible. Further, the data is often low quality due to low sensor resolutions and low frame rates. If the face is visible, it varies greatly in pose, facial expressions, and illumination conditions. All these factors make capturing reliable facial data and subsequent face recognition very difficult.

Even though the state-of-the-art face recognition techniques yield high recognition rates it is important to note these results are obtained on high resolution data captured under controlled lighting and pose settings. Automated facial recognition on low-resolution images under variations in pose, age and illumination conditions is still an open problem [6, 148]. Gait is a behavioral biometric that has been effective for human identification [63]. Gait is especially suited for Re-ID as it can be extracted by non-obtrusive methods and does not require co-operative subjects. Figure 5.1 shows images and corresponding silhouettes of the same person taken from different cameras. In figure 5.1 (a) the person's face is visible but not in figure 5.1 (b). This situation occurs frequently in surveillance videos as the camera angles are uncontrolled. Further, we can see that in sequence (a) the face resolution is really low and hence extracting usable facial features is challenging. However, the detected silhouettes can be used to extract gait features for Re-ID, even with low resolution and low frame rate data. The availability of video data makes gait feature extraction feasible as gait is extracted over multiple frames. On the other hand, gait is

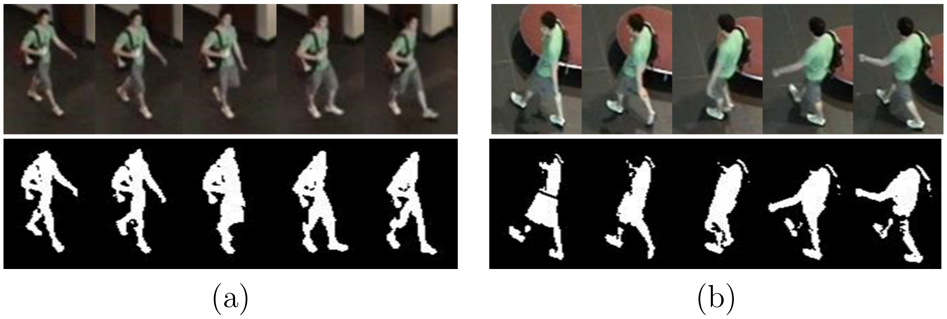

(a) (b)

Figure 5.1: Images from SAIVT SoftBio dataset and extracted silhouettes of the same person from different cameras, (a) images with visible face regions, (b) images without visible face regions.

sensitive to view angles and walking poses. Surveillance cameras usually have a wide field-of-view (FOV) and people often tend to change their walking pose during the duration of observation. This greatly increases the probability of common walking views across different camera views, which can be leveraged for gait recognition. Nonetheless, silhouette extraction errors due to occlusions, illumination variations as shown in figure 5.2, can affect gait feature potency.

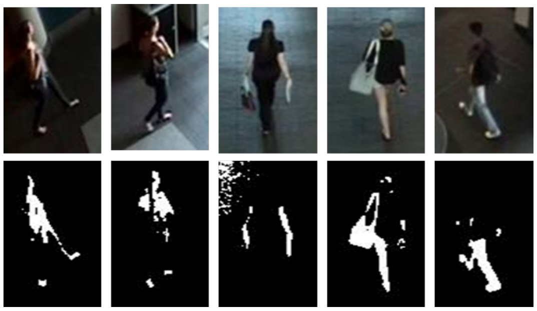


Figure 5.2: Example images from SAIVT SoftBio dataset and erroneous extracted silhouettes.

In this work, we study the impact of incorporating gait features extracted from real world surveillance videos along with color (clothing-based appearance) features on Re-ID performance. If the number of frames available for a subject are not enough or silhouette extraction is faulty, then the gait features are not used and the Re-ID is only based on color feature.

5.1 Gait Features

Gait features are divided into two categories, *model-based* and *model-free* [129]. Model-based features like stride and cadence require an explicit model construction, hence, are more sensitive to the accuracy of silhouette extraction techniques and model fitting requires large computational cost. On the other hand, they are invariant to view angle changes and scale. Model-free features capture changes in silhouette shapes or body motion over time. This makes them robust to errors in silhouette extraction process but are more sensitive to variations due to pose and scale changes. Since silhouette extraction on surveillance video can be very challenging due to illumination variations, complex backgrounds, occlusions and unconstrained environments [30], model-free gait features are well-suited for Re-ID.

Han *et al.* [63] extract a gait feature called Gait Energy Image (GEI) that is robust to silhouette errors and computationally easy to extract. GEI captures the spatio-temporal description of person's walking pattern into a single image template by averaging silhouette over time, however it fails to retain the dynamic changes in the pattern. In order to incorporate the temporal information for gait recognition, Gait History Image (GHI) was proposed in [88]. It used difference between consecutive frames to retain frequency of motion within a GEI. All of these features are robust to silhouettes shapes but not as much to incomplete silhouette detections. The effect of incomplete silhouettes is partially alleviated by the Frame Difference Energy Image (FDEI) [35] which retains only positive portions from consecutive frame differences and combines them with GEIs to retain both static and dynamic portions of the person's walking pattern. In this work, GEI and FDEI features are selected for gait representation.

5.2 Cross-view Gait Recognition

Gait recognition across different view angles is also an active area of research [69, 16, 80]. A view transformation model (VTM) is learnt in [80] that learns a mapping between different view angles to transform data from probe and gallery sequences in the same view angle. A model that projects both the gallery and probe features into a subspace where they are highly correlated is learned to improve gait recognition in [16]. The model is learnt using canonical correlation analysis and combined with a Gaussian process classifier trained for view angle recognition for effective cross-view gait recognition. Gait dynamics from different views are synchronized into fixed length stances using non-linear similarity based HMMs in [69] and multi-linear projection model is learned to help gait recognition. All of these methods require some way to quantify the view angles from either gallery sequence, probe sequence or both, and involves a projection stage that introduces gait feature errors. Due to low quality data available for gait feature extraction such methods prove challenging to leverage for Re-ID. We propose using sparsified representation based gait recognition technique that requires no explicit common view angle synthesis or projection subspace learning and we discuss it in detail in the next section.

5.3 Gait Features Model

Gait is described as the walking characteristic of a person and in order to understand the applicability of gait features to Re-ID they are combined with color features to create a person model for recognition. A person's walking pattern is periodic in nature, thus motion patterns during a walking sequence repeat themselves at regular time intervals. Thus, the features are extracted over a sequence of frames over a gait period. Gait period is defined as the time between two successive strikes of the same foot, or twice the time between successive strikes of opposite feet. We adopt the method proposed in [130] for gait period estimation of arbitrary walking sequences. The aspect ratio of silhouette over a sequence of frames is used to estimate the gait period. Hence the assumption is that silhouettes are already extracted and pre-processed. Pre-processing involves size normalization and horizontally aligning the silhouettes. Figure 5.3 depicts the step by step results of the gait period estimation method, which is as follows:

- The aspect ratio of the silhouette over a sequence of frames is represented as a 1D temporal signal. The signal is z-normalized (subtracting the signal mean and dividing by signal's standard deviation) and smoothed using a moving average filter.

- Peaks in the aspect ratio signal are magnified by computing its auto-correlation sequence and the first derivative of the auto-correlation signal is used to detect zero-crossings.

- Zero-crossing positions of positive and negative peaks are used to compute

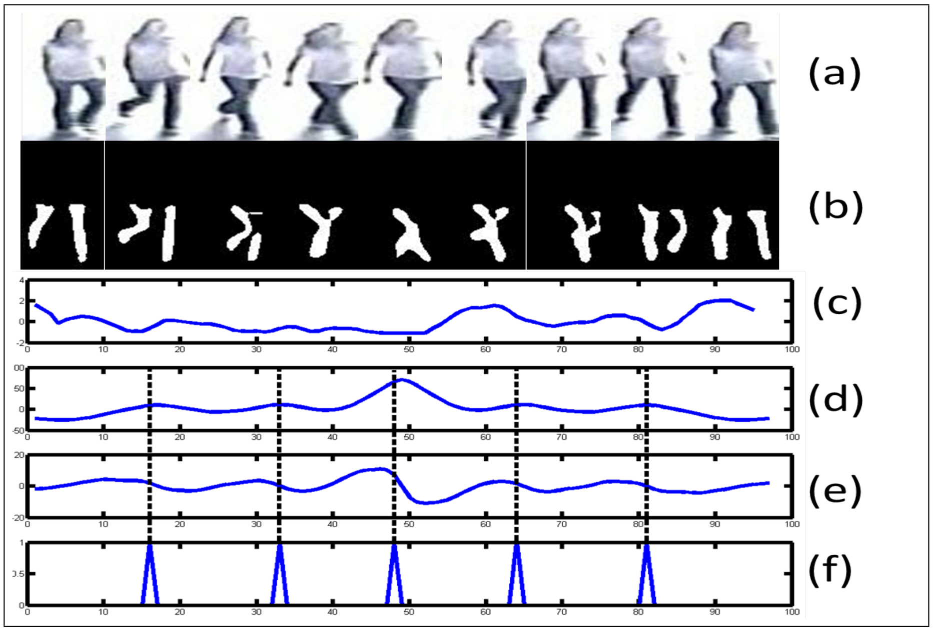


Figure 5.3: Gait Period Estimation: (a) Sample image sequence. (b) Sample silhouette sequence. (c) Bounding box aspect ratio after average filtering. (d) Auto-correlation signal of aspect ratio signal. (d) First order derivative of auto-correlation. (f) Detected peak positions.

distance between prominent peaks, average of distances between consecutive peaks result in the gait period in number of frames.

Most of the dynamic information about walking motion can be deduced from the lower half of the silhouette region [63], hence gait period estimation is performed using aspect ratio of lower body silhouettes.

5.3.1 Gait Energy Image (GEI)

The first of the two gait features used is called the Gait Energy Image (GEI) first proposed in [63]. GEI is a spatio-temporal representation of a person's walking characteristic condensed into a single image and it captures the changes in the shape of the silhouette over a sequence of images. GEI is computed by averaging the silhouettes in the spatial domain over a gait cycle. Given, a sequence of silhouettes, the GEI is given by,

$$GEI(i,j) = \frac{1}{N}\sum_{t=1}^{N} S_t(i,j) \tag{5.1}$$

where, S_t is the silhouette at frame t, (i,j) are the spatial image co-ordinates and N is the estimated gait cycle period. A given sequence of images can contain multiple gait cycles and hence can be represented by multiple GEIs. It has been shown that GEIs are robust to errors in silhouettes in individual frames [63] and are computationally efficient. Figure 5.4 shows GEI extracted from a sample sequence from both the SAIVT SoftBio dataset and the dataset we acquired, which we refer to as the multi-camera Re-ID (MCID) dataset. For the MCID dataset, we only used lower body silhouettes to extract gait features due to challenging conditions for silhouette extraction.

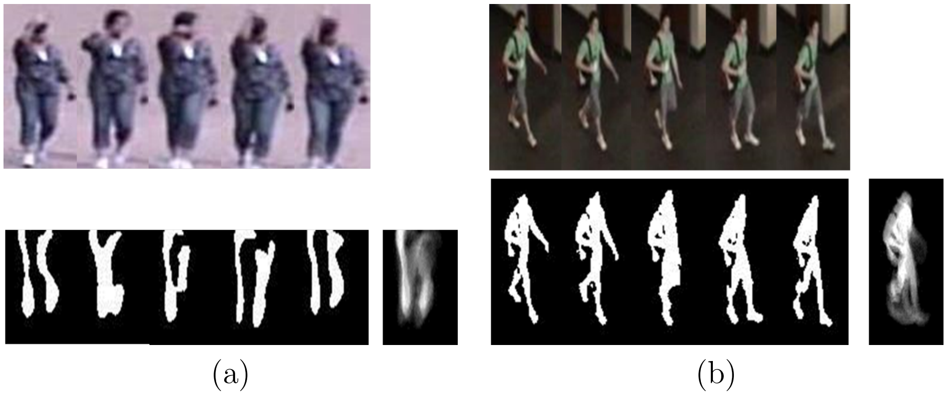


Figure 5.4: GEI features extracted from sample sequences from the 2 datasets, (a) Lower body GEI from MCID dataset and (b) GEI from the SAIVT Soft-Bio dataset.

