

# **Resource-constrained re-identification in camera networks**

by

Syed Fahad Tahir

BS in Computer Sciences 2004

MS in Computer System Engineering 2006

A dissertation submitted to

The School of Electronic Engineering and Computer Science

in partial fulfilment of the requirements for the Degree of

Doctor of Philosophy

in the subject of

Interactive and Cognitive Environments

Queen Mary University of London

Mile End Road

E1 4NS, London, UK

February 2015



## Acknowledgements

This PhD Thesis has been developed in the framework of, and according to, the rules of the Erasmus Mundus Joint Doctorate in Interactive and Cognitive Environments EMJD ICE [FPA n° 2010-0012] with the cooperation of the following Universities:



Alpen-Adria-Universität Klagenfurt – AAU



Queen Mary, University of London – QMUL



Technische Universiteit Eindhoven – TU/e



Università degli Studi di Genova – UNIGE



Universitat Politècnica Catalunya – UPC

According to ICE regulations, the Italian PhD title has also been awarded by the Università degli Studi di Genova.

## **Acknowledgements**

I would like to first thank *Almighty Allah* for all his blessings and successes bestowed in my life.

I am grateful to my parents *Mrs. Ismat Tahir and Mr. S.R. Tahir* for their endless love and support that have enabled me to chase my goals in life.

I am thankful to my wife *Labiba Fahad* who fully supported me in all my decisions.

I would like to express my gratitude to *Prof. Andrea Cavallaro* (my primary supervisor) for his guidance, support, valuable suggestions and advices, which helped me to perform this research.

I would like to thank *Prof. Bernhard Rinner* (my secondary supervisor) for his guidance and support throughout this research and during my stay in Alpen-Adria Universität Klagenfurt.

I am thankful to all my *colleagues* for the great time we spent together and in the discussions during the course of my PhD studies.

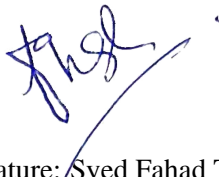
I, Syed Fahad Tahir, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without the prior written consent of the author.



Signature: Syed Fahad Tahir

Date: 14/2/2015



First supervisor

**Professor Andrea Cavallaro**

Second supervisor

**Professor Bernhard Rinner**

Author

**Syed Fahad Tahir**

## **Resource-constrained re-identification in camera networks**

### **Abstract**

In multi-camera surveillance, association of people detected in different camera views over time, known as person re-identification, is a fundamental task. Re-identification is a challenging problem because of changes in the appearance of people under varying camera conditions. Existing approaches focus on improving the re-identification accuracy, while no specific effort has yet been put into efficiently utilising the available resources that are normally limited in a camera network, such as storage, computation and communication capabilities. In this thesis, we aim to perform and improve the task of re-identification under constrained resources. More specifically, we reduce the data needed to represent the appearance of an object through a proposed feature selection method and a difference-vector representation method.

The proposed feature-selection method considers the computational cost of feature extraction and the cost of storing the feature descriptor jointly with the feature's re-identification performance to select the most cost-effective and well-performing features. This selection allows us to improve inter-camera re-identification while reducing storage and computation requirements within each camera. The selected features are ranked in the order of effectiveness, which enable a further reduction by dropping the least effective features when application constraints require this conformity. We also reduce the communication overhead in the camera network by transferring only a difference vector, obtained from the extracted features of an object and the reference features within a camera, as an object representation for the association.

In order to reduce the number of possible matches per association, we group the objects appearing within a defined time-interval in un-calibrated camera pairs. Such a grouping improves the re-identification, since only those objects that appear within the same time-interval in a camera pair are needed to be associated. For temporal alignment of cameras, we exploit differences between the frame numbers of the detected objects in a camera pair. Finally, in contrast to pairwise camera associations used in literature, we propose a many-to-one camera association method for re-identification, where multiple cameras can be candidates for having generated the

previous detections of an object. We obtain camera-invariant matching scores from the scores obtained using the pairwise re-identification approaches. These scores measure the chances of a correct match between the objects detected in a group of cameras.

Experimental results on publicly available and in-lab multi-camera image and video datasets show that the proposed methods successfully reduce storage, computation and communication requirements while improving the re-identification rate compared to existing re-identification approaches.

# Contents

<b>Acknowledgements</b>	<b>3</b>
<b>Abstract</b>	<b>5</b>
<b>Publications</b>	<b>10</b>
<b>Glossary of symbols</b>	<b>11</b>
<b>1 Introduction</b>	<b>14</b>
1.1 Motivation . . . . .	14
1.2 Problem formulation . . . . .	16
1.3 Challenges . . . . .	17
1.3.1 Re-identification challenges . . . . .	17
1.3.2 Resource constraints . . . . .	19
1.4 Contributions . . . . .	19
1.5 Organisation of the thesis . . . . .	20
<b>2 Related work</b>	<b>22</b>
2.1 Introduction . . . . .	22
2.2 Object acquisition . . . . .	23
2.3 Feature descriptors . . . . .	25
2.3.1 Colour . . . . .	25
2.3.2 Texture . . . . .	27
2.3.3 Shape . . . . .	28
2.3.4 Grouping . . . . .	28
2.4 Dimensionality reduction . . . . .	29
2.4.1 Feature extraction . . . . .	29
2.4.2 Feature selection . . . . .	30
2.5 Data compression . . . . .	32
2.6 Cross-camera calibration . . . . .	33

2.6.1	Colour calibration . . . . .	34
2.6.2	Spatio-temporal calibration . . . . .	35
2.7	Object association . . . . .	36
2.7.1	Distance minimisation . . . . .	36
2.7.2	Learning classifiers . . . . .	37
2.7.3	Optimisation approaches . . . . .	38
2.8	Datasets . . . . .	38
2.9	Discussion . . . . .	42
<b>3</b>	<b>Cost-effective features</b>	<b>44</b>
3.1	Introduction . . . . .	44
3.2	Upper-body image representation . . . . .	45
3.3	Feature sets . . . . .	46
3.4	Feature performance . . . . .	47
3.5	Feature cost . . . . .	48
3.6	Feature selection . . . . .	49
3.7	Discussion . . . . .	51
3.8	Summary . . . . .	54
<b>4</b>	<b>Association for re-identification</b>	<b>55</b>
4.1	Introduction . . . . .	55
4.2	Association using difference-vectors . . . . .	56
4.2.1	Histogram equalisation . . . . .	56
4.2.2	Difference-vector representation . . . . .	58
4.2.3	Temporal alignment . . . . .	58
4.2.4	Object association . . . . .	60
4.3	Association using camera-invariant scores . . . . .	60
4.3.1	Training . . . . .	62
4.3.2	Testing . . . . .	63
4.4	Summary . . . . .	66
<b>5</b>	<b>Experimental evaluation</b>	<b>68</b>
5.1	Introduction . . . . .	68

5.2	Re-identification with cost-effective representations . . . . .	69
5.2.1	Feature sets . . . . .	69
5.2.2	CoPE with varying parameters . . . . .	71
5.2.3	CoPE vs all-features . . . . .	73
5.2.4	CoPE vs feature selection methods . . . . .	75
5.2.5	CoPE with learning models . . . . .	79
5.2.6	CoPE and re-identification approaches . . . . .	82
5.2.7	CoPE and feature budgeting . . . . .	85
5.3	Re-identification with difference-vector representation . . . . .	86
5.3.1	Data reduction . . . . .	87
5.3.2	Re-identification rate . . . . .	87
5.4	Re-identification with camera-invariant scores . . . . .	90
5.4.1	Three-camera setting . . . . .	92
5.4.2	Variable cameras setting . . . . .	94
5.5	Summary . . . . .	94
<b>6</b>	<b>Conclusions</b>	<b>97</b>
6.1	Summary of achievements . . . . .	97
6.2	Future directions . . . . .	99
	<b>Bibliography</b>	<b>100</b>

## Publications

The following publications are part of this thesis:

### Journal papers

- [J1] Syed Fahad Tahir and Andrea Cavallaro. Cost-effective features for re-identification in camera networks. *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 24, Issue 8, pp. 1362-1374, 2014.
- [J2] Riccardo Mazzon, Syed Fahad Tahir and Andrea Cavallaro. Person re-identification in crowd. *Elsevier Pattern Recognition Letters*, Vol. 33, Issue 14, pp. 1828-1837, 2012.

### Conference papers

- [C1] Syed Fahad Tahir, Andrea Cavallaro and Bernhard Rinner. Re-identification with multiple source-cameras. In *Proc. of IEEE International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP)*, Singapore, April 2015.
- [C2] Syed Fahad Tahir and Andrea Cavallaro. Low-cost multi-camera object matching. In *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy, May 2014.

## Glossary of symbols

$\mathbf{C}$	Camera network
$N$	Total number of cameras in the network
$C_n$	Destination camera
$\mathcal{N}_n$	Set of source-cameras of $C_n$
$C_{n_q}$	Source-camera for $C_n$ from $\mathcal{N}_n$
$\hat{N}$	Number of source-cameras in $\mathcal{N}_n$
$\mathbf{P}_n$	Set of objects detected in $C_n$
$P_n^m$	$m^{th}$ object in $\mathbf{P}_n$
$M_n$	Number of detected objects in $\mathbf{P}_n$
$\mathbf{P}_{n_q}$	Set of objects detected in $C_{n_q}$
$P_{n_q}^k$	$k^{th}$ object in $\mathbf{P}_{n_q}$
$M_{n_q}$	Number of detected objects in $\mathbf{P}_{n_q}$
$\mathbf{F}_n^m$	Extracted feature set from $P_n^m$
$\mathbf{f}_n^{mr}$	$r^{th}$ feature in $\mathbf{F}_n^m$
$\mathbf{F}_{n_q}^k$	Extracted feature set from $P_{n_q}^k$
$\mathbf{f}_{n_q}^{kr}$	$r^{th}$ feature in $\mathbf{F}_{n_q}^k$
$\mathbf{f}^r$	$r^{th}$ type of feature
$R$	Total number of feature types
$d^{mkr}$	Distance between $P_n^m$ and $P_{n_q}^k$ using $\mathbf{f}^r$
$\mathbf{E}^{mr}$	Set of distances for incorrect matches
$med(\mathbf{E}^{mr})$	Median of values in $\mathbf{E}^{mr}$
$\Pi^{mr}$	Performance score of $\mathbf{f}^r$ for $P_n^m$
$\Pi^r$	Performance vector of $\mathbf{f}^r$ for camera pair $(C_n, C_{n_q})$
$\Delta$	performance matrix for $R$ features and $M$ objects in training
$\chi^m$	$m^{th}$ column of $\Delta$ : performance of $P_n^m$ for $\hat{R}$ features
$\beta_n^{mr}$	Storage size of $\mathbf{f}_n^{mr}$

$\Gamma_n^{mr}$	Computational time of $\mathbf{f}_n^{mr}$
$\Psi^r$	Cost vector of $\mathbf{f}^r$ for camera pair $(C_n, C_{n_q})$
$V$	Sorted vector containing $\Pi^{mr} \leq 1$ from $\Delta$
$\Phi_i^r$	Set containing $\Pi^{mr}$ within bin $I_i$ for each $\mathbf{f}^r$ .
$A_i^r$	Combined importance score of $\mathbf{f}^r$
$\hat{r}$	ID of the selected feature with the highest $A_i^r$
$\hat{\mathbf{f}}$	Selected feature type
$Y_{n_q n}$	List of selected features for camera pair $(C_n, C_{n_q})$
$Z$	Set containing objects already taken part in the selection
$\mathcal{E}$	Normalised cost of selected feature sets (used in evaluation)
$Q_n$	Computation and storage by a camera (used in evaluation)
$\kappa^j$	$j^{\text{th}}$ reference feature set
$\Omega_n^{mj}$	Difference between $\mathbf{F}_n^m$ and $\kappa^j$
$\mathbf{\Omega}_n^m$	$J$ dimensional difference-vector representation of $P_n^m$
$t_n^{m(s)}$	Index of first tracking frame of $P_n^m$ in $C_n$
$t_n^{m(e)}$	Index of last tracking frame of $P_n^m$ in $C_n$
$w_n^m$	Number of frames during $P_n^m$ is tracked in $C_n$
$\tilde{w}_n$	Average of values in $\{w_n^m\}_{m=1}^{M_n}$
$t_{n_q}^{k(s)}$	Index of first tracking frame of $P_{n_q}^k$ in $C_{n_q}$
$\varphi$	Ratio of the frame rates of $C_n$ and $C_{n_q}$
$\Lambda_{n_q n}^m$	Set of $M_{n_q}$ difference between $t_n^{m(s)}$ and $\{t_{n_q}^{k(s)}\}_{k=1}^{M_{n_q}}$
$D_{n_q n}$	Difference frame number matrix containing $\{\Lambda_{n_q n}^m\}_{m=1}^{M_n}$
$\delta_{n_q n}$	Time shift between $C_n$ and $C_{n_q}$
$W_{n_q}^m$	Temporal search window of $P_n^m$ in $C_{n_q}$
$\hat{M}_{n_q}$	Number of detected objects in $C_{n_q}$ within $W_{n_q}^m$
$\mathbf{B}_n^m$	Set of $\hat{M}_{n_q}$ differences from $P_n^m$
$H$	$M_n \times M_{n_q}$ matrix containing $\{\mathbf{B}_n^m\}_{m=1}^{M_n}$
$S_{n_q n}^{mk}$	Similarity score between $\mathbf{F}_n^m$ and $\mathbf{F}_{n_q}^k$
$\mathcal{S}_{n_q n}$	Set containing $S_{n_q n}^{mk}$ between objects common in $(C_n, C_{n_q})$
$\mathcal{S}$	Hypothesis: Two images are of same object in $(C_n, C_{n_q})$
$\bar{\mathcal{S}}$	Hypothesis: Two images are of different objects in $(C_n, C_{n_q})$



$\mathcal{S}_{n_q n}^+$	$\mathcal{S}_{n_q n}$ of $\mathcal{S}$
$\mathcal{S}_{n_q n}^+$	$\mathcal{S}_{n_q n}$ of $\bar{\mathcal{S}}$
$T_{n_q n}^+$	Nearest parametric distribution model to $\mathcal{S}$
$T_{n_q n}^-$	Nearest parametric distribution model to $\bar{\mathcal{S}}$
$Pr(\mathcal{S}   \mathcal{S}_{n_q n}^{mk})$	Probability of $\mathcal{S}$ , given $\mathcal{S}_{n_q n}^{mk}$
$Pr(\bar{\mathcal{S}}   \mathcal{S}_{n_q n}^{mk})$	Probability of $\bar{\mathcal{S}}$ , given $\mathcal{S}_{n_q n}^{mk}$
$\mathcal{L}_{n_q n}^{mk}$	Matching score between $P_n^m$ and $P_{n_q}^k$
$\mathcal{L}_n$	Matching score matrix containing $\mathcal{L}_{n_q n}^{mk}$ between $\mathbf{P}_n$ and $\{P_{n_q}^k\}_{k=1}^{M_{n_q}}$

# Chapter 1

## Introduction

---

### 1.1 Motivation

Networks of cameras are deployed for the surveillance of wide areas, such as airports, train stations and shopping malls. These cameras may have disjoint Fields-of-View (FoV). Re-identification of the same person in a camera network by human operators is a tiresome and costly job and depends upon individuals' consistent attention and experience [67]. This task is crucial for activities like long-term tracking and forensic search. The widespread increase in large camera networks makes automated re-identification a fundamental requirement for surveillance systems [54]. Re-identification is typically performed by comparing the image(s) of a person from one camera to the images of multiple persons from another camera or a set of cameras [71, 111, 136, 179]. Other sources of information, such as inter-camera relations and environmental constraints, are also exploited for the task [89, 100, 120, 123]. In the majority of surveillance systems, cameras are centrally connected, and both communication and processing are done by the central node. Given that visual data processing has large computational and storage requirements, and the number of cameras is increasing, a high data transfer rate within the network and a high processing power of the central node is required [156]. It becomes highly important to devise solutions, which could improve the re-identification task by utilising less resources so that the task can be completed within an acceptable time frame and remains useful for the applications.

*Multiple cameras* are deployed with different viewing configurations depending upon the surveillance requirements. Cameras can have overlapping or non-overlapping FoV of the scene.

In *overlapping* FoV the association process can be considered as long-term tracking, consistent labelling or multi-view object matching [7, 32, 96]. It may be performed by estimating the object position in the scene. Geometric properties, such as homography matrices [27], camera projection matrices [61, 76, 141] and epipolar constraints [28], can be exploited based on inter-camera relations and camera calibration. Cameras deployed for the surveillance of larger areas may have blind regions in between resulting in *non-overlapping* FoV. In such a case, *appearance* information, such as colour and texture descriptors [71, 112, 178], their relative positioning [101], and high-dimensional feature point descriptors like SIFT and HOG [122], are extracted from single [71, 136] or multiple images [122] of an object and communicated across the network for association. Inter-camera *transition times* can also be exploited [89, 100, 123], which requires *network synchronisation* to identify the time of occurrence of the same event across the network. Network Time Protocol (NTP) can be applied for synchronisation in wired camera networks, where the central node communicates the time information to the clients [124]. In the case of a wireless network, Global Positioning System (GPS) based synchronisation can be applied [9]. However, it requires a GPS receiver on each wireless device and the line of site communication with the satellite, which are not always possible. A more feasible approach is based on relative timing i.e, to keep track of the order of occurrence of events [102]. Visual events can also be used in such a synchronisation. Synchronisation approaches based on visual information are mostly applied to fixed cameras [30, 120, 127] with prior knowledge of the scene under observation. Approaches also exist for moving cameras [31], which exploit the known object-association information.

Object *association* is generally performed between pairs of cameras. Learning approaches, such as AdaBoost [71] and RankSVM [136], may be applied for object association when sufficient training data is available. In the case of insufficient training data, Direct Distance Minimisation (DDM) approaches, such as those based on the Kullback-Leibler [92], Bhattacharyya [135] or Euclidean distance, are applied. However, these methods are in general less robust to illumination changes, which can be compensated for by learning inter-camera colour calibration [89, 92, 135].

In order to improve scalability, a smart-camera network can be deployed. *Smart cameras* are able to perform image processing locally and aim at transferring the minimum amount of data over the network to accomplish collaborative tasks such as object detection, tracking and

re-identification [3, 139, 154]. Such a network can enable a continued surveillance of the environment by using local storage and computation capabilities along with intelligent processing of the data. However, smart cameras have limited resources, which is also common for battery-powered devices such as smartphones and wireless smart cameras [33, 152]. We achieve the same or improved results compared to the existing approaches for the task of re-identification while adapting to the constraints in a smart camera network such as reducing the amount of data to be processed and shared across the network, real-time operations and energy efficiency.

In Sec. 1.2 we present our problem formulation. Sec. 1.3 discusses the challenges involved in the problem of re-identification. We then highlight in Sec. 1.4 the specific contribution of this thesis.

## 1.2 Problem formulation

Let  $\mathbf{C} = \{C_n\}_{n=1}^N$  be a network of  $N$  cameras. Re-identification is performed in the destination-camera referred as  $C_n$ . A camera,  $C_{n_q}$ , is a source-camera for  $C_n$  if an object exiting  $C_{n_q}$  is expected to enter  $C_n$  without passing through FoV of other cameras. Destination and source-cameras together form a camera pair  $(C_n, C_{n_q})$ . Source-cameras can be variable in number. Each  $C_n$  has a set of source-cameras  $\mathcal{N}_n = \{C_{n_q}\}_{q=1}^{\hat{N}_n}$ , where  $\mathcal{N}_n \subseteq \mathbf{C}$  and  $\hat{N}_n \leq N$ . We assume a detection method that returns a single image of an object (person) on its first appearance in a camera. Each detected object in  $C_n$  is represented by a single image-patch  $P_n^m$ , and  $C_{n_q}$  has the previous instance (image-patch) of that object. The overall set of  $M_n$  object-images extracted in camera  $C_n$  is defined as  $\mathbf{P}_n = \{P_n^m\}_{m=1}^{M_n}$ . A feature set  $\mathbf{F}_n^m = \{\mathbf{f}_n^{mr}\}_{r=1}^R$  containing  $R$  feature types  $\mathbf{f}^r$  is extracted from  $P_n^m$ , where  $r = 1 \dots R$ . Since all features may not have equal importance in re-identification, we aim to perform feature selection while considering both the features' performance and cost. The performance vector,  $\mathbf{\Pi}^r$ , measures the discriminating ability of  $\mathbf{f}^r$  in person re-identification, whereas the cost vector,  $\mathbf{\Psi}^r$ , measures the extraction time and the storage size associated with  $\mathbf{f}^r$ . Selected features are communicated over the network for association.

Re-identification is considered as an information retrieval problem, where for a given query (object information) the most relevant matches are retrieved/ranked. Association is performed by comparing the selected features of  $P_n^m$  with those from a generic  $P_{n_q}^k$  in  $C_{n_q}$ , where  $k = 1 \dots M_{n_q}$  and  $M_{n_q}$  objects are detected in  $C_{n_q}$ . The inter-camera space, time and appearance relations between camera pairs can also be exploited, which requires cross-camera calibration. In the case

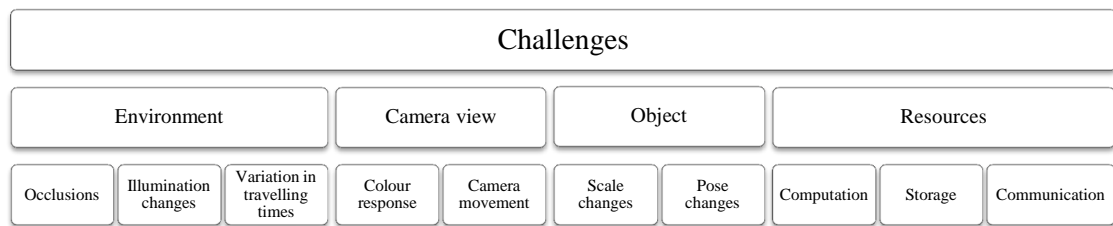


Figure 1.1: Categorisation of challenges involved in person re-identification.

of multiple source-cameras, for each  $P_n^m$  we aim to identify its other instance,  $P_{n_q}^k$ , detected in an unspecified source-camera  $C_{n_q}$  in  $\mathcal{N}_n$  by performing a many-to-one camera association.

### 1.3 Challenges

Person re-identification is a challenging problem. Challenges can be broadly categorised based on: environment, camera view, and object (Fig. 1.1). Since visual data processing also involves large computation, communication and storage resources, we discuss these as challenges under resource constraints.

#### 1.3.1 Re-identification challenges

Challenges because of the environment/surroundings in which an object is observed include: occlusions, illumination changes and variations in travelling time. *Occlusion* is the partial/complete obstruction between camera and the object. It occurs when multiple objects are not sufficiently far apart from each other, or when stationary or other moving objects in the environment are between the camera and the observed object. For example, in a crowd an object's full body is often not visible because of occlusions. Occlusion modifies how an object is viewed across cameras [72]. *Illumination* changes are due to variations in the lighting conditions across camera FoV. Indoor or outdoor settings, or partially reflective areas and objects in a camera view can cause changes in illumination [118]. Moreover, people exiting a camera may enter the next camera in different *regions* of its view so the time needed to travel across cameras and the area of re-appearance are variable and hence difficult to model [89].

Camera *colour response* may be inconsistent across cameras because of variations in aperture size, electrical noise and camera model even under the same illumination conditions [85]. Camera *movements* result in jitter and blurring in the recordings, and a continuous cross-camera spatio-temporal re-calibration may be required [138].



Figure 1.2: Appearance of people across cameras. Column 1: camera 1, full frame; Column 2: corresponding crop of a person of interest; Column 3: camera 2; Column 4: corresponding crop of a person of interest. People appearance under different illumination conditions, as shown in (b) and (d), and under different poses and levels of occlusion, as shown in (f) and (h).

The *pose* of the object can vary across cameras, since an object can be moving and deformable as in the case of people. Depending on where an object enters a camera, the viewing *angle* and the distance of the object from the camera may result in variation in the object's *scale* and orientation [100]. Fig. 1.2 shows typical challenges of re-identification in a surveillance system.

Finally, the challenges related to the acquired data include heterogeneity, inconsistency and limits on the amount acquired [174]. Incorrect data is more frequent when a re-identification approach is applied to real-world environments where there is no control on the way an object moves. In order to overcome this challenge large numbers of features are extracted [71, 112, 136, 178]. A large feature set may not always compensate for the data errors; however, the performance of discriminating features can be suppressed because of redundant features in the feature set. Noise because of sensor or transmission channel can cause distortion in an image, which degrades the quality of the image itself. Noise makes the re-identification a more challenging task, since a good feature representation becomes difficult to extract from a degraded image.

### 1.3.2 Resource constraints

Real-time, continuous and reliable re-identification, along with scalability, as the number of cameras increases, are the main goals of a re-identification system, while processing of visual data inherently involves large amount of computational time and storage requirements [156]. In a typical RGB, VGA-resolution surveillance setting, a single detection results in an object image of size  $128 \times 48 \times 3$  pixels and requires 18 *KB* for storage, which becomes nearly 2 *KB* after compression [67]. The extraction of information in the form of the most commonly used 2784-dimensional feature set, containing histograms of colour and texture [71, 112, 136, 179], requires 1 *KB* for storage, while its extraction time is nearly 2 *sec* on a 3.0 *GHz* desktop computer. If we also include the object detection time, it will be 0.2 *sec* per person per frame [172]. Given these resource requirements per object, nearly 1 *MB* of storage and 8 *min* of processing time of the central node is required to perform re-identification in 2-min surveillance videos from a pair of cameras in a moderately crowded scene (with  $\sim 100$  objects). These storage and computational requirements can drastically increase as the duration of the videos and the number of cameras increase. Smart cameras can be useful in achieving scalability, where most of the processing can be performed locally in a distributed fashion [139]. However, an additional overhead of communicating the locally stored data to a central server is involved, which can be typically in the order of  $\sim 30$  *MB* per camera per hour.

Furthermore, the resources, such as computation, storage and communication, are limited - especially in the case of smart cameras. A typical smart camera<sup>1</sup> has 1.6 *GHz* of processor, 30 *GBs* of storage capacity, and 2 *GB* of memory (RAM). In order to achieve a real-time performance, computation and storage requirements need to be reduced. If object-images could be discarded after the extraction of required information (feature sets), half of the storage and communication requirements can be reduced. However, the extracted information still needs to be stored and communicated. This demands the development of efficient and resource-aware information extraction/representation algorithms for re-identification.

## 1.4 Contributions

The main contributions of this thesis are the following:

1. A cost-effective feature selection approach for re-identification is proposed by taking into

---

<sup>1</sup>Smart-camera information obtained from SLR engineering, <http://www.slr-engineering.at/smart-camera/>

account the performance of a feature jointly with its cost. The performance is measured by the ability of a feature to discriminate an object from others. Cost combines storage size and the computation time required to extract a feature. We select each feature based on its individual importance and rank them based on their contribution to the task. This makes the approach adaptable for resource-constrained environments [J1].

2. An object representation approach for association is proposed that minimises the inter-camera information sharing for the re-identification in a smart-camera network. Each object is represented as a difference vector between the extracted features and the locally stored set of reference features. In the association phase, instead of transferring the extracted features, only the obtained difference vectors are communicated over the network, which minimises the inter-camera information sharing for re-identification [C2].
3. A multi-camera object association approach for re-identification is proposed that extends the existing appearance based re-identification approaches to the case when multiple source-cameras can exist. We analyse variations in the matching distances between objects in each camera pair in order to estimate the probability of a correct match when multiple source-cameras exist. These probabilities generate camera-invariant matching scores for re-identification [C1].

## 1.5 Organisation of the thesis

This thesis is organised as follows:

*Chapter 1:* Introduction to person re-identification and its applications in surveillance and camera networks are discussed in Sec. 1.1. We formulate the re-identification problem in Sec. 1.2, and define the challenges that can be encountered in real-world scenarios (Sec. 1.3). The contributions of this thesis are listed in Sec. 1.4.

*Chapter 2:* Existing related work on person re-identification is organised based on object acquisition (Sec. 2.2), feature descriptors (Sec. 2.3), dimensionality reduction (Sec. 2.4), data compression (Sec. 2.5), cross-camera calibration (Sec. 2.6) and object association (Sec. 2.7). A summary of the datasets used for validating the state of the art is presented in Sec. 2.8. A brief discussion on the existing approaches is also provided along with their limitations (Sec. 2.9).



*Chapter 3:* The proposed feature selection method for re-identification that combines both the performance (Sec. 3.4) and cost (Sec. 3.5) of a feature in the selection (Sec. 3.6) is presented in this chapter.

*Chapter 4:* This chapter discusses the two proposed association approaches for re-identification. The first approach minimises the information required to be communicated over the network for both camera calibration and re-identification (Sec. 4.2). The second approach extends the re-identification from pairwise association to multiple source-cameras (Sec. 4.3).

*Chapter 5:* Experimental evaluation of the proposed feature selection method (Sec. 5.2) and association methods (Sec. 5.3 and Sec. 5.4) for re-identification using challenging publicly available and in-lab people datasets is presented, followed by a summary of the results (Sec. 5.5).

*Chapter 6:* This chapter presents a summary of the achievements of this thesis (Sec. 6.1) and the possible future directions (Sec. 6.2).

## Chapter 2

### Related work

---

#### 2.1 Introduction

Person re-identification has been the focus of interest in multi-camera surveillance for the last fifteen years [162]. In this chapter, we present a unifying overall structure and an in-depth survey of the state-of-the-art for person re-identification methods and the datasets used for evaluation.

We can identify four main phases of re-identification, namely object acquisition, feature extraction, cross-camera calibration and object association (Fig. 2.1). The first phase, object *acquisition*, identifies the image regions corresponding to the object [57]. The second phase is the acquisition of *information* from the acquired object-images. The information includes feature extraction and feature selection. Appearance features include colour, texture and shape. Moreover, temporal concatenations of appearance features can also be used [71]. The third phase is the cross-camera *calibration*, namely the establishment of the colour and spatio-temporal relationship across cameras that allows to account for the variability of observations of the same object across different FoV [89]. Finally, the fourth phase is the *association* of candidates across cameras to match different instances of the same object using the information extracted in the previous phases. Existing re-identification methods are validated on snapshot-based or video-based datasets [48, 70, 122, 148].

This chapter is organised to group the concepts according to the main contributions of the thesis (Sec.1.4). First we discuss, in Sec. 2.2, how an object can be acquired from a video frame. The existing research related to the first contribution, cost effective feature selection,

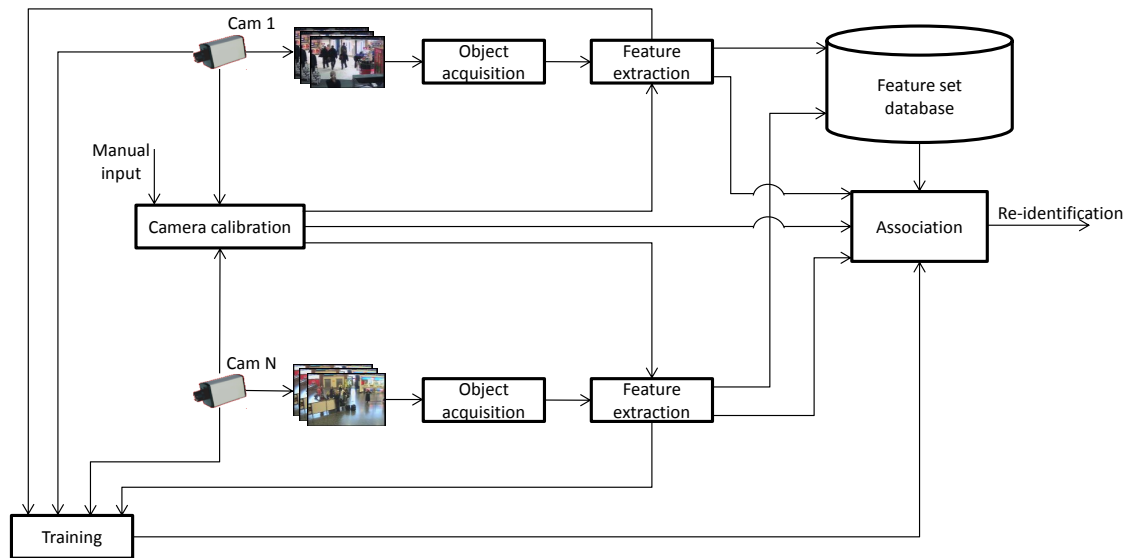


Figure 2.1: General block diagram of existing re-identification approaches.

is grouped into Feature descriptors (Sec. 2.3) and Dimensionality reduction (Sec. 2.4). For the second contribution, minimisation of information sharing and improvement of association, the related work is grouped into Data compression (Sec. 2.5) and Cross-camera colour calibration (Sec. 2.6.1). Finally, Spatio-temporal calibration (Sec. 2.6.2) and object association (Sec. 2.7) discuss the existing approaches to support the third contribution, multi-camera object association. We then give an overview of the available and proposed datasets in Sec. 2.8. The conclusions about the described approaches are drawn in Sec. 2.9.