5.3.2 Frame Difference Energy Image (FDEI)

The second gait feature used is called Frame Difference Energy Image (FDEI) and was proposed in [35]. This feature is designed to deal with incomplete silhouettes at the same time retaining the shape and motion changes. This is done by constructing multiple thresholded GEIs from a single gait cycle and then summing these with differences between consecutive silhouettes along the sequence. The difference is taken in such a manner that only the positive values are retained and negative values are discarded so that only parts of the silhouette that are missing or contain motion information is retained. Having estimated the gait cycle, it is further divided into sub-cycles with equal number of frames. The silhouettes within each sub-cycle are averaged using equation 5.2 as follows:

$$GEI_c(i,j) = \frac{1}{N_c} \sum_{t=1}^{N_c} S_t(i,j), \tag{5.2}$$

where, N_c denotes the period of the sub-cycle. This creates GEIs per sub-cycle of every gait cycle. These GEIs are thresholded, only retaining the pixels with value greater than or equal to 80% of the maximum value, in order to remove noise due to silhouette errors. These are referred to as dominant GEIs (DGEIs) for each sub-cycle and are computed using equation 5.3 as follows:

$$DGEI_c(i,j) = \begin{cases} GEI_c(i,j), & \text{if} \quad GEI_c(i,j) \geq (0.8 * max(GEI_c)) \\ 0, & \text{otherwise} \end{cases} \tag{5.3}$$

The frame difference is computed by subtracting consecutive silhouettes and only the positive portion of the difference is included in the feature as follows:

$$DS_t(i,j) = \begin{cases} 0, & \text{if} \quad S_t(i,j) \geq S_{t-1}(i,j) \\ S_{t-1}(i,j) - S_t(i,j), & \text{otherwise} \end{cases} \tag{5.4}$$

The FDEI is generated by summation of the positive frame difference and corresponding sub-cycles dominant GEI as follows,

$$FDEI(i,j) = DS_t(i,j) + DGEI_c(i,j) \tag{5.5}$$

If the silhouette in frame t is incomplete, and one at frame $t-1$ is complete, then DS_t contains the incomplete portion of the silhouette and summation of difference image with the dominant GEI compensates for missing portion. Figure 5.5 shows FDEI extracted from a sample sequence from both the datasets.

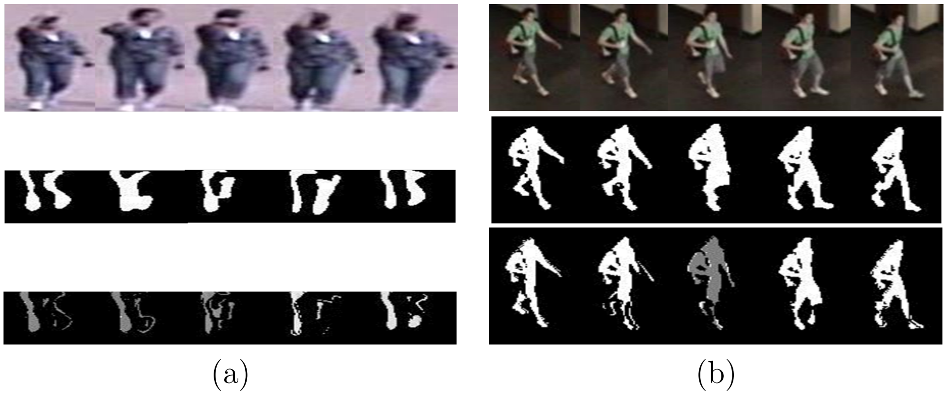


Figure 5.5: GEI features extracted from sample sequences from the 2 datasets, (a) Lower body FDEI from MCID dataset and (b) FDEI from the SAIVT Soft-Bio dataset.

5.4 Feature matching for Re-ID

The distribution of colors of the person's clothing is characterized using weighted HSV histogram proposed in [19]. The silhouette is divided into three body parts corresponding to head, torso and legs regions by detection of one vertical axis of symmetry and two horizontal axes of asymmetry. The histogram for each body part is weighted by the distance from the axis of symmetry. The histograms from each body part are concatenated channel-wise to generate a single color feature descriptor. The similarity measure between a gallery subject G and a probe subject P can be defined as:

$$dist(G, P) = w_{color}.d_{color}(G, P) + w_{gait}.d_{gait}(G, P) \tag{5.6}$$

where, d_{color} is the color features based similarity and d_{gait} is the gait features based similarity. If good quality silhouette sequences are available for both gallery and probe subject and gait features can reliably be extracted, gait similarity is incorporated in the overall similarity measure. In this work, we define the acceptable quality of the silhouettes simply as to whether or not the number of positive pixels is greater than 40% of the total pixels in the image. Gait period estimation process provides another gait feature selection mechanism. If the silhouettes are too noisy or the sequences are too short, then gait period cannot be reliably estimated. Gait similarity is combined with color similarity using w_{color} and w_{gait}. This means that in the absence of gait model, Re-ID is established using only color features. The color similarity is obtained using the Bhattacharyya distance [25]. Since, a sequence of frames are available for every subject, color similarity for a given probe-gallery pair is simply the minimum cost among all the probe-gallery frame pairs.

5.4.1 Gait Recognition by Sparsified Representation

In order to compute the gait similarity/recognition we utilize a sparsified representation [137]. The underlying implication is that given a dictionary matrix built using labeled training features of several subjects, a test feature of i-th subject is the linear combination of the training images of only the same subject from the dictionary. Following this implication, we construct a dictionary which is simply a matrix $V = [v_1, ..., v_n]$, where each column is obtained by vectorizing the gait features belonging to all the gallery subjects. Since, each gallery subject can have multiple gait features, either GEIs or FDEIs, multiple columns in the matrix V may belong to the same gallery subject. Given a dictionary matrix V, if the probe subject is as close as possible the gallery subject in both identity and view angle, the probe image will lie approximately in a linear span defined by only a subset of the gait features that constitute the matrix V. This implies that given a probe gait feature, for example, GEI_P, it can be expressed as $GEI_P = V.\alpha$ and the intent is to find the sparest α that generated GEI_P in V. Thus, among all possible solutions of α, we want the sparsest. This amounts to solving the following ℓ_1-minimization:

$$\hat{\alpha} = \arg\min \|\alpha\|_1 \quad s.t. \quad GEI_P = V.\alpha \tag{5.7}$$

This optimization is solved using linear programming that leverages augmented Lagrange multipliers [142]. Thus, d_{gait} is given by equation 5.8 and is an estimate of how well $\hat{\alpha}$ reproduces GEI_P.

$$d_{gait}(G, P) = \|GEI_P - V.\hat{\alpha}\|_2 \tag{5.8}$$

Further, for a given probe subject, there can be multiple GEIs/FDEIs and the gait similarity is simply the minimum among all the probe GEIs/FDEIs.

5.5 Experimental Results

The combined gait and color features based Re-ID is tested on two datasets, our multi-camera Re-ID (MCID) dataset and SAIVT SoftBio dataset [26]. The SAIVT SoftBio dataset was selected as it provides real word multi-camera surveillance videos with multiple frames per person from different camera views.

5.5.1 SAIVT SoftBio dataset

The dataset contains 150 subjects in up to 8 camera views. Only 122 subjects are seen in at least 2 camera views. In our experiments, the gallery is formed by different cameras and so are the probe subjects, so it is as close to the real world Re-ID scenario as possible. The only constraint is that the same subjects gallery and probe camera views are different. The dataset provides background images for each subject per camera view, hence simple background subtraction followed by low level image processing are sufficient to extract silhouette sequences for gait feature extraction. For each subject, in a given camera view, multiple frames are available. Only 5 randomly selected frames from a sequence are used to extract the color features, while all frames are used to extract gait features. Table 5.1 shows the Re-ID performance of only color, color & GEI (color+GEI) and color & FDEI

	Rank 1(%)	Rank 5(%)	Rank 20(%)	nAUC(%)
Color Model	0.82	3.28	18.85	56.55
Color(0.8)+GEI(0.2)	0.82	3.28	21.31	59.09
Color(0.5)+GEI(0.5)	0.82	4.92	27.05	62.34
Color(0.1)+GEI(0.9)	**4.10**	**9.84**	**36.06**	**64.75**
Color(0.8)+FDEI(0.2)	0.82	3.28	21.31	58.98
Color(0.5)+FDEI(0.5)	0.82	5.74	27.05	62.24
Color(0.1)+FDEI(0.9)	**4.10**	**11.48**	**36.06**	**64.82**

Table 5.1: Rank 1, 5 and 20 matching accuracy and nAUC measures for color, color+GEI and color+FDEI with varying weights.

(color+FDEI) in terms of rank 1, 5, and 20 Re-ID accuracy and normalized AUC extracted from cumulative matching characteristic (CMC) curves using all the 122 subjects. When gait features cannot be extracted, Re-ID relies only on color features. If gait is available, then a weighted sum of color and gait similarity is used to establish a match. The table shows the performance variations with different color and gait weight combinations. We can see that overall using combination of color and gait performs better than only color. As we increase the weight assigned to gait features, we notice a significant boost in performance. The improvement in rank 1 matching accuracy is most notable with changes in weights suggesting that the gait features can be more effective than color features. Figure 5.6 shows the CMC curves obtained by using only color, combined color and GEI and combined color and FDEI. Figure 5.6 (a), (b) and (c) show the CMC curves obtained by varying weights given to color and GEI/FDEI. We see that the difference in performance using GEI and FDEI is negligible. Thus, gait features even when extracted from imperfect silhouettes and varying viewpoints provide discriminative value. They add value to color features

	Rank 1(%)	Rank 5(%)	Rank 20(%)	nAUC(%)
Color Model	0	17.40	78.26	49.72
GEI	17.40	**56.52**	95.65	72.78
FDEI	**30.43**	47.83	**95.65**	**73.16**

Table 5.2: Rank 1, 5 and 20 matching accuracy and nAUC measures for color, GEI and FDEI features.

even in short period Re-ID where clothing color is a reasonable feature.

In order to better understand the potential of gait features, we performed another experiment with the same dataset. Re-ID was performed using only gait features without the help of color and then compared with pure color based performance. Since only gait features are used the number of subjects for which usable gait features available across views were limited to 23 subjects. Figure 5.7 shows the CMC curves obtained using color, GEI and FDEI features. From the figure, the power of gait features is even more prominent. Either of the gait features outperform the color features significantly. FDEI features yeild a much better Rank 1 matching accuracy than GEI features, yet overall their performance is still comparable. Table 5.2 summarizes the rank 1, 5, and 20 matching accuracy and nAUC measures for the 3 features. From the table we can see that in terms of rank 1 and nAUC, FDEI features perform better than GEI. These results can be viewed as performance of gait features for long period Re-ID as they do not utilize any clothing or appearance features. This speaks to the ability of gait features for long period Re-ID and also demonstrates the robustness of the proposed gait recognition method for cross view angle gait matching. Figure 5.8 shows the first 5 ranked images from the gallery set given two different probes using the 3 models, we can see that in both cases the color only Re-ID cannot find an ID match to the probe in the top 5 ranked IDs. In the probe ID shown in the left column, the color+GEI Re-ID located the true ID in the top 5 ranked IDs. On the other hand, the color+FDEI Re-ID rank 1 ID is the correct ID as the probe.

5.5.2 MCID dataset

As the Re-ID is in the context of multi-camera tracking, Re-ID is established for each camera pair. Consistent with the previous dataset experiments, only 5 randomly selected frames from a sequence are used to extract the color features, while all frames are used to extract gait features. Due to complex and uncontrolled conditions of acquisition, traditional background modeling techniques did not perform well for silhouette extraction. Gray level thresholding combined with morphological operations were used to extract lower body silhouettes. Thus, both gait features were extracted only from lower body dynamics. For this dataset, we empirically set $w_{gait} = 0.2$ and $w_{color} = 0.8$ as this gives the best possible Re-ID accuracy. Since Re-ID is performed between camera pairs, all the subjects in the probe set might not be included in the gallery set. Thus, for each camera pair, only a subset of the probe set or closed probe set, i.e. an intersection between the probe and gallery set IDs is used to establish Re-ID. This is a closed set experiment, in the sense that it is only possible to have correct matches or mismatches and results are evaluated using matching accuracy, i.e. number of probe subjects matched correctly. Again, closed set experiment is consistent with our experiments on the SAIVT SoftBio dataset. Figure 5.9(a) shows the Re-ID performance using all 3 models: one based on only color, color+GEI and color+FDEI. The value of adding gait features to color is very evident, as we observe a significant improvement in the Re-ID accuracy. To better understand the role of gait features in Re-ID performance, we perform another experiment, using only gait features without the help of color and then compared with pure color based performance. Since only gait features are used, only subjects from each camera pair for which usable gait features are available for Re-ID. The results of this experiment is shown in figure 5.9(b). This figure provides balanced view of performance of all the features, alone and combined. Both purely GEI and FDEI perform reasonably well as compared to the color even though they do not outperform color. It should be noted that only lower body silhouettes are used to generate gait features. The Re-ID accuracy using FDEI comes within 4% of the color accuracy. Combining either GEI or FDEI boosts the Re-ID accuracy significantly. Incorporating FDEI with color helps enhance Re-ID performance to 81% from 66%, a considerable boost. Thus, we argue that combining gait features with color is an effective strategy to improve Re-ID performance. Figure 5.10 shows the first 5 ranked images from the gallery set given two different probes using the 3 models, we can see that in both cases the color only Re-ID finds the true ID match to the probe in the top 5 ranked IDs at ranks 3 and 2 for the 2 IDs, respectively. In the probe ID shown in the left column, the color+GEI Re-ID located the true ID at rank 2. On the other hand, the color+FDEI Re-ID rank 1 ID is the correct ID as the probe in both probe IDs.

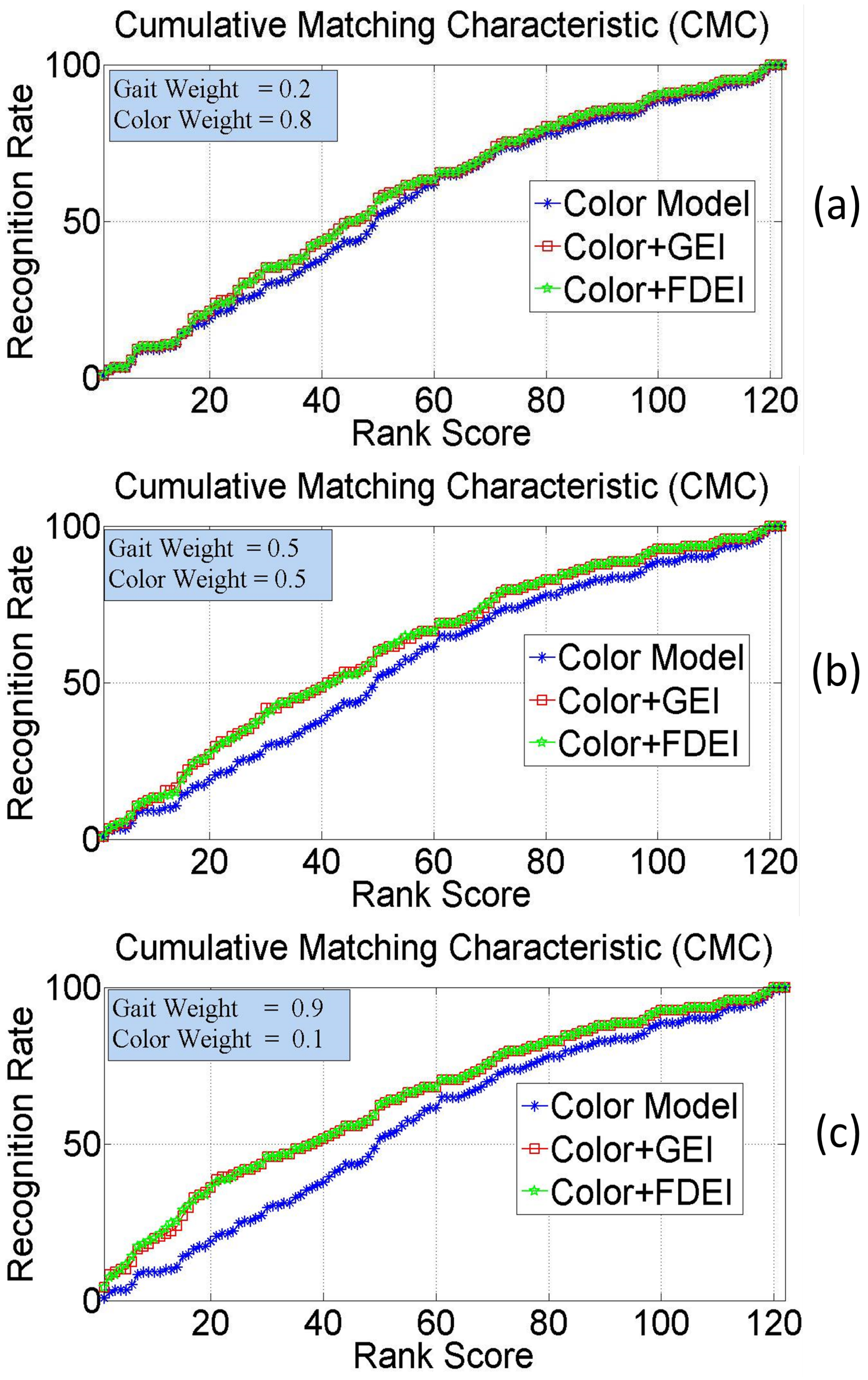


Figure 5.6: CMC Curves on the SAIVT SoftBio dataset, combining color and gait features with varying weights for Re-ID.

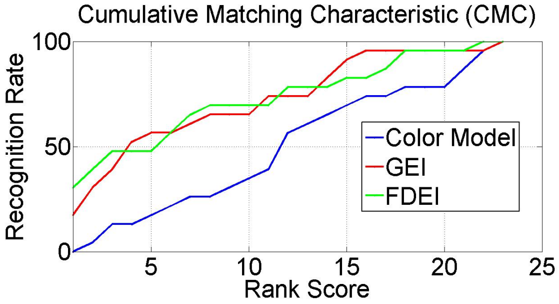


Figure 5.7: CMC Curves on the SAIVT SoftBio dataset, using only color, only GEI and only FDEI features for Re-ID.

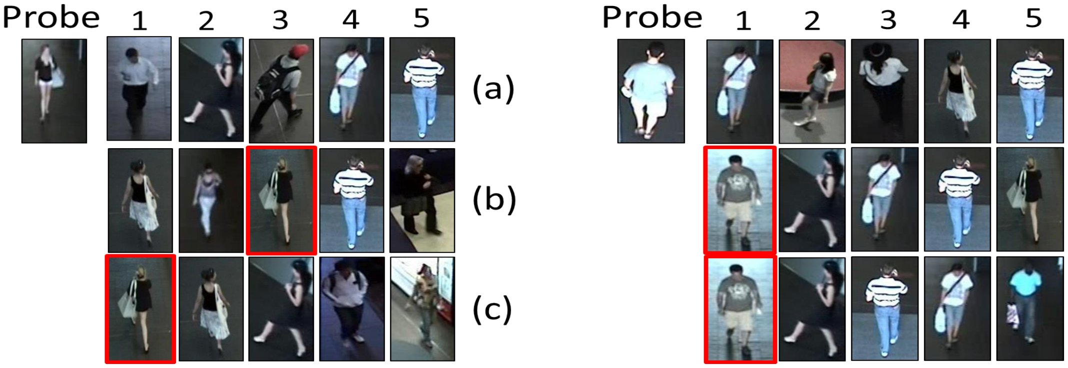


Figure 5.8: Examples of Re-ID gallery ranking, showing the first 5 ranked on SAIVT SoftBio dataset using: (a) only color, (b) color+GEI and (c) color+FDEI, for two different IDs. The images highlighted by the red border denote the true match.

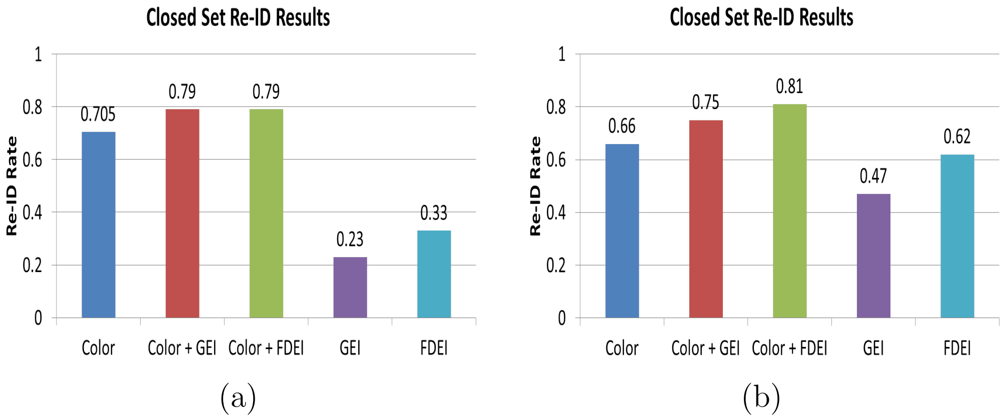


Figure 5.9: Closed set Re-ID performance: (a)Bar graph shows the results obtained by our color, color+GEI and color+FDEI on the entire MCID dataset and (b) Bar graph shows the results obtained by our color, color+GEI, color+FDEI, only GEI and only FDEI on subset of MCID dataset.

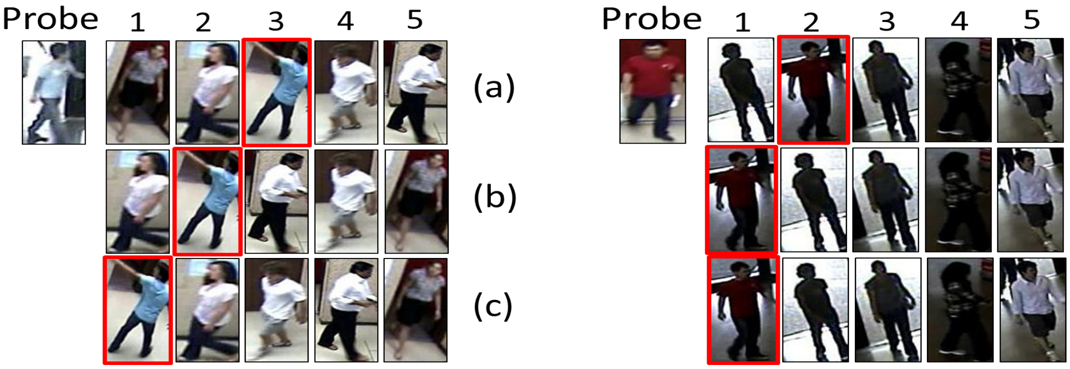


Figure 5.10: Examples of Re-ID gallery ranking, showing the first 5 ranked on MCID dataset using: (a) only color, (b) color+GEI and (c) color+FDEI for two different IDs. The images highlighted by the red border denote the true match.

Chapter 6

Open-Set Person Re-ID

6.1 Why is Re-ID an open set recognition problem?

A recognition problem is termed as open set when the probe presented to the system is not assumed to have match in the gallery set. In other words, the gallery is a set of *known* classes and the incoming probe may be of a *known* class (found in the gallery) or an *unknown* class. The same is the scenario for Re-ID mainly in multi-camera tracking. Figure 6.1 shows a typical multi-camera tracking scenario that depends on Re-ID to assign consistent IDs across cameras. Here, Re-ID is done across each camera pair. The first time a person is seen, his/her appearance model is learned, and the subject is enrolled in the gallery set. Gallery is a set of people IDs previously seen. Thus, all people observed in the second camera form the probe set. Usually,

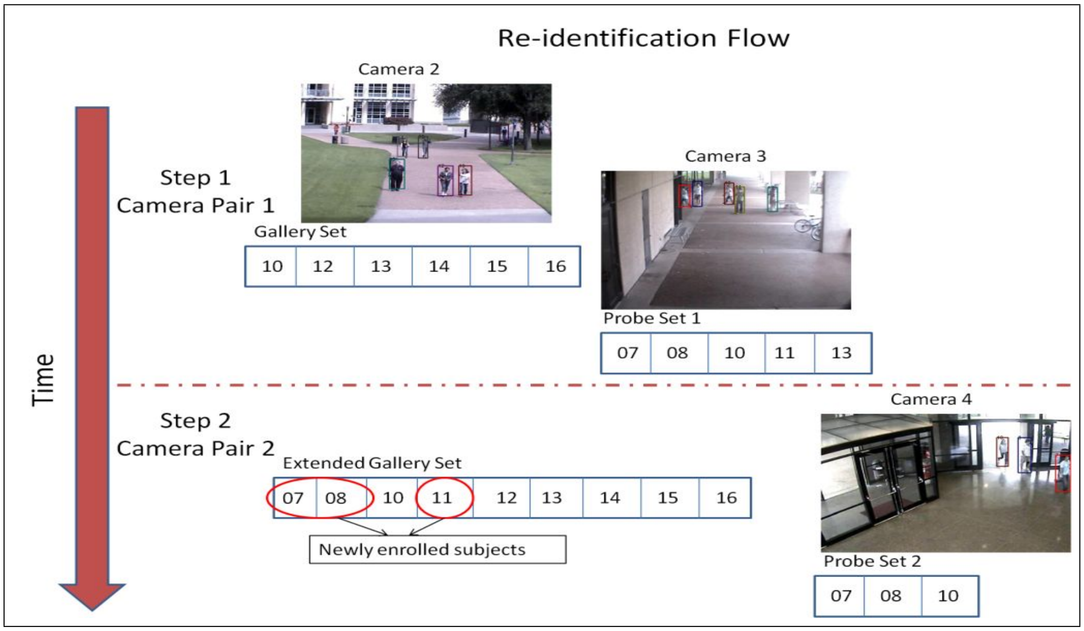


Figure 6.1: Re-identification setup and gallery generation in a multi-camera scenario.

surveillance scenes are uncontrolled and hence the an assumption that people seen in the second camera only come from the first camera is unrealistic. Ideally, after Re-ID, all the people observed in the second camera who were previously unseen are enrolled into the gallery. As the Re-ID moves to the next camera pair the gallery set is extended. However, enrolling unseen IDs in the gallery post Re-ID requires the ability to first determine whether the presented probe ID is a part of the gallery i.e. is a known ID or a *novel* ID. This process of identification of a new or unknown ID (class) is known as *novelty detection* and it is an indispensable part of an open set Re-ID system.