## 2.2 Object acquisition

Object acquisition is the extraction of image parts that correspond to the object of interest (people or moving objects) in video frames [57]. Object acquisition can yield a single cropped image of an object in the case of detections only – single-shots [71, 111, 136, 179], and it can also yield multiple images in a camera when intra-camera tracking information is available – multi-shots [12, 19, 41, 64, 112, 145]. In order to acquire the object information, object detection can be applied.

*Object detection* provides the position of the object in an image or video frame [57, 68, 86]. The position can be the head location [172], the feet location [123] or a point in (the centre of) the object [57]. The detection is solved using a trained classifier [45], a motion detector [60] or a combination of both [57]. From the detections we extract a cropped image or a bounding box

around the object. Obtained images using object detection also contain background along with the object, which can be removed by applying background subtraction.

The pixels with minimum change in their values for a range of frames can be considered as the background, while pixels with varying values over time represent the foreground or objects of interest [142]. *Background subtraction* is applied to extract regions corresponding to moving objects in an image [40, 87, 142]. Background subtraction can be done by considering an empty frame with minimum to no foreground clutter as the background and then applying frame differencing [87]. The background can also be estimated by taking the median of a set of video frames where the foreground exists; however, the background pixels must be visible in at least half of the frames [40]. Background subtraction is not suitable in crowded scenes with frequent foreground clutter, since multiple objects are grouped into a single moving region [116]. Constantly moving backgrounds like billboards, leaves of trees, and shadows also become moving objects, which are non-trivial to segment [72, 142]. In such cases object detection without background subtraction can be used to improve the results.

In *single-shot* approaches, the single most representative image of the object can be selected from a group of images, such as when the object is detected for the first/last time in a camera [J2]. Methods using single-shot representation associate pairs of images obtained from two cameras. For *multi-shot* approaches, images obtained from tracking are grouped over time. Gheissari *et al.* [64] use spatio-temporal over-segmentation of cropped images from ten frames to create a signature for each person. In [41], ten key frames are selected. The most common approach is to keep all the object images grouped over time and then perform association by analysing the similarity among the features extracted from all available object images [20, 89, 100]. Another multi-shot approach combines the relevant information from multiple instances of the same object using image-epitome, which groups the extracted patches from multi-images with similar properties [19]. Single-shot approaches have a wider range of applications than multi-shot approaches. However, Multi-shot images can also provide spatio-temporal information of the object and hence remain a more invariant representation than single-shot [13]. Since multiple representative images for each object are available, the effects of illumination changes because of light variations within the same camera and short-term occlusions, are reduced. However, multi-shot approaches are computationally more expensive compared to single-shot approaches.

The acquired object-image is usually divided into two [129], three [5], five [13], six [179]

or ten [22] horizontal slices, roughly separating into head, torso and legs structure. Feature descriptors, as discussed in the next section, are extracted from each slice resulting in a large set of features.

## 2.3 Feature descriptors

Features describe an object so as to allow re-identification in the next camera. Features should be robust in identifying the same object, while able to discriminate between different objects. Features can be extracted after identifying different parts of the object or may represent a holistic view [145]. Various feature types have been used for appearance-based person re-identification [54, 162]. The appearance features commonly used in the state-of-the-art methods for person re-identification can be classified into colour, texture and shape. Multiple features are usually combined in order to obtain a more representative descriptor of the object [70, 111, 136]. Furthermore, temporal consistency of features can be exploited to merge the available information of a person over time (multi-shot) [14]. Table 2.1 summarises different feature descriptors that are extracted from the object images in the existing approaches.

### 2.3.1 Colour

The most commonly used appearance feature is the colour extracted in the form of histograms [39, 50, 59, 64, 71, 89, 100, 115, 136, 178, 179]. Normalised histograms of colours are scale invariant. Different colour channels and their combination are used: the Hue channel from the HSV colour space [50]; the Hue and Saturation channels jointly [64]; or the three channels of the HSV colour space [24, 59]. Also, the histogram of the RGB colour space is widely used [20, 39, 89, 135]. A concatenation of histograms from RGB, YCbCr, and HS (from HSV) colour channels is adopted in [71, 136, 178]. An analysis by boosting classifier using image dataset VIPeR shows how, for the re-identification task, the Hue channel is the most discriminative followed by Saturation, Blue, Red, and Green channels [71]. This analysis is limited to scenes where people are fully visible. Alternatively, the two chrominance channels from the YUV space are used in [92], where a Gaussian Mixture Model is applied to find the most relevant colour clusters, whose centres are adopted as descriptors. The Dominant Colour Descriptor (DCD) [13] and the Major Colour Spectrum Histogram Representation (MCSHR) [116] compute the most recurrent RGB colour values that are then used to represent an image. Moreover, Maximally

Table 2.1: Object acquisition and feature extraction methods applied to datasets with different camera settings. Key: Single - Single shot, Multi. - Multi-shot, Uncal. - Uncalibrated, Disj. - Disjoint, LTH - Leg Torso Head, MRCG - Mean Riemannian Co-variance Grid, BTF - Brightness Transfer Function, LBP - Local Binary Pattern, SIFT - Scale Invariant Feature Transform, ISM - Implicit Shape Model, HOG - Histogram of Oriented Gradients, SURF - Speeded Up Robust Features, GLOH - Gradient Location and Orientation Histogram, DCD - Dominant Colour Descriptor, RHSP - Recurrent High-Structured Patches, Hist. - Histogram, P.V. - Personal video dataset

Ref	Camera settings	Images	Shape	Features	Datasets
[4]	Calib. Disj. Indoor/Outdoor	Single	2D-Grid	Colour, Shape, Texture & Position	VIPeR, P.V.
[7]	Overlapping Indoor	Single	Bounding box	Position on ground plane & colour	Sports videos
[11]	Uncal. Disj. Indoor	Single	LTH	SIFT,SURF & Spin	P.V.
[13]	Uncal. Disj. Indoor	Single	2D body parts	Haar & DCD	iLIDS
[14]	Uncal. Disj. Indoor	Multi.	Bounding box	MRCG	iLIDS, ETHZ
[16]	Uncal. Disj. Indoor/Outdoor	Multi.	3D model	Colour hist.	ViSOR, Sarc3D
[18]	Uncal. Disj. Indoor	Multi.	Bounding box	SIFT,SURF,SC & GLOH	Caviar
[19]	Uncal. Disj. Indoor/Outdoor	Multi.	LTH	Colour hist. & Epitome	iLIDS,ETHZ, CAVIAR
[27]	Overlapping Outdoor	Single	Bounding box	Feet head positions & vertical axis	ViSOR
[32]	Overlapping Outdoor	Single	Bounding box	Feet position	PETS2001
[41]	Uncal. Disj. Indoor	Multi.	LTH	Colour hist. & Positions	P.V.
[42]	Uncal. Disj. Outdoor	Multi.	Bounding box	Colour & appearance mask	PETS2010
[47]	Uncal. Disj. Indoor	Single	Bounding box	Colour & Texture	Feret
[51]	Uncal. Disj. Indoor	Multi.	LTH	Colour hist., height of LTH	PETS2006
[59]	Uncal. Disj. Indoor/Outdoor	Multi.	LTH	Colour hist., RHSP	iLIDS, VIPeR, ETHZ
[77]	Uncal. Disj. Indoor/Outdoor	Single	2D-Grid	LBP hist. & Mean colour	VIPeR, ETZH, Prid2011
[91]	Calib. Disj. Outdoor	Multi.	Bounding box	Feet position & BTF	Online cameras
[94]	Uncal. Disj. Indoor	Multi.	Bounding box	SIFT & ISM	Casia infrared dataset
[93]	Uncal. Disj.Indoor	Multi.	Bounding box	SIFT on infra-red images	Casia infrared dataset
[100]	Calib. Disj.	Multi.	Bounding box	Colour hist., Co-variance & HOG	CAVIAR, TRECVID08
[103]	Uncal. Disj. Indoor/Outdoor	Single	Bounding box	High level attributes	iLIDS, ViPeR, ETZH
[105]	Overlapping Indoor	Single	Bounding box	Vertical axis & homography	P.V.
[106]	Disj. Outdoor	Single	2D Grid	Colour hist., Gabor & HOG	VIPeR, CAVIAR
[107]	Overlapping Outdoor	Single	Bounding box	Adaptive homographies	P.V.
[112]	Uncal. Disj. Indoor/Outdoor	Single	6 stripes	Colour hist. , Schmid & Gabor	iLIDS, VIPeR
[115]	Uncal. Disj. Indoor/Outdoor	Single	6 stripes	Colour hist. , Schmid & Gabor	GRID, VIPeR
[117]	Overlapping Outdoor	Multi.	Bounding box	Colour hist. & feet position	PETS2009, Caviar
[136]	Uncal. Disj. Indoor/Outdoor	Single	6 stripes	Colour hist., Schmid & Gabor	iLIDS, VIPeR
[150]	Calib. Disj.	Single	Bounding box	Colour hist. and shape	CAVIAR, Video Web
[157]	Disj. Indoor/Outdoor	Single	2D Grid	Colour hist. & LBP descriptors	VIPeR, ETHZ, iLIDS
[161]	Calib. Disj. Outdoor	Single	Bounding box	Colour hist. & position	ViSOR
[166]	Disj. Outdoor	Single	2D Grid	Colour hist. & LBP descriptors	CUHK, PRID, VIPeR
[167]	Uncal. Disj. Indoor	Single	Upper Lower	Colour hist.	P.V.
[178]	Uncal. Disj. Indoor/Outdoor	Single	6 stripes	Colour hist., Schmid and Gabor	iLIDS, VIPeR
[179]	Uncal. Disj. Indoor/Outdoor	Single	6 stripes	Colour hist., Schmid & Gabor	iLIDS, VIPeR, ETZH

Stable Colour Regions (MSCR) [59] extract the homogeneous colour in the object image by grouping neighbouring colour blobs. Finally, camera parameters and reflectance of the objects' surface can be studied to obtain the main appearance characteristic of the object [89]. Camera parameters refer to exposure time, focal length, and aperture size of each camera, which may vary from one camera to the other, and they also depend upon camera settings. DCD, MCSHR, MSCR and object reflectance are applicable only when an object image is obtained at medium/high resolution (i.e. larger than  $100 \times 40$  pixels) and there is a full-body visibility [89].

### 2.3.2 Texture

Texture represents the spatial distribution of the intensities in an object image and can be a key feature for person re-identification. Gabor and Schmid filters define two kernels for texture extraction applied to the luminance channel [71, 115, 136, 178, 179]. Gabor filters are linear filters used for edge detection. Frequency and orientation representations of Gabor filters are similar to those of the human visual system. Schmid filters are rotational invariant Gabor-like filters. HAAR-like features can be used to extract relevant textural information from the object image with the aim to find recurrent colour distributions [13]. Furthermore, the ratios between different regions in an image can be used as a discriminative feature. Ratios of colours, ratios of oriented gradients and ratios of saliency maps can also be used as textural features [20]. Similarly, Recurrent High-Structured Patches (RHSP) extract the most common blobs from the image [59]; in addition to this, salient spatio-temporal edges (edgels) obtained from watershed segmentation carry information of the dominant boundary and of ratios between RGB channels [64].

The distribution of spatial patches can be directly extracted in the frequency domain, for example, Discrete Cosine Transform (DCT) coefficients can be used as textural features [17]. Spatial patch distribution can be extracted by computing the first and the second derivatives of the person patch resulting in a covariance matrix [100, 165]. The symmetry within an object-image is exploited in the extraction of local features, by weighting features based on their position with respect to the symmetric part. In particular, Gabor and Schmid filters, and HAAR-like features are local descriptors suitable for small patches, while the ratios, RHSP, salient edgels, DCT coefficients, and covariance matrix can only be applied to images with medium/high resolution. Furthermore, the Histogram of Oriented Gradients (HOG) gives information on the orientation of the edges in an object image [100, 106, 121, 165], thus creating a feature that models the shape of the object by its edge distribution. However, HOG features are only invariant to changes in illumination and not to changes in pose and scale. Local Binary Patterns (LBP) are used to describe spatial patterns using normalised colour intensities [77, 157, 166]. LBP can be combined with HOG to extract the spatial information robust to illumination changes [157]. The Mean Riemannian Covariance Grid [14] is used to generate a human signature from the detected objects using LBP on the head regions [43]. Finally, interest points can be used for re-identification in the case of variations in scale, pose and illumination [18]. Examples are SIFT [11, 14, 94, 93, 158], SURF-like features [50, 75] and the Hessian Affine invariant operator [64].

### 2.3.3 Shape

Shape can be used in the form of object representation (Sec. 2.2), and features are extracted from the defined shape, while it can also be used as a feature itself [54, 143, 162]. A bounding box can be an essential or minimal representation of the object. The bounding box around each object is also exploited by extracting the angle formed by the vertical edge and the diagonal of the bounding box [39]. A more general feature is the height of the object when calibration information is available [20]. Another method defines the principal axis as the height of the object based on the camera spatial information [80].

A silhouette containing the pixels belonging to an object is also used to represent the shape of the object [39, 54]. A silhouette is obtained by background subtraction and is based on the salient edges [64]. The symmetry within the obtained silhouette can also be exploited in the extraction of local features and weighting of features based on their position with respect to the symmetric part [59]. These methods are robust to illumination changes but cannot deal with large pose changes. In another approach, the silhouette is divided into decomposable triangulated graph structures to represent more localised body parts [60]. However, silhouette segmentation requires high-resolution frontal or back images of the full-object body and without occlusions. Alahi *et al.* [4] define shape as rectangular regions starting from the centre and progressively moving outwards. Features can be extracted from the image using a Region Covariance Descriptor (RCD) [160], which aims to preserve shape, location and colour information. RCD is used in a multi-scale quadtree descriptor [10].

Gait is a feature from the class of soft biometrics, which can be used along with other appearance features [17]. Gait is obtained by background subtraction and by accumulation of the silhouette over time, which requires multiple high-resolution images containing side pose of objects. However, cameras may be located far from the objects in video surveillance, resulting in low-resolution images. The requirements to obtain gait may not always be fulfilled because of unavailability of full-body visibility or occlusions. Therefore gait remains an unsuitable feature in re-identification approaches [23, 35].

### 2.3.4 Grouping

Multiple features are grouped with the intention to maximise the inter-object discrimination. Appearance features are extracted from single [71, 136] or multiple images [122] in the form



of colour, texture and shape of an object [20, 54, 59, 71, 89, 100, 136, 179]. Histograms of colour-channels RGB [20, 39, 89], HSV [59], and YUV are used as colour features. Gabor and Schmid filters are applied to one image channel and the histograms of convolved images are used as texture information [71, 179]. LBP [108] and feature point descriptors like SIFT [121] are also applied to extract textures. Features from colour spaces (RGB, YCbCr, and SV) and texture types (Gabor and Schmid) are concatenated to increase their discriminative power [71, 136, 179]. The shape of an object can be preserved by RCD [160] and HOG [122]. HOG, SIFT and HSV colour histograms can be used for shape, texture and chromatic content to build a discriminative signature [121]. The mean colour values from small image regions can be combined with the histogram of LBP to represent the image, and then pairwise sample differences are learned for re-identification [79].

Features are also combined over time when extracted from multiple images of the same object. Features extracted from single images can be grouped over time either by temporal accumulation [75] or by clustering [59]. Features can also be incrementally updated over time, for example using Incremental MCSHR that updates MCSHR in order to increase robustness to abrupt changes in illumination [116]. Features extracted from images of the same person over time can be used as a set of positive samples for training a learning model-based method [13]. Satta *et al.* [145] divide images into small components and the difference is found from an existing bag of components, where the difference vector is represented as a descriptor of the image.

## 2.4 Dimensionality reduction

Dimensionality reduction is the process of retaining the relevant information by describing most but not all of the variance within the data. Dimensionality reduction reduces the amount of information necessary to represent data. Two general approaches for dimensionality reduction are: feature extraction and feature selection.

### 2.4.1 Feature extraction

Feature extraction is a transformation of data into a new feature space with lower dimensions. The most well-known feature extraction method is Principal Component Analysis (PCA) [151]. PCA projects a dataset to a new coordinate system by determining the eigenvectors and eigen-

values of a matrix. It calculates a covariance matrix of a dataset to minimise redundancy and to maximise variance. PCA and Kernel PCA are used for feature extraction in face recognition, where K-nearest neighbour is applied for classification [56]. PCA is suitable for data representation; however, it does not perform well in classification problems [151], such as re-identification.

Another feature extraction method, Fischer's Linear Discriminant Analysis (LDA) is more suitable for classification in the lower dimensional space [153]. LDA is a supervised technique to classify samples of different classes by transforming the data to a different space. LDA tries to find a line that best separates the two classes. A person re-identification approach combines the parametric and non-parametric representation of colour as features followed by the feature extraction using unsupervised PCA and supervised local LDA [130].

Moreover, Factor Analysis (FA) reduces the number of features by combining two or more features into a single factor [171]. FA is useful in identifying groups with similar features. In person re-identification, features are represented using local maximal occurrence, which analyses the occurrence of local features to form a stable representation against viewpoint changes, while a discriminant metric is learned by cross-view quadratic discriminant analysis [110]. K-means clustering can also be considered as a feature extraction method [173]. It is an unsupervised method, where features can be clustered and the centroid of the cluster can be considered as a transformed feature.

Feature extraction requires the complete initial feature set for the feature transformation. The extracted features can be altered if the initial feature set is unavailable/changed. Therefore, the reduction in the feature cost because of less number of features, compared to the initial set, may not be possible; however, the benefit in the performance gain can be achieved.

#### **2.4.2 Feature selection**

Feature selection aims at finding the most important features and their combinations for effectively describing and matching objects [46, 74]. Feature selection is an NP-hard problem. Approaches based on heuristics exist, which approximate the solution by exploiting problem-specific properties. Selection approaches produce a subset of features [97] and reduce redundancies among features [62]. Feature selection is also an important pre-processing step in machine learning that avoids over-fitting and increases the effectiveness of learning. Features can be selected either based on *group performance* or on their *individual performance* [175]. The set of individually selected features may not collectively provide good classification performance

because of lack of information about feature correlation, while individual weak features may provide strong discriminatory power in a group [82]. However, individually selected features can perform well in constrained environments, for example if features need to be adaptively discarded because of user requirements, application constraints or resource-constrained devices; whereas in the case of feature grouping, the removal of a single feature may significantly reduce the effectiveness of the whole feature set.

A method for *ranking* features according to their contribution to the task is presented by Wei *et al.* [168]. The similarity between features is measured to generate a score for each feature. The highest-scoring feature is selected and the process is repeated to choose the next relevant feature. The feature importance and similarity between features can be exploited with a greedy selection method [62], or boosted regression trees can be applied [128]. A hierarchical feature selection method is developed by using RankSVM along with a quality measure to predict the number of selected features [81]. The best-first search can be used to partition the features into subsets that are then combined to maximise the defined information-retrieval measures [46]. The coherence between subgroups of data can also be used to rank features [82]. An approach based on cooperative game theory evaluates the performance of each feature individually and within groups to achieve a single collaborative goal [155]. The structural similarity between data before and after feature selection is maintained and topological neighbourhood information is used for computing the structural similarity [125]. An unsupervised feature-ranking algorithm discovers Bi-clusters (subsets of data exhibiting similar behaviour for a subset of features) that are used to evaluate feature inter-dependencies, separability of instances and feature ranking [82]. This approach inherits some characteristics from ranking and wrappers. Wrappers use learning methods for feature selection and are classifier-dependent. A minimum-redundancy maximum-relevance (mRMR) based approach can be combined with a wrapper method to select a compact subset from the candidate features [131]. A kernel-based feature selection criterion incorporates the kernel trick with the class separability measures [163], where the kernel parameters are automatically tuned by maximising kernel class separability criteria. Feature selection based on a distance discriminant method converts the search problem of feature selection into feature ranking. The approach achieves feature selection performance comparable to exhaustive-search methods with a lower computational complexity [109]. The hierarchical clustering is applied to select the optimal feature subset [175]. Features can also be selected based on improved per-

Table 2.2: State-of-the-art feature selection methods.

		[62]	[125]	[175]	[81]	[155]	[82]	[46]	[163]	[109]	[131]	[140]	[111]	CoPE
<b>Selection approach</b>	Best First Search	✓						✓						✓
	Structural Similarity		✓											
	Feature Cooperation			✓										
	Hierarchical Clustering			✓	✓									
	Game Theory					✓								
	Co-ordinate Ascent							✓						
	Kernel Class Separability								✓					
	Random forest													✓
	Bi-clusters						✓							
	mRMR										✓			
	ReliefF											✓		
Distance Discriminant										✓			✓	
<b>Dataset</b>	Text Retrieval	✓												
	Medical Data	✓												
	UCI ML Benchmarks		✓	✓		✓	✓		✓	✓	✓			
	LETOR 4.0				✓									
	Handwriting Images								✓					
	Carnegie Mellon Datasets									✓				
	Bio-Informatics						✓				✓			
	UCI regression											✓		
Surveillance Videos												✓	✓	
<b>Evaluation criteria</b>	Performance	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Cost													✓

formance in sub-groups [112] of data. Recently, Ensemble of decision trees known as random forests are used to group images of persons into sub-clusters based on similarity in the colour and texture information, and the features relevant to each sub-clusters are selected by weighting to improve the re-identification rate [111, 112]. Table 2.2 summarises and compares state-of-the-art feature selection methods.

## 2.5 Data compression

Data compression refers to the process of reducing the amount of data needed to represent the information. The reduced information can be stored and communicated over the network; however, data compression involves the additional overhead of encoding and decoding [69]. In a re-identification scenario, the information can refer to the object representation in the form of images and feature descriptors. Three types of redundancy can be identified and exploited in the compression of object representation, namely, coding redundancy, spatial and temporal redundancy, and irrelevant information. *Coding redundancy* occurs when more bits than needed are used for the representation. *Spatial and temporal redundancy* occurs when information is repeated, for example a pixel similar to the neighbouring pixel, or pixels similar to those in the next frame. *Irrelevant information* is the information which, if removed, does not affect the original

information.

Compression can be of two types: (i) lossy compression, where the compressed information cannot be completely recovered after the decompression, and (ii) lossless compression, where there is no information loss. Removal of irrelevant information results in a lossy compression, for example in the case of quantisation. Run Length Encoding (RLE) [69] is a lossless compression technique suitable for spatial and temporal redundancy. RLE groups the similar data values as run-length pairs, where each pair contains the data value and the number of times it is repeated. Huffman coding [83] is another lossless compression method. Huffman coding generates the smallest possible number of code symbols to represent the source symbol. A code lookup table needs to be maintained and communicated for the decoding of Huffman code. Lampel-Ziv-Welch (LZW) coding [170], on the other hand, does not require the codebook to be communicated. An identical code book is generated while decoding, which removes the requirement of storing and sharing the codebook along with the compressed data. In arithmetic coding, [1] an entire sequence of source symbols is assigned to a single arithmetic codeword. As the information increases, the interval (arithmetic value) used to represent it becomes smaller, while its storage requirement increases.

Compression approaches can be applied in parallel to the existing object representation and feature selection methods to further achieve data reduction in storage and communication. The performance of compression methods improve if the data is provided in a batch, since more redundancy is expected.

## **2.6 Cross-camera calibration**

Cross-camera calibration uses scene and context information to assist the feature descriptors and improve re-identification. It includes colour-calibration and spatio-temporal calibration. Colour calibration maps the colour information from one camera to the other [116, 132, 135]. Spatio-temporal calibration encapsulates information about the camera deployment, the spatial relation between cameras, the entry/exit points in the scene and the travelling time across cameras [58, 89, 120, 123].

### 2.6.1 Colour calibration

Cameras are deployed at different positions relative to the light conditions in a scene, which results in variation in illumination conditions across cameras. The colour responses of individual cameras may also differ from each other. Cross-camera colour calibration models the colour relationship between pairs of cameras to compensate for illumination variations and cameras' colour responses. Similar colour responses can be achieved by iteratively tuning the camera hardware parameters [85]. Madden *et al.* [116] perform the intensity transformation by a cumulative histogram equalisation of the data from each camera view. A colour mapping is obtained by finding the minimum cost path from the correlation matrix between two colour histograms [132]. Black *et al.* [24] perform the colour calibration by minimising the inter-bin distances between object histograms across cameras. Javed *et al.* [89, 90] perform a direct mapping of brightness value from one view to another, while assuming that a certain percentage of an object image in a camera view has the brightness less than or equal to  $B_i$  and this percentage is equal to the percentage of brightness less than or equal to  $B_j$  in another camera view. This approach requires a learning stage, where for each camera-pair a relationship must be found, which needs to be updated over time to cope with changes in the lighting conditions throughout the day. It is demonstrated that all brightness transfer functions (BTF) lie in a low-dimensional space that is discovered using PCA on RGB colour intensities [89]. Clustering on the chromaticity space can also be used to find an affine colour calibration [92].

Colour mapping models trained for a single illumination condition need to be re-trained as the illumination of the scene varies over time because of variations in the lighting conditions, weather and the sun position. Gilbert *et al.* [65] introduce an online learning method for inter-camera illumination changes. Objects are tracked across camera view, and RGB transformations are obtained using singular value decomposition. However, the approach requires good inter-camera correspondence for the training. Chen *et al.* [37] exploit probable matches to calculate the BTFs. The BTF subspace is updated over time by merging the new BTFs into the already learned BTF subspace. An improvement of this approach is the use of Cumulative BTF (CBTF), where the contribution of less common training samples is taken into account [135]. An Adaptive-CBTF exploits the background information to estimate changes in the illumination conditions of the foreground over time [134]. The colour calibration can perform well in the case of large inter-camera illumination changes; however, it can only be applied to scenes where abrupt illumination

changes are unlikely to happen.

### 2.6.2 Spatio-temporal calibration

Spatio-temporal calibration exploits the knowledge of the environment in which cameras are deployed to estimate when and where objects can reappear in the next camera, thus restricting the re-identification task within a certain time interval and certain regions of the monitored scene [120]. Spatio-temporal calibration performs a key role in the case of multi-camera association. Learning the travelling time across cameras can be complemented by the learning of probable entry/exit regions in the camera network [58]. Kuo *et al.* [100] combine the information of travelling time across cameras and the expected entry/exit points in the scene, with the appearance model to obtain a probability of matching. When the relative camera positions are known, people location and speed can also be discriminative features for each person [39].

Kettner *et al.* [95] propose a Bayesian approach for disjoint camera views, where it is specified that a person can be in one camera at a time. The approach requires manual input such as expected transition time between cameras, whereas entry/exit regions are selected manually in [135]. Javed *et al.* [89] learn the inter-camera transition time using known object-association information, the exit velocity obtained from object tracking within a camera, and known entry/exit points. The probability density functions of transition times combined with the appearance information are used in the object association. A Markov model can be applied to understand the discontinuities in the tracking between cameras [52]; however the approach requires an identifiable object to be manually traversed through the network, for example a red ball in this case. Another approach exploits the trajectory of an object passing through different camera views as a Markov chain model, where the position of the object is updated over time using the velocity [137]. In this approach the object information is used to estimate the camera position in a global space. However, the approach assumes that the image plane and the ground plane are parallel, which is not common in surveillance settings. An extension to this approach uses the Kalman filters to estimate the tracks between the cameras [8]; however, the approach relies on the assumption of linear motion, and therefore cannot predict the obstacles in between, such as walls or other objects. On the other hand, in [123], hypotheses about locations of objects in non-observed regions are generated based on the velocity of the objects, their position in the observed regions, and by using an area map.

The approaches that exploit the spatial information are suitable for scenarios where non-

observed regions are easy to model and people always follow the same paths. Makris *et al.* [120] instead exploit the temporal transitions to create a topological map of the network. Each camera's entry/exit points are clustered using expectation maximisation [25, 119]. The peak in the time differences between disappearance of objects from exit nodes and reappearance in the entry nodes defines a temporal link between the two nodes. Gilbert *et al.* [66] extend [120] by incorporating online recursive topology learning, and combine it with the appearance model from [88]. The approach needs a light training on good initial tracking to start with. In [120], inter-camera transition time is considered as a simple Gaussian distribution, which is extended to a multi-model distribution in [159] by employing Markov Chain Monte Carlo process. Moreover, Cai *et al.* [26] extend the Gaussian distribution by applying k-means clustering to separate the transition times based on slow medium and fast object movements. Another approach suitable for busy scenes, such as tube/train stations, exploits activity correlation to estimate camera transition times. Activities that are repeated in time, for example train arrival, are identified to develop a temporal link between cameras.

## 2.7 Object association

Association is the comparison of the extracted information (Sec. 2.2 and Sec. 2.3) from objects across cameras to identify different instances of the same object. In order to perform the association, we can measure the feature (dis)similarity using distance minimisation, learned classifier, or by optimisation process (Table 2.3).

### 2.7.1 Distance minimisation

Person association using distance minimisation estimates the point-to-point dissimilarity between feature vectors. The Euclidean distance is applied on feature vectors representing colour values [59], interest points, or hypotheses about the locations [64, 123]. The Euclidean distance between two colours is also included in an ad-hoc similarity measure created to compare two DCD feature sets [13]. Alternative measures are the sum of quadratic distances [50] and the sum of absolute differences [75]. Other distance measures include the Kullback-Leibler Distance [20, 92], the Bhattacharyya Distance [59, 135] and the Mahalanobis distance [59]. An additional measure derived from the Kolmogorov distance is introduced by Madden *et al.* to compare IMCSHR features [116]. If features do not belong to the Euclidean space, the Euclidean distance cannot



Table 2.3: State-of-the-art methods for person re-identification. Legend: Spatio-temp = Spatio-temporal, Distance = Distance based, Learning = Learning based, Optim = Optimisation based.

Ref.	Appearance features			Temporal grouping	Calibration		Association		
	Colour	Texture	Shape		Colour	Spatio-temp	Distance	Learning	Other
[13]	✓	✓		✓			✓		
[17]		✓						✓	
[20]	✓	✓	✓	✓			✓		
[39]	✓	✓	✓			✓			✓
[50]	✓	✓					✓		
[59]	✓	✓		✓			✓		
[64]	✓	✓		✓			✓		
[71]	✓	✓						✓	
[75]		✓		✓			✓		
[89]	✓			✓	✓	✓			✓
[92]	✓				✓		✓		
[100]	✓	✓	✓	✓		✓			✓
[116]	✓			✓			✓		
[123]						✓	✓		
[132]	✓				✓				✓
[135]	✓				✓	✓	✓		
[136]	✓	✓						✓	
[158]		✓		✓				✓	
[165]	✓	✓	✓					✓	
[178]	✓	✓						✓	

be used [143]. For example, a covariance distance metric is used for covariance descriptor [160]. Correlation between colour histograms and HOGs of the objects is also used in [100]. If features do not belong to the Euclidean space, the Euclidean distance cannot be used [143]. For example a covariance distance metric is used for covariance descriptor [160]. In distance minimisation methods, the most challenging part is the selection of the best distance for the specific set of features usually performed by trial and error. One common aspect of these measuring methods is non-discrimination between features. Approaches based on measuring similarity between feature sets are not robust to illumination changes unless cross-camera colour transformation is performed. For example, a spatio-temporal relationship is used by Chen et al. [37] to find the probability of matching a person from one camera to another, coupled with an adaptive BTF to handle illumination changes.