However, current solutions to Re-ID always assume that Re-ID is a closed set problem which can be addressed by effective ranking of the gallery IDs based on a feasible notion of similarity with probe. This is an incomplete solution based on an unreasonable assumption. Hence, a mechanism to classify an incoming probe as *unknown* or *known* is important. Novelty detection is a fundamental aspect of any classification system and hence many models have been proposed [97, 98]. However, this is non-trivial and some important issues need to be considered:

- Effective features that are capable of distinguishing between *known* and *unknown* classes.
- Not possible to present all possible *unknown* classes to a system beforehand.
- As the list of *known* classes increase exponentially, how should the system scale up to unknown classes.

Most often novelty detection is treated as a one-class classification problem. Machine learning models are trained using all possible instances of the *known* class. However, with a problem like Re-ID the *known* class consists of multiple IDs, hence multiple classes are *known.* Thus, *known* class (gallery) is a meta class consisting of all the *known* (IDs seen before) classes. So to train a one-class system all the *known* IDs are labeled as positive instances. One-class novelty detection techniques using one-class SVMs [120] and Gaussian process classifiers [78] have been proposed. A one-class SVM arranges all the data relative to the origin of the feature space and attempts to detect a classification hyperplane that maximizes the distance between the training data points and the origin. Other one-class classifiers attempt to construct a hypersphere around the *known* data such that the hypersphere volume is minimized. However, such a representation is unable to take advantage of distinct classes within the *known* meta class to better generalize the classifier which leads to poor novelty detection performance. An improvement to this approach in attempted in [139], where a small number of outliers within the *known* classes is used to better refine the boundary of the hypersphere. This approach is closer to binary classification. To capitalize on the variability within the *known* classes to better characterize the *known* space the method proposed in [28] attempts to learn a joint subspace where all samples of each known classes are mapped to a single point. Thus, novelty detection is simply the process of computing a distance measure between the test sample in the learnt subspace.

Scheirer *et al.* [119] proposed a one-vs-set SVM designed to establish a boundary around the *known* class to better define a threshold beyond which a test sample is no longer considered from the *known* class. This is done to strike a balance between generalization and specialization of the *known* space. We adopt this approach to better characterize the space defined by the gallery IDs and minimize the probability of incorrectly labeling an *unknown* ID with a gallery ID. For Re-ID, if a novel ID is incorrectly labeled with a gallery ID (false positive) the effect is 3 fold: 1)Obvious false positive, 2)mis-label the probe ID which is the true match of the gallery ID (mis-match), and, 3) as the novel ID was not detected as such it will not be enrolled into the gallery, causing gallery misrepresentation. We also utilize a binary SVM for novelty detection to better understand the effect of limiting the known space.

Re-ID challenges to novelty detection are as follows:

- Known IDs is a meta class (collection or previously seen IDs).
- Known and Unknown categorization is relative to the Re-ID system (does not

align with fundamental differences in feature space).

- The probe ID always comes from another camera view: drastic appearance changes.
- Gallery ID descriptions are low quality.

6.2 Open Set Re-ID: Proposed Approach

Typically, an Re-ID system not only has to classify the presented probe as unknown or known, but also if it is known it has to determine which of the gallery IDs should be assigned to it. The proposed method to open set Re-ID tackles both these aspects separately. When a probe is presented to the Re-ID system it is first subjected to novelty detection i.e. classified as either known or unknown. If the probe is classified as a known ID, then gallery ranking is performed to assign it the most likely of the gallery IDs. Figure 6.2 depicts the proposed Re-ID setup as an open set system with novelty detection. Given a probe, the gallery ranking based Re-ID does not have a mechanism to assign an unknown label to the probe. The proposed method

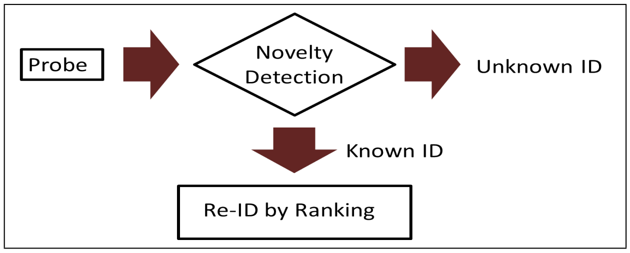


Figure 6.2: Open Set Re-ID: Novelty Detection followed by Gallery Ranking.

addresses this by first presenting the probe to the novelty detection stage. This helps the system to avoid labeling a non-gallery ID incorrectly with the gallery ID, which implies that false positive ID matches will be reduced. This also provides a means to enroll a newly seen ID into the gallery. The importance of such a stage is more evident as the number of unknown IDs encountered increases, which is the case in multi-camera tracking. For example, let us consider a camera network placed at an airport. On any given day, the number of distinct people that go through an airport for the very first time is extremely large as compared to the people seen over and over again. The number of unknown IDs a system encounters compared to the number of gallery IDs and number of probe IDs can be thought of as the openness of the problem. Thus, for a trained novelty detection system the openness is expressed in [119] as follows:

$$opennness = 1 - \sqrt{\frac{2 * |num_{gallery}|}{|num_{probe}| + |num_{target}|}} \tag{6.1}$$

where, $num_{gallery}$ is the number of IDs that exits in the gallery and used to train the novelty detection stage, num_{probe} is the number of probe IDs and num_{target} is the subset of probe IDs that exist in the gallery. As the openness increases the importance of the novelty detection stage for Re-ID performance becomes more pronounced. The gallery and probe IDs are represented using clothing color features. We leverage the weighted HSV histograms described in the previous chapter as the color features. The weighted HSV histograms were proposed in [19]. The body region is divided into two parts: upper and lower, HSV histograms are extracted for each part. They are formed using 16 bins for the H and S channels and 4 for the V channel. Each pixels contribution to the histogram is weighted by it's distance from the vertical dividing line. The histograms from the two body parts are concatenated to for the final feature vector.

6.2.1 Binary SVM for novelty detection

The gallery set is represented as: G_i, for $i = 1, ..., num_{gallery}$. All the gallery IDs are used as the training samples of the positive (*known*, $y_i = +1$) class for the SVM. For every ID that exists in the gallery, a corresponding negative ID that does not exist in the gallery set G is used as the sample of the negative (*unknown*, $y_i = -1$)class. The non-gallery set is represented as: NG_i, for $i = 1, ..., num_{gallery}$. Thus, the training (IDs) set for the SVM is $x = G \cup NG$. A binary SVM attempts to learn a classifier $f(x)$ such that,

$$f(x_i)\begin{cases} \geq 0, & y_i = +1 \\ < 0, & y_i = -1 \end{cases} \tag{6.2}$$

The binary SVM adjusts the hyperplane separating the gallery and non-gallery classes in order to achieve maximum separation between them as shown in figure 6.3. The hyperplane is detected by solving the optimization $\min \frac{1}{2}\|w\|^2 + C\|\xi\|$ subject to $y_i(wx_i + b) \geq 1 - \xi_i$. Here w is the normal to the hyperplane and ξ_i is the slack variable, which can be thought of as a measure of minimum classification error that is tolerable. It allows room for error in hyperplane detection when training data is non-separable. By introducing lagrange multiplier the above quadratic optimization, the weight vector w can be expressed as the linear combination of the training data. Thus, $w = \sum_{i=1}^{2*num_{gallery}} \alpha_i x_i y_i$. Thus, the SVM optimization problem is transformed

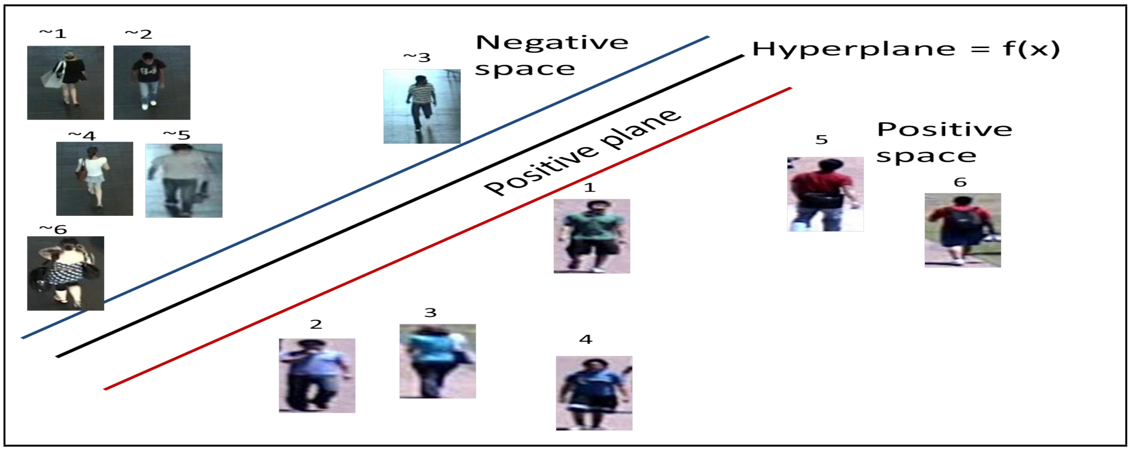


Figure 6.3: Binary SVM: separating hyperplane.

into the following form:

$$\begin{aligned} \max_{\alpha_i \geq 0} \quad & \sum_i \alpha_i - \frac{1}{2} \sum_{(i,j)} \alpha_i \alpha_j y_i y_j x_i^T x_j \\ \text{subject to} \quad & 0 \leq \alpha_i \leq C, and \sum_i \alpha_i y_1 = 0. \end{aligned} \tag{6.3}$$

However, if the training data is not linearly separable, the data must be mapped into some other Hilbert Space (complete inner product space) using a mapping function. The SVM is then trained in the transformed space, i.e. the training data is given by, $\phi(x_i)$, for $i = 1, ..., 2 * num_{gallery}$ and the hyperplane is represented by $f(x) = (< w, \phi(x) > +b)$. If there is a *kernel function* K such as $K(x_i, x_j) = \phi(x_i^T).\phi(x_j)$, we do not compute explicit transform ϕ. Incorporating the kernel function in equation 6.3 becomes:

$$\begin{aligned} \max_{\alpha_i \geq 0} \quad & \sum_i \alpha_i - \frac{1}{2} \sum_{(i,j)} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \\ \text{subject to} \quad & 0 \leq \alpha_i \leq C, and \sum_i \alpha_i y_1 = 0. \end{aligned} \tag{6.4}$$

In this work, we utilize a linear kernel function. This separation hyperplane is well suited to classify the two classes during training but it does not establish hyperplanes that can classify further additional known or unknown classes. The hyperplane location is refined by classification error (empirical error, $R_\epsilon(f)$) minimization on the training data, hence the SVM specializes at classifying IDs that lie on either side of the hyperplane and close to the training IDs in feature space. However, it does not address the open space on positive side of the hyperplane and thus a probe ID will be classified as belonging to the gallery, no matter how far it is from the decision boundary. One-vs-set (oneSet) SVM attempts to limit the distance a probe ID can be from the hyperplane and still be considered as belonging to the gallery.

6.2.2 One-vs-Set SVM for novelty detection

Unlike a binary object recognition problem, the distinction between classes (IDs) in Re-ID is bit more subtle as the boundary between known and unknown IDs is difficult to generalize. For instance, in figure 6.4 the probe IDs denoted by green circles can be classified correctly into either *known* or *unknown* with respect to the gallery IDs used for training as they conform nicely to the detected decision boundary. However, the probe denoted by the red circle is placed on the positive side of the hyperplane in the feature space. The placement in the feature space is correct given the probe feature(color), but incorrect with respect to the learned decision boundary (in ID space). The one-vs-set SVM offers an elegant solution to this problem by better specialization of the space defined by the gallery IDs. The gallery space specialization is treated as a problem of minimization of the risk of placing a gallery ID outside

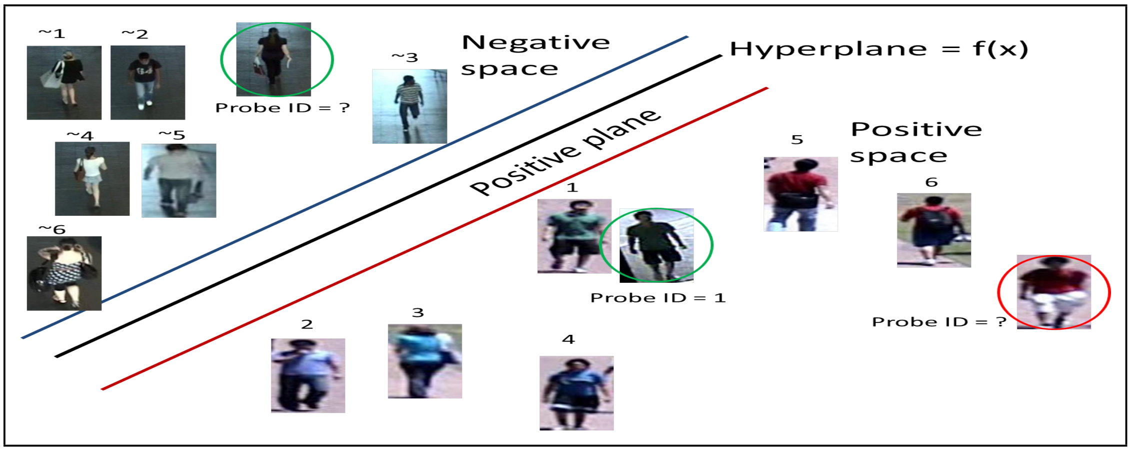


Figure 6.4: Positive open space: all the space on the right side of the positive plane is the positive open space, PO and the space denoted by the red circle is the positively labeled space, PT (defined by the positive labeled training data).

the gallery space. Figure 6.4 better illustrates the proposed solution. A boundary is established around the gallery space to better define it. Only those probe IDs that are placed between *positive plane 1* and *positive plane 2* as classified as *known* IDs that exist in the gallery. As the gallery increases these boundaries are continually readjusted to handle increasing number of unknown class classification and to better generalize the gallery space. Generalization/Specialization of the gallery space is not handled as error minimization of the SVM training function but the after the SVM is trained, the positive hyperplane placements are adjusted by minimization of an error function that combines empirical risk over training data with open space risk. The open space risk is modeled as follows:

$$Risk_{open} = \frac{\delta_\beta - \delta_\alpha}{\delta^+} + \frac{\delta^+}{\delta_\beta - \delta_\alpha} \tag{6.5}$$

where, δ_α and δ_β is the distance from the decision hyperplane to *positive plane* α and *positive plane* β. δ_+ is the separation needed to encompass all the gallery data. During optimization the open space risk is balanced with the empirical risk determined by classifying training data with the separating hyperplane. The optimization process is described in algorithm 1. Here, λ is a regularization constant. This results in both the boundaries of the gallery space on a decision score generated with respect to the hyperplane $f(x)$.

Algorithm 1 Optimization procedure: Gallery Boundary placement

```
procedure TRAIN($\lambda_r, G, NG$)
    train a linear binary SVM using G and NG
    for $x_i$ do
        classify$\hat{y_i} = f(x_i)$                    ▷ Generate decision scores
    end for
    $\hat{s} = sort(\hat{y})$                           ▷ Sort decision scores
    $s_\alpha = min(\forall \hat{s_i} \in f(G))$
    $s_\beta = max(\forall \hat{s_i} \in f(G))$
    positive plane $\alpha$ = marginal plane of $f$
    positive plane $\beta$ = plane parallel to positive plane $\alpha$ at $s_\beta$
    Greedy optimization iteratively to move positive plane $\alpha$ to $s_{\alpha-1}$ or $s_{\alpha+1}$
  and positive plane $\beta$ to $s_{\beta-1}$ or $s_{\beta+1}$ to minimize $Risk_{open}(f) + \lambda R_\epsilon(f(N \cup NG))$.
end procedure
```

6.3 Experimental Results

In order to test the proposed framework we utilize the SAIVT SoftBio dataset. The datasets provides 122 subjects that appear in at least 2 distinct camera views. The reason for selecting this dataset is that along with large number of IDs it also provides multiple frames per subject per camera, which is important for training the novelty

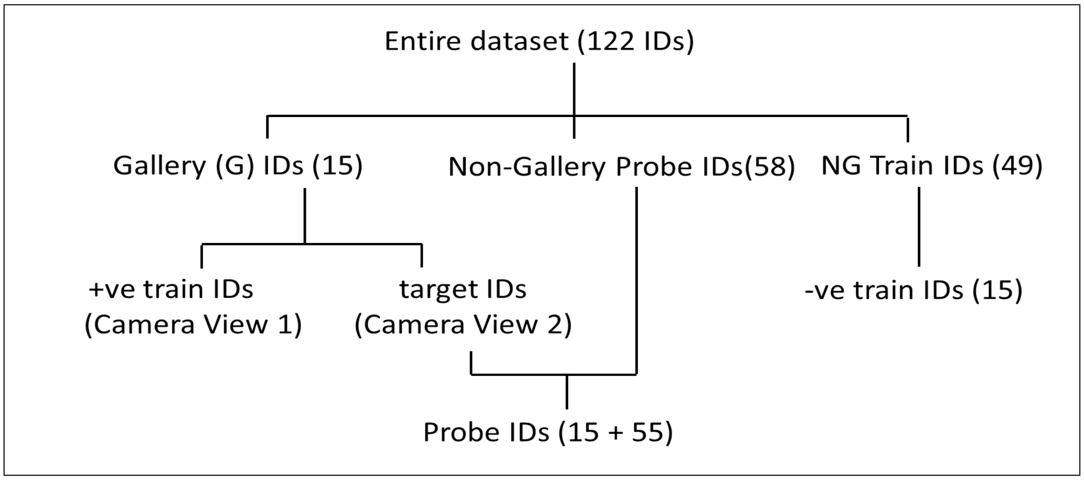


Figure 6.5: Dataset split to train and test the proposed novelty detection SVMs.

detection stage. For each ID, one camera view is used strictly to generate test features and the training features come from a distict camera view.

First, the entire dataset is split into 3 sets to generate the set of negative training IDs(49), non-gallery probe IDs(58) and gallery IDs (15). Out of the 49 negative training IDs, 15 (one negative ID per gallery ID) are used as the negative IDs to train the novelty detection stage. Thus, the training set (x) consists of 30 IDs from one camera view: 15 IDs that belong to the gallery (positive) and 15 negative IDs that are not part of the gallery. For each ID, 35 images are used for training, this means that per gallery ID 35 positive and negative samples are available for training. Both the binary SVM and oneSet SVM are trained using the same training data. Figure 6.5 pictorially depicts the dataset split.

The probe set always consists of the exact same gallery (known) IDs, i.e., the target IDs, except the probe images come from a different camera view compared to the gallery view. This probe set is used to test the closed set Re-ID scenario. However, the probe set increases in openness as progressively more probe IDs are added from the non-gallery probe IDs set. The number of probe IDs goes from 15 (same as gallery IDs) to 70 (15 gallery and 55 non-gallery IDs). Thus, the probe openness goes 0%, 7% to 40%. The gallery set is always fixed at 15 IDs. This experiment is randomly repeated 5 times with distinct 3 way split of the 122 IDs. For every probe ID, 20 images are available to the novelty detection stage for classification. A probe ID is label as belonging to a particular class, if more than 80% of the probe frames are classified by the novelty detection stage as belonging to that class.

The novelty detection stage is evaluated using two metrics: accuracy and F-measure as per [119]. Accuracy is defined as the correctly classified IDs divided by the total IDs presented to the novelty detection stage. True positives (TPs) are defined as the number of probe IDs (that truly exist in gallery) that are detected as gallery IDs. True negatives (TNs) are defined as the number of probe IDs (that do not exist in gallery) that are detected as non-gallery IDs. False positives (FPs) are defined as the number of probe IDs (that do not exist in gallery) that are classified as gallery IDs. False negatives (FNs) are defined as the number of probe IDs (that truly exist in gallery) that are classified as non-gallery IDs. The accuracy and F-measure are given as: $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$ and $F - measure = 2\frac{Precision*Recall}{Precision+Recall}$, $Precision = \frac{TP}{TP+FP}$ and $Recall = \frac{TP}{TP+FN}$. Figure 6.6 shows the changes in novelty detection performance with increasing probe set openness. The overall accuracy of classification using oneSet SVM is slightly better than binary SVM. Binary SVM gives better accuracy when the probe set of closed, which implies that it has ab better true positive rate. However, as the probe openness increases the accuracy of oneSet SVM increases slighly faster than binary SVM. This implies that the oneSet SVM is more robust to probe openness than binary SVM as it is better at true negatives or identifying *novel* IDs. The F-measure of both SVMs are comparable but as the probe openness increases, we can see the same pattern as in the accuracy graph.

The objective of novelty detection stage is to correctly identify if the probe ID exist in the gallery or not, in order to understand this better figure 6.7 shows the TNs(accuracy of identifying a *novel* ID) and FPs (error in *novel* ID identification) of the oneSet and binary SVM with respect to gallery openness. The oneSet SVM does consistently better in both metrics with probe openness. Thus, it is reasonable to conclude that the bounding of the gallery space does result in better novelty detection performance. Figure 6.8 examples TP, TN and FP IDs with respect to the shown gallery IDs using both SVMs. One important observation is that the TPs of the novelty detection stage can be improved significantly by using more discriminative features other than color to better specialize the SVMs. In order to understand the impact of the novelty detection stage on Re-ID performance we take the probe set IDs classified as known by both SVMs and present them to gallery ranking stage to achieve Re-ID. The rank 1 matched gallery ID is assigned to each probe ID. We compare the performance with Re-ID by ranking without the novelty detection stage to further highlight the need for novelty detection as the probe set openness affects the Re-ID rate (accuracy). For completeness purposes, we also compare a method that incorporates a threshold on the gallery-probe similarity to reject false matches. The threshold is varied from 0.1 to 0.95 in increments of 0.5, we only show the Re-ID performance with the threshold that performs the best, which empirically was found to be 0.2. Figure 6.9 shows the Re-ID performance with and without the novelty detection stage. For Re-ID, true positives (TPs) are the number of probe IDs that are correctly matched (Rank 1 matching rate). Mismatches (MMs) are the number of probe IDs that are incorrectly matched to gallery, when that probe ID does exist in the gallery. False positives (FPs) are the number of probe IDs that are matched to the gallery when the probe ID does not exist in the gallery. The accuracy (Re-ID rate) and FAR are defined as $Accuracy = \frac{TPs+TNs}{N_P}$ and $FAR = \frac{(MMs+FPs)}{N_P}$, respectively, where N_P denotes the total number of probe subjects. From the figure it is evident that incorporating the proposed (binary or oneSet) novelty detection boosts the Re-ID rate significantly. For the closed set scenario, gallery ranking based Re-ID rate with or without a threshold is better than Re-ID with novelty detection. However, as probe openness increases this performance advantage is reversed significantly.