### 2.7.2 Learning classifiers

An alternative to direct measures and colour calibrations, a classifier can be trained to learn the changes between cameras using pairs of features labelled for positive (same objects) or negative (different objects) classes. Support Vector Machines (SVM) can be employed with DCT features [17] and SIFT [158]. An improvement to SVM is the Ensemble SVM, which reduces the computational cost of RankSVM for high-dimensional feature spaces besides converting the

re-identification problem into a ranking problem [136]. Furthermore, AdaBoost is applied for person re-identification to learn weak classifiers based on different feature sets and to identify the most discriminative features [13, 71]. In an unsupervised learning approach, appearance attributes are used to mine Attribute Sensitive Feature Importance (ASFI), which is then combined with global features [111, 112].

A learning approach Large Margin Nearest Neighbour (LMNN) performs a linear transformation to minimise distances between a feature point and its K-neighbours with the same label, and maximise distances from those with different label [53]. Another approach based on LMNN, Probabilistic Relative Distance Comparison (PRDC) [178], maximises the probability of correct matches while minimising that of wrong matches by learning the best distance measure for the association. Unlike direct distances, these methods are less sensitive to feature selection. However, their results can be biased by the selection of the classifier parameters, thus making the methods less flexible in different scenarios.

### 2.7.3 Optimisation approaches

Other approaches use optimisation-based algorithms. The concept of belief/uncertainty assignment can be exploited and the decision for the association problem can be made on specific ad-hoc rules [39]. An alternative approach finds the maximum likelihood Probability Density Functions (PDF) for the appearance and spatio-temporal features of different observations of the same object. The final decision is made by split graph [89]. Re-identification can also be performed by Hungarian algorithm using colour, texture, and spatio-temporal features [100], where the 'potentially' correct matches are selected by Multi Instance Learning boosting on the spatio-temporal features. Finally, dynamic programming is used to fit body models across cameras [64]. The main drawback of optimisation-based approaches is that they operate in a batch mode and cannot run on-line.

## 2.8 Datasets

Snapshot-based and video-based datasets are used for the evaluation of person re-identification methods. Some of the datasets have become standard by their extensive use, such as VIPeR and iLIDS. Some datasets are created to evaluate the specific scenarios and challenges of re-identification. In addition, we present two self-generated datasets: iLIDS-TC dataset [J1] having

Table 2.4: Dataset for person re-identification

Ref	Name	Type	No. of cams	Scenario	Video resolution	FPS	No of persons	Image size
[13]	iLIDS-MA	Images	-	Indoor	-	-	40	21x53 to 176x326
[14]	iLIDS-AA	Images	-	Indoor	-	-	119	21x53 to 176x326
[15]	3DPeS	Video	8	Outdoor	704x576	15	200	31x100 to 176x267
[16]	Visor	Images	-	Outdoor	704x576	-	50	54x187 to 149x306
[21]	SAIVT	Video	8	indoor	640x480	30	150	30x95 to 30x156
[38]	CAVIAR4REID	Images	-	Indoor	384x288	-	72	17x39 to 72x144
[48]	RAiD	Video	4	Outdoor	640x480	30	16	32x62 to 86x170
[70]	VIPeR	Images	-	Outdoor	-	-	632	128x48 to 176
[121]	WARD	Images	3	Outdoor	-	-	70	20x85 to 30x156
[126]	iLIDS	Video	5	Indoor	640x480	25	1000	21x53 to 176x326
[148]	PETS2009	Video	8	Outdoor	768x576	7	40	26x67 to 57x112
[146]	ETHZ	Video	1	Outdoor	640x480	15	146	13x30 to 158x432
[149]	TRECvid2008	Video	5	Indoor	640x480	25	300	21x53 to 176x326

occluded instances of objects in a camera-pair, and the Torch dataset [C2] with five cameras that view the same objects overtime. Table 2.4 summarises the datasets used to evaluate the state-of-the-art approaches.

Common snapshot-based datasets are: iLIDS [13, 59, 136, 177], VIPeR [71], WARD [122], UnderGround-GRID [111, 114] and CAVIAR4REID [38] (Fig. 2.2). These datasets are used to validate appearance-based methods mostly containing people with full-body visibility. In *iLIDS* images of people taken from four cameras at London Gatwick airport represent an indoor setting in a crowded environment. Four datasets extracted from the iLIDS are: iLIDS-MCTS [177], iLIDS-MA [13], iLIDS-AA [14] and iLIDS-MTC [J2]. *iLIDS-MCTS* [177] contains 476 images of 119 people in four cameras. *iLIDS-MA* [13] contains multiple images of 44 people manually extracted from video frames. *iLIDS-AA* [14] contains multiple images of 100 people automatically extracted using a HOG detection algorithm. *iLIDS-MTC* [J2] contains manually cropped multiple images of 60 pairs of persons in two cameras. *VIPeR* [71] contains 632 image-pairs of people taken from two outdoor arbitrary viewpoints [71] and presents significant pose changes. A more recently introduced dataset *WARD* [121] contains 70 persons from three non-overlapping fixed-cameras with the challenges of illumination changes, and variations in pose and size. *UnderGround-GRID* [111, 114] dataset has eight cameras with non-overlapping FoV in an underground train station. The dataset contains 250 pairs of images between two cameras and 775 images of people in a single camera. A multi-camera tracking dataset *CAVIAR4REID* [38] represents an indoor shopping mall with two partially overlapping camera views. The dataset contains multiple images of 72 pedestrians, where 50 people appear in both cameras and 22

people remain in one only.

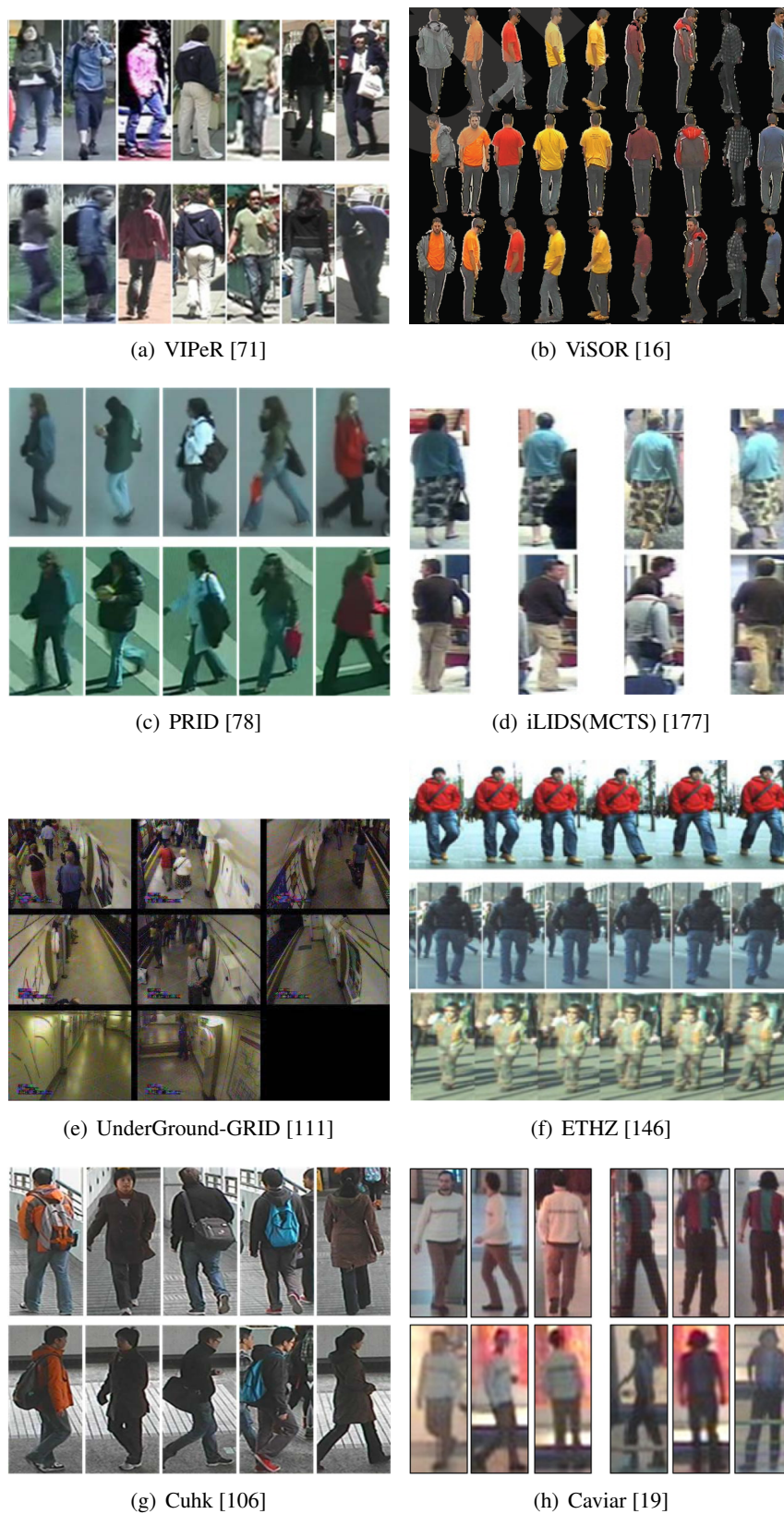


Figure 2.2: Datasets used in the evaluation of person re-identification approaches.

Video-based datasets include: Terrascope [92], V-47 [164], SAIVT-SoftBio [21], ETHZ dataset [146], and two more datasets from Javed *et al.* [89] and Kuo *et al.* [100]. The *Terrascope* dataset [92] consists of nine indoor cameras where eight people walk and act in an office environment. *V-47* [164] dataset contains videos of 47 pedestrians in two cameras in an indoor setting. This dataset is not crowded and has no illumination changes. *SAIVT-SoftBio* [21] consists of 150 people moving through an indoor building environment, recorded through eight calibrated fixed cameras. The challenges of illumination, appearance and pose changes exist in these datasets; however, the objects are without occlusions. *ETHZ* dataset [146] was originally designed for object detection. Dataset is gathered using moving cameras, which makes it more challenging than the datasets obtained using fixed cameras. Javed *et al.* [89] present video sequences from three cameras in indoor and outdoor scenarios with large illumination changes and up to four fully visible people. A more challenging dataset in terms of occlusions and with three outdoor cameras is presented by Kuo *et al.* [100], where up to 10 people walk alone or in small groups.

Self-generated datasets include iLIDS-TC [J1] and Torch [C2]. *iLIDS-TC* dataset contains 348 people that transit from camera 1 ( $C_1$ ) to camera 3 ( $C_2$ ) in iLIDS. These two cameras represent non-overlapping views with considerable illumination changes in an indoor camera setting. Images of 124 people are manually extracted. Each person is represented by a pair of images; one while exiting a camera,  $C_1$ , when the person is completely in  $C_1$  and the other on reappearance when (s)he is completely in  $C_2$ . People can be only partially visible because of occlusions. For the remaining 224 people, we utilise the four existing datasets iLIDS-MA [13], iLIDS-AA [14], iLIDS-MTC [J2] and iLIDS-MCTS [178] extracted from iLIDS videos and select one image per person per camera, such that no person is repeated. All images are normalised to 128 x 64 pixels. The *Torch* dataset is recorded during the Olympics 2012 torch relay passing through Mile End road in London, UK. The dataset represent an outdoor crowded scene. Five partially overlapping hand-held smartphones are used thus leading to occasional jitters and blurring (Fig. 2.3) in addition to changes in illumination, size and pose of people, and occlusions. Single images of 50 people common in all cameras are manually extracted on their first appearance in each camera and their detection frames are stored.

Challenges of illumination and view point changes are common in all datasets. The additional challenge of occlusion can be observed in iLIDS, ETHZ and Torch datasets, where ETHZ and Torch datasets also include the challenges related to camera-movement as discussed in Sec. 1.3.



Figure 2.3: Two sample frames from (a)  $C_3$  and (b)  $C_2$  in the Torch dataset. These frames are captured almost at the same time instance and represent two very different views of the same scene.

## 2.9 Discussion

Re-identification approaches focus on different phases of the problem from object acquisition to label assignment as discussed in this chapter. Re-identification algorithms solely based on appearance usually achieve an accuracy of less than 40-50% [136] for the first ranking position (the real re-identification score) because of challenges related to changes in pose and illumination conditions, positions of the cameras and occlusions. Re-identification algorithms that operate in batch mode also exploit spatio-temporal features, achieving results usually over 90% [89] for the first ranking position in scenes with full-body visibility and uninterrupted straight-line transition of people in non-observed regions (using a self-generated dataset). Nevertheless, methods solely based on appearance can be tested using single snapshots of people and they become very important when cameras are located far apart, where cross-camera calibration is very challenging and spatio-temporal relations become unreliable.

In order to improve the recognition rate, a large number of features are extracted, combined and communicated over the network [54, 71, 89, 136, 179]. After a certain number of feature concatenations, additional features might decrease the re-identification performance while large feature sets demand high processing, storage and transmission capabilities [154]. The choice of the useful information, such as via feature selection, has not been explicitly applied in re-identification. Until very recently a feature selection approach has been proposed with the motivation of improving the re-identification [111] with no constraints on the resources, such as, for example, the computational time for extraction of information and the amount of data generated for the storage. The cost constraint in feature selection becomes particularly important when the cost varies significantly across features to be shared among nodes of a smart camera network.

We argue that the cost of a feature should be considered jointly with its performance, measured as its ability to represent and discriminate an object, for the feature selection.

Another important aspect in the existing re-identification approaches is that persons are associated in a camera-pair, which assumes that the camera with the previous detections of the same person (source-camera) is known [20, 59, 71, 136, 179]. Spatio-temporal calibration information can be used for the source-camera selection if the paths to be followed can be identified/known [120, 100, 135, 89]. In the case of surveillance of open public areas with multiple entry/exit points and varying persons' movements, a person detected in one camera can have for its previous instance multiple candidate cameras with varying views, illumination and appearance settings. The pairwise-associations required for re-identification need to be extended for multiple cameras taking into account both the inter-camera and inter-person variations. In such a case, a many-to-one camera association can be performed for person re-identification.

Based on this survey, we propose a feature selection method that represents the object with a selected set of cost-effective and well-performing features and this reduces the storage and computation requirements within a camera (Chapter 3). Next, we minimise the requirement of communication between cameras for temporal alignment and object association (Sec. 4.2). Finally, we extend from pairwise to multiple source-cameras the associations for re-identification (Sec. 4.3).

## Chapter 3

### Cost-effective features

---

#### 3.1 Introduction

Multiple types of feature are exploited in existing re-identification approaches (Sec. 2.3) for improving the re-identification rate without considering constraints on resource utilisation, which significantly vary between the features. We propose a Cost-and-Performance-Effective (CoPE) feature selection method that selects features which are both well-performing and inexpensive such that feature selection and object association can be performed within smart cameras.

In this chapter, we first describe the object acquisition method from the head bounding box that minimises the number of pixels required for object representation [J2]. We also define the type of features that are used throughout this research. Next, we discuss the proposed CoPE approach that combines the cost (computational time and storage size) of using features with their performance in re-identification to identify the most appropriate feature subset for the task of person re-identification in a smart camera network [J1]. Instead of optimising the combined contribution of the best set of features, the most discriminative, well-performing and cost-effective features are selected by evaluating each feature individually. Selected features are ranked in accordance with their added contribution to the task.

In Sec. 3.2, we describe the object acquisition step. The features used in our work are discussed in Sec. 3.3. We measure the performance of a feature for re-identification in Sec. 3.4 and in Sec. 3.5 cost of feature is estimated. Sec. 3.6 discusses the combining strategy to select a subset of the best performing features. In Sec. 3.7, we discuss the application of the CoPE in a



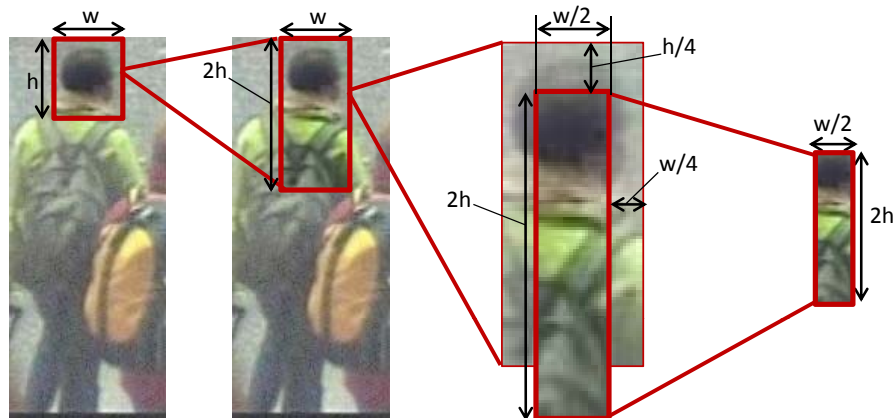


Figure 3.1: Selection of the support for person representation. Starting from the bounding box resulting from head detection, a stripe whose height is twice the height ( $h$ ) and half the width ( $w$ ) of the bounding box of the head is selected. The stripe is shifted downward by  $h/4$  to reduce the presence of background pixels in the features used for the association.

camera network. Finally, Sec. 3.8 summarises the chapter.

### 3.2 Upper-body image representation

We introduce a person representation model for crowded scenarios that is defined as a vertical stripe around the head location (Fig. 3.1) [J2]. The head and the upper-body are the most frequently visible and recognisable parts of a person in the case of surveillance settings with multiple people in the scene [176]. We assume that the person detection phase is solved using a head detector [68, 172] resulting in a bounding box  $b_n^m = (x, y, w, h)$  for the head of a person,  $P_n^m$  in  $C_n$ , where  $x$  and  $y$  are the x-y coordinates of the top left corner,  $w$  is the width and  $h$  is the height of the bounding box. From a given bounding box  $b_n^m$ , a vertical stripe of the upper-body is generated as:

$$P_n^m = f(b_n^m) = [x + w/4, y + h/4, w/2, h * 2]. \quad (3.1)$$

A set of features as discussed in Sec. 3.3 is extracted from the upper-body shape. We compare the re-identification performance of the proposed upper-body shape with the existing approaches using full-body object representation that is divided into multiple horizontal stripes (Sec. 5.2). Full-body image representation is an image obtained in object acquisition that includes all the body parts of an object.

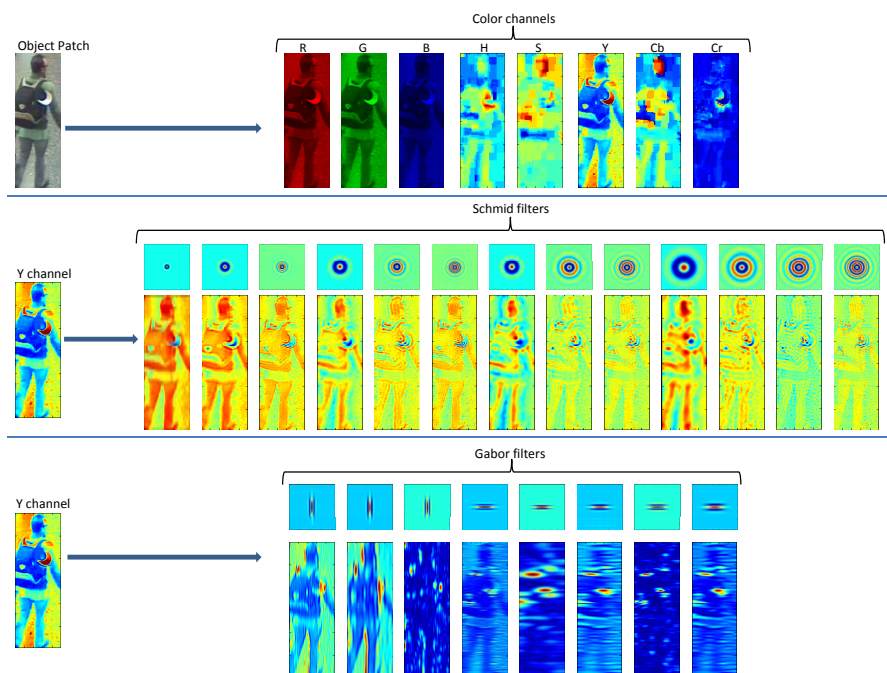


Figure 3.2: Colour and texture features extraction. Colour features are extracted from different channels (top row). Texture features are extracted by applying Schmid (middle row) and Gabor (bottom row) filters on the Y channel [71, 112, 136, 179].

### 3.3 Feature sets

We use the most commonly employed colour and texture features in re-identification<sup>1</sup> [71, 112, 136, 179]. These features remain suitable for extraction in most surveillance settings because they do not depend strongly on the object appearance (Fig. 3.2). Each feature is a 12-bin histogram of a colour channel or a filtered image. Nine colour channels (R, G, B, Y, Cb, Cr, H, S, V) from RGB, YCbCr and HSV colour spaces are used. For texture, Gabor and Schmid filters are applied on the Y-colour channel of the image. Eight Gabor filters are applied with the following parameters:  $(\gamma, \theta, \lambda, \sigma^2) = (0.3, 0, 4, 2), (0.3, 0, 8, 2), (0.4, 0, 4, 1), (0.3, \pi/2, 4, 2), (0.3, \pi/2, 8, 1), (0.3, \pi/2, 8, 2), (0.4, \pi/2, 4, 1), (0.4, \pi/2, 8, 2)$ , where  $\gamma$  is the aspect ratio,  $\theta$  is the angle in radian,  $\lambda$  is the wavelength of sinusoidal factor and  $\sigma^2$  is the variance. Thirteen Schmid filters are applied with the following parameters:  $(\sigma, \tau) = (2, 1), (4, 1), (4, 2), (6, 1), (6, 2), (6, 3), (8, 1), (8, 2), (8, 3), (10, 1), (10, 2), (10, 3), (10, 4)$ , where  $\sigma$  is the standard deviation and  $\tau$  is the number of cycles.

<sup>1</sup>It is to be noted that the proposed approach to select cost effective features does not depend on the type of features extracted.

### 3.4 Feature performance

The feature performance is measured for each pair of the destination and source-cameras  $(C_n, C_{n_q})$ . The performance vector  $\mathbf{\Pi}^r$  represents the performance of a feature  $\mathbf{f}^r$  on  $M$  persons visible in two cameras. The training set is composed of  $\mathbf{P}_n = \{\mathbf{P}_n^m\}_{m=1}^{M_n}$  and  $\mathbf{P}_{n_q} = \{\mathbf{P}_{n_q}^k\}_{k=1}^{M_{n_q}}$ , where  $M_n = M_{n_q} = M$  in the case of two cameras, and same value of  $k$  and  $m$  represents the same person. We extract the feature sets  $\mathbf{F}_n^m = \{\mathbf{f}_n^{mr}\}_{r=1}^R$  and  $\mathbf{F}_{n_q}^k = \{\mathbf{f}_{n_q}^{kr}\}_{r=1}^R$  from each object-image in  $\mathbf{P}_n$  and  $\mathbf{P}_{n_q}$ , respectively. We measure the performance of  $\mathbf{f}^r$  by analysing the similarity between the two views of the same person as well as the similarity with the other  $M - 1$  people, using  $\mathbf{f}^r$ .

The similarity between two instances  $\mathbf{f}_n^{mr}$  and  $\mathbf{f}_{n_q}^{kr}$  of  $\mathbf{f}^r$  is obtained by a relative matching distance function<sup>2</sup>  $g(\cdot, \cdot)$ , which receives as input a feature pair and returns the feature similarity  $d^{mkr}$  between  $\mathbf{P}_n^m$  and  $\mathbf{P}_{n_q}^k$  as

$$d^{mkr} = g(\mathbf{f}_n^{mr}, \mathbf{f}_{n_q}^{kr}). \quad (3.2)$$

For each  $\mathbf{P}_n^m$ , we have  $M$  distances, each from  $\mathbf{P}_{n_q}^k$  in  $\mathbf{P}_{n_q}$ . Each  $d^{mkr}$  is normalised ( $0 \leq \hat{d}^{mkr} \leq 1$ ) as

$$\hat{d}^{mkr} = \frac{d^{mkr} - \min_k d^{mkr}}{\max_k d^{mkr} - \min_k d^{mkr}}, \quad (3.3)$$

where  $\min_k d^{mkr}$  and  $\max_k d^{mkr}$  are, respectively, the minimum and the maximum distances of  $\mathbf{P}_n^m$  from  $\mathbf{P}_{n_q}$  using  $\mathbf{f}^r$ . The set of  $M$  normalised distances  $\hat{\mathbf{d}}^{mr} = \{\hat{d}^{mkr}\}_{k=1}^M$  contains one distance corresponding to the same person in  $C_n$  and  $C_{n_q}$  ( $\hat{d}^{mnr}$ : distance for correct match) and  $M - 1$  distances of  $\mathbf{P}_n^m$  from the instances of other persons in  $C_{n_q}$  ( $\mathbf{E}^{mr} \subset \hat{\mathbf{d}}^{mr}$ : the set of distances for incorrect matches).

In the ideal case, a feature  $\mathbf{f}^r$  is considered *well-performing* for  $\mathbf{P}_n^m$  if the distance between the correct matching pair is smaller than the minimum value of distances in  $\mathbf{E}^{mr}$  using  $\mathbf{f}^r$ . However, the ideal condition cannot be satisfied for a real-world re-identification scenario with the currently available features in the state of the art (Sec. 2.3) and the challenges involved (Sec. 1.3). Therefore, we need to relax the criterion, which requires an averaging method that could represent the whole mass of the data. Thus, the median of the incorrect distances,  $med(\mathbf{E}^{mr})$ , is selected because, compared to other averaging approaches, median gives the most central tendency of a set, and remains least effected by outliers and measurement errors. Although median ensures that the selected feature performs better for the majority of the cases, i.e  $> 50\%$ , it is still

<sup>2</sup> $g(\cdot, \cdot)$  refers to Bhattacharyya, L1-Norm and Chi-square distances (Table 5.1).

an empirical choice. In the case of an easy set of data, the median should be replaced with that of an ideal case discussed earlier, or a near ideal case (e.g. the average could be defined as greater than 95%, or  $2\sigma$ ). The performance score  $\Pi^{mr}$  is measured as

$$\Pi^{mr} = \frac{\hat{d}^{mmr}}{\text{med}(\mathbf{E}^{mr})}. \quad (3.4)$$

The condition  $\hat{d}^{mmr} < \text{med}(\mathbf{E}^{mr})$  leads to  $0 \leq \Pi^{mr} < 1$  in Eq. 3.4. The smaller  $\Pi^{mr}$ , the better the performance. The condition  $\Pi^{mr} \geq 1$  indicates that  $\mathbf{f}^r$  performs poorly. For each  $\mathbf{f}^r$ , we define the performance vector  $\mathbf{\Pi}^r$  using  $M$  persons as

$$\mathbf{\Pi}^r = [\Pi^{mr}]_{m=1}^M, \quad (3.5)$$

where each element  $\Pi^{mr}$  corresponds to the performance score of  $\mathbf{f}^r$  for a single person in the training data. The feature  $\mathbf{f}^r$  with  $\min_m \Pi^{mr} \geq 1$  for all  $M$  persons are discarded before performing the feature selection thus resulting in  $\hat{R} \leq R$  remaining features. We then define the  $M \times \hat{R}$  performance matrix  $\mathbf{\Delta}$  as

$$\mathbf{\Delta} = \left[ \Pi^{mr} \right]_{M \times \hat{R}}, \quad (3.6)$$

where  $m = 1, \dots, M$  and  $r = 1, \dots, \hat{R}$ . The  $r^{\text{th}}$  row of  $\mathbf{\Delta}$  represents the performance vector  $\mathbf{\Pi}^r$  of  $\mathbf{f}^r$  for  $P^n$ , while  $\boldsymbol{\chi}^m$  is the  $m^{\text{th}}$  column representing the performance comparison of  $P_n^m$  for  $\hat{R}$  features. The performance matrix  $\mathbf{\Delta}$  is further analysed jointly with the cost of features discussed in the next section.

### 3.5 Feature cost

We define the cost vector  $\boldsymbol{\Psi}^r$  of  $\mathbf{f}^r$  by considering two independent components: the computational time for feature extraction,  $\Gamma_n^{mr}$ , and the feature storage size,  $\beta_n^{mr}$ . The range of cost components can vary considerably across different  $\mathbf{f}^r$ , i.e.  $\Gamma_{mn}^r$  and  $\beta_n^{mr}$  can have extremely large or small values of time and size, respectively. In the case of irregular distribution of values, these extreme cases can result in making a single feature dominate others in the feature selection process. Since normalisation alone is not sufficient for such scenarios, we explicitly define the upper bound of the cost components by assuming that  $\min \beta_n^{mr} = 1 \text{ byte}$  and  $\min \Gamma_n^{mr} = 1 \text{ ms}$ . We aim to define and compare the basic units with useful information and least measurement errors. For size we select the smallest unit which retains the meaningful information, i.e. byte. For time we

select the unit millisecond (ms). While the minimum units may depend upon the measurement system, it is observed that under the given experimental conditions and measurement tools the ranges below ms are found to be less accurate and more sensitive to noise. The lower bound of cost components is obtained by taking the inverse of the average of the two cost components. Since we aim to select features with smaller cost components, the inverse average becomes useful by suppressing the large values while magnifying the small ones. The cost vector  $\Psi^r$  is given as

$$\Psi^r = (\Psi_\beta^r, \Psi_\Gamma^r) = \left( \frac{\alpha MN}{\sum_{n=1}^N \sum_{m=1}^M \Gamma_n^{mr}}, \frac{(1-\alpha)MN}{\sum_{n=1}^N \sum_{m=1}^M \beta_n^{mr}} \right), \quad (3.7)$$

where  $\alpha \in [0, 1]$  is a weight that accounts for the generalisation of the approach to different scenarios, where one constraint may be more important than the other. For example,  $\alpha = 0$  when a limited storage space is available with no constraints on the computational time, and  $\alpha = 1$  for vice versa.

In order to combine the two cost components, we perform scaling, which standardise the ranges of the two independent values. The scale factor may vary depending upon the system requirements. In this research, the scale factor is defined as  $1 \text{ byte} = 1 \text{ ms}$ . We measure the magnitude of the cost vector by calculating the Euclidean norm  $\|\Psi^r\|$ , where  $0 < \|\Psi^r\| \leq \sqrt{2}$ . The larger  $\|\Psi^r\|$ , the cheaper the feature. Since this cost score is obtained by combining two independent components in Eq. 3.7, new cost constraints can be included as additional independent components of the vector.

### 3.6 Feature selection

We perform a competitive feature selection by exploiting  $\Psi^r$  and  $\Pi^r$  of each  $\mathbf{f}^r$  (Fig. 3.3) such that the least costly features exhibiting the best performance are selected. We define a vector  $\mathbf{V}$  that contains the elements  $\Pi^{mr} \leq 1$  from  $\mathbf{\Delta}$  sorted in ascending order. We divide  $\mathbf{V}$  into  $\hat{R}$  bins where each bin  $I_i$  contains  $M$  performance scores sorted in decreasing order such that in the best case the feature with the best performance for all the  $M$  persons can be selected in a single iteration. A set  $\Phi_i^r$  is defined which contains the performance scores  $\Pi^{mr}$  within  $I_i$  for each  $\mathbf{f}^r$ .