Another note worthy point is, as the FAR obtained using the threshold based false rejection technique is much lower than without it. However, either of the novelty detection SVMs produce a significantly lower FAR. Thus, we can safely deduce that either binary or oneSet SVM based novelty detection is a good first step towards open set Re-ID. Further, the oneSet SVM based novelty detection better handles probe openness than binary SVM and hence enables better open set Re-ID performance.

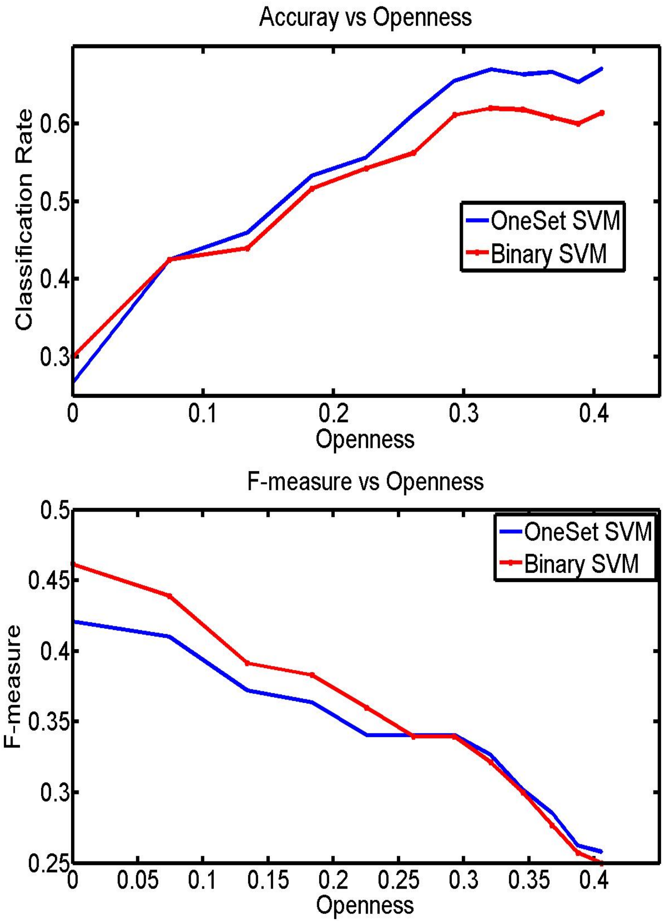


Figure 6.6: Novelty Detection Stage Performance: Top row shows the changes in classification accuracy with the probe set openness and the bottow row images shows the F-measure.

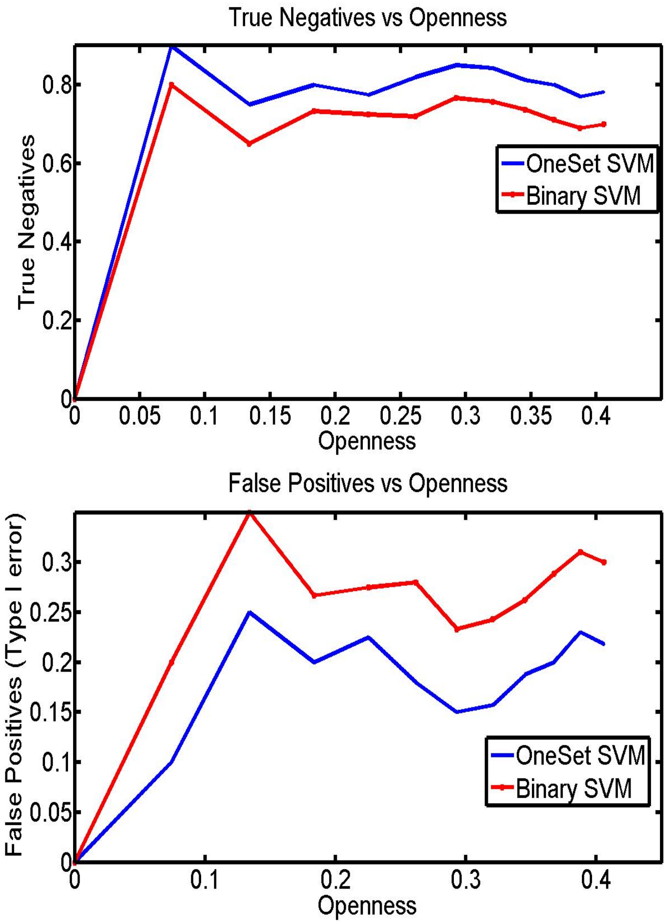


Figure 6.7: Novelty Detection Stage Performance: Top and bottom row show evolution of TNs and FPs, respectively, with the probe set openness.

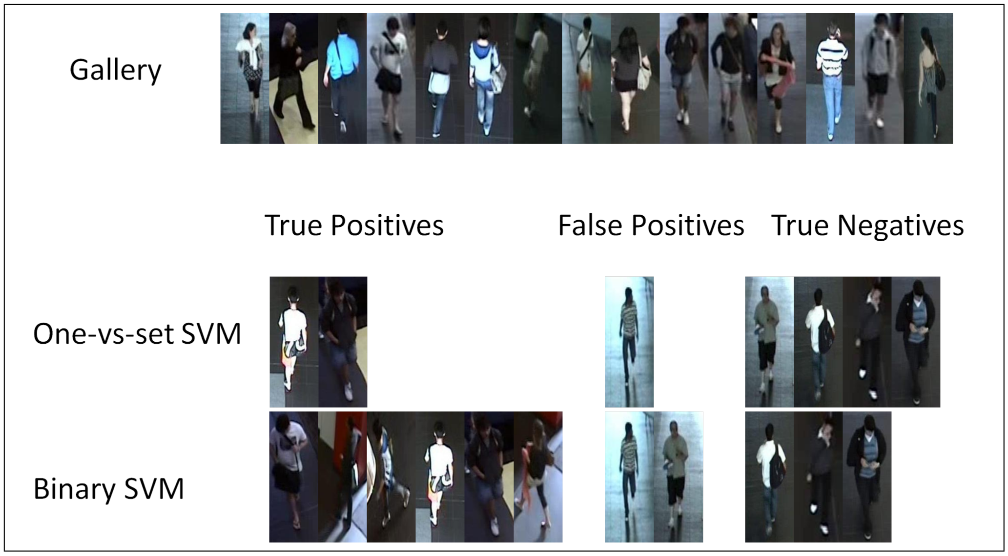


Figure 6.8: Novelty Detection Output: Comparison of the classification performance of both the oneSet and binary SVMs for probe openess = 7% case, i.e. probe set consists of all 15 gallery IDs plus 5 unknown IDs.

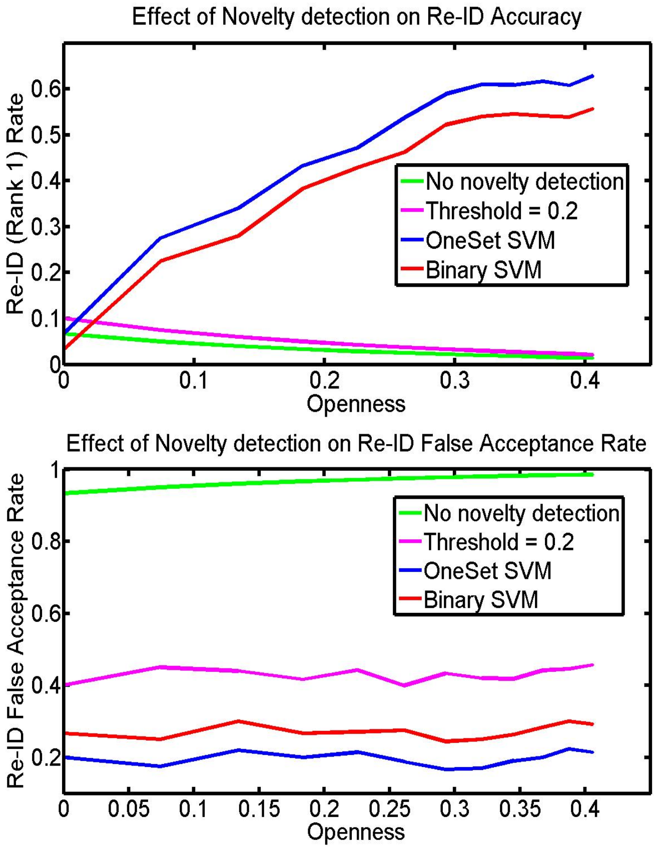


Figure 6.9: Open-Set Re-ID: Top row shows the changes in Re-ID rate (accuracy) with the probe set openness and the bottom row images shows the Re-ID false acceptance rate.

Chapter 7

Conclusion and Future Work

7.1 Summary of Current Work

We have proposed two person models for Re-ID: 1) A spatio-temporal model based on color and facial features that captures complementary aspects of a person's appearance, and 2) A combined color and gait features model that attempts to characterize a person with behavioral features. Both these models are based on the assumption that multiple frames of a person are available. The color model is suited for both single and multi-shot Re-ID. A strategy for multiple person re-identification based on the rectangular assignment problem was presented. Two strategies to address open set Re-ID were also proposed: 1) Threshold applied to person model similarity, and 2) SVM based novelty detection. The key contributions of our work are as follows:

1. An adaptive part-based spatio-temporal model based on color and facial features was proposed.

2. A principled strategy for multiple-person re-identification based on rectangular assignment problem was proposed.

3. A person-model similarity based false acceptance reduction criteria was developed.

4. Investigated if gait extracted from real world video sequences can be successfully leveraged as an additional feature along with appearance features for person Re-ID in the context of surveillance.

5. Identified robust gait features that can be extracted from noisy and incomplete silhouettes and yet retain discriminative capability for short and long Re-ID.

6. Proposed a SVM-based novelty detection technique to address open set Re-ID. The novelty detection incorporated Re-ID significantly boosts open set Re-ID performance.

We would also like to point out that, to the best of our knowledge, this is the first proposed work that offers strategies for principled multiple person re-identification, and novelty detection for open set Re-ID. This is also the first proposed work that combines facial and gait features with clothing-based appearance features to boost Re-ID accuracy. The potential of gait as an independent feature for long period Re-ID was also demonstrated.

7.2 Future Work

In this work, we demonstrated the potential of facial and gait features as a means to boost Re-ID performance. Thus, they are still applicable to short period Re-ID. In order to better understand the conditions under which these features can be reliably applied to Re-ID some more work is required. Further, we have simply scratched the surface in terms of open set Re-ID. Following are some of the future directions we take from this work:

- Use machine learning techniques to better estimate the contribution of color, face and gait features to Re-ID accuracy for effective combination of these features.
- Explore strategies for fusion of all three features into a single combined model.
- Quantify the conditions like pose changes, view angle variations, scale and resolution that make features like face or gait reliable for Re-ID.
- Use non-clothing based gait and facial features to train the novelty detection stage to improve open set Re-ID.

7.2.1 Open Issues in Person Re-ID

As is evident, most of the work on person Re-ID leverages clothing appearance based features designed for short-period Re-ID and is evaluated in closed set Re-ID scenarios. The issue of long-period Re-ID is entirely unexplored and open set Re-ID is not completely tackled. In order to boost Re-ID performance, different sensors like RGB-D [3, 15] and infrared [76] that capture soft biometric cues insensitive to appearance variations are being explored. Some other variations of Re-ID are verification and searching of people based on textual queries [117, 55]. For the verification task, the system is presented by a probe that claims an ID and the system has to decide if the probe ID is the same ID that is claimed [150]. Large amounts of video data can be searched at high speeds using textual queries. Law enforcement agencies can utilize such systems for surveillance or forensic purposes. For evaluation of these Re-ID applications more specialized measures might be necessary.

The focus of current work in Re-ID is geared towards robust descriptors and effective matching schemes but the issue of scalability is often overlooked. Scalability refers to the ability of the system to adapt itself to realistically varying factors while maintaining the performance. The following scalability issues need further research to address the specified shortcomings:

- In real-world applications the gallery size is large and constantly increasing. The common similarity based ranking techniques do not scale well and hence efficient matching schemes need to be explored.
- As the gallery is ever changing, new models are added, learning based Re-ID techniques like classifiers, bag-of-words or distance metric optimization need to be re-calibrated in order to incorporate the variability in the gallery set to maintain their performance.
- In order to maximize uniqueness, descriptors are often complex, high-dimensional

and expensive to extract. This also makes the recognition process compute intensive and complicated. These factors affect the temporal complexity of the system making real time performance difficult to achieve.

- Large gallery sizes and high-dimensional models require large amounts of storage space and computational resources to effectively analyze the data.

- Automated video analytics can be simplified by on-camera data processing (smart cameras) and communications between cameras. However, storage and computational resource intensive Re-ID systems cannot be easily scaled to work with low power processors and narrow bandwidth transmission channels.

- All of the current approaches to Re-ID assume accurate person detection/tracking prior to feature extraction. A rigorous analysis of effects of detection/t racking errors on Re-ID performance has, to the best of our knowledge, not been performed.

- Following novelty detection, enrolling new subjects in the gallery is non-trivial. Issues like quality of model, reconciling models from several cameras and effect of Re-ID errors on gallery enrollment require further investigation.

Consideration of scalability issues within Re-ID research can lead to better designed and more efficient systems. Most Re-ID systems produce a ranked list of gallery, but this list might need to be refined by a human to boost the accuracy of the ranking. As the gallery size increases this becomes more difficult to achieve. Thus, efficient re-ranking schemes based on human input need to be addressed [87].

Person Re-ID is a very challenging task with wide ranging application in numerous fields. It has received a lot of attention lately and the Re-ID models and recognition techniques have come a long way but are still very narrow and specific in their application to real world problems. The most obvious next step in development of unique models is the incorporation of biometric cues. Semantic information involving human visual system based perceptual attributes can provide valuable descriptive ability to the models. Building hierarchical models that incorporate relationships between low level features and high level semantics would yield more coherent descriptors. The models should be designed keeping in mind the complexity of feature extraction and their storage footprints. Hierarchical models can be used to significantly alleviate the search space within the gallery. This will greatly ease out the scalability issues as well as reduce the compute intensive nature of recognition.

In summary, person Re-ID is a broad and challenging field with vast opportunities for improvements and research. This work attempts to provide an overview of the Re-ID problem, its challenges, propose solutions to short and long period Re-ID, provide a starting point for open set Re-ID and present areas of future exploration.

Bibliography

[1] Video surveillance and monitoring homepage.

[2] Britain is 'surveillance society'. BBC News, 2006.

[3] A. Albiol, A. Albiol, J. Oliver, and J. Mossi. Who is who at different cameras: people re-identification using depth cameras. *IET Computer Vision*, 6(5):378–387, September 2012.

[4] M. Andriluka, S. Roth, and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2008.

[5] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009*, 2009.

[6] M. Ao, D. Yi, Z. Lei, and S. Li. Face recognition at a distance: System issues. In M. Tistarelli, S. Li, and R. Chellappa, editors, *Handbook of Remote Biometrics*, chapter 6, pages 155–167. Springer London, 2009.

[7] T. Avraham, I. Gurvich, M. Lindenbaum, and S. Markovitch. Learning implicit transfer for person re-identification. In *Proceedings of the 12th European conference on Computer Vision. ECCV Workshops*, pages 381–390, 2012.

[8] S. Bak, G. Charpiat, E. Corvée, F. Brémond, and M. Thonnat. Learning to match appearances by correlations in a covariance metric space. In *Proceedings of the European conference on Computer Vision*, pages 806–820, 2012.

[9] S. Bak, E. Corvee, F. Bremond, and M. Thonnat. Person re-identification using haar-based and dcd-based signature. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 1–8, 2010.

[10] S. Bak, E. Corvee, F. Bremond, and M. Thonnat. Person re-identification using spatial covariance regions of human body parts. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 435–440, 2010.

[11] S. Bak, E. CorvéE, F. BréMond, and M. Thonnat. Boosted human re-identification using riemannian manifolds. *Image Vision Computing*, 30(6-7):443–452, June 2012.

[12] D. Baltieri, R. Vezzani, and R. Cucchiara. 3dpes: 3d people dataset for surveillance and forensics. In *Proceedings of the joint ACM workshop on Human gesture and behavior understanding*, pages 59–64, 2011.

[13] D. Baltieri, R. Vezzani, and R. Cucchiara. Sarc3d: A new 3d body model for people tracking and re-identification. In *International Conference on Image Analysis and Processing*, pages 197–206, 2011.

[14] D. Baltieri, R. Vezzani, R. Cucchiara, . Utasi, C. Benedek, and T. Szirnyi. Multi-view people surveillance using 3d information. In *The Eleventh International Workshop on Visual Surveillance (in conjunction with ICCV 2011).*, 2011.

[15] I. Barbosa, M. Cristani, A. D. Bue, L. Bazzani, and V. Murino. Re-identification with rgb-d sensors. In *Proceedings of the 12th European conference on Computer Vision. ECCV Workshops*, pages 433–442, 2012.

[16] K. Bashir, T. Xiang, and S. Gong. Cross view gait recognition using correlation strength. In *Procedings of the British Machine Vision Conference 2010*, 2010.

[17] M. Bauml, K. Bernardin, M. Fischer, H. Ekenel, and R. Stiefelhagen. Multi-pose face recognition for person retrieval in camera networks. In *Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2010*, 2010.

[18] M. Bauml and R. Stiefelhagen. Evaluation of local features for person re-identification in image sequences. In *International Conference on Advanced Video and Signal-Based Surveillance*, pages 291–296, 2011.

[19] L. Bazzani, M. Cristani, and V. Murino. Symmetry-driven accumulation of local features for human characterization and re-identification. *Computer Vision and Image Understanding*, 117(2):130–144, 2013.

[20] L. Bazzani, M. Cristani, A. Perina, M. Farenzena, and V. Murino. Multiple-shot person re-identification by hpe signature. In *International Conference on Pattern Recognition*, pages 1413–1416, 2010.

[21] L. Bazzani, M. Cristani, A. Perina, and V. Murino. Multiple-shot person re-identification by chromatic and epitomic analyses. *Pattern Recognition Letters*, 33(7):898–903, May 2012.

[22] A. Bedagkar-Gala and S. K. Shah. Part-based spatio-temporal model for Multi-Person re-identification. *Pattern Recognition Letters*, September 2011.

[23] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997. Special Issue on Face Recognition.

[24] N. Bellotto, E. Sommerlade, B. Benfold, C. Bibby, I. Reid, D. Roth, C. Fernández, L. V. Gool, and J. Gonzàlez. A distributed camera system for multi-resolution surveillance. In *Proceedings of the 3rd ACM/IEEE Int. Conf. on Distributed Smart Cameras (ICDSC)*, 2009.

[25] A. Bhattacharyya. On a measure of divergence between two multinomial populations. *Sankhy: The Indian Journal of Statistics (1933-1960)*, 7(4):401–406, 1946.

[26] A. Bialkowski, S. Denman, S. Sridharan, C. Fookes, and P. Lucey. A database for person re-identification in multi-camera surveillance networks. In *Digital Image Computing: Techniques and Applications*, pages 1–8, 2012.

[27] A. Bialkowski, S.Denman, S. Sridharan, C. Fookes, and P. Lucey. A database for person re-identification in multi-camera surveillance networks. In *International Conference on Digital Image Computing Techniques and Applications*, pages 1–8, 2012.

[28] P. Bodesheim, A. Freytag, E. Rodner, M. Kemmler, and J. Denzler. Kernel null space methods for novelty detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013*, pages 3374–3381, June 2013.

[29] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *International Conference on Computer Vision*, 2009.

[30] T. Bouwmans. Recent advanced statistical background modeling for foreground detection: A systematic survey. *Recent Patents on Computer Science*, 4(3), 2011.

[31] Y. Brand, T. Avraham, and M. Lindenbaum. Transitive re-identification. In *British Machine Vision Conference*, 2013.

[32] L. Brun, D. Conte, P. Foggia, and M. Vento. People re-identification by graph kernels methods. In *Proceedings of the 8th international conference on Graph-based representations in pattern recognition*, pages 285–294, 2011.

[33] Y. Cai and M. Pietikinen. Person re-identification based on global color context. In *The Tenth International Workshop on Visual Surveillance (in conjunction with ACCV 2010).*, pages 205–215, 2010.

[34] 2004.

[35] C. Chen, J. Liang, H. Zhao, H. Hu, and J. Tian. Frame difference energy image for gait recognition with incomplete silhouettes. *Pattern Recognition Letters*, 30(11):977–984, August 2009.

[36] D. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murinog. Custom pictorial structures for re-identification. In *Proceedings of the British Machine Vision Conference*, pages 68.1–68.11, 2011.

[37] D.-N. T. Cong, C. Achard, and L. Khoudour. People re-identification by classification of silhouettes based on sparse representation. In *2nd International Conference onImage Processing Theory Tools and Applications*, pages 60–65, 2010.

[38] D. N. T. Cong, L. Khoudour, C. Achard, C. Meurie, and O. Lezoray. People re-identification by spectral classification of silhouettes. *Signal Processing*, 90(8):2362–2374, Aug. 2010.

[39] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 1998.

[40] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005.

[41] A. D'angelo and J.-L. Dugelay. People re-identification in camera networks based on probabilistic color histograms. In *Electronic Imaging Conference on 3D Image Processing and Applications*, volume 7882, pages 78820K–78820K–12, 2011.

[42] H. Detmold, A. van den Hengel, A. Dick, A. Cichowski, R. Hill, E. Kocadag, K. Falkner, and D. Munro. Topology estimation for thousand-camera surveillance networks. In *First ACM/IEEE International Conference on Distributed Smart Cameras*, pages 195–202, 2007.

[43] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja. Pedestrian recognition with a learned metric. In *Asian Conference in Computer Vision (ACCV), 2010*, 2010.

[44] A. Ess, B. Leibe, and L. V. Gool. Depth and appearance for mobile scene analysis. In *International Conference on Computer Vision*, pages 1–8, 2007.

[45] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results. http://www.pascal-network.org/challenges/VOC/voc2008/workshop/index.html.

[46] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1778–1785, 2009.

[47] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61:2005, 2003.

[48] P. F. Felzenszwalb, D. A. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[49] M. A. T. Figueiredo and A. K. Jain. Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):381–396, 2002.

[50] P.-E. Forssén. Maximally stable colour regions for recognition and matching. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2007.

[51] N. Gagvani. Challenges in video analytics. In *Embedded Computer Vision*, Advances in Pattern Recognition. Springer London, 2009.

[52] A. Gallagher and T. Chen. Clothing cosegmentation for recognizing people. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[53] T. Gandhi and M. M. Trivedi. Person tracking and reidentification: Introducing panoramic appearance map (pam) for feature representation. *Machine Vision and Applications*, 18, May 2007.

[54] N. Gheissari, T. Sebastian, and R. Hartley. Person reidentification using spatiotemporal appearance. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1528–1535, 2006.

[55] D. Grangier and S. Bengio. A discriminative kernel-based approach to rank images from text queries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(8):1371–1384, August 2008.

[56] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, 2007.

[57] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European Conference on Computer Vision*, pages 262–275, 2008.

[58] M. W. Green. The appropriate and effective use of security technologies in u.s. schools, a guide for schools and law enforcement agencies. National Institue of Justice, Research Report, September 1999.

[59] L. Greenemeier. The apple of its eye: Security and surveillance pervades post-9/11 new york city. Scientific American, 2011.

[60] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux. Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In *2nd ACM/IEEE International Conference on Distributed Smart Cameras*, pages 1–6, 2008.

[61] A. Hampapur, L. Brown, J. Connell, S. Pankanti, A. Senior, and Y. Tian. Smart surveillance: Applications, technologies and implications. In *In IEEE Pacific-Rim Conference On Multimedia*, 2003.

[62] A. Hampapur, S. Pankanti, A. Senior, Y.-L. Tian, L. Brown, and R. Bolle. Face cataloger: Multi-scale imaging for relating identity to location. In *Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS) 2003*, 2003.

[63] J. Han and B. Bhanu. Individual recognition using gait energy image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:316–322, February 2006.

[64] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang. Face recognition using laplacianfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence,*, 27(3):328–340, March 2005.

[65] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *Proceedings of the 17th Scandinavian conference on Image analysis*, pages 91–102, 2011.

[66] M. Hirzer, P. Roth, and H. Bischof. Person re-identification by efficient impostor-based metric learning. In *IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance*, pages 203–208, 2012.

[67] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *Proceedings of the 12th European conference on Computer Vision. ECCV Workshops*, pages 780–793, 2012.

[68] H. Hotelling. Relations between two sets of variates. In S. Kotz and N. L.Johnson, editors, *Breakthroughs in Statistics*, Springer Series in Statistics, pages 162–190. Springer New York, 1992.