Figure 3.4 shows an example of performance matrix  $\mathbf{\Delta}$ , and highlights the vector  $\mathbf{V}$ , the bin  $I_i$  and the set  $\Phi_i^r$ . We iteratively traverse each bin  $I_i$  until all performance scores in  $\mathbf{V}$  are exploited for feature selection. Cost is considered jointly with performance to select a cheaper feature

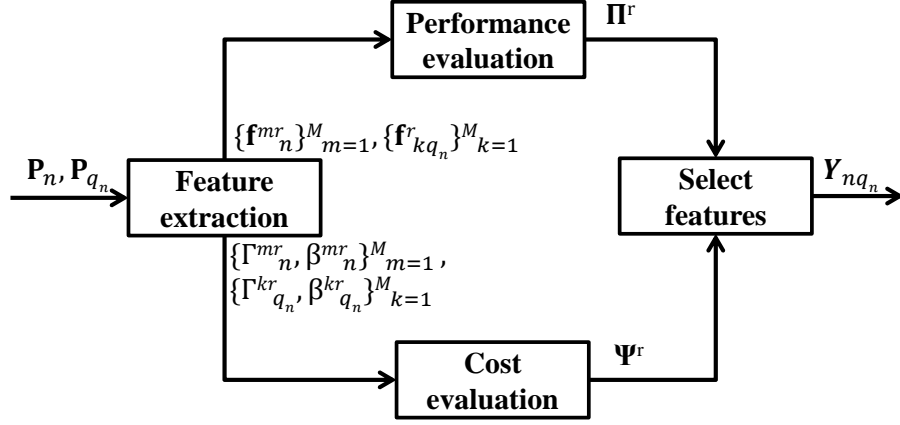


Figure 3.3: Block diagram of the proposed Cost-and-Performance-Effective (CoPE) feature selection approach.

when comparable results can be obtained by the features in the set. We calculate the combined importance score  $A_i^r$  of each  $\mathbf{f}^r$  within  $I_i$  as

$$A_i^r = \frac{|\Phi_i^r| \|\Psi^r\|}{\text{med}(\Pi^r)}, \quad (3.8)$$

where  $\|\Psi^r\|$  is the Euclidean norm,  $|\Phi_i^r|$  is the cardinality of  $\Phi_i^r$  that represents the number of persons for which  $\mathbf{f}^r$  has the performance scores within  $I_i$ , and  $\text{med}(\Pi^r)$  is the median of  $\Pi^r$  representing the overall performance of  $\mathbf{f}^r$  in the whole data set. The importance score  $A_i^r$  gets the maximum value for the feature  $\mathbf{f}^r$ , which has the least cost, the maximum number of  $\Pi^{mr}$  within the interval  $I_i$  (highest performance in  $I_i$ ), and the highest average performance. The best performing feature can be selected as

$$\hat{r} = \arg \max_r A_i^r, \quad (3.9)$$

where  $\hat{r}$  is the ID of the feature with the highest combined importance score  $A_i^r$ .

Let  $\mathbf{Y}_{n_{q_n}}$  be the list of selected features for  $C_n$  and  $C_{n_q}$ . If  $\mathbf{f}^{\hat{r}} \notin \mathbf{Y}_{n_{q_n}}$  then  $\mathbf{f}^{\hat{r}}$  is appended in  $\mathbf{Y}_{n_{q_n}}$ . The set  $\mathbf{Z}$  contains the list of persons from the dataset that have already taken part in the selection of  $\mathbf{f}^{\hat{r}}$ , given as  $\mathbf{Z} \cup P_n^m \forall \Pi^{mr} \in \Phi_i^{\hat{r}}$ . Once  $\mathbf{f}^{\hat{r}}$  is selected, we remove from  $\mathbf{V}$  the performance scores  $\Pi^{\hat{r}}$ , and  $\chi^m$  for which  $\chi^m \cap \Phi_i^{\hat{r}} \neq \emptyset$ . This removal avoids the selection of a feature that has the same performance as that of an already selected feature. Each selected feature is now representative of a unique subset of data, thus increasing the diversity in the feature set by covering a wider range of data. We then repeat the process for selecting the next best feature.

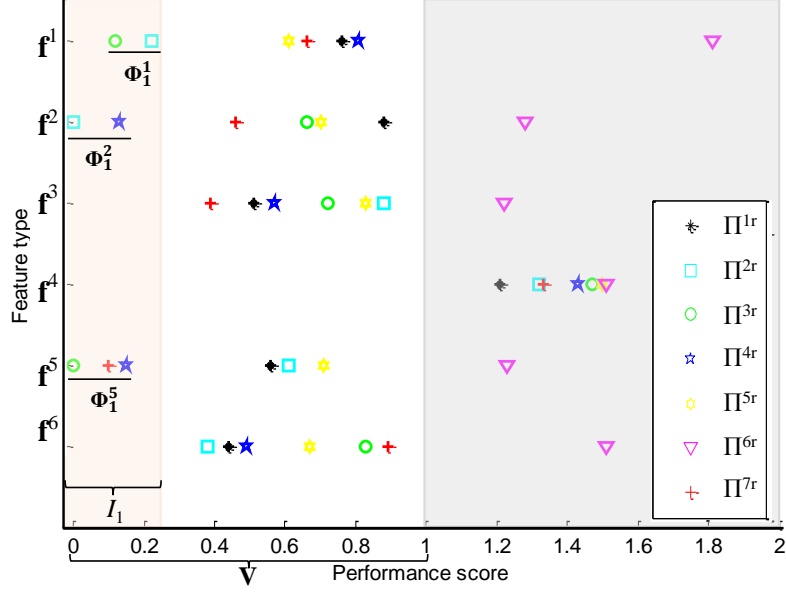


Figure 3.4: Example of performance matrix  $\Delta$  containing performance score values  $\Pi^{mr}$  (horizontal axis) obtained for  $R$  features (vertical axis) and  $M$  persons (colour coded), where  $r = 1 \dots 6$ ,  $m = 1 \dots 7$ ,  $R = 6$  and  $M = 7$ .  $\mathbf{V}$  contains  $\Pi^{mr}$  in the range  $[0, 1]$ . The bin  $I_i$  within  $\mathbf{V}$  contains  $M$  values of  $\Pi^{mr}$ . The values within  $I_i$  are spread among  $R$  features such that each feature has a set  $\Phi_i^r$  containing  $\Pi^{mr}$  and  $\sum_{r=1}^R |\Phi_i^r| = M$ . The bin  $I_1$  for  $i = 1$  is illustrated in the figure. For each feature  $\mathbf{f}^r$  we define a set  $\Phi_1^r$ , e.g. when  $r = 5$ ,  $\mathbf{f}^5$  contains  $\Phi_1^5 = \{\Pi^{35}, \Pi^{75}, \Pi^{45}\}$ ,  $|\Phi_1^5| = 3$  and  $\sum_{r=1}^6 |\Phi_1^r| = 7$ ;  $\mathbf{f}^4$  is discarded because  $\min_m \Pi^{m4} \geq 1$ ; person  $m = 6$  is discarded because  $\min_r \Pi^{6r} \geq 1$ .

Feature selection continues within the same bin  $I_i$  until all performance scores have been utilised for the selection. Then we move to the next bin in  $\mathbf{V}$ . The list  $\mathbf{Y}_{nqn}$  is progressively filled with  $\mathbf{f}^r$  in order of importance. The algorithm stops when all persons in the training data are exhausted ( $|\mathbf{Z}| = M$ ) or when all features are selected ( $\langle \mathbf{Y}_{nqn} \rangle = \hat{R}$ , where  $\langle \cdot \rangle$  counts the elements in the list). In the former case we obtain a subset of features. In the latter case the method returns the complete feature set with features ranked in order of importance.

Note that because the selected features in  $\mathbf{Y}_{nqn}$  are ranked by decreasing importance, the feature set can be further reduced by dropping the IDs of the least important features should the constraints of the application become more restrictive. CoPE is summarised in Algorithm 1.

### 3.7 Discussion

Typical steps that can be performed for object re-identification in a smart camera are shown in Fig. 3.5. Detection is performed in a video frame and the image corresponding to an object is obtained. From the acquired object-image a set of features can be extracted. The features are encoded for temporary storage within the camera and for communicating over the network

**Algorithm 1** CoPE feature selection

---

$M$  : total number of persons;  
 $\hat{R}$  : number of features;  
 $C_n$  : Source-camera in the network;  
 $C_{n_q}$  : Destination camera;  
 $\mathbf{f}^r$  :  $r^{th}$  feature in the feature set;  
 $P_n^m$  :  $m^{th}$  person in  $C_n$ ;  
 $\Psi^r$  : cost vector of  $\mathbf{f}^r$ ;  
 $\Pi^{mr}$  : performance score value for  $P_n^m$  using  $\mathbf{f}^r$ ;  
 $\Pi^r$  : performance vector for  $\mathbf{f}^r$ ;  
 $\Delta$  : performance matrix;  
 $\chi^m$  :  $\hat{R}$  performance scores for  $P_n^m$ ;  
 $\mathbf{V}$  : vector containing sorted values of  $\Pi^{mr} \leq 1$  from  $\Delta$ ;  
 $I_i$  :  $i^{th}$  bin with values from  $\mathbf{V}$ ;  
 $A_i^r$  : combined importance score of  $\mathbf{f}^r$  in  $I_i$ ;  
 $\mathbf{Y}_{n_qn}$  : list of selected features for  $C_n$  and  $C_{n_q}$ ;  
 $\mathbf{Z}$  : set of people taking part in the selection;  
 $\Phi_i^r$  : set of  $\Pi^{mr}$  in  $\mathbf{V}$  within  $I_i$  for  $\mathbf{f}^r$ ;  
 $\langle \cdot \rangle$  : number of elements in the list;  
 $|\cdot|$  : cardinality of a set;

```

1:  $\mathbf{Z} = \phi, \mathbf{Y}_{n_qn} = \phi$ 
2: while  $|\mathbf{Z}| \leq M$  or  $\langle \mathbf{Y}_{n_qn} \rangle \leq \hat{R}$  do
3:   while  $1 \leq i \leq \hat{R}$  do
4:     for  $r = 1$  to  $\hat{R}$  do
5:        $\Phi_i^r = \Pi^{mr}$  in  $\mathbf{V}$  within  $I_i$  for  $\mathbf{f}^r$ 
6:     end for
7:     for  $r = 1$  to  $\hat{R}$  do
8:       calculate  $A_i^r$  using Eq. 3.8
9:     end for
10:    get  $\hat{r}$  using Eq. 3.9 ▷ ID of selected feature
11:    if  $\mathbf{f}^{\hat{r}} \notin \mathbf{Y}_{n_qn}$  then
12:      append  $\mathbf{f}^{\hat{r}}$  to  $\mathbf{Y}_{n_qn}$ 
13:    end if
14:    remove  $\chi^m$  from  $\mathbf{V}$ ;  $\forall \Pi^{mr} \in \Phi_i^{\hat{r}}$ 
15:     $\mathbf{Z} = \mathbf{Z} \cup P_n^m$ ;  $\forall \Pi^{mr} \in \Phi_i^{\hat{r}}$ 
16:    remove  $\Pi^{\hat{r}}$  from  $\mathbf{V}$ 
17:  end while
18: end while

```

---

for association. In order to perform the re-identification, association is performed between the received and extracted features to obtain the correspondences.

Feature selection using CoPE is performed once using training data when a camera network is set-up. Then each camera locally stores the list of selected features  $\mathbf{Y}_{n_qn}$  for each source-camera  $C_{n_q}$ . If a new camera is added to the network, the training is performed for the new camera in pair-wise manner [96]. Note that features selected for a camera pair may not always be appropriate for another camera pair because of differences in illumination conditions and camera pose with respect to the objects. This approach, developed for camera pairs, is appropriate for distributed multi-camera settings where cameras communicate with each other without a central control unit. CoPE feature selection reduces the storage and computational requirements for re-identification. A performance matrix  $\Delta$  is generated for each camera pair  $(C_n, C_{n_q})$ , while the



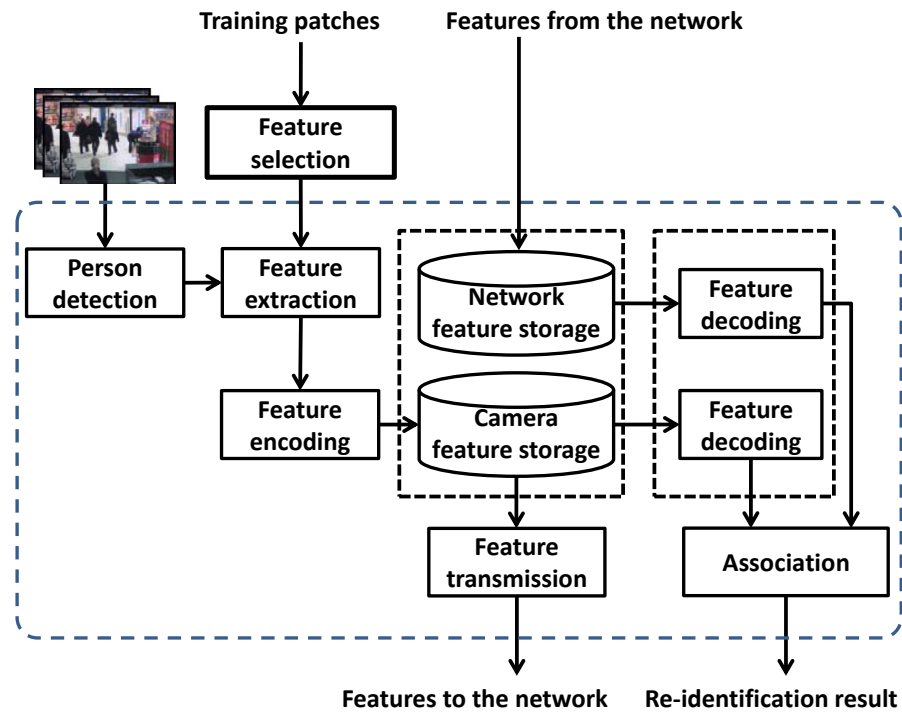


Figure 3.5: Block diagram representing the processing steps of person re-identification within a smart camera.

cost vector  $\Psi^r$  already takes into account  $N$  cameras and therefore remains the same.  $N$  cameras, in the extreme case, form a complete graph, where each camera has  $N - 1$  source-cameras. The time complexity of feature selection for such a network is  $N(N + 1)/2$  times that of the feature selection for a camera pair.

Object association can be performed by direct distance minimisation or a learning method can be applied. When machine learning is used for object association, two training phases are involved, namely the training for feature selection and the training for learning the re-identification model. Each destination-camera stores the trained models (weights) in addition to the selected feature IDs for each of its source-cameras. The inclusion of learning models with the CoPE feature selection is independent of the feature selection itself, and there is an increase in the storage cost (fixed) because of the local storing of the trained models (and not because of the feature selection).

In a surveillance system resources are mainly deployed to perform three main tasks for re-identification, namely: object detection, feature extraction, and association. The computational

time and storage size of a complete system can be calculated as

$$\begin{aligned} Time_{sys} &= 2 \times M(T_{detect} + T_{feats} + T_{assoc}), \\ Storage_{sys} &= 2 \times M(B_{img} + B_{feats}) + B_{model}, \end{aligned} \quad (3.10)$$

where  $M$  is the number of objects detected in each of  $C_n$  and  $C_{n_q}$ ,  $T_{detect}$ ,  $T_{feats}$  and  $T_{assoc}$  are the times required for detection, feature extraction and object association, respectively. In storage requirements  $B_{img}$ ,  $B_{feats}$  and  $B_{model}$  are the bytes required for storing image, feature sets and weights of a learning model (if any), respectively. Given that  $T_{detect} = 0.2 \text{ sec}$  (with 2.66 GHz processor on PETS dataset) [172] and  $T_{feats} = 2 \text{ sec}$  (Sec. 1.3.2). For associations we require  $M$  comparisons, where  $M$  is the total number of object-images detected in a source-camera. If time for one comparison is  $T_{assoc} = 1 \text{ ms}$ , we require 316 comparisons in VIPeR dataset for a single association. Thus,  $Time_{sys} = 2 \times 316(0.2 + 2 + 0.001) = 1391 \text{ sec}$ , where the major contribution is due to  $T_{feats}$ . In order to measure the system storage requirement we assume that the object detection part returns an object image with  $B_{img} = 2 \text{ KB}$  and the feature set has  $B_{feats} = 1 \text{ KB}$  (Sec. 1.3.2). If we ignore  $B_{model}$ , the  $Storage_{sys} = 2 \times 316(2 + 1) = 1896 \text{ KB}$ . If object-image can be discarded after the feature extraction then  $B_{feats}$  has the sole contribution in the storage requirements of the system. With the proposed CoPE features, the storage and computation requirements can be reduced by 80% of the complete feature set. Further details are discussed in experimental evaluation (Sec. 5.2.3).

### 3.8 Summary

In this chapter, we proposed a feature selection approach that identifies the most appropriate features for person re-identification. The amount of data stored for each feature and the computational time for its extraction are used jointly with their performance to generate an overall feature score. The best features are selected in a defined range of scores to reduce the performance overlap; a measure of similarity among features. We also discussed how the proposed approach can be applied in a camera network of  $N$  nodes and the setup requirements.

In the next chapter we discuss our two association approaches, which require less information sharing, and can perform the re-identification with multiple source-cameras.

## Chapter 4

### Association for re-identification

---

#### 4.1 Introduction

Extracted features from the objects detected in a source-camera need to be communicated to the destination-camera to perform re-identification. This communication can be limited by the bandwidth constraints. The number of possible matches for association within a camera pair and the number of source-cameras also affect the re-identification rate. In this chapter we propose two methods of object association that improve the re-identification rate with minimum information sharing and extends the association to multiple source-cameras.

The first association approach (Sec. 4.2) minimises the information sharing requirements between the cameras for association [C2]. The approach exploits the well-known concept of difference from reference features for object representation. The generated difference vectors are communicated over the network for association. We perform temporal alignment same as in [120] to restrain object assignments within the defined temporal boundaries. The association is performed by the optimal assignment using the Hungarian algorithm.

In the second approach (Sec. 4.3), unlike existing appearance based association methods specific to camera pairs, we perform association in a more generalised case, where a camera detects objects that can come from an unspecified source-camera of the network [C1]. We estimate the distributions of matching scores obtained by association of objects (using appearance information) in each camera pair. Using these distributions we measure probabilities of a correct match from the objects detected in a group of source-cameras.

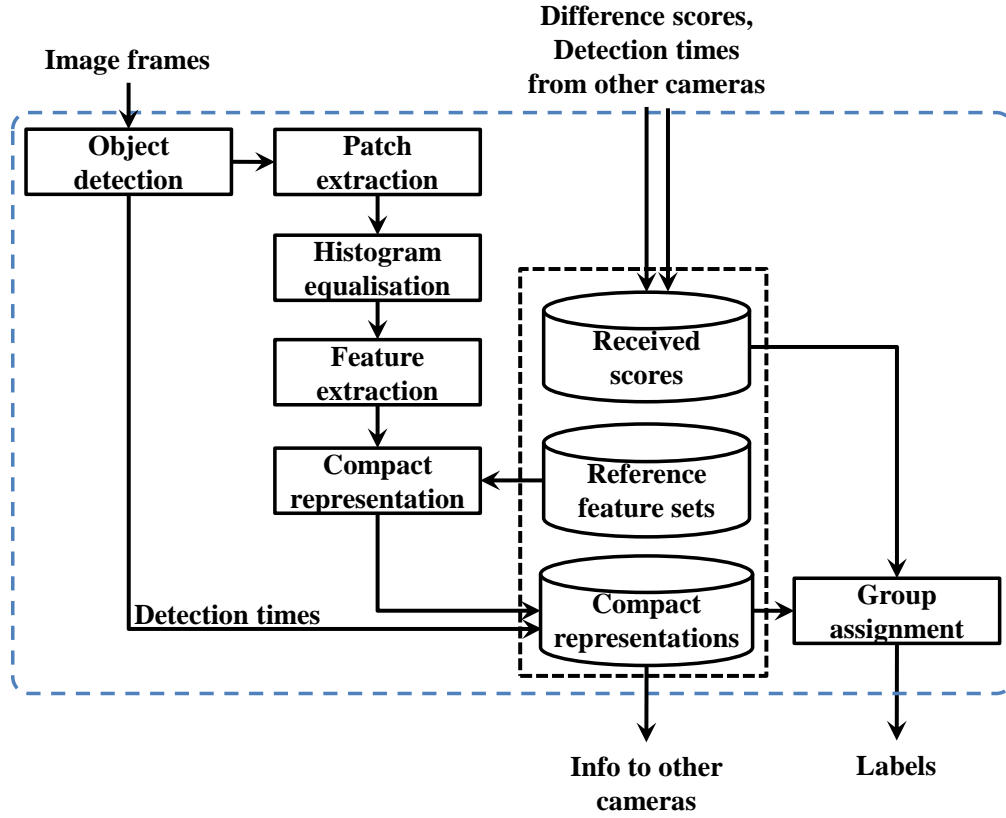


Figure 4.1: Block diagram of the proposed object association approach using difference-vectors (Sec. 4.2).

## 4.2 Association using difference-vectors

The block diagram of the proposed difference-vector based association approach is shown in Fig. 4.1. We have a set of  $N$  smart cameras  $\mathbf{C} = \{C_n\}_{n=1}^N$  with partially overlapping FoV. We assume that the object detection and tracking have been solved [68, 86] within each camera independently. A set of  $M_n$  objects  $\mathbf{P}_n = \{P_n^m\}_{m=1}^{M_n}$  is detected in destination-camera  $C_n$ . Each object is represented with a cropped image.

### 4.2.1 Histogram equalisation

In order to minimise the effect of illumination variations and contrast adjustment for each  $P_n^m$ , we perform histogram equalisation within each camera [2]:

$$hist_{eq}^{(i)} = \left\lfloor \frac{(U \times hist_{cf}^{(i)}) - (h \times w)}{(h \times w)} \right\rfloor, \quad (4.1)$$

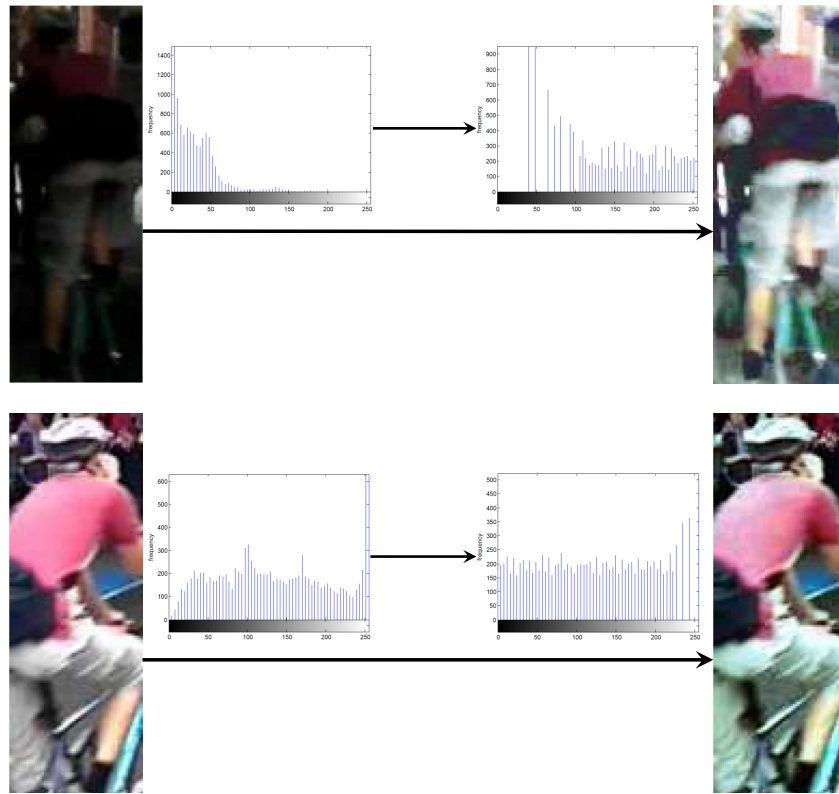


Figure 4.2: Example of histogram equalisation. The colour contrast between the same person detected in two cameras can be observed. After histogram equalisation the colour of the images are transformed into the same range.

where  $U$  is the number of intensity levels,  $h$  and  $w$  are the height and the width of the object image  $P_n^m$  in pixels, and  $hist_{cf}^{(i)}$  is the cumulative sum of the histogram until the bin with intensity value  $i$  in  $P_n^m$ . For each R, G, and B colour plane, the intensity value  $i$  of the image is replaced with  $hist_{eq}^{(i)}$  and a potentially narrow band of colours is spread over the whole available intensity range. Unlike colour-calibration approaches (Sec. 2.6), histogram equalisation neither require training data nor any information shared from other cameras. Fig. 4.2 shows an example of histogram equalisation applied to each of the RGB colour channels of the extracted images of the same person detected in two cameras. It can be qualitatively noted in the example images that after the histogram equalisation the colour of images obtained from the cameras with low and high contrast settings become near similar. However, since histogram equalisation does not use cross-camera colour information for adjusting brightness, the results, in general, are less accurate compared to the colour-calibration [133]. Histogram equalisation is indiscriminate between background and the object, and spreading of intensities can lead to the loss of information.

### 4.2.2 Difference-vector representation

From each histogram-equalised image, we extract  $R$  colour and texture features<sup>1</sup> as a feature vector  $\mathbf{F}_n^m = [\mathbf{f}_n^{mr}]_{r=1}^R$  [71, 136, 178]. In order to reduce the cost of storage and transfer of  $\mathbf{F}_n^m$ , we generate for each  $P_n^m$  an object representation  $\mathbf{\Omega}_n^m = [\Omega_n^{mj}]_{j=1}^J$ , which is a difference-vector obtained by measuring differences between the extracted feature vector  $\mathbf{F}_n^m$  and  $J$  reference-feature vectors  $\{\boldsymbol{\kappa}^j\}_{j=1}^J$  within each camera as

$$\mathbf{\Omega}_n^m = [ \|\boldsymbol{\kappa}^j - \mathbf{F}_n^m\| ]_{j=1}^J, \quad (4.2)$$

where  $\|\cdot\|$  is the Euclidean norm. We normalise  $\mathbf{\Omega}_n^m$  such that  $\sum_{j=1}^J \Omega_n^{mj} = 1$ . In order to obtain  $\boldsymbol{\kappa}^j$ , we use an image dataset [71] for reference images. Unlike other methods for image retrieval and classification [36], we have no scene dependency requirements. The only requirement is that the features extracted from the detected object is the same as the one extracted from the reference images. The extracted feature vectors from the reference dataset are clustered for data reduction to their centroids. We use the Lloyd's k-mean clustering algorithm [113] because of being the simplest, less computationally expensive, and requiring the least parameter-adjustments, in grouping similar data elements. The clustering returns  $J$  clusters of feature vectors, where  $J$  is fixed to the number of features, i.e.  $R$ , and the centroid of each cluster represents one reference-feature vector  $\boldsymbol{\kappa}^j$ . Similarly to the bag-of-words model, each camera locally stores  $\{\boldsymbol{\kappa}^j\}_{j=1}^J$ . We use the obtained object representation  $\mathbf{\Omega}_n^m$  (Eq. 4.2) for associating objects across cameras that reduces the amount of data required for communication.

### 4.2.3 Temporal alignment

We perform the temporal alignment of cameras for object association. The concept of temporal alignment is adapted from [120]. Temporal alignment allows us to perform the group assignment using  $\mathbf{\Omega}_n^m$  of the detected objects. Let  $P_n^m$  be detected and tracked between frames  $t_n^{m(s)}$  and  $t_n^{m(e)}$  in  $C_n$ , where  $s$  and  $e$  indicate the start and the end frames of a tracked object. The number of frames  $w_n^m$  during which  $P_n^m$  is tracked are  $w_n^m = t_n^{m(e)} - t_n^{m(s)} + 1$ .

For each  $P_n^m$ , we define a temporal search window  $\mathbf{W}_{n_q}^m$  in  $C_{n_q}$  representing the time interval in which  $P_n^m$  is likely to be observed in source-camera  $C_{n_q}$ . In order to select  $\mathbf{W}_{n_q}^m$ , we apply a plesiochronous approach to perform the temporal alignment of the cameras. Let  $\mathbf{P}_{n_q} = \{\mathbf{P}_{n_q}^k\}_{k=1}^{M_{n_q}}$

<sup>1</sup>Features are discussed in Sec. 3.3.

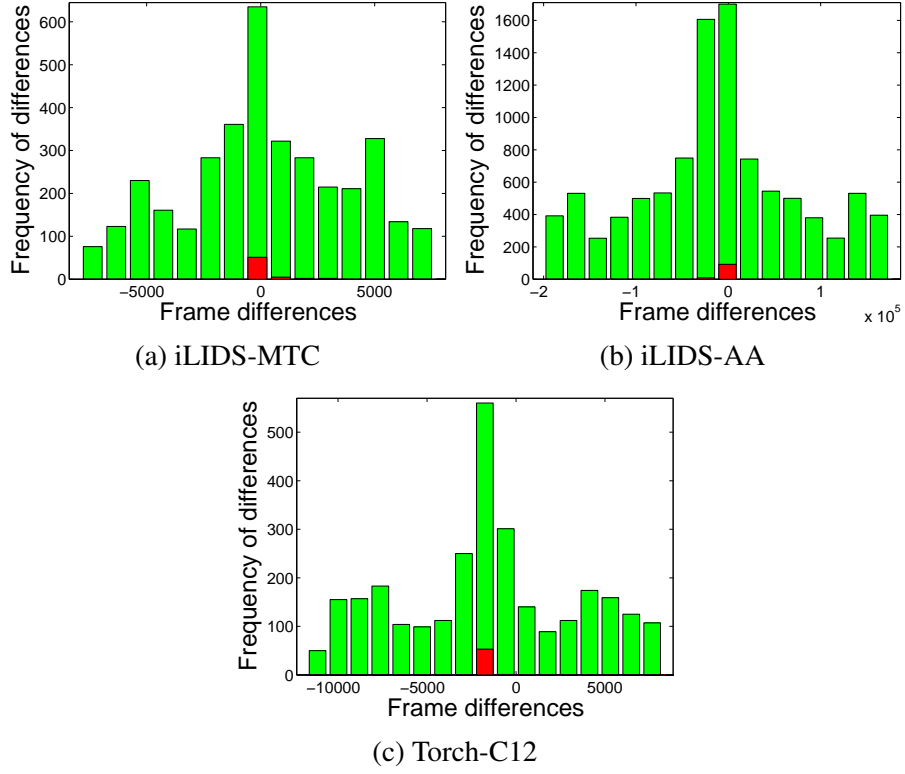


Figure 4.3: Histograms of differences of the detection frame-numbers. The green bars show all the possible differences between detection pairs across two cameras. The red bars show the differences in detections of the same object in two cameras.

be  $M_{n_q}$  objects in  $C_{n_q}$ , first detected in frames  $\{t_{n_q}^{k(s)}\}_{k=1}^{M_{n_q}}$ . For each  $t_n^{m(s)}$ , we obtain a set  $\Lambda_{n_q n}^m$  of  $M_{n_q}$  differences from  $\{t_{n_q}^{k(s)}\}_{k=1}^{M_{n_q}}$  in  $C_{n_q}$  as

$$\Lambda_{n_q n}^m = \{\varphi t_n^{m(s)} - t_{n_q}^{k(s)}\}_{k=1}^{M_{n_q}}, \quad (4.3)$$

where  $\varphi$  is the ratio of the frame rates of  $C_n$  and  $C_{n_q}$ . For  $M_n$  detected objects in  $C_n$ , we obtain an  $M_n \times M_{n_q}$  difference matrix  $\mathbf{D}_{n_q n} = \{\Lambda_{n_q n}^m\}_{m=1}^{M_n}$ . By analysing the distribution of values in  $\mathbf{D}_{n_q n}$ , we can observe that the difference of frame numbers of the first frames of two different tracked objects detected in  $C_n$  and  $C_{n_q}$  can vary significantly, while the difference between the first frames of the same objects detected in two cameras consistently remains within a narrow range. In order to identify that range, we take the histogram of values in  $\mathbf{D}_{n_q n}$  (Fig. 4.3). The bin size of the histogram depends on the average number of frames during which an object remains visible in  $C_n$ , measured as  $\tilde{w}_n = \frac{1}{M_n} \sum_{m=1}^{M_n} w_n^m$ . Mean of the bin with the most frequently occurring values represented as  $\delta_{n_q n}$  is the time shift (in number of frames) between  $C_n$  and  $C_{n_q}$ . Using  $\delta_{n_q n}$ ,

we estimate the temporal search window  $\mathbf{W}_{n_q}^m$  for  $P_n^m$  as

$$\left[ \varphi_n^{m(s)} + \delta_{n_q n} - \varphi \tilde{w}_n \right] < \mathbf{W}_{n_q}^m \leq \left[ \varphi_n^{m(s)} + \delta_{n_q n} + \varphi \tilde{w}_n \right]. \quad (4.4)$$

The objects detected in  $C_{n_q}$  within  $\mathbf{W}_{n_q}^m$  are the candidates for matching with  $P_n^m$ .

#### 4.2.4 Object association

In order to find the association between the objects, we measure the Bhattacharyya distances  $\mathbf{B}_n^m = \{B_n^{mk}\}_{k=1}^{M_{n_q}}$  between  $\Omega_n^m$  and  $\{\Omega_{n_q}^k\}_{k=1}^{M_{n_q}}$ , where  $M_{n_q}$  objects are detected in  $C_{n_q}$ . The distance  $B_n^{mk}$  is given as

$$B_n^{mk} = \begin{cases} -\ln \left( \sum_{j=1}^J \sqrt{\Omega_n^{mj} \cdot \Omega_{n_q}^{kj}} \right) & \text{for } P_{n_q}^k \text{ within } \mathbf{W}_{n_q}^m \\ \infty & \text{otherwise.} \end{cases} \quad (4.5)$$

In order to avoid the association with objects detected outside the temporal window  $\mathbf{W}_{n_q}^m$ , we assign  $B_n^{mk} = \infty$ . The assignment of  $P_n^m$  to  $P_{n_q}^k$  with the minimum distance  $B_n^{mk}$  from  $P_n^m$  results in multiple assignments to a single object because  $P_{n_q}^k$  can also have minimum matching distance from another object in  $\mathbf{P}_n$  (Fig. 4.4). The problem of multiple assignments can be solved by performing group assignment in which no object is assigned more than once. We perform the group assignment using the Hungarian algorithm [99], which takes as input an  $M_{n_q} \times M_n$  distance matrix  $\mathbf{H} = [\mathbf{B}_n^m]_{m=1}^{M_n}$  and assigns labels to the objects in two cameras without repetition.

### 4.3 Association using camera-invariant scores

Appearance based re-identification approaches perform association between a camera pair. In a multi-camera system many-to-one camera associations are needed, since targets may transit from different source-cameras to a destination-camera (Fig. 4.5). In such a scenario existing approaches require spatio-temporal calibration information (Sec. 2.6.2), such as paths to be followed and entry/exit regions, for camera selection. Spatio-temporal information may not always be available and can be difficult to model. For example, open spaces such as parks and halls without fixed paths and entry/exit points; and closed spaces with obstructions such as doors where multiple exits may converge to a single entry point. The difference in the appearance of an object also varies from one camera pair to the other because of variations in source-cameras' positions and illumination conditions (Sec. 1.3). To address these challenges, we propose a per-



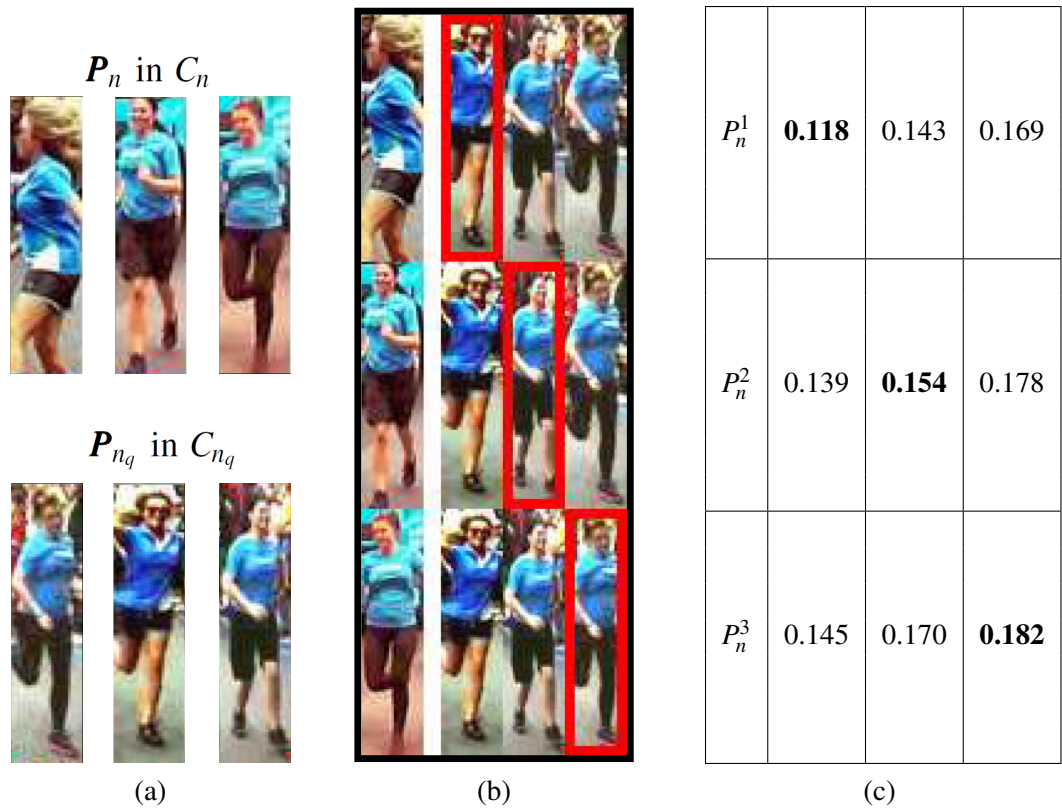


Figure 4.4: Example of object association for re-identification for (a) the three objects detected in cameras  $C_n$  and  $C_{n_q}$ . (b) In each row, the detected object-image  $P_n^m$  in  $C_n$  is in the left most column, while the next three columns show  $P_{n_q}$  detected in  $C_{n_q}$  sorted based on (c) the Bhat-tacharyya distances between  $P_n^m$  and  $P_{n_q}$ . Red boxes show the optimal assignment using the Hungarian algorithm [99]. The algorithm selects those matches that have the minimum distances while avoiding multiple assignments.

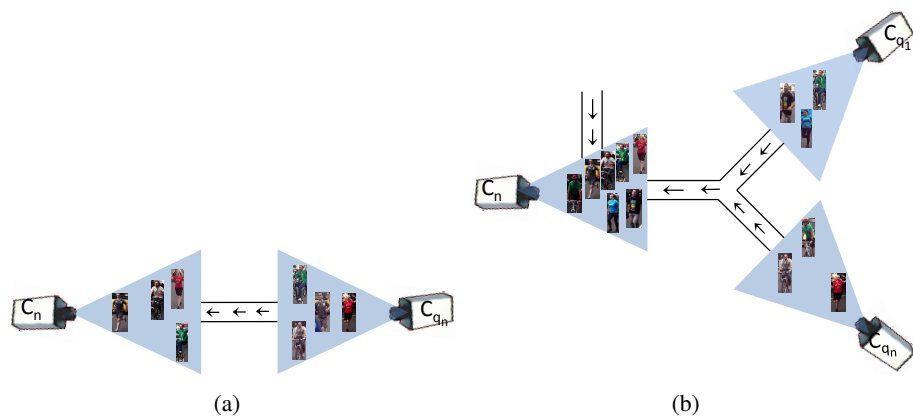


Figure 4.5: Person re-identification performed in camera  $C_n$  in (a) state-of-the-art and (b) the proposed approach. Arrows represent the considered direction of movement of people.

son re-identification approach that generates camera-invariant matching scores by exploiting the variations in the appearance of objects in camera pairs (Fig. 4.6). Each camera pair is represented with two parametric distribution models obtained by curve fitting on intra-class and inter-class

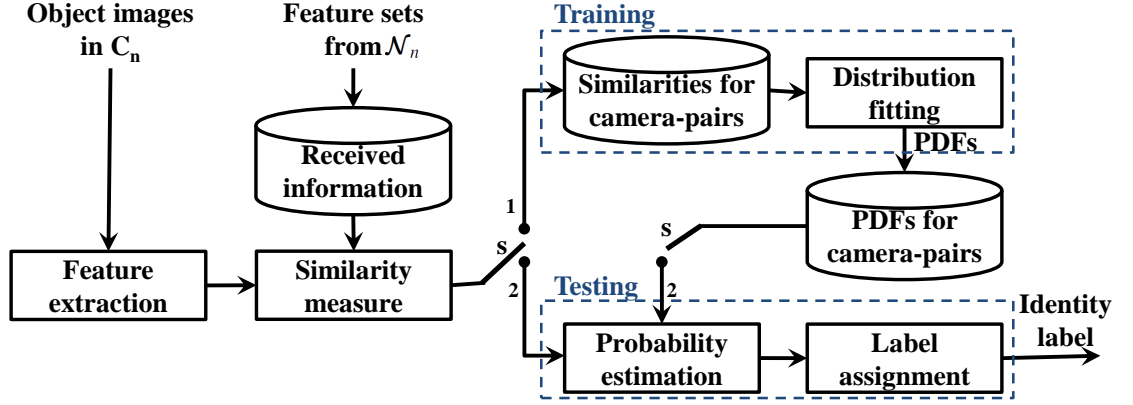


Figure 4.6: Block diagram of the proposed multiple source-cameras re-identification approach (Sec. 4.3). Switch  $s = 1$  is Training and  $s = 2$  is Testing,  $\mathcal{N}_n$  is the set of source-cameras of  $C_n$ .

similarity scores. These two models are combined to generate the probability of a correct match between a new target in the destination-camera and those in all source-cameras. Since the proposed approach relies only on the appearance information to perform both the camera selection and the object association, it can perform the re-identification even in the case when spatio-temporal calibration information is not available/reliable.

For each object-image  $P_n^m$  in  $C_n$  we aim to identify its other instance detected in an unspecified source-camera  $C_{n_q}$  in  $\mathcal{N}_n$ . If  $M_{n_q}$  objects are detected by each source-camera  $C_{n_q}$  that go to  $C_n$ , the number of objects  $M_n$  detected in  $C_n$  is given as

$$M_n = \sum_{q=1}^{\hat{N}_n} M_{n_q}. \quad (4.6)$$

We extract the feature set  $\mathbf{F}_n^m = \{\mathbf{f}_n^{mr}\}_{r=1}^R$  containing  $R$  features from each  $P_n^m$  in  $C_n$ . We obtain  $M_{n_q}$  similarity scores  $\{S_{n_q n}^{mk}\}_{k=1}^{M_{n_q}}$  between  $\mathbf{F}_n^m$  and the obtained feature sets  $\{\mathbf{F}_{n_q}^k\}_{k=1}^{M_{n_q}}$  from  $C_{n_q}$ . A similarity score can be obtained by measuring distances such as Bhattacharyya distance and L1-Norm [70]. It can also be obtained as a probability such as PRDC [179], or as confidence scores obtained by learning methods such as RSVM [136] and ASFI [111].

### 4.3.1 Training

We exploit the set of similarity scores  $\mathbf{S}_{n_q n}$  between  $M_{n_q}$  objects detected in a camera pair  $(C_n, C_{n_q})$  given as

$$\mathbf{S}_{n_q n} = \left\{ \left\{ S_{n_q n}^{mk} \right\}_{m=1}^{M_{n_q}} \right\}_{k=1}^{M_{n_q}}. \quad (4.7)$$

The set  $\mathbf{S}_{nqn}$  contains  $M_{nq} \times M_{nq}$  elements. We divide  $\mathbf{S}_{nqn}$  in two subsets  $\mathbf{S}_{nqn}^+$  and  $\mathbf{S}_{nqn}^-$ , which contain the similarity scores for the same and the different objects, respectively. Training for re-identification suffers from under-sampling because of the availability of few object-images and many pose and illumination changes [179]. In addition,  $|\mathbf{S}_{nqn}^+| \ll |\mathbf{S}_{nqn}^-|$  results in an unbalanced class problem ( $|\cdot|$  is the cardinality of a set). In order to compensate for the under-sampled and unbalanced data, we include more related-samples, generated by applying the Synthetic Minority Oversampling Technique SMOTE [34] on the scores in  $\mathbf{S}_{nqn}^+$  and  $\mathbf{S}_{nqn}^-$ . SMOTE solves the class imbalance problem by generating new samples of the minority class. In SMOTE, the difference between the sample and its nearest neighbour(s) is measured. The difference is added to the sample under consideration to generate similar synthetic examples. Next, we normalise the histograms of  $\mathbf{S}_{nqn}^+$  and  $\mathbf{S}_{nqn}^-$  to obtain their corresponding PDFs (Fig. 4.7). We characterise the PDFs by fitting the existing parametric distribution models [63] (Table 4.1). For each camera pair  $(C_n, C_{nq})$ , we obtain two models  $G_{nqn}^+$  and  $G_{nqn}^-$  that best fit the PDFs of similarity scores  $\mathbf{S}_{nqn}^+$  and  $\mathbf{S}_{nqn}^-$ , respectively. Models are selected by applying Bayesian Information Criterion (BIC) [147] given as

$$BIC = -2.\ln\hat{L} + d.\ln U, \quad (4.8)$$

where  $U$  is the sample size of the training set, and  $d$  is the number of parameters.  $\hat{L}$  is the maximised value of the likelihood function of the model  $G$ , i.e.  $\hat{L} = p(x|\theta, G)$ , where  $\theta$  are the parameter values that maximise the likelihood function. BIC avoids over-fitting through the penalty term  $d.\ln U$  that increases with the number of parameters. The lower the BIC, the better the model. The parameters of  $G_{nqn}^+$  and  $G_{nqn}^-$  are noted as  $\theta_{nqn}^+$  and  $\theta_{nqn}^-$ , respectively.

### 4.3.2 Testing

In the testing phase, a new object is detected in  $C_n$ . Feature sets of the objects detected in the set of source-cameras  $\mathcal{N}_n$  are also received. We measure the similarity score  $S_{nqn}^{mk}$  between  $\mathbf{F}_n^m$  and each obtained feature set  $\mathbf{F}_{nq}^k$  from  $C_{nq}$ . From the given similarity score,  $S_{nqn}^{mk}$ , we measure the probability that  $P_n^m$  and  $P_{nq}^k$  are instances of the same objects represented as  $Pr(\mathcal{S}|S_{nqn}^{mk})$ , and the probability that  $P_n^m$  and  $P_{nq}^k$  are instances of two different objects represented as  $Pr(\bar{\mathcal{S}}|S_{nqn}^{mk})$ . Using the two PDFs,  $G_{nqn}^+$  and  $G_{nqn}^-$ , and their parameters,  $\theta_{nqn}^+$  and  $\theta_{nqn}^-$ , (Table 4.1) from the

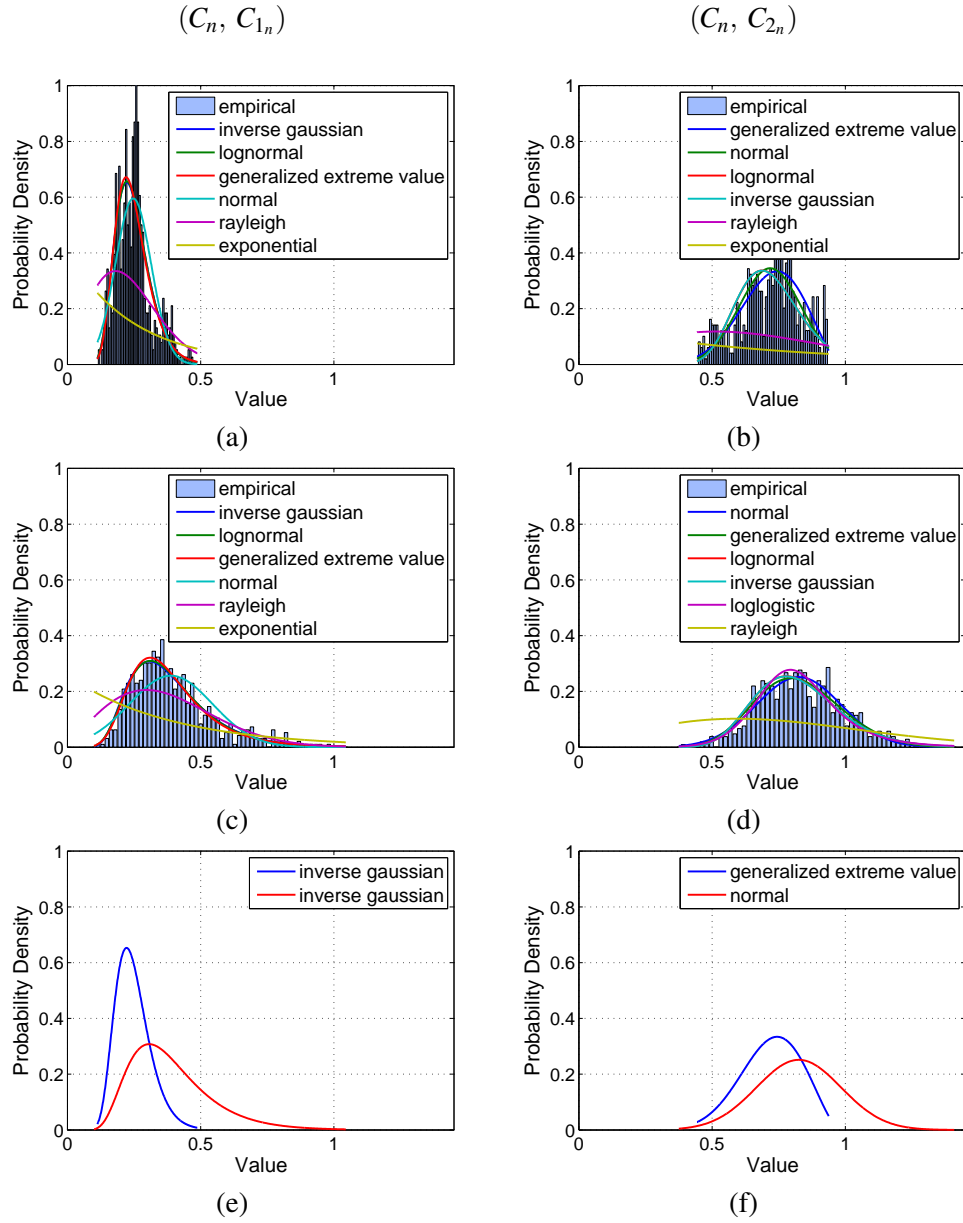


Figure 4.7: An example of distributions of matching distances between objects detected in  $C_n$  and two of its source-cameras (left-column)  $C_{1_n}$  and (right-column)  $C_{2_n}$ . Legends are sorted in the order of the nearest distribution to the data, identified using Bayesian information criterion. Distances are between (a,b) same and (c,d) different objects in each camera pair. (e,f) Two selected distributions for a camera pair (blue same and red different objects).

training of corresponding camera pairs  $(C_n, C_{n_q})$ , the probabilities can be obtained as

$$\begin{aligned}
 Pr(\mathcal{S}|S_{n_q}^{mk}) &= G_{n_q}^+(S_{n_q}^{mk}, \theta_{n_q}^+), \\
 Pr(\bar{\mathcal{S}}|S_{n_q}^{mk}) &= G_{n_q}^-(S_{n_q}^{mk}, \theta_{n_q}^-).
 \end{aligned}
 \tag{4.9}$$

Table 4.1: Probability density functions of the continuous parametric distributions used in the curve fitting. key:  $\mu$  - location or mean,  $\sigma$  - scale or standard deviation,  $\lambda$  - shape,  $B(\cdot)$  - Beta function,  $\Gamma(\cdot)$  - Gamma function,  $s$  - Non-centrality,  $b$  - positive scalar value,  $v$  - degree of freedom.

Distributions	Formula
Inverse Gaussian	$G(x \mu, \lambda) = \sqrt{\frac{\lambda}{2\pi x^3}} e^{-\frac{\lambda}{2\mu^2 x}(x-\mu)^2}; x > 0$
Logistic	$G(x \mu, \sigma) = \frac{e^{-\frac{x-\mu}{\sigma}}}{\sigma \left(1 + e^{-\frac{x-\mu}{\sigma}}\right)^2}$
Log-logistic	$G(x \mu, \sigma) = \sigma^{-1} x^{-1} \frac{e^{\frac{\log(x)-\mu}{\sigma}}}{\left(1 + e^{\frac{\log(x)-\mu}{\sigma}}\right)^2}; x \geq 0$
Normal	$G(x \mu, \sigma) = (\sigma\sqrt{2\pi})^{-1} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
Lognormal	$G(x \mu, \sigma) = (x\sqrt{2\pi\sigma})^{-1} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}$
Extreme value	$G(x \mu, \sigma) = \sigma^{-1} e^{\frac{x-\mu}{\sigma}} e^{-e^{\frac{x-\mu}{\sigma}}}$
Generalised extreme value	$G(x 0, \mu, \sigma) = \sigma^{-1} e^{-e^{-\frac{x-\mu}{\sigma} - \frac{(x-\mu)}{\sigma}}}$
Generalised Pareto	$G(x \lambda, \sigma, \theta) = \sigma^{-1} \left(1 + \lambda \frac{x-\theta}{\sigma}\right)^{-1-\frac{1}{\lambda}}$
Beta	$G(x a, b) = \frac{1}{B(a, b)} x^{a-1} (1-x)^{b-1} I_{(0,1)}(x)$
Exponential	$G(x \mu) = \frac{1}{\mu} e^{-\frac{x}{\mu}}$
Gamma	$G(x a, b) = \frac{1}{b^a \Gamma(a)} x^{a-1} e^{-\frac{x}{b}}$
Nakagami	$G(x \mu, \sigma) = 2 \left(\frac{\mu}{\sigma}\right) \frac{1}{\Gamma(\mu)} x^{2\mu-1} e^{-\frac{\mu}{\sigma} x^2}; x > 0$
Rician	$G(x s, \sigma) = I_0\left(\frac{xs}{\sigma^2}\right) \left(\frac{x}{\sigma^2}\right) e^{-\frac{x^2+s^2}{2\sigma^2}}; x > 0$
t location-scale	$G(x \mu, \sigma, v) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sigma\sqrt{v\pi}\Gamma\left(\frac{v}{2}\right)} \left[\frac{v + \left(\frac{x-\mu}{\sigma}\right)^2}{v}\right]^{-\frac{v+1}{2}}$
Weibull	$G(x \sigma, \lambda) = \frac{\lambda}{\sigma} \left(\frac{x}{\sigma}\right)^{\lambda-1} e^{-\left(\frac{x}{\sigma}\right)^\lambda}; x \geq 0$

The two events  $\mathcal{S}$  and  $\bar{\mathcal{S}}$  are independent events, since their PDFs are obtained from learning on two different sets of data,  $\mathbf{S}_{nqn}^+$  and  $\mathbf{S}_{nqn}^-$ . The problem is similar to tossing two coins and measuring the probability that the first coin will have head and the second will be tail, referred as compound probability. We measure the compound probability to obtain the matching score  $\mathcal{L}_{nqn}^{mk}$ , given as

$$\mathcal{L}_{nqn}^{mk} = Pr(\mathcal{S}|\mathbf{S}_{nqn}^{mk}) \cdot (1 - Pr(\bar{\mathcal{S}}|\mathbf{S}_{nqn}^{mk})). \quad (4.10)$$

The larger  $\mathcal{L}_{nqn}^{mk}$ , the higher the chances for the pair to be a correct match. Since we exploit the distributions of similarity scores in each camera pair to generate the probability of a correct

match while taking into account the variations between camera pairs, the assignment between  $P_n^m$  and  $P_{n_q}^k$  coming from any source-camera,  $C_{n_q}$ , becomes possible. This makes  $\mathcal{L}_{n_q n}^{mk}$  camera invariant. For each  $P_n^m$ , we get  $M_n$  matching scores from all the objects detected in the set of source-cameras  $\mathcal{N}_n$  forming a matching-score matrix  $\mathcal{L}_n$  for the set of  $\mathbf{P}_n$  objects given as

$$\mathcal{L}_n = \left[ \left[ \left[ \mathcal{L}_{n_q n}^{mk} \right]_{m=1}^{M_n} \right]_{k=1}^{M_{n_q}} \right]_{q=1}^{\hat{N}_n}. \quad (4.11)$$

Finally, we select the correct match from the obtained camera-invariant matching scores  $\mathcal{L}_n$  by optimal assignment using the Hungarian algorithm [99] as discussed in Sec. 4.2.4.

#### 4.4 Summary

We proposed two association methods, where the first one reduces the amount of information needed to communicate for re-identification, and the second improves the re-identification rate in the case of multiple source-cameras.

Association using difference-vectors (Sec. 4.2) is a simple yet effective object association approach that minimises the amount of data to be shared among cameras. The approach requires limited information for re-identification, thus permitting association during short temporal intervals – typical to the videos recorded using smartphone cameras. Optimal assignment using Hungarian algorithm improves the object association.

In the second association approach (Sec. 4.3), we are able to extend the pairwise re-identification methods to multiple cameras. Because of the differences in camera view and environment settings (Sec. 1.3), the pairwise association approaches cannot be directly applied for association in the case of multiple source-cameras. The proposed approach estimates a compound probability of a correct match in a camera network by exploiting similarity scores in camera pairs. Thus the approach makes it possible to perform many-to-one camera association for retrieving the correct match from a group of source-cameras. The performance of the proposed approach can be improved by increasing the number of objects detected in each camera pair and available for training, which needs to be performed only once during the camera network set-up. If  $M_{n_q}$  objects are required for the training of a camera pair  $(C_n, C_{n_q})$ , the addition of a new destination-camera  $C_n$  to  $N$  cameras of the network would require  $M_{n_q} \times \hat{N}_n$  objects that move from  $\hat{N}_n$  source-cameras to the new camera. In the worst-case scenario when every destination

camera has  $N - 1$  source-cameras, the training required for the  $N$  cameras network is increased by  $N(N - 1)/2$ .

In the next chapter, we discuss the evaluation of the presented association approaches for re-identification, using the initial set of all features and the proposed CoPE features, (Chapter 3). Five publicly available and one self-generated multi-camera challenging datasets are used for the evaluation.

## Chapter 5

### Experimental evaluation

---

#### 5.1 Introduction

In this chapter, we evaluate the proposed object representation and association methods for person re-identification. The results are compared with state-of-the-art re-identification approaches [49, 59, 71, 73, 98, 111, 112, 136, 157, 169, 179]. We use the validation criteria based on the cost of features and the re-identification rate. In order to measure the *cost of features*, the average storage size (in bytes) and computational time (in secs.) per object is calculated for each camera. The *re-identification rate* is measured using the Cumulative Matching Characteristics (CMC) curves [71]. CMC curves show the ranked matching rates i.e. the number/percentage of persons correctly matched at each rank. Matching at first rank refers to the true re-identification rate. The overall performance is also evaluated using the Area Under the CMC Curves (AUC).

For the evaluation of the proposed approaches, we use both publicly available datasets from VIPeR [71], iLIDS [84] and WARD [122] and an in-house generated Torch [C2] dataset (see Sec. 2.8 for datasets). The selected datasets present a mix of characteristics such as outdoor and indoor settings, variations in viewing angle, occlusions and illumination changes. We assume that the person detection problem is solved and the results generated by a person detector are available as input to our pipeline. We apply two-fold cross validation using half of the data for training and the remaining for testing the approaches. The experiments are carried out using Matlab 7.11 on a 3.3 GHz dual core desktop system with 3 GB of RAM.

We group the experiments into three categories based on the proposed methods. In Sec. 5.2,



we evaluate the performance of the proposed cost-effective feature selection method for re-identification and compare the results with existing state-of-the-art approaches in terms of computation and storage cost, and re-identification rate. Sec. 5.3 shows the evaluation and results of the proposed difference-vector representation for re-identification and amount of data needed for communication. In Sec. 5.4 the results for the proposed camera association approach for re-identification in multiple source-cameras are compared with the existing association approaches. Finally, Sec. 5.5 summarises the chapter.

## 5.2 Re-identification with cost-effective representations

We evaluate CoPE feature selection method on datasets VIPeR [71] and iLIDS-TC [J1] (Sec. 2.8), using the defined object-shape (Sec. 3.2) [J2] and initial feature set as in [71, 111, 112, 136, 179] (Sec. 5.2.1). The re-identification capabilities are measured using Direct Distance Minimisation (DDM) [59, 71] such that the two objects detected in a pair of cameras are assigned the same label if they have the minimum matching distance between their features. We compare DDM and learning approaches (RankSVM [136] and AdaBoost [71]) using CoPE and the existing feature selection methods: Fisher score [55], Information gain [44], mRMR [131], ReliefF [140] and Bi-clusters [82]. CoPE with DDM is further compared with PRDC [179], ASFI [111], KISSME [98], KISS-RS [157], LDML [73], LMNN [169] and ITML [49].

We consider three validation criteria, namely cost of features, re-identification rate and feature budgeting. The *cost of features* is calculated for the initial feature set and then for the selected features to analyse improvements in data reduction and computational time. The data generated by each object representation is encoded using the lossless data compression algorithm ‘deflate’ [144], which combines LZ77 and Huffman coding. In addition, we evaluate the training time for feature selection. The *re-identification rate* for the association methods is compared with the initial feature set and then with the selected features using the CMC curves [71] and AUC. Finally, we consider *feature budgeting* in constrained environments to analyse the scalability of CoPE and the effects in terms of cost and performance of further feature reductions.

### 5.2.1 Feature sets

We obtain the histograms of colour and texture (discussed in Sec. 3.3). Existing approaches use a 2784-dimensional feature vector by dividing the full-body person image into a set of stripes (6)

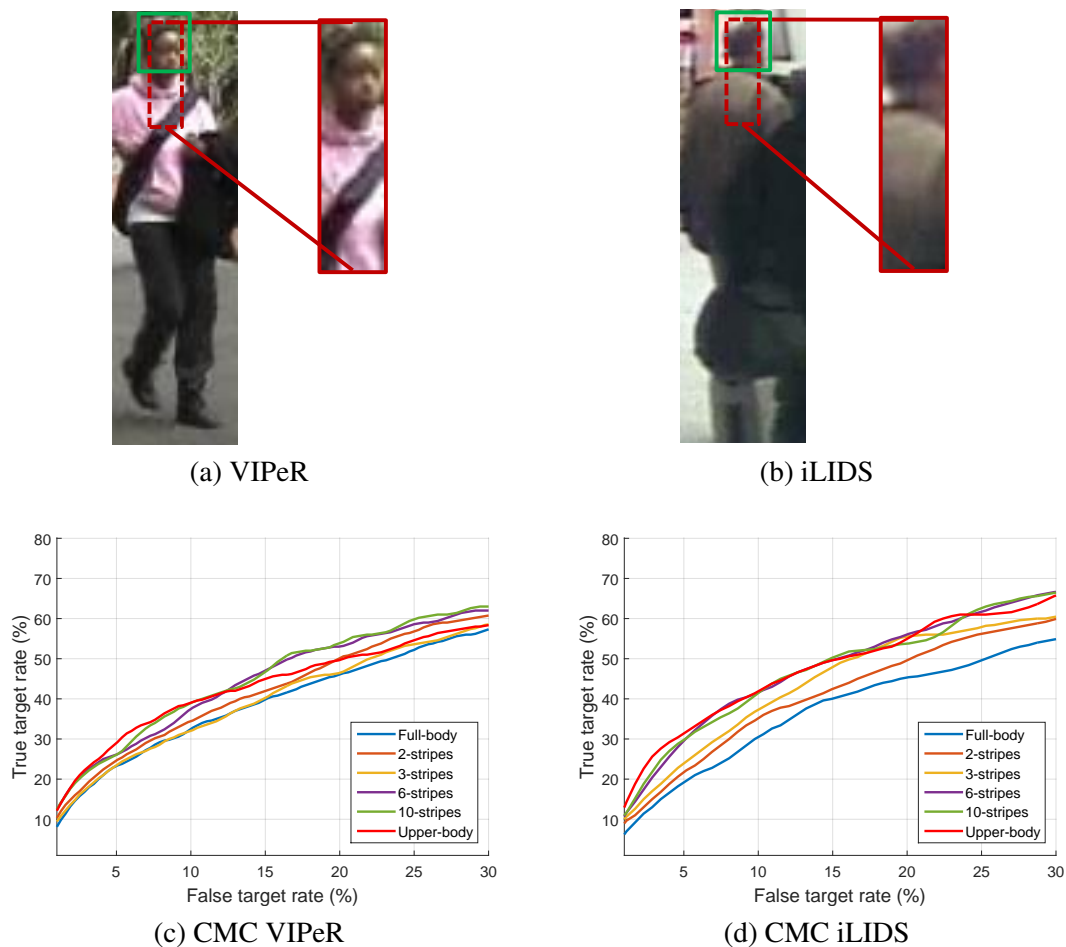


Figure 5.1: Examples of proposed upper-body image representation (cropped region) extracted based on the location of the head (green bounding box) using Eq. 3.1 in (a) VIPeR and (b) iLIDS. Re-identification rates using Bhattacharyya distance on the initial feature set extracted from the full-body image divided into two [129], three [5], six [179] and ten [22] horizontal slices and the upper-body image representation, in (c) VIPeR ( $M = 174$ ) and (d) iLIDS ( $M = 174$ ) datasets.

and then concatenating the corresponding features from each stripe [71, 111, 112, 136, 179]. We reduce the size of the object representation by extracting features from the defined upper-body shape (discussed in Sec. 3.2) as a single stripe (better suited for crowded scenes). In order to extract the colour features, the upper-body image is divided into upper and lower half. The upper half representing the head bounding box is given double the weight<sup>1</sup> compared to the lower half, since that is the most visible and least occluded part of the defined image shape. The weighted histogram of the upper half is added to that of the lower half.

Fig. 5.1 shows the re-identification results in VIPeR and iLIDS datasets for the complete feature set extracted from the defined upper-body and the full-body images, without feature se-

<sup>1</sup>Each pixel is considered twice in the upper half of the defined shape while computing the histogram.

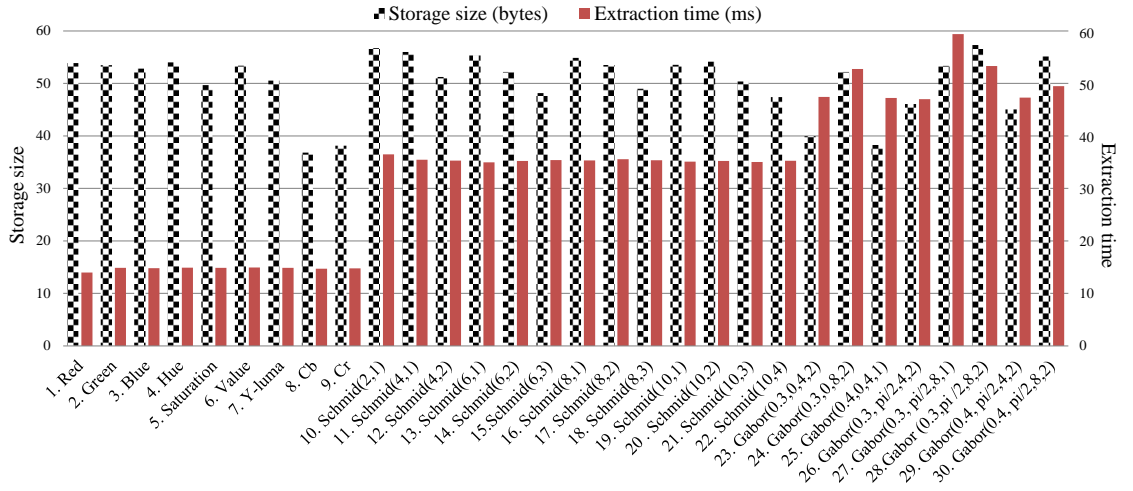


Figure 5.2: Storage size and the extraction time for each feature in the initial feature set (listed on the horizontal axis with their IDs and parameters). For Gabor ( $\gamma, \theta, \lambda, \sigma^2$ ) and Schmid ( $\sigma, \tau$ ) filters,  $\gamma$  is the aspect ratio,  $\theta$  is the angle in radians,  $\lambda$  the wavelength of the sinusoidal factor,  $\sigma$  the standard deviation and  $\tau$  the number of cycles. The vertical axes represent the storage size (bytes) and the feature extraction time (ms) for a single person within a camera.

lection. The full-body image is also divided into two [129], three [5], six [179] and ten [22] horizontal slices. In the case of occlusions and crowd, a better re-identification rate can be achieved with the upper-body images compared to the full-body images (divided into one, two and three stripes). The results of upper-body images are comparable to that of full-body images divided into six and ten stripes. However, by using the upper-body image representation, we are able to reduce the storage requirements of the extracted features by 6 and 10 times compared to six and ten stripes representations.

### 5.2.2 CoPE with varying parameters

Fig. 5.2 shows the storage size  $\beta^r$  and the extraction time  $\Gamma^r$  of the 30 features used. The total count of bins is fixed; however,  $\beta^r$  varies between 29 and 56 bytes because the data encoding is applied before the feature storage. The extraction time  $\Gamma^r$  of the feature extraction varies between 16 and 60 ms. We obtain the overall computational time and the storage size required by a single camera  $C_n$  for  $\hat{R}$  features by summation of individual feature's  $\Gamma_n^{mr}$  and  $\beta_n^{mr}$  over  $M$  persons. Since  $\Psi^r$  is the cost of a single feature, where the higher the  $\Psi^r$  the better it is. The cost of the selected feature set cannot be obtained by simple addition of  $\Psi^r$  of each feature, since it would not be able to differentiate if the cost is high because of higher  $\Psi^r$  or large number of selected features. Thus, for the comparisons, we measure the normalised cost  $\mathcal{E}$  of the selected feature

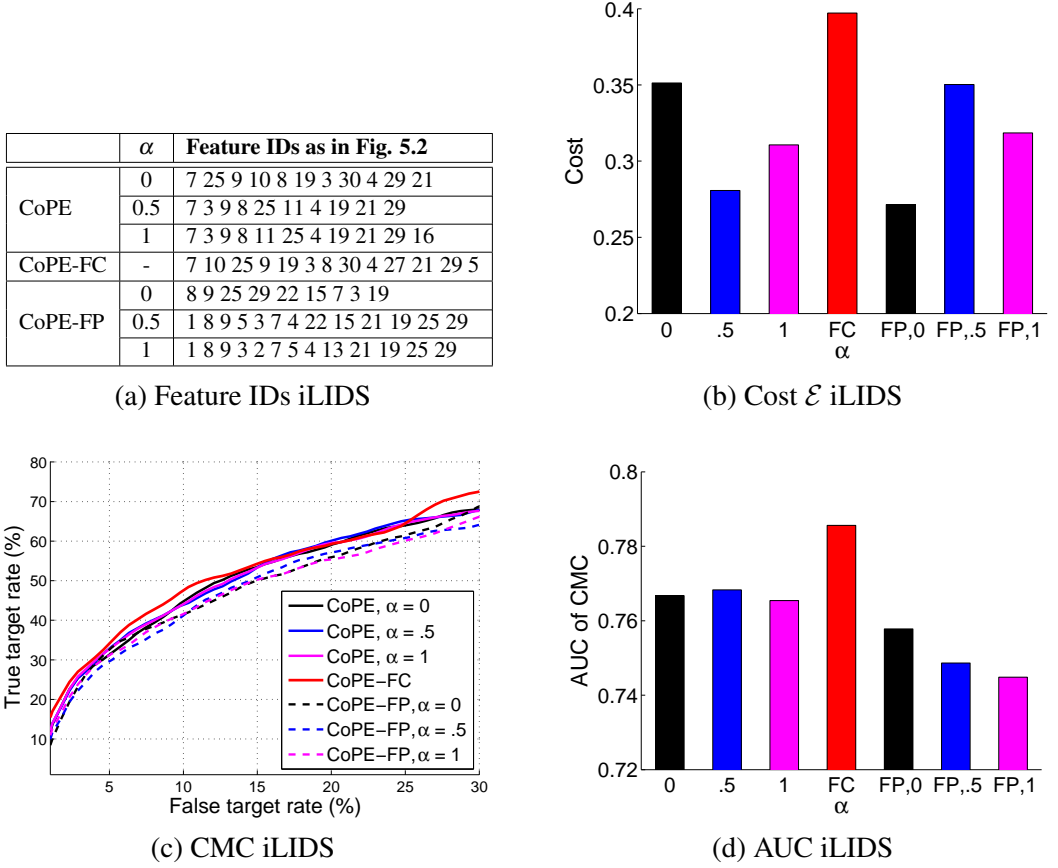


Figure 5.3: Analysis of feature selection with varying selection criteria: CoPE-FC (fixed cost, performance only), CoPE-FP (fixed performance, cost only), and CoPE (both cost and performance), in iLIDS ( $M = 174$ ) in terms of (a) selected features, (b) normalised cost  $\mathcal{E}$ , (c) CMC curves, and (d) AUC of CMC curves. The cost parameter  $\alpha$  in Eq. 3.7 is varied for CoPE-FP and CoPE as:  $\alpha = 0$  (black),  $\alpha = 0.5$  (blue) and  $\alpha = 1$  (magenta), while  $\alpha = n/a$  (red) in CoPE-FC, since the cost component is not included.

sets using the cost score  $\Psi^r$  (Eq. 3.7) as

$$\mathcal{E} = \frac{\sum_{\mathbf{f}^i \in \mathbf{Y}_{nqn}} 1/\|\Psi^{\hat{\mathbf{f}}}\|}{\sum_{r=1}^{\hat{R}} 1/\|\Psi^r\|}, \quad (5.1)$$

where  $\mathbf{Y}_{nqn}$  is the list of selected features. The set of 30 features has the maximum cost  $E_{max} = 1$ . We consider 316 and 174 persons in VIPeR and iLIDS, respectively.

Fig. 5.3 shows the analysis of CoPE on iLIDS with three selection criteria: (i) feature selection as a function of performance only keeping the cost of all features fixed (CoPE-FC); (ii) feature selection as a function of cost only while varying  $\alpha$  for the two components of cost in Eq. 3.7 and keeping the performance of all features fixed (CoPE-FP); and (iii) feature selection

considering both performance and cost (with varying  $\alpha$ ) of a feature (CoPE). The cost of the selected features is the highest for CoPE-FC [Fig. 5.3 (b)] (red), since the selection is carried out based on performance only and the cost component is fixed for all features (thus variation in  $\alpha$  not required). However, CoPE-FC is also able to achieve the highest re-identification rate (using DDM) in the absence of cost constraints [Fig. 5.3 (c)] (red). Both the cost and the re-identification rate of the selected features are reduced with CoPE. When varying the cost parameter  $\alpha$  in CoPE, while the composition of the selected features remains similar, their order changes [top three rows of the table in Fig. 5.3 (a)]. Since cost and performance are independent in the feature selection, varying  $\alpha$  does not affect the performance of a feature. The selected features may vary based on the requirement of a system controlled by  $\alpha$ , i.e. for well-performing features with a limited extraction time  $\alpha = 1$ , and for limited storage size  $\alpha = 0$ . Note that a limited extraction time may not imply a higher storage size (and vice versa). The smallest cost for CoPE is obtained when there is an equal contribution of computational time and storage size ( $\alpha = 0.5$ ).

In CoPE-FP, although performance is not used for feature selection, in order not to obtain a sorted list of all features based on cost, we remove the people from the training data for which the selected minimum cost feature has good performance so that the algorithm stops when all the people in the training data are exhausted. For  $\alpha = 0$  the order of selection is controlled by the storage size, while for  $\alpha = 1$  the features with the shortest extraction times (colour features) are selected first (see Fig. 5.2 for time and size). An interesting case is when the features with IDs 8 and 9 are selected for all three values of  $\alpha$ , since these features have both the shortest computational time and the smallest storage size. In contrast, the feature with  $ID = 1$  has the shortest extraction time and a large storage size. This makes it the first feature with  $\alpha = 0.5$  and 1, while it is not selected with  $\alpha = 0$ . The order of performance of the selected features for the three criteria is as follows  $CoPE-FP < CoPE < CoPE-FC$ . The rest of the evaluation is performed for  $\alpha = 0.5$  to have an equal contribution from the storage size and the extraction time.

### 5.2.3 CoPE vs all-features

Table 5.1 shows the storage size and the computational time for the features extracted from each person observed in one camera. We compare the results of the initial feature set with that of the three non-unique sets of selected features obtained by CoPE using three similarity measures: Bhattacharyya distance, L1-Norm and Chi-square distance in Eq. 3.2. For VIPeR, the number

Table 5.1: Storage size, computational time and normalised cost of the initial feature set per camera used in existing re-identification approaches compared with CoPE features obtained for three similarity measures in Eq. 3.2.

Dataset	Distance as in Eq. 3.2	Total features	Feature IDs as in Fig. 5.2	Size (KB)	Time (sec)	Cost $\mathcal{E}$ (Eq. 5.1)
VIPeR (M=316)	-	30	1-30	466.43	314.83	1.00
	Bhattacharyya	6	4 8 9 5 7 25	82.46	38.34	0.16
	L1-Norm	6	8 9 4 7 5 12	86.46	34.56	0.16
	Chi-Square	8	8 9 4 5 7 18 17 25	114.07	60.75	0.23
iLIDS (M=174)	-	30	1-30	256.83	173.35	1.00
	Bhattacharyya	10	7 3 9 8 25 11 4 19 21 29	80.74	47.70	0.29
	L1-Norm	8	3 9 7 25 8 4 18 29	61.90	35.48	0.24
	Chi-Square	9	7 9 3 8 25 6 15 29 16	70.97	41.64	0.26

of selected features are 6, 6 and 8, respectively, for the three similarity measures that reduce the storage size per camera to 11%, 18% and 24% of the total size (466.43 KB) of the initial 30 features. In iLIDS, 10, 8 and 9 features are selected for the three similarity measures that respectively reduce the storage size to 31%, 23% and 27% of the storage requirement for the initial feature set (256.83 KB). Similarly, the computational time of feature extraction per camera is reduced significantly. In the VIPeR dataset, the computational time is reduced to 12%, 10% and 19% for the three similarity measures, respectively. In the case of iLIDS, the computational time is reduced to 27%, 20% and 23%. It can also be observed that the normalised cost  $\mathcal{E}$  of the selected CoPE features is reduced more in VIPeR than in iLIDS because mostly the colour features are selected in VIPeR. The colour features are fast to extract with less or comparable storage size (Fig. 5.2) and perform better than texture features. In VIPeR, we reduce the cost  $\mathcal{E}$  of the feature set to 20%, while in iLIDS we reduce it to 33% of the initial feature set.

Fig. 5.4 compares the re-identification rate for the three DDM approaches with the state of the art. In the DDM approach, two persons are considered correctly matched for re-identification, if their obtained feature sets have the minimum matching distance between them. The performance of the selected features is measured in terms of improvement of the re-identification rate of DDM approaches compared to that of using the initial feature set. CMC curves highlight the true target rate for the first 30% of false target rates (the most important part of CMC for evaluation). In VIPeR, a higher re-identification rate is obtained using the selected features. For example, at 20% false target rate in the CMC curves, the true target rate is above 65% for selected features compared to the initial feature set with true target rates between 40% to 50% for all the three measures. Because of the limited illumination changes between cameras, mostly colour features are selected (Table 5.5). In iLIDS, both colour and texture features are selected.

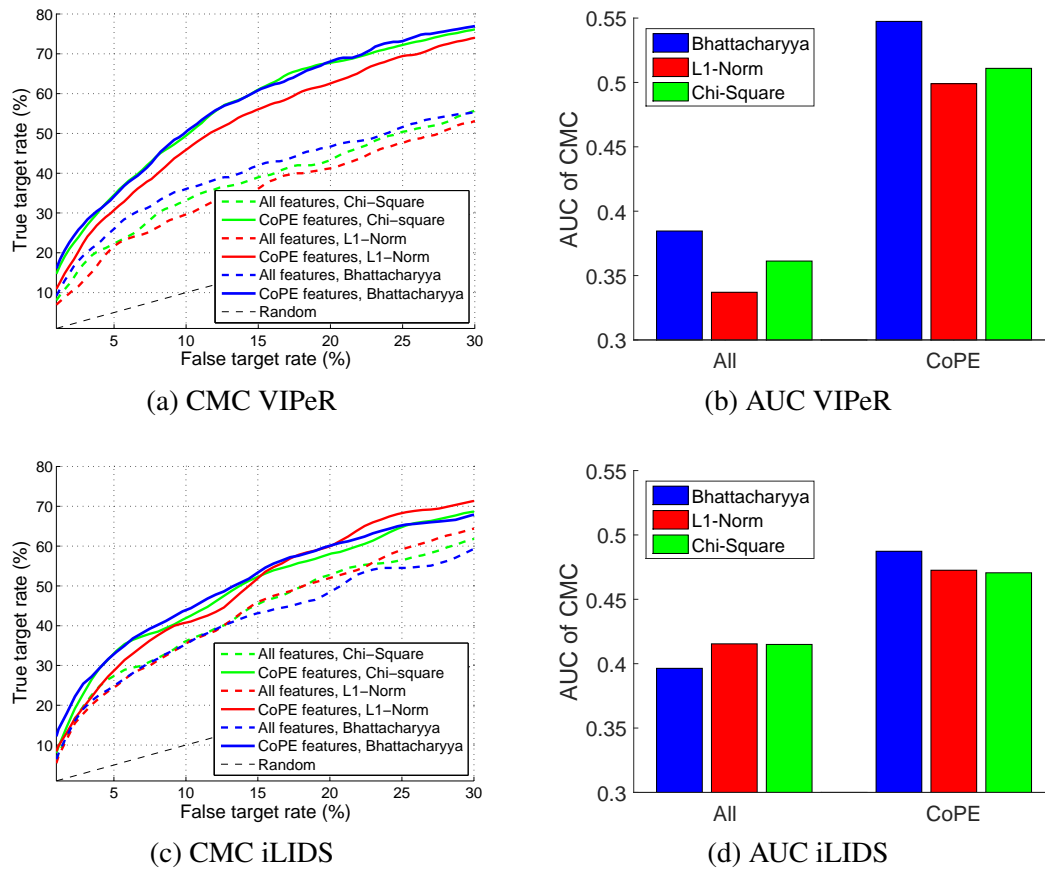


Figure 5.4: Person re-identification comparison for CoPE features selected using three similarity measures, namely Bhattacharyya distance (blue), L1-Norm (red) and Chi-square (green) in Eq. 3.2, compared with the initial complete feature set using DDM approaches for re-identification, using CMC curves representing true target rate for the top 30% of false target rate, and the AUC of the CMC curves in VIPeR ( $M = 316$ ) and iLIDS ( $M = 174$ ) datasets.

The re-identification results for association using the selected features are improved and in some points are comparable to that of using all features. The AUC shows that the features selected using all the three similarity measures have overall better performance than that of the initial feature set. The highest re-identification rate is obtained when the features are selected using the Bhattacharyya distance. Therefore, in the following experiments we use the Bhattacharyya distance as a similarity measure while comparing with existing re-identification and feature selection approaches.

#### 5.2.4 CoPE vs feature selection methods

We compare CoPE with five existing feature selection and ranking methods, namely Fisher score [55], Information gain [44], mRMR [131], ReliefF [140] and Bi-clusters [82]. Since these are single-objective feature selection approaches, for comparison we perform feature selection using

Table 5.2: Training times and ranking orders of features for re-identification using Fisher score [55], Information gain [44], mRMR [131], ReliefF [140] and Bi-clusters [82] as feature selection methods compared with CoPE and CoPE-FC using VIPeR ( $M = 316$ ) and iLIDS ( $M = 174$ ).

Feature selection	Training time (sec)		Ranking order	
	VIPeR	iLIDS	Ratio	
<b>Fisher score</b>	3.67	0.15	24.47	VIPeR: 4 8 9 5 1 2 6 7 3 17 14 13 11 10 23 30 16 12 18 21 26 27 24 28 22 29 19 25 15 20 iLIDS: 3 4 7 6 2 10 11 9 1 13 19 30 16 20 17 26 14 27 12 8 28 23 21 24 18 25 15 29 22 5
<b>Information gain</b>	14.10	4.99	2.82	VIPeR: 4 8 9 5 6 2 1 3 7 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 iLIDS: 1 7 3 6 11 10 20 13 9 16 4 17 2 14 19 12 30 26 27 15 18 28 21 23 8 25 24 29 22 5
<b>mRMR</b>	29.06	9.54	3.04	VIPeR: 4 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 9 8 5 7 6 3 2 1 iLIDS: 1 30 26 23 7 6 2 24 28 3 27 29 21 17 14 22 15 18 20 12 10 19 16 11 13 25 8 4 9 5
<b>ReliefF</b>	124.70	44.98	2.77	VIPeR: 4 8 9 5 27 25 24 2 13 29 6 1 28 30 10 7 3 26 23 22 11 15 16 19 12 21 20 18 17 14 iLIDS: 8 4 9 25 13 22 23 16 7 26 1 5 18 2 29 15 21 24 19 28 30 6 27 14 12 20 3 11 17 10
<b>Bi-clusters</b>	-	72000	-	- iLIDS: 9 8 24 25 12 14 5 11 4 15 10 21 17 16 6 1 28 13 26 23 7 3 20 18 2 19 30 27 29 22
<b>CoPE</b>	0.76	0.30	2.53	VIPeR: 4 8 9 5 7 25 iLIDS: 7 3 9 8 25 11 4 19 21 29
<b>CoPE-FC</b>	0.52	0.20	2.60	VIPeR: 4 8 9 5 7 25 iLIDS: 7 10 25 9 19 3 8 30 4 27 21 29 5

the performance only while keeping the cost fixed (CoPE-FC). The similarities between the feature pairs obtained using Eq. 3.2 along with the assigned labels as correct/incorrect matches are given as input to the feature selection methods. Feature selection methods return a ranked list of features and a weight vector in the case of Fisher score, Information gain and ReliefF methods, while mRMR and Bi-clusters return only a ranked feature list.

Table 5.2 shows the training time for feature selection and the obtained features ranked in order of importance for re-identification. Note that the training time of feature selection does not include the time required for other steps involved in a re-identification system, such as object detection, image representation and feature extraction. Training time is useful to understand the feasibility of the single time set-up off-line process and becomes crucial as the size of the network increases. The training time is measured using 316 and 174 people in the VIPeR and iLIDS, respectively. With VIPeR, CoPE takes 0.76 seconds, 5 times less than the next shortest training time by the Fisher score. ReliefF requires the maximum time (124.70 seconds) for training, while Bi-cluster could not be trained for VIPeR even after 25 days. With iLIDS, the training time of Bi-clusters is nearly 20 hours. Therefore, in a larger camera network Bi-clusters may not be applicable for feature selection. CoPE and CoPE-FC take 0.30 and 0.20 seconds, respectively. The Fisher score takes 0.15 seconds. As the dataset size almost doubles from iLIDS to VIPeR, the time requirement for Fisher Score is increased by nearly 24 times, whereas others are only 3 times longer. With the smallest ratio and minimum training time, CoPE is desirable for feature selection in a camera network.

In Table 5.2, each selection approach returns a different ranking order of features, since



there exists no unique feature subset to solve the same task. If two features show an identical performance, either of the two can be selected. In performing a cost-aware feature selection, CoPE returns a subset of well-performing cost-effective features until any further addition in the cost of features does not improve performance. In VIPeR, most feature selection methods, including CoPE and CoPE-FC, return similar sets with colour features in the top ranks. CoPE and CoPE-FC returns the same set of 6 features because of the similarity in the selection procedure. In iLIDS, 10 features are selected by CoPE, while 13 features are selected by CoPE-FC. We fix the number of selected features for the existing methods to be equal to the number of features selected by CoPE-FC (a comparison with varying number of selected features can be seen in Fig. 5.11). We pick the top 13 features in iLIDS and the top 6 features in VIPeR from the ranked features of the existing approaches.

Fig. 5.5 (a, b) shows the normalised cost  $\mathcal{E}$  (Eq. 5.1) of the obtained selected features. Even after fixing the number of selected features,  $\mathcal{E}$  for CoPE features remains the smallest. mRMR features show the highest cost in both datasets, while those of Fisher score, Information gain and Bi-clusters have costs comparable with that of CoPE-FC. In VIPeR, the CoPE feature set contains all colour features because of the limited illumination changes, while in iLIDS both colour and texture features are selected. CoPE selects the colour features first and then the texture, resulting in the lowest  $\mathcal{E}$  of 0.15 and 0.30 in VIPeR and iLIDS, respectively.

Fig. 5.5 (c-f) shows the re-identification performance of the selected features using DDM (Bhattacharyya) as the association method. In both VIPeR and iLIDS, the selected features using CoPE and CoPE-FC reach the highest re-identification rate. In iLIDS, CoPE-FC reaches the highest performance in the absence of the cost constraints. Unlike the existing approaches based on overall performance only, CoPE selects features by iteratively relaxing the performance score  $A_i^r$ , thus achieving cost as well as performance advantages.

Fig. 5.6 shows the cross-data robustness of selected features. Features are selected on one dataset and tested on the other to analyse the amount of degradation in the results. We compare CoPE with two feature selection approaches, namely Fisher score and Information gain, which have the highest performance in the cross validation within the same dataset. In VIPeR, the performance of CoPE is degraded less compared to the other two methods. The performance of features selected using VIPeR deteriorates at a greater rate in iLIDS, which is a more challenging dataset. The results are degraded at a comparable rate for all the feature selection approaches,

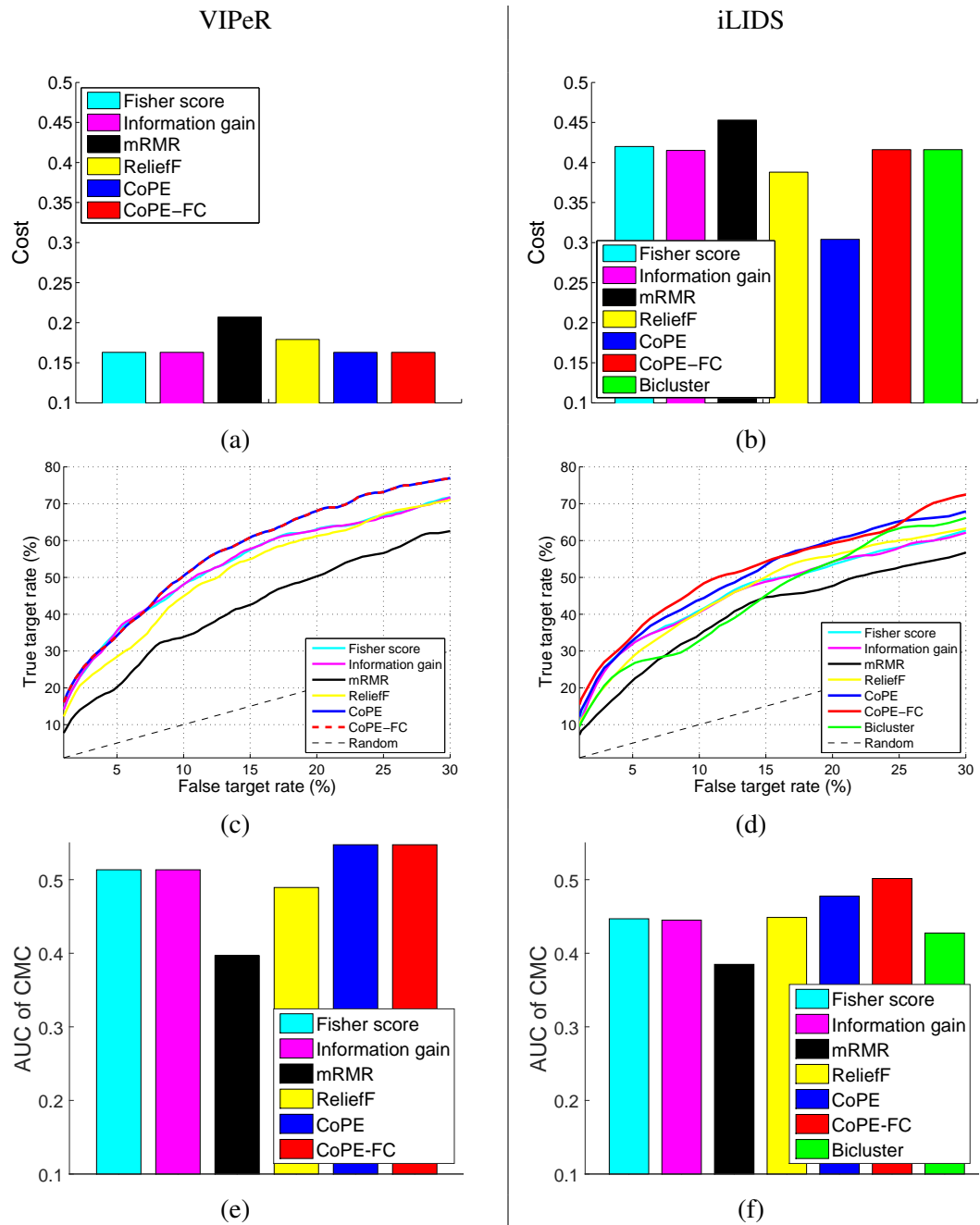


Figure 5.5: (a, b) Normalised cost, (c, d) CMC curves and (e, f) AUC of CMC curves obtained for re-identification by applying DDM to the features selected using the Fisher score [55] (cyan), Information gain [44] (magenta), mRMR [131] (black), ReliefF [140] (yellow) and Bi-clusters [82] (green), CoPE (blue) and CoPE-FC (red) on (left-column) VIPeR ( $M = 316$ ) and (right-column) iLIDS ( $M = 174$ ).

since almost the same 6 colour features are selected by the three feature selection approaches.

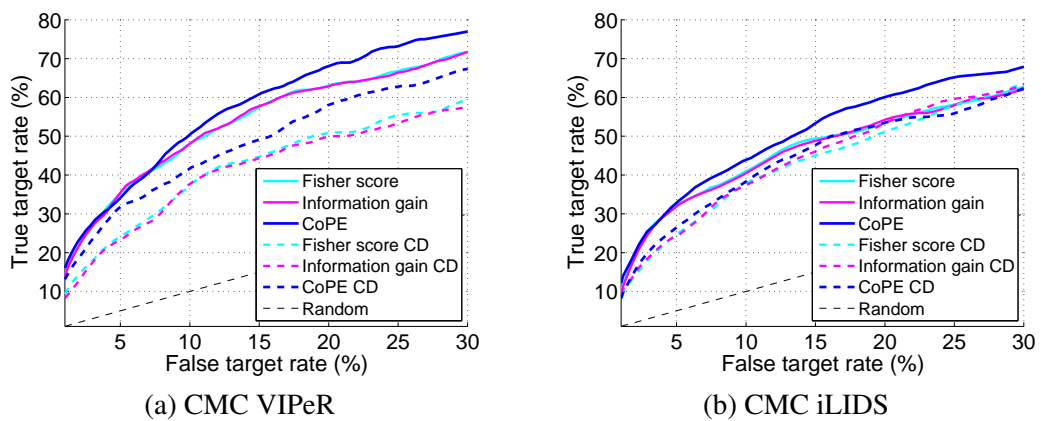


Figure 5.6: Cross Data (CD) performance comparison in re-identification for the top-two performing existing feature selection approaches, namely Fisher score [55] (cyan) and Information gain [44] (magenta); and CoPE (blue). The CMC curves are obtained by (a) feature selection on iLIDS ( $M = 174$ ) and testing on VIPeR ( $M = 316$ ) and by (b) feature selection on VIPeR ( $M = 316$ ) and testing on iLIDS ( $M = 174$ ).

### 5.2.5 CoPE with learning models

The top-ranked selected features are used as input to the two learning methods, namely RankSVM [136] and AdaBoost [71] for re-identification, which apply implicit feature selection by weighting the feature set. In these cases, feature selection may be used to remove poorly performing features as a pre-processing step to improve the effectiveness of learning methods. Since the features are rearranged and weighted within the specific learning method, the order of selection is not important and only the difference in the selected features affects the performance. We compare the performance of RankSVM and AdaBoost with their default settings. The comparisons are performed with and without feature selection keeping the same settings, which may not be optimal. However, the improvement in the results can be observed after the feature selection by the proposed approach.

RankSVM assigns relative weights to the input features based on the combined contributions in the feature set. Fig. 5.7 shows that RankSVM has a better re-identification rate for both VIPeR and iLIDS using the features selected by CoPE compared to those from existing feature selection methods. The variation in re-identification rates using the selected features from different approaches is smaller in VIPeR than in iLIDS because mostly the same colour features are selected (Table 5.2). With iLIDS, the features selected by different methods (and the re-identification rate) vary in their composition. The best performance of CoPE-FC in the true target rate (CMC curves) is almost 15% higher than that of mRMR at the same false target rate, followed by CoPE

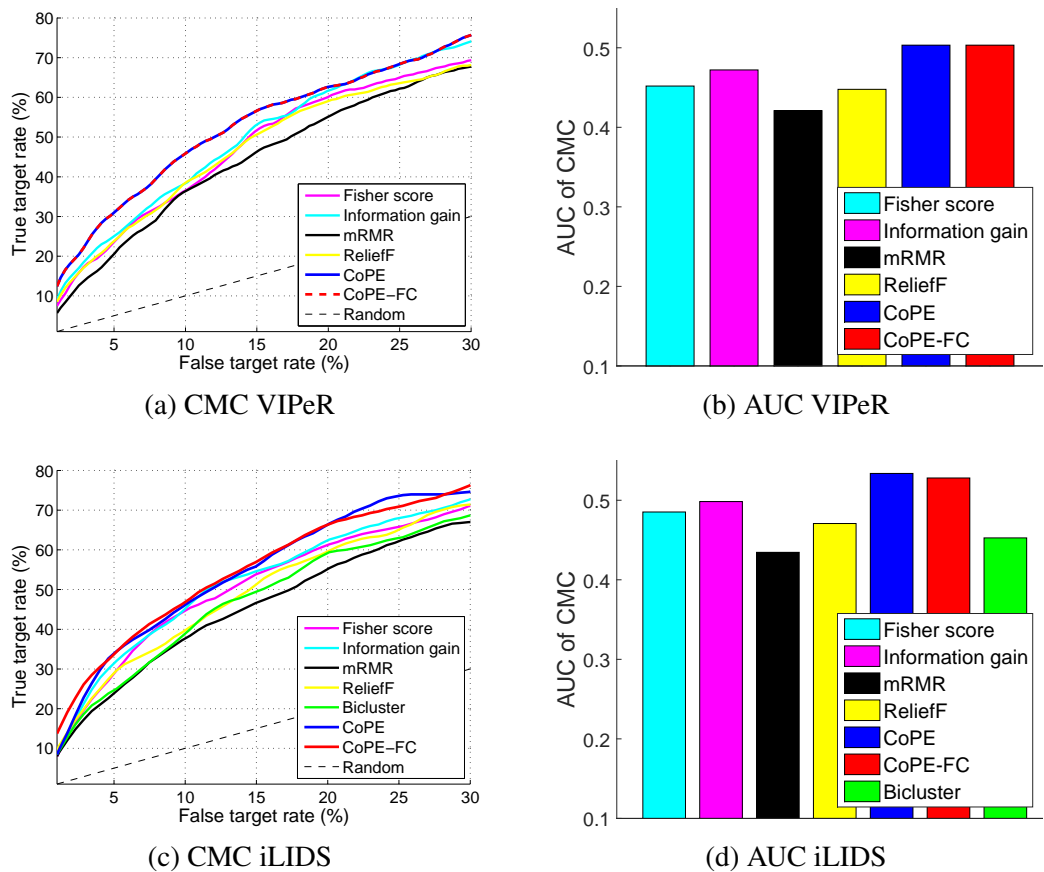


Figure 5.7: CMC curves and AUC of CMC curves for re-identification using the learning method RankSVM [136] applied to the features selected by CoPE (blue) and CoPE-FC (red) and existing methods: Fisher score [55] (cyan), Information gain [44] (magenta), mRMR [131] (black), ReliefF [140] (yellow) and Bi-clusters [82] (green), using (a, b) VIPeR ( $M = 316$ ); and (c, d) iLIDS ( $M = 174$ ).

with a slightly smaller re-identification rate because of the additional cost constraints. However, CoPE remains higher than existing feature selection approaches. Also, the obtained AUCs are highest for CoPE and CoPE-FC.

AdaBoost combines multiple weak classifiers/features to improve the matching performance. Fig. 5.8 shows the performance for AdaBoost. In both VIPeR and iLIDS, the features selected by CoPE have an overall better or comparable re-identification rate than existing feature selection methods. In VIPeR, similarly to the RankSVM, CMC curves show a smaller re-identification rate variation among existing methods because of the limited number of selected features (i.e. only 6). In iLIDS, the variation in performance between CoPE and existing feature selection methods becomes high as the number of selected features is increased (up to 13). AdaBoost has a better learning ability in iLIDS than in VIPeR. The performance on the CMC curves, especially in the starting part, shows that CoPE and CoPE-FC are able to remove noisy features

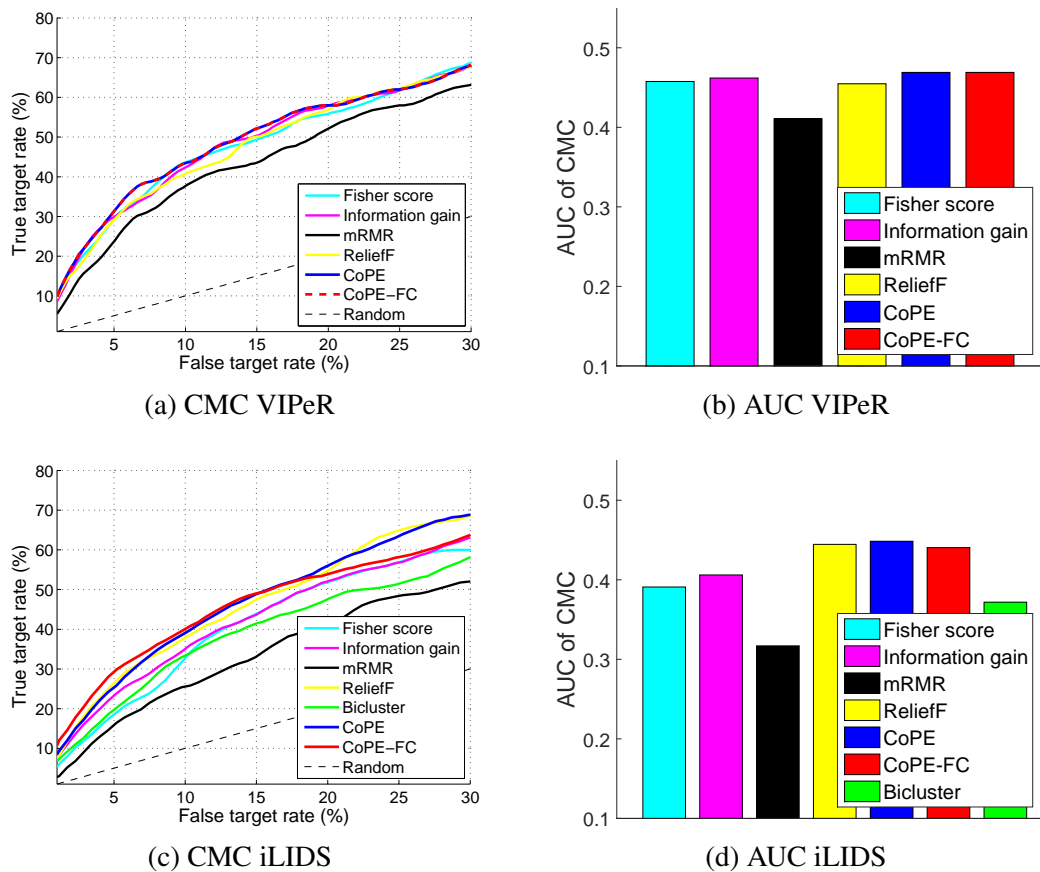


Figure 5.8: CMC curves and AUC of CMC curves for re-identification using the learning method AdaBoost [71] applied to the features selected by CoPE (blue) and CoPE-FC (red) and existing methods: Fisher score [55] (cyan), Information gain [44] (magenta), mRMR [131] (black), ReliefF [140] (yellow) and Bi-clusters [82] (green), using (a, b) VIPeR ( $M = 316$ ); and (c, d) iLIDS ( $M = 174$ ).

more effectively than existing feature selection methods thus resulting in a better re-identification rate. In Fig. 5.8 (c), the CMC curve for ReliefF shows a marginal improvement of up to 2% in true target rate between 20% and 25% of false target rates at the expense of more costly features than that of CoPE (Table: 5.2). In CoPE because of the cost constraints, we may observe a drop in the performance in a few instances in favour of cost reduction and an overall performance improvement. Overall CoPE-FC remains the highest (AUC) followed by CoPE and ReliefF features.

Since learning algorithms are dependent on the training data in addition to the selected features, in challenging scenarios the performance of learning methods can be reduced. A single person may exhibit several pose and illumination changes, while we can only extract a few patches for re-identification thus resulting in an under sampled data representation [179]. For example, in VIPeR [CMC curves in Fig. 5.4 (a) in comparison with Fig. 5.7 (a) and Fig. 5.8 (a)],

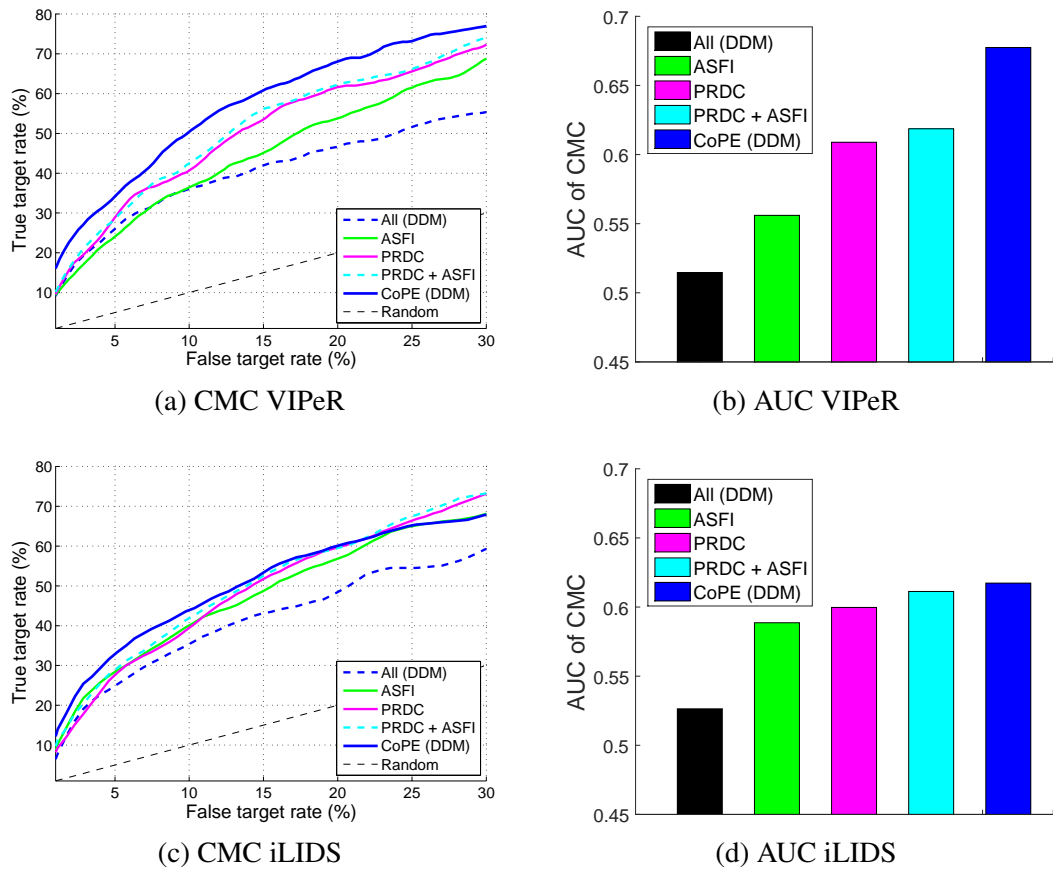


Figure 5.9: Person re-identification comparison of existing re-identification approaches: PRDC [179] (magenta) and ASFI [111] (green) using the complete feature set compared with DDM (Bhattacharyya) [71] using CoPE features (blue) and the complete feature set (black), using CMC curves representing true target rate for the top 30% of false target rate, and the AUC of the CMC curves in VIPeR ( $M = 316$ ) and iLIDS ( $M = 174$ ) datasets.

the performance of learning methods is slightly reduced. In Fig. 5.4 (a) we can see that after CoPE feature selection the performance is improved (almost double compared to using the initial feature set). A further improvement through a learning method will require a more robust training set.

### 5.2.6 CoPE and re-identification approaches

Fig. 5.9 shows the performance comparison of DDM (Bhattacharyya) using CoPE with two recent state-of-the-art re-identification approaches: PRDC [179] and ASFI[111]. The extracted features from the upper-body patch are given as input to PRDC and ASFI. In both iLIDS and VIPeR, a better or comparable re-identification performance is achieved by CoPE with less storage and computational requirements. CMC curves show a higher re-identification rate for CoPE especially at lower false target rates. CoPE outperforms PRDC and ASFI (AUC in the case of

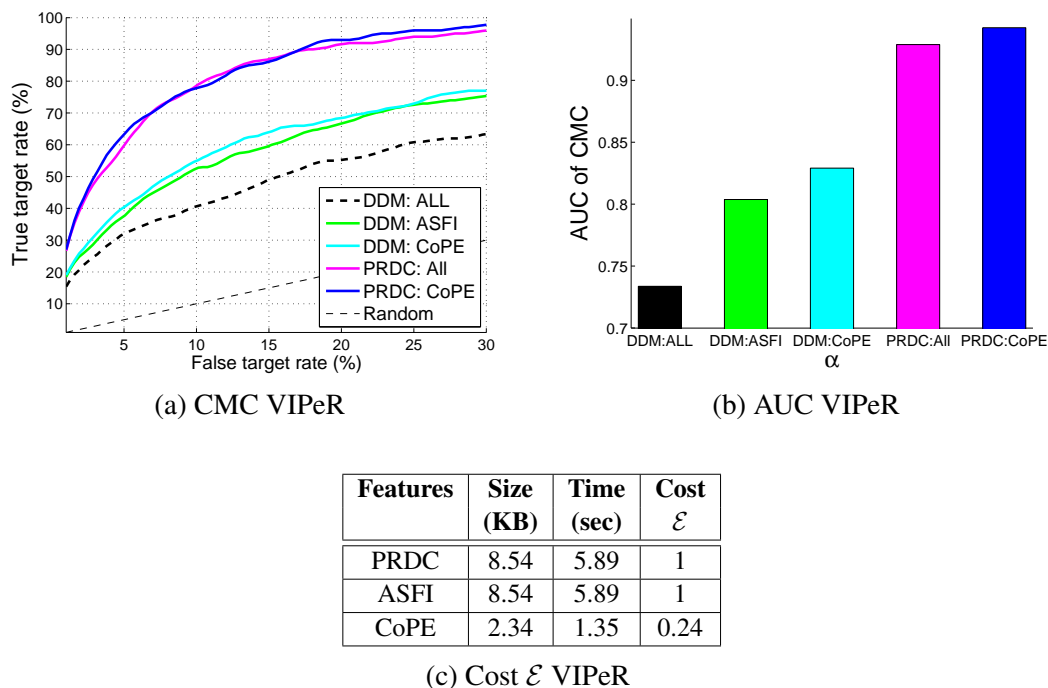


Figure 5.10: Person re-identification comparison using full-body patches and a 2784-dimensional feature vector on VIPeR ( $M = 316$ ). (a) CMC curves and (b) AUC of CMC curves obtained by existing re-identification approaches: PRDC [179] (magenta) and ASFI [111] (green). The results are compared with CoPE using DDM (cyan) and PRDC (blue) for association. (c) The storage size and the computational time of the extracted features.

Table 5.3: Comparison of average training times of existing re-identification approaches compared with CoPE on VIPeR ( $M = 316$ ) dataset. The compared results are from [98]

Approaches	KISSME [98]	KISS-RS [157]	LDML [73]	LMNN [169]	ITML [49]	CoPE	CoPE-FC
Training time (sec)	0.1	0.027	0.72	27.56	8.6	0.76	0.52

VIPeR [Fig. 5.9 (f)], with a cost of 20% of the initial feature set used in these methods. In iLIDS comparable results can be observed at 33% of the cost.

PRDC and ASFI approaches reported their results using the full-body patches and a large 2784-dimensional feature set. Therefore we also include a comparison while performing the CoPE feature selection on the larger feature set and full-body patches. Fig. 5.10 shows the cost-performance comparison on VIPeR, which has fewer occlusions and thus justifies the use of the full patch for person description. It can be observed from the CMC curves and the AUC that CoPE selected features with DDM show a better re-identification rate than ASFI with a 73% reduction in the storage size and a 77% reduction in the extraction time. Finally, the use of the CoPE features as input to PRDC further improves the re-identification rate at 24% of the cost [Fig. 5.10 (c)] of feature sets of PRDC and ASFI.

Table 5.4: Comparison of re-identification rates of existing re-identification approaches at different ranks compared to CoPE using VIPeR ( $M = 316$ ). The compared results are from [6].

Rank $\rightarrow$	1	10	20	50	100
LDML [73]	5	21	30	51	71
ELF [71]	12	43	60	81	93
ITML [49]	14	52	71	90	98
LMNN [169]	18	59	75	91	97
LMNN-R [53]	20	68	80	93	99
KISSME [98]	20	62	77	92	98
KISS-RS [157]	24	66	84	93	-
SDALF [59]	20	50	65	85	98
Bhat. [59]	5	17	24	45	60
RSVM [136]	13	50	67	85	94
PRDC [178]	16	54	70	87	97
CoPE-Bhat.	19	30	40	60	79
CoPE-PRDC	29	60	70	88	98

Table 5.3 shows the initial offline training time of existing re-identification approaches: KISSME [98], KISS-RS [157], LDML [73], LMNN [169] and ITML [49], compared to CoPE and CoPE-FC. KISS-RS has the least training time because of incremental learning. CoPE with 0.72 sec remains comparable with LDML, while LMNN has the highest training time of 27.56 sec. Unlike CoPE, which includes the feature selection step, the learning based re-identification process does not perform an explicit feature selection, while a feature weighting for association is performed. Thus, the feature storage and extraction cost for the existing re-identification approaches remain maximum (Fig. 5.10).

Finally Table 5.4 provides the comparison between CoPE and existing re-identification approaches: KISSME [98], KISS-RS [157], LDML [73], LMNN [169], LMNN-R [169] and ITML [49], in the VIPeR dataset. The results are shown in terms of ranks. The table shows the number of persons correctly re-identified till a particular rank, where rank 1 represents the true re-identification rate. CoPE features used with Bhattacharyya distance shows comparable re-identification rate to existing re-identification approaches. 19 people are correctly matched at rank 1 compared to 5 in the case of only Bhattacharyya distance. CoPE features when used as input to PRDC show the highest number of correct matches at rank 1, i.e. 29 compared to 16 of PRDC when applied on the complete feature set.



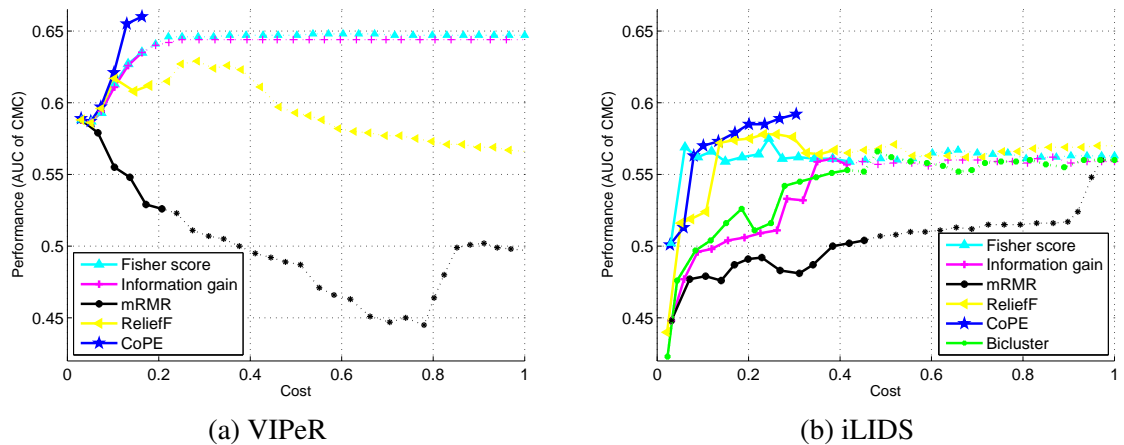


Figure 5.11: Cost vs performance analysis of CoPE (blue star) and existing feature selection methods: Fisher score [55] (cyan plus), Information gain [44] (magenta triangle), mRMR [131] (black circle), ReliefF [140] (yellow triangle) and Bi-clusters [82] (green circle) using DDM (Bhattacharyya) for re-identification on (a) VIPeR and (b) iLIDS. Features are added in order of decreasing performance. Solid lines show the number of selected features equal to those generated with CoPE. Dotted lines extend the cost vs performance comparison for the feature excluded by CoPE. The cost (horizontal axis) is measured using (Eq. 5.1) and the performance (vertical axis) is measured as the area under the first-half of the CMC curves. At each marker point a new feature is added.

### 5.2.7 CoPE and feature budgeting

In a constrained environment, a further reduction of the feature set might be necessary. In such cases the performance needs to be reduced in a predictable manner (feature budgeting). Fig. 5.11 shows the cost vs performance comparison of feature selection methods for re-identification using DDM. The performance is measured using the area under the first half of the CMC curves and the cost is measured using Eq. 5.1. At the beginning, all feature selection methods selected the same feature or the one with comparable performance (Table 5.2), thus resulting in the similar performance at the start of the selection. In CoPE, a consistent increase in performance and cost can be observed with the addition of each new feature. The rate of improvement in the performance is high at the beginning, since the most important features for re-identification are selected first. The performance keeps increasing monotonically as the cost increases, the most desirable behaviour in feature budgeting.

In VIPeR [Fig. 5.11 (a)], the performance of Fisher score and Information gain becomes constant after selection of up to 9 features because of minimal weighting to the lower ranked features; however, the low ranked features keep increasing the cost of the feature set. Such feature selection represents the majority of data with similar properties while neglecting the features

with discriminating capability for small amounts of data. The mRMR feature selection produces a monotonically decreasing performance after reaching a high performance point because of the ranking only strategy. Since VIPeR requires up to 6 discriminant features as selected by CoPE, the additional features result in redundant information and the performance decreases (mRMR) or remains constant (Fisher and Information gain), while the cost increases. In the iLIDS dataset [Fig. 5.11 (b)], the Fisher score shows a non-monotonically increasing performance at the start and, while selecting the second feature, shows a higher performance than CoPE because of the selection of a comparatively costly feature (with Feature ID=4). However, as new features are added the performance starts decreasing, while CoPE preserves a balance between cost and performance, which results not only in a monotonically increasing performance but also in the highest performance with the smallest cost when the same number of features are used. The specific feature (with ID=4) is selected by CoPE at a later stage when its cost justifies the performance. A non-monotonically increasing performance is observed in the Information gain and Bi-clusters; however, their performance is lower than that of CoPE as the cost increases.

This evaluation shows how CoPE can select, in the correct order, less expensive and well-performing features. Improved or comparable performance than the existing selection approaches is achieved by DDM and learning methods for re-identification with cost-effective features.

### 5.3 Re-identification with difference-vector representation

We compare the proposed difference-vector representation and association approach with the following DDM, learning and probabilistic methods: the Bhattacharyya distance [71], RankSVM [136], Attribute-Sensitive Feature Importance (ASFI) [111], Probabilistic Relative Distance Comparison (PRDC) [178] and Landmark Based Model (LBM) [J2]. We use the validation criteria based on the amount of data (in bytes) to be communicated among cameras, and the re-identification rate using CMC curves. The datasets used for the evaluation are Torch [C2], iLIDS-AA [14] and iLIDS-MTC [J2] (Sec. 2.8).

Each object-image is histogram equalised (Sec. 4.2.1), and we extract colour and texture features, where each feature is a 16-bin histogram of a colour channel or a filtered image (Fig. 3.2), extracted from each of the 6 horizontal stripes of the person image as in [71, 136, 178]. We apply the 2-fold cross validation for the evaluation.

Table 5.5: Comparison of the amount of data per person needed to be stored within the camera for object matching.

Dataset	Number of features	Number of people	Bytes per person	
			$\mathbf{F}_n^m$	$\mathbf{\Omega}_{mn}$
Torch	29	54	7539	64
iLIDS-MTC [J2]	29	60	7415	64
iLIDS-AA [14]	29	100	6422	62

### 5.3.1 Data reduction

Table 5.5 shows the amount of data that needs to be stored and communicated per person between the cameras. It can be observed that the storage size per person is reduced to 1% using the compact representation  $\mathbf{\Omega}_{mn}$  of the proposed approach as compared to that of the initial feature set  $\mathbf{F}_n^m$ , since  $\mathbf{F}_n^m$  for each person contains 2784 elements for 29 features extracted from 6 stripes of the image, whereas  $\mathbf{\Omega}_n^m$  contains only  $K = 29 \times 6$  elements. In addition, we require 170 KB per camera for storing the reference-feature vectors  $\{\mathbf{k}^j\}_{j=1}^J$ . The size of the additional storage requirement is a constant that is not affected by the observed number of persons and can be pre-allocated.

Table 5.6 shows the comparison of existing data compression methods: Run-length encoding (RLE) [69], Lempel-Ziv-Welch coding (LZW) [170], Deflate [144], GZip and jpg, applied to the object representations: images, feature sets and difference vectors, in three datasets. In the case of less data available for communication, these compression methods do not perform well because of less redundancy available to exploit (Sec. 2.5). GZip is able to reduce the data, on average, up to 90% of the original size in the case of image representation, while up to 40% in the cases of feature sets and difference vector representations. LZW and Deflate perform better than GZip in the cases of feature sets and difference vector representations, while their performance is less than GZip in the case of image representation. RLE performs the worst because repetitive data is doubled when not placed consecutive to each other. The size of difference vector representation remains significantly less compared to all the compression methods.

### 5.3.2 Re-identification rate

Fig. 5.12 shows the performance gain in re-identification with the addition of each step in the proposed approach. Histogram equalisation improves the performance compared to applying the Bhattacharyya distance on the initial feature sets on both Torch and iLIDS-MTC datasets, since illumination conditions vary in the camera pairs. Improvement can be observed when

Table 5.6: Comparison of different compression methods applied to object images, feature sets and difference-vector representations

Datasets	Information type	Original size (bytes)	Compressed size (bytes)				
			LZW [170]	RLE [69]	Deflate [144]	GZip	jpg
Torch	Object image	156702	109502	313403	123269	9735	9855
	Feature set $F_n^m$	2784	1655	4445	1851	2024	-
	Difference vector $\Omega_{mn}$	64	46	114	55	186	-
iLIDS-MTC [J2]	Object image	63201	37465	126402	42310	3260	3380
	Feature set $F_n^m$	2784	1553	4005	1761	1935	-
	Difference vector $\Omega_{mn}$	64	45	112	54	180	-
iLIDS-AA [14]	Object image	18000	13708	36000	16449	2314	2434
	Feature set $F_n^m$	2784	1795	4754	1986	2159	-
	Difference vector $\Omega_{mn}$	62	44	110	55	178	-

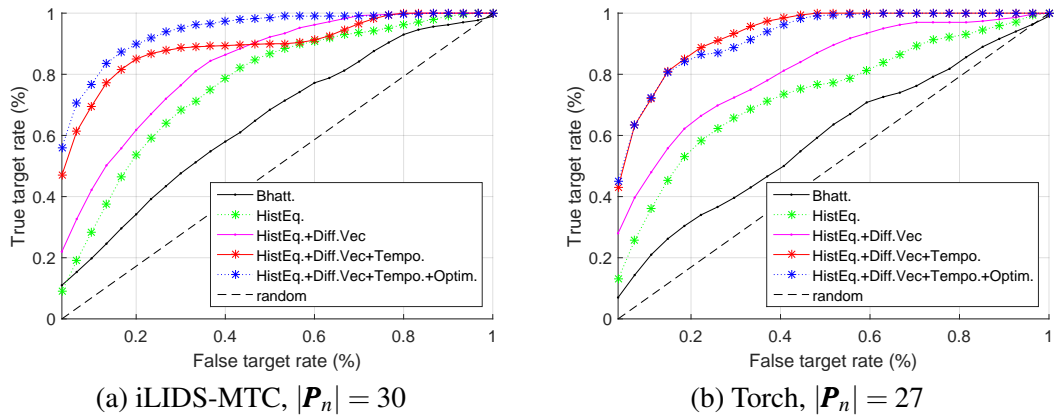


Figure 5.12: CMC curves obtained for matching while adding each step of the proposed approach in two datasets: (a) iLIDS-MTC [J2], (b) Torch [C2]. Key: Bhatt. - Bhattacharyya distance; HistEq. - Histogram Equalisation; Diff.Vec - Difference Vector representation; Tempo. - Temporal alignment; Optim. - Optimal assignment.

difference-vector representation is combined with the histogram equalisation because of the selection of a good set of reference features stored locally. Temporal grouping further improves the re-identification rate, reaching above 40% of the true target rate for zero false target, since the number of matches required for the association is reduced. With the inclusion of optimal assignment in the proposed approach, results in iLIDS-MTC are further improved while no significant improvement compared to temporal grouping is observed in the Torch dataset.

Fig. 5.13 shows the re-identification rate of objects from three camera pairs in the Torch dataset. The proposed approach shows the highest matching results between 50% and 75% true target rate for zero false targets, as compared to the existing approaches, which show a maximum of 40% true target rate for zero false targets in all three pairs of camera settings. In data gatherings using hand-held smart cameras, a sufficient data from the same scene may not always be possible that limits the training of the learning methods and their performance is compromised. The DDM

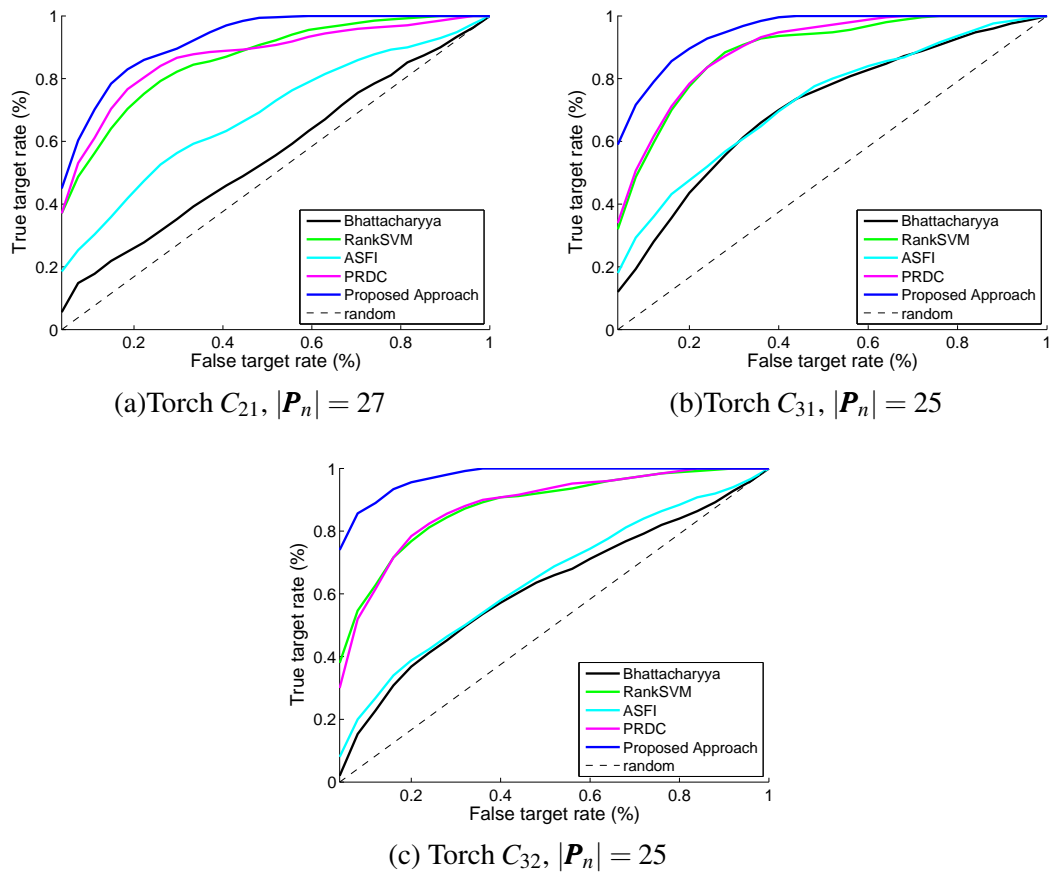


Figure 5.13: CMC curves obtained for matching using the existing approaches: PRDC [178], ASFI [111] and RankSVM [136] compared with the proposed approach on the new Torch dataset with 3 hand-held cameras. The matching is performed pairwise when an object is observed in (a) C2 and C1, (b) C3 and C1, and (c) C3 and C2.

approach shows the minimum performance in the absence of illumination and contrast handling. Additionally, the proposed approach effectively reduces the search space for matching by locally estimating the inter-camera temporal shift, which results in a higher matching rate.

Fig. 5.14 shows the evaluation results of the proposed approach on two datasets obtained from a pair of cameras in iLIDS. The proposed approach shows the higher matching rates with 60% and 45% true target rates at zero false target rate in iLIDS-MTC and iLIDS-AA respectively as compared to the existing approaches. In iLIDS-MTC, we also compare the proposed approach with LBM, a spatio-temporal and appearance approach requiring the actual map and the location of people in the scene along with the appearance information. By only utilising the detection frame numbers and without the requirement of the spatial information of the scene, our approach outperforms LBM, thus allowing the proposed approach to be applied in devices which vary their locations, such as in Torch dataset. The performance of the learning methods is again affected

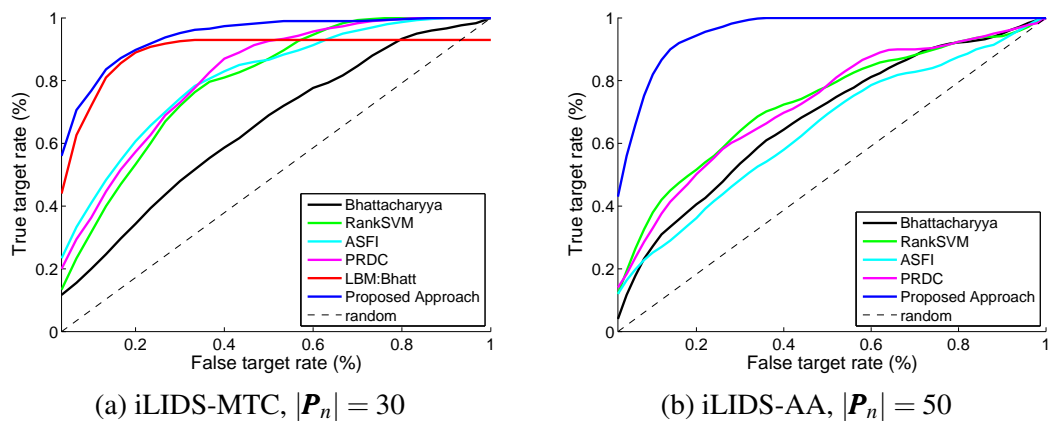


Figure 5.14: CMC curves obtained for matching using existing approaches: PRDC [178], ASFI [111], RankSVM [136] and LBM [J2] compared with the proposed approach in two existing datasets extracted from iLIDS: (a) iLIDS-MTC [J2], (b) iLIDS-AA [14].

by the amount of training data. In the iLIDS-AA dataset, since the objects are extracted after applying the HoG detection algorithm, even in the case of true detections, the extracted image may not have the complete representation of the object. In such scenarios the temporal grouping of the proposed approach improves the overall performance as compared to the approaches based only on appearance information.

This evaluation shows that the proposed approach reduces the amount of data needed to be transferred while improving the re-identification accuracy with respect to existing approaches.

#### 5.4 Re-identification with camera-invariant scores

We evaluate the proposed approach with learning, probabilistic and DDM based pairwise association approaches: PRDC [179], RankSVM [136], ASFI [111] and Bhattacharyya distance, using rank-ROC curves. Unlike CMC [71], rank-ROC curves explicitly show at each rank the false positive rate (FPR) [1-specificity] along with the true positive rate (TPR) [sensitivity].

Single images of persons per camera are manually extracted. We extract colour and texture features. Each feature is a 16-bin histogram of a colour channel or a filtered image, extracted from each of the 6 horizontal stripes of the person image. We use eight colour channels (R, G, B, H, S, Y, Cb, Cr) from RGB, HSV, and YCbCr colour spaces, and for texture, eight Gabor and 13 Schmid filters are applied on the Y channel of the image as in [179, 71, 111].

We also apply the proposed association approach with the features selected by CoPE [J1]. CoPE returns a list  $\mathbf{Y}_{nqn}$  of selected features for each camera pair such that  $P_n^m$  and  $P_{n_q}^k$  are represented by  $\mathbf{F}_n^m = \{\mathbf{f}_n^{m\hat{r}}\}_{\hat{r} \in \mathbf{Y}_{nqn}}$  and  $\mathbf{F}_{n_q}^k = \{\mathbf{f}_{n_q}^{k\hat{r}}\}_{\hat{r} \in \mathbf{Y}_{nqn}}$ , respectively. Using each selected feature

Table 5.7: Experimental setup used for the evaluation.

Datasets	Network size ( $ \mathbf{C} $ )	Persons in $C_n (M_n)$	Source cameras ( $ \mathcal{N}_n $ )	Persons in $C_{n_q} (M_{n_q})$	False positives
WARD	2	30	1	30	-
	3	30	2	15	-
	3	38	2	15	8
Torch	2	24	1	24	-
	3	24	2	12	-
	3	30	2	12	6
	4	24	3	8	-
	5	24	4	6	-

$\hat{\mathbf{f}}^i$ , we apply DDM (Bhattacharyya) to obtain a similarity score matrix  $\mathbf{S}_{nqn}^{\hat{\mathbf{f}}}$  (Eq. 4.7) and estimate  $T_{nqn}^{(\hat{\mathbf{f}})+}$  and  $T_{nqn}^{(\hat{\mathbf{f}})-}$ , for the camera pair  $(C_n, C_{n_qn})$ . For the new objects  $P_n^m$  and  $P_{n_q}^k$ , we obtain the matching score  $\mathcal{L}_{nqn}^{mk(\hat{\mathbf{f}})}$  of a match (Eq. 4.10) using each selected feature separately, and obtain a combined matching score  $\mathcal{L}_{nqn}^{mk}$  for CoPE as

$$\mathcal{L}_{nqn}^{mk} = \frac{\sum_{\hat{\mathbf{f}} \in \mathbf{Y}_{nqn}} \mathcal{L}_{nqn}^{mk(\hat{\mathbf{f}})}}{|\mathbf{Y}_{nqn}|}. \quad (5.2)$$

Finally, we obtain the matrix  $\mathbf{L}_n$  (Eq. 4.11) for assignments.

We use two publicly available person datasets: the WARD dataset [122] and the Torch dataset [C2]. *WARD* contains 70 persons from three non-overlapping fixed-cameras with the challenges of illumination changes, and variations in pose and size. *Torch* contains 50 persons from five hand-held smartphone cameras representing an outdoor crowd scene with additional challenges of occlusions, occasional jitters and blurring. We assume that the person detection problem is solved [57].

We apply two-fold cross validation such that half of the dataset is used for training (Table 5.7). The number of persons detected in  $\mathcal{N}_n$  are fixed to 30 in WARD, and 24 in the Torch dataset. We also perform the experiments with 25% added false positives (FPs), i.e. persons detected in  $C_n$  that do not appear in  $C_{n_q}$ . Note that while we do not perform experiments with false negative detections (i.e. persons detected in  $C_{n_q}$  that do not appear in  $C_n$ ) because of the limited size of the dataset; using the optimal assignment used in the proposed approach, we would expect that the influence of false negatives would be comparable to that of additional false positives in terms of re-identification performance. Finally, we analyse the changes in re-identification rate by increasing the number of cameras.

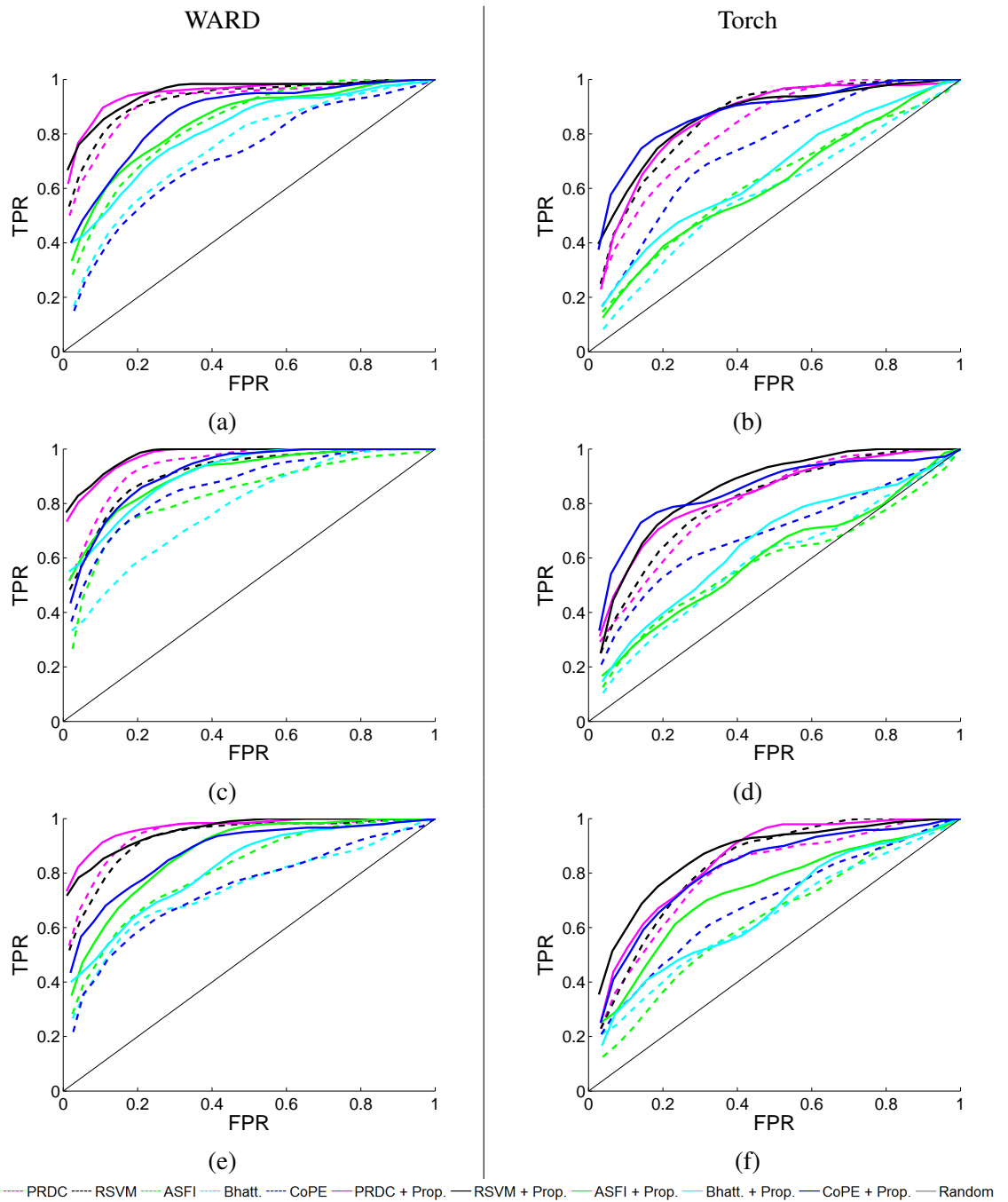


Figure 5.15: Ranked ROC curves for re-identification in three-camera settings using existing approaches: PRDC [179], ASFI [111], RankSVM [136] and CoPE [J1] compared with the proposed association approach using, (left-column) WARD [122] ( $|\mathbf{P}_n| = 30$ ), and (right-column) Torch [C2] ( $|\mathbf{P}_n| = 24$ ) datasets. Each camera detects persons that come from the other two as source-cameras such that persons appear in (a, b)  $C_1$  from  $C_2$  and  $C_3$ , (c, d)  $C_2$  from  $C_1$  and  $C_3$ , and (e, f)  $C_3$  from  $C_1$  and  $C_2$ . Key: Bhatt. - Bhattacharyya distance, Prop. - Proposed approach.

#### 5.4.1 Three-camera setting

Fig. 5.15 shows the ROC curves for re-identification in three-camera settings, i.e. persons detected in  $C_n$  can be from any of the two cameras. Because of the probability estimation in



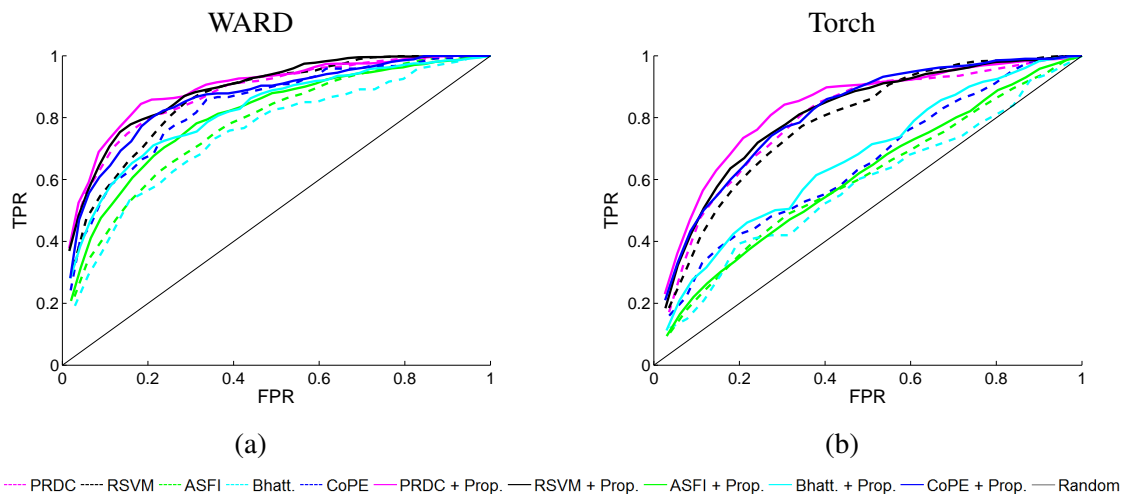


Figure 5.16: Ranked ROC curves for re-identification in three-camera settings with 25% additional persons (FPs) using the existing approaches: PRDC [179], ASFI [111], RankSVM [136] and CoPE [J1] compared with the proposed association approach on (a) WARD dataset [122] and (b) Torch dataset [C2]. Key: Bhatt. - Bhattacharyya distance, Prop. - Proposed approach.

Eq. 4.10, the performances of the existing pairwise methods are improved by the proposed association approach in the two datasets. In WARD dataset, the less illumination changes resulting in more inter-camera similarities make it challenging to learn differences between camera pairs. The proposed approach increases TPR of PRDC and RSVM, from the range between 0.25 and 0.55, up to 0.75 in the start of the curves. In Torch dataset, the re-identification rate is less compared to the WARD for all approaches because of the additional challenges of occlusions and blur; however, compared to the existing approaches, improvement in the re-identification rate can be observed by the proposed approach. CoPE using the proposed approach shows the highest improvement in the performance because of the probability estimation at the feature level (Eq. 5.2), while ASFI and DDM remain the least.

Fig. 5.16 shows the ROC curves for re-identification with added FPs in three-camera settings (8 in WARD and 6 in Torch). In both datasets the re-identification rate is improved with the proposed association approach. In WARD, TPRs for the proposed approach with PRDC and RSVM remain higher starting at 0.3 and reach to 1 at 80% of the FPR. In the Torch dataset, because of the more challenging settings, ASFI and Bhattacharyya do not perform well; however, the learning methods RSVM and PRDC with the proposed approach remain least effected and show improvement in TPR. CoPE with the proposed association approach shows the highest rate of improvement in the re-identification. It is to be noted that the experiments with the False

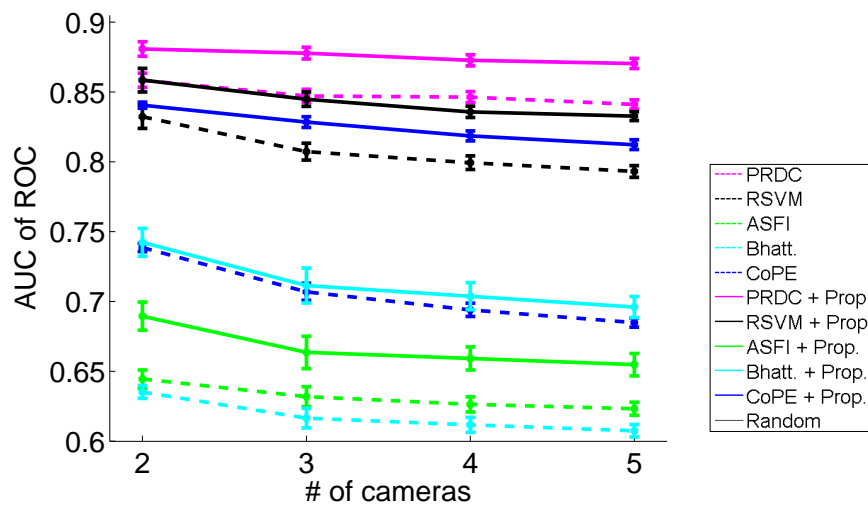


Figure 5.17: AUC of ROC for re-identification as the number of cameras are increased from two to five in Torch dataset. Key: Bhatt. - Bhattacharyya distance, Prop. - Proposed approach.

Negatives (FNs) – persons detected in  $C_{n_q}$  that do not appear in  $C_n$  – are not performed because of the limited availability of data; however, because of the optimal assignment in the proposed approach, we can expect that FNs will have a similar effect on the re-identification as that of additional FPs.

#### 5.4.2 Variable cameras setting

Finally, we analyse using the AUC of ROC how the re-identification performance varies as the number of cameras increases (Fig. 5.17). We use five cameras from the Torch dataset. Keeping the total number of persons detected in  $C_n$  fixed to 24, the experiments are performed for all combinations containing one, two, three and four source-cameras (Table 5.7). The proposed approach improves the re-identification performance for all combinations of cameras in the network. As the number of cameras increases the performance decreases gradually; however, the rate of decrease in performance is relatively small, especially when the proposed association approach is applied with PRDC, RankSVM and CoPE. DDM shows the highest improvement with the proposed approach.

### 5.5 Summary

The experimental evaluation of the proposed feature selection and association methods shows improved performance in terms of re-identification rate with reduced storage, computation and communication costs compared to the existing re-identification approaches.

Unlike existing feature selection methods based on the performance only, the proposed CoPE feature selection approach explicitly incorporates the cost of the feature extraction in the selection process to generate a combined importance score (Eq. 3.8). CoPE decreases both the amount of data generated per feature set and the amount of time needed for the extraction of the selected feature set by up to 80% in VIPeR and up to 70% in the iLIDS dataset, without compromising on the re-identification rate compared with the existing re-identification approaches. We also demonstrated that, compared to the existing feature selection methods, CoPE improves the performance of other learning based re-identification approaches such as those based on RankSVM [136] and AdaBoost [71] by reducing the feature dimensions and the training time, and by improving their effectiveness. A further reduction of the selected features is made possible to account for additional operational constraints (e.g, limited resources). However, we see a limitation of the proposed approach, such as in the case of multiple source-cameras it is possible that the locally stored lists of selected features may together result in the extraction of the complete feature set. Such a case may occur if source-cameras are located far away and reduces the benefits of feature selection.

The proposed difference-vector representation approach exploits and combines already existing concepts such as histogram equalisation, difference from reference sets and temporal alignment so that most of the computation can be performed locally. Thus the approach considerably reduces network traffic because of less inter-camera feature sharing. With the inclusion of the temporal alignment [120], the proposed approach also achieves a higher matching rate compared to the existing re-identification approaches – PRDC [179], ASFI [111], RankSVM [136] and LBM [J2]. We use both outdoor and indoor datasets for the evaluation, and the results show that the proposed method reduces up to 95% the amount of information to be communicated – less than 100 bytes per person and a fixed local storage required for the reference-feature vectors. We achieved up to 75% re-identification accuracy in the Torch dataset. However, the re-identification performance is highly dependant on the selection/generation of the reference feature sets. The task of sharing the reference feature sets with each device taking part in the re-identification can also be seen as a limitation.

The proposed multi-camera re-identification approach extends the pairwise re-identification methods to multiple source-cameras. Matching scores are generated by measuring the probability of a correct match, which makes it independent of the camera pair, where the object appears.

Results from two multi-camera datasets, WARD [122] and Torch [C2], show that the proposed approach improves the re-identification rate by 20% on average, while the degradation in the performance as the number of cameras increases is less compared to existing approaches, namely: PRDC [179], ASFI [111]) and RankSVM [136]. The proposed approach also supplements the CoPE feature selection [J1] by proposing a suitable association method for the individually selected good performing features and results show better performance than the existing distance minimisation approaches. However, as the number of cameras increases the availability of sufficient data required for the training can become a limitation. In order to overcome this limitation, approaches based on transfer learning for training, such as [104], can be exploited.

## Chapter 6

### Conclusions

---

#### 6.1 Summary of achievements

In this thesis, we addressed the problem of person re-identification with the aim of reducing computational, storage and communication resources, while maintaining or improving the re-identification performance. The main applications for our work could be identified in video surveillance and smart-camera networks. Changes in illumination, pose and scale of the object, variability in inter-camera travelling times and the location of re-appearances of objects in the next camera, make re-identification a challenging problem. In order to improve the re-identification accuracy, multiple types of object descriptors are developed and combined [12, 18, 59, 64, 136, 179]; however, these representations may become highly computationally extensive. From the survey of the state of the art in Chapter 2, it is observed that the existing re-identification approaches exploit these features for improving the re-identification rate without considering constraints on resource utilisation thus limiting the usability and scalability of the approaches in real-world applications. This problem can be addressed by feature selection; however, very little work has been done to consider the cost of a feature such as, for example, the computational time for its extraction and the amount of data that is necessary for its storage. Existing feature selection approaches [82, 131, 140] focus on reducing the number of features. This implicitly implies that all features belong to the same class and thus reducing the number of features will reduce the cost. However, the cost of feature extraction and storage varies significantly across different topology of features to be shared among cameras, and it is independent of

the feature's performance.

To this end, we proposed a cost-effective feature selection method that selects, the feature set which is computationally less expensive to extract and requires less storage, while having performance comparable or better than other features available [J1]. We evaluated the approach on challenging person datasets to analyse the performance gain with less costly features. In particular, we extended the iLIDS dataset to get up to 348 pairs of person images in two cameras and used the existing VIPeR dataset. Our approach reduces computation and storage requirements within a camera by 70% of the initial feature set, while the re-identification performance is better than the existing feature selection methods and re-identification approaches. The individually selected features perform well in constrained environments, thus making the approach scalable for transmitting a selection of data over the network.

In addition to this, we discussed how the association of objects can be performed with the minimum amount of information shared between cameras in order to address bandwidth constraints [C2]. Our approach maximises the dependencies on information available locally within the camera. The objects are also grouped within a defined time-interval, which further improves the re-identification by reducing the number of candidates for matching. For evaluation, we generated a multi-camera outdoor crowded dataset of short-duration videos using hand-held smartphones (Torch dataset). The dataset contains 50 people seen by five cameras over time. The communication cost is considerably reduced, while the re-identification rate is maintained better than existing re-identification approaches.

Furthermore, the existing re-identification approaches perform re-identification in a pair of cameras with the assumption that a source-camera with prior detections of the object is known. However, we can have more than one source-camera in real-world scenarios. We discussed our object association approach that relaxes the assumption of pairwise association [C1]. The approach extends the existing pairwise methods to a multiple source-camera scenario by generating camera-invariant matching scores for association. The evaluation on two multi-camera datasets shows a better re-identification rate, and less performance degradation as the number of cameras is increased, compared to the existing re-identification approaches.

In summary, we have demonstrated that cost and performance effective solutions for the open problem of person re-identification can be designed to improve the re-identification rate while efficiently utilising the existing resources, thus improving the applicability of re-identification

approaches to real-world environments.

## 6.2 Future directions

Possible directions from this research are:

1. In the re-identification task, manual annotations were used in the evaluations and detection and tracking of people are assumed to be solved. One extension of the work could be the use of actual detection and tracking outputs, which would involve the estimation and separation of missed and false detections before applying the re-identification methods.
2. CoPE feature selection method in Chapter 3 uses distances from the same and different objects to measure the discriminating ability of a feature jointly with the extraction cost. This makes the approach suitable for different classes of features having varying costs and extendible to problems other than the re-identification, such as information retrieval, medical image analysis and diagnostics, text classification, data mining and big data analysis.
3. CoPE feature selection needs a training between each camera pair. The concept of transfer learning can be exploited for this task, such that features learned for one scenario/camera pair can be mapped to the others. Initial research has been done in this area [104]; however, the variations in challenges, for re-identification, from one camera pair to another needs to be addressed.
4. In the evaluation of our multi-camera association approach (Sec. 5.4), we explored the concepts of an open system [29] by including False Positives FPs (new persons detected in a camera) and False Negatives (FNs) (persons that exit a camera network) in the framework of re-identification. The evaluation remained limited up to five cameras because of a limited data. This concept requires further exploration, which could include the proposal of new or modified measures to quantitatively evaluate an open system. A larger dataset involving a network of multiple cameras viewing a large number of objects is also required to measure the scalability of the approach, since this is currently not available.

## Bibliography

- [1] Norman Abramson. *Information Theory and Coding*. Mc Graw-Hill, New York, USA, 1963.
- [2] Tinku Acharya and Ajoy Kumar Ray. *Image Processing - Principles and Applications*. Wiley-Interscience, USA, 2005.
- [3] Hamid Aghajan and Andrea Cavallaro. *Multi-Camera Networks: Principles and Applications*. Elsevier, Netherlands, 2009.
- [4] Alexandre Alahi, Pierre Vanderghenst, Michel Bierlaire, and Murat Kunt. Cascade of descriptors to detect and track objects across any network of cameras. *Computer Vision and Image Understanding*, 114(6):624–640, 2010.
- [5] Alexandra Branzan Albu, Denis Laurendeau, Sylvain Comtois, Denis Ouellet, Patrick Hebert, Andre Zaccarin, Marc Parizeau, Robert Bergevin, Xavier Maldague, Richard Drouin, Stephane Drouin, Nicolas Martel-Brisson, Frederic Jean, Helene Torresan, Langis Gagnon, and France Laliberte. Monnet: Monitoring pedestrians with a network of loosely-coupled cameras. In *Proc. of IEEE Int. Conf. on Pattern Recognition*, Hong Kong, China, August 2006.
- [6] Le An, Mehran Kafai, Songfan Yang, and Bir Bhanu. Reference-based person re-identification. In *Proc. of IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*, Krakow, Poland, August 2013.
- [7] Nadeem Anjum and Andrea Cavallaro. Trajectory association and fusion across partially overlapping cameras. In *Proc. of IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*, Genova, Italy, September 2009.
- [8] Nadeem Anjum, Murtaza Taj, and Andrea Cavallaro. Relative position estimation of non-overlapping cameras. In *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Honolulu, USA, April 2007.
- [9] Neil Ashby. Relativity and the global positioning system. *Physics Today*, 55(1):41–47, 2003.
- [10] Walid Ayedi, Hichem Snoussi, and Mohamed Abid. A fast multi-scale covariance descriptor for object re-identification. *Pattern Recognition Letters*, 33(14):1902–1907, 2012.



- [11] Kheir-Eddine Aziz, Djamel Merad, and Bernard Fertil. People re-identification across multiple non-overlapping cameras system by appearance classification and silhouette part segmentation. In *Proc. of IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*, Klagenfurt, Austria, August 2011.
- [12] Slawomir Bak, Guillaume Charpiat, Etienne Corvee, Francois Bremond, and Monique Thonnat. Learning to match appearances by correlations in a covariance metric space. In *Proc. of IEEE Int. Conf. on Computer Vision*, Firenze, Italy, October 2012.
- [13] Slawomir Bak, Etienne Corvee, Francois Bremond, and Monique Thonnat. Person re-identification using haar-based and dcd-based signature. In *Proc. of IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*, Boston, USA, August 2010.
- [14] Slawomir Bak, Etienne Corvee, Francois Bremond, and Monique Thonnat. Multiple-shot human re-identification by mean riemannian covariance grid. In *Proc. of IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*, Klagenfurt, Austria, September 2011.
- [15] Davide Baltieri, Roberto Vezzani, and Rita Cucchiara. 3dpes: 3d people dataset for surveillance and forensics. In *Proc. of Joint ACM Workshop on Human Gesture and Behavior Understanding*, Arizona, USA, November 2011.
- [16] Davide Baltieri, Roberto Vezzani, and Rita Cucchiara. Sarc3d: A new 3d body model for people tracking and re-identification. In *Proc. of IEEE Int. Conf. on Image Analysis and Processing*, Ravenna, Italy, April 2011.
- [17] Martin Bauml, Keni Bernardin, Mika Fischer, Hazim Kemal Ekenel, and Rainer Stiefelhagen. Multi-pose face recognition for person retrieval in camera networks. In *Proc. of IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*, Boston, USA, August 2010.
- [18] Martin Bauml and Rainer Stiefelhagen. Evaluation of local features for person re-identification in image sequences. In *Proc. of IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*, Klagenfurt, Austria, August 2011.
- [19] Loris Bazzani, Marco Cristani, Alessandro Perina, and Vittorio Murino. Multiple-shot person re-identification by chromatic and epitomic analyses. *Pattern Recognition Letters*, 33(7):898–903, 2012.
- [20] Guy Berdugo, Omri Soceanu, Yair Moshe, Dmitry Rudoy, and Itsik Dvir. Object re-identification in real world scenarios across multiple non-overlapping cameras. In *Proc. of Europ. Conf. on Signal Processing*, Aalborg, Denmark, August 2010.

- [21] Alina Bialkowski, Simon Denman, Sridha Sridharan, Clinton Fookes, and Patrick Lucey. A database for person re-identification in multi-camera surveillance networks. In *Proc. of IEEE Int. Conf. on Digital Image Computing Techniques and Applications*, Fremantle, Australia, December 2012.
- [22] Nathaniel Bird, Osama Masoud, Nikolaos Papanikolopoulos, and Aaron Isaacs. Detection of loitering individuals in public transportation areas. *IEEE Transactions on Intelligent Transportation Systems*, 6(2):167–177, 2005.
- [23] Alessandro Bissacco and Stefano Soatto. Hybrid dynamical models of human motion for the recognition of human gaits. *International Journal of Computer Vision*, 85(1):101–114, 2009.
- [24] James Black, Tim Ellis, and Dimitrios Makris. Wide area surveillance with a multi camera network. In *Proc. of Intelligent Distributed Surveillance Systems*, London, UK, March 2004.
- [25] James Black, Dimitrios Makris, and Tim Ellis. Hierarchical database for a multi-camera surveillance system. *Pattern Analysis and Applications*, 7(4):430–446, 2004.
- [26] Yinghao Cai, Wei Chen, Kaiqi Huang, and Tieniu Tan. Continuously tracking objects across multiple widely separated cameras. In *Proc. of Asian Conf. on Computer Vision*, Tokyo, Japan, November 2007.
- [27] Simone Calderara, Rita Cucchiara, and Andrea Prati. Bayesian-competitive consistent labeling for people surveillance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):354–360, 2008.
- [28] Simone Calderara, Andrea Prati, and Rita Cucchiara. Hecol: Homography and epipolar-based consistent labeling for outdoor park surveillance. *Computer Vision and Image Understanding*, 111(1):21–42, 2008.
- [29] Brais Cancela, Timothy Hospedales, and Shaogang Gong. Open-world person re-identification by multi-label assignment inference. In *Proc. of British Machine Vision Conference*, Nottingham, UK, September 2014.
- [30] Yaron Caspi, Denis Simakov, and Michal Irani. Feature-based sequence-to-sequence matching. *International Journal of Computer Vision*, 68(1):53–64, 2006.
- [31] Luca Zini Andrea Cavallaro and Francesca Odone. Action-based multi-camera synchro-

- nization. *IEEE Transactions on Circuits and Systems for Video Technology*, 3(2):165–174, 2013.
- [32] Jing-Ying Chang, Tzu-Heng Wang, Shao-Yi Chien, and Liang-Gee Chen. Spatial-temporal consistent labeling for multi-camera multi-object surveillance systems. In *Proc. of IEEE Int. Symp. on Circuits and Systems*, Seattle, USA, May 2008.
- [33] Youssef Charficonjunc, Naoki Wakamiya, and Masayuki Murata. Challenging issues in visual sensor networks. *IEEE Wireless Communications Magazine*, 16(2):44–49, 2009.
- [34] Nitesh Chawla