[69] M. Hu, Y. Wang, Z. Zhang, and Z. Zhang. Multi-view multi-stance gait identification. In *IEEE International Conference on Image Processing (ICIP), 2011*, pages 541–544, Sept 2011.

[70] W. Hu, M. Hu, X. Zhou, T. Tan, J. Lou, and S. Maybank. Principal axis-based correspondence between multiple cameras for people tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28, April 2006.

[71] O. Javed, K. Shafique, Z. Rasheed, and M. Shah. Modeling inter-camera space-time and appearance relationships for tracking across non-overlapping views. *Computer Vision and Image Understanding*, 2008.

[72] O. Javed, K. Shafique, and M. Shah. Appearance modeling for tracking in multiple non-overlapping cameras. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 26–33, 2005.

[73] C. Jaynes, A. Kale, N. Sanders, and E. Grossmann. The terrascope dataset: scripted multi-camera indoor video surveillance with ground-truth. In *Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 309–316, 2005.

[74] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 133–142, 2002.

[75] K. Jungling, , C. Bodensteiner, and M. Arens. Person re-identification in multi camera networks. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2011.

[76] K. Jungling and M. Arens. Local feature based person reidentification in infrared image sequences. In *Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 448–455, 2010.

[77] J. Kang, I. Cohen, and G. Medioni. Object reacquisition using invariant appearance model. In *Proceedings of International Conference on Pattern Recognition*, volume 4, pages 759–762, 2004.

[78] M. Kemmler, E. Rodner, E.-S. Wacker, and J. Denzler. One-class classification with gaussian processes. *Pattern Recognition*, 46(12):3507–3518, 2013.

[79] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *IEEE International Conference on Computer Vision*, pages 365–372, 2009.

[80] W. Kusakunniran, Q. Wu, H. Li, and J. Zhang. Multiple views gait recognition using view transformation model based on optimized gait energy image. In *IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops), 2009*, 2009.

[81] M. Lantagne, M. Parizeau, and R. Bergevin. Vip: Vision tool for comparing images of people. *Vision Interface*, 2003.

[82] R. Layne, T. M. Hospedales, and S. Gong. Re-identification by attributes. In *Proceedings of the British Machine Vision Conference*, pages 24.1–24.11, 2012.

[83] R. Layne, T. M. Hospedales, and S. Gong. Towards person identification and re-identification with attributes. In *Proceedings of the 12th European conference on Computer Vision. ECCV Workshops*, pages 402–412, 2012.

[84] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *European Conference on Computer Vision*, 2004.

[85] W. Li and X. Wang. Locally aligned feature transforms across views. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3594–3601, 2013.

[86] C. Liu, S. Gong, C. C. Loy, and X. Lin. Person re-identification: What features are important? In *Proceedings of the 12th European conference on Computer Vision. ECCV Workshops*, pages 391–401, 2012.

[87] C. Liu, C. C. Loy, S. Gong, and G. Wang. Pop: Person re-identification post-rank optimisation. In *IEEE International Conference on Computer Vision*, December 2013.

[88] J. Liu and N. Zheng. Gait history image: A novel temporal template for gait recognition. In *IEEE International Conference on Multimedia and Expo, 2007*, pages 663–666, July 2007.

[89] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004.

[90] C. Loy, T. Xiang, and S. Gong. Multi-camera activity correlation analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1988–1995, June 2009.

[91] C. Loy, T. Xiang, and S. Gong. Time-delayed correlation analysis for multi-camera activity understanding. *International Journal on Computer Vision*, 90(1):106–129, October 2010.

[92] C. Loy, T. Xiang, and S. Gong. Incremental activity modeling in multiple disjoint cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1799–1813, 2012.

[93] C. C. Loy, C. Liu, and S. Gong. Person re-identification by manifold ranking. In *IEEE International Conference on Image Processing*, 2013.

[94] B. Ma, Y. Su, and F. Jurie. Bicov: a novel image representation for person re-identification and face verification. In *Proceedings of the British Machive Vision Conference*, pages 57.1–57.11, 2012.

[95] B. Ma, Y. Su, and F. Jurie. Local descriptors encoded by fisher vectors for person re-identification. In *Proceedings of the 12th European conference on Computer Vision. ECCV Workshops,*, pages 413–422, 2012.

[96] D. Makris, T. Ellis, and J. Black. Bridging the gaps between cameras. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2004.

[97] M. Markou and S. Singh. Novelty detection: a reviewpart 1: statistical approaches. *Signal Processing*, 83(12):2481–2497, 2003.

[98] M. Markou and S. Singh. Novelty detection: a reviewpart 2:: neural network based approaches. *Signal Processing*, 83(12):2499–2521, 2003.

[99] R. Mazzon, S. F. Tahir, and A. Cavallaro. Person re-identification in crowd. *Pattern Recognition Letters*, 33(14):1828–1837, October 2012.

[100] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 27(10):1615–1630, 2005.

[101] J. Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society of Industrial and Applied Mathematics*, 5(1):32–38, 1957.

[102] M. Naylor and C. Atwood. Final report, annotated digital video for intelligent surveillance and optimised retrieval (advisor), (ist1999-11287), advisor consortium., 2007.

[103] U. H. Office. i-lids multiple camera tracking scenario definition, 2007.

[104] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions onPattern Analysis and Machine Intelligence*, 24(7):971–987, jul 2002.

[105] K. Okuma, A. Taleghani, N. D. Freitas, J. J. Little, and D. G. Lowe. A boosted particle filter: Multitarget detection and tracking. In *European Conference on Computer Vision*, 2004.

[106] D. Parikh and K. Grauman. Relative attributes. In *IEEE International Conference on Computer Vision*, pages 503–510, 2011.

[107] U. Park, A. Jain, I. Kitahara, K. Kogure, and N. Hagita. Vise: Visual search engine using multiple networked cameras. In *18th International Conference on Pattern Recognition (ICPR) , 2006*, 2006.

[108] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *IEEE International Conference on Computer Vision*, 2009.

[109] F. Porikli. Inter-camera color calibration by correlation model function. In *International Conference on Image Processing (ICIP), 2003*, 2003.

[110] S. J. D. Prince, J. Elder, Y. Hou, M. Sizinstev, and E. Olevsky. Towards face recognition at a distance. In *The Institution of Engineering and Technology Conference on Crime and Security.*, pages 570–575, June 2006.

[111] B. Prosser, S. Gong, and T. Xiang. Multi-camera matching using bi-directional cumulative brightness transfer functions. In *Proceedings of the British Machive Vision Conference*, pages 64.1–64.10, 2008.

[112] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by support vector ranking. In *Proceedings of the British Machine Vision Conference*, pages 21.1–21.11, 2010.

[113] A. Rahimi, B. Dunagan, and T. Darrell. Simultaneous calibration and tracking with a network of non-overlapping sensors. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 187–194, 2004.

[114] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth movers distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[115] R. Satta. Appearance descriptors for person re-identification: a comprehensive review. *CoRR*, 2013.

[116] R. Satta, G. Fumera, and F. Roli. Fast person re-identification based on dissimilarity representations. *Pattern Recognition Letters*, 33(14):1838–1848, October 2012.

[117] R. Satta, G. Fumera, and F. Roli. A general method for appearance-based people search based on textual queries. In *Proceedings of the 12th European conference on Computer Vision. ECCV Workshops*, pages 453–461, 2012.

[118] R. Satta, G. Fumera, F. Roli, M. Cristani, and V. Murino. A multiple component matching framework for person re-identification. In *Proceedings of the 16th international conference on Image analysis and processing*, pages 140–149, 2011.

[119] W. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, July 2013.

[120] B. Schölkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor, and J. C. Platt. Support vector method for novelty detection. In S. Solla, T. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 582–588. MIT Press, 2000.

[121] D. Schorn. We're watching. CBS News, 2009.

[122] W. R. Schwartz and L. S. Davis. Learning discriminative appearance-based models using partial least squares. In *Proceedings of the XXII Brazilian Symposium on Computer Graphics and Image Processing*, pages 322–329, 2009.

[123] D. Simonnet, M. Lewandowski, S. A. Velastin, J. Orwell, and E. Turkbeyler. Re-identification of pedestrians in crowds using dynamic time warping. In *Proceedings of the 12th European conference on Computer Vision. ECCV Workshops*, pages 423–432, 2012.

[124] J. Sivic, C. L. Zitnick, and R. Szeliski. Finding people in repeated shots of the same scene. In *Proceedings of the British Machine Vision Conference*, 2006.

[125] P. Tu, G. Doretto, N. Krahnstoever, A. G. A. Perera, F. Wheeler, X. Liu, J. Rittscher, T. Sebastian, T. Yu, and K. Harding. An intelligent video framework for homeland protection. In *Proceedings of SPIE Defence and Security Symposium - Unattended Ground, Sea, and Air Sensor Technologies and Applications IX*, 2007.

[126] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, Jan. 1991.

[127] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010.

[128] J. Wang, K. Markert, and M. Everingham. Learning models for object recognition from natural language descriptions. In *Proceedings of British Machine Vision Conference*, pages 1–11, 2009.

[129] J. Wang, M. She, S. Nahavandi, and A. Kouzani. A review of vision-based gait recognition methods for human identification. In *International Conference on Digital Image Computing: Techniques and Applications (DICTA), 2010*, pages 320–327, December 2010.

[130] L. Wang, T. Tan, H. Ning, and W. Hu. Silhouette analysis-based gait recognition for human identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1505 – 1518, December 2003.

[131] S. Wang, M. Lewandowski, J. annesley, and james Orwell. Re-identification of pedestrians with variable occlusion and scale. In *The Eleventh International Workshop on Visual Surveillance (in conjunction with ICCV 2011).*, 2011.

[132] X. Wang. Intelligent multi-camera video surveillance: A review. *Pattern Recognition Letters*, 34(1):3–19, 2013.

[133] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu. Shape and appearance context modeling. In *International Conference on Computer Vision*, pages 1–8, 2007.

[134] Y. Wang and G. Mori. A discriminative latent model of object classes and attributes. In *Proceedings of the 11th European Conference on Computer Vision*, pages 155–168, 2010.

[135] F. Wheeler, R. Weiss, and P. Tu. Face recognition at a distance system for surveillance applications. In *Fourth IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS), 2010*, 2010.

[136] Y. Wong, S. Chen, S. Mau, C. Sanderson, and B. C. Lovell. Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition. In *Computer Vision and Pattern Recognition Workshops*, 2011.

[137] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, 2009.

[138] B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *International Journal of Computer Vision*, 75:247–266, 2007.

[139] M. Wu and J. Ye. A small sphere and large margin approach for novelty detection using training data with outliers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):2088–2092, Nov. 2009.

[140] Y. Wu, M. Minoh, M. Mukunoki, and S. Lao. Set based discriminative ranking for recognition. In *Proceedings of the 12th European conference on Computer Vision. ECCV Workshops*, pages 497–510, 2012.

[141] A. Yang, J. Wright, Y. Ma, and S. Sastry. Feature selection in face recognition: A sparse representation perspective. Technical report, University of Illinois, 2007.

[142] J. Yang and Y. Zhang. Alternating direction algorithms for l1-problems in compressive sensing. Technical report, Rice University, 2009.

[143] L. Yang and R. Jin. Distance metric learning: a comprehensive survey. Technical report, Michigan State University, 2006.

[144] A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Computing Surveys*, 38(4):13+, Dec. 2006.

[145] Y. Zhang and S. Li. Gabor-lbp based region covariance descriptor for person re-identification. In *Sixth International Conference on Image and Graphics*, pages 368–371, 2011.

[146] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In *IEEE International Conference on Computer Vision*, December 2013.

[147] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3586–3593, 2013.

[148] W. Zhao, R. Chellappa, J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, 35(4):399–458, Dec. 2003.

[149] W.-S. Zheng, S. Gong, and T. Xiang. Associating groups of people. In *Proceedings of the British Machive Vision Conference*, pages 23.1–23.11, 2009.

[150] W.-S. Zheng, S. Gong, and T. Xiang. Transfer re-identification: From person to set-based verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2650–2657, 2012.

[151] W.-S. Zheng, S. Gong, and T. Xiang. Reidentification by relative distance comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):653–668, 2013.

[152] S. Zhou, V. Krueger, and R. Chellappa. Face recognition from video: A condensation approach. *Computer Engineering*, 2002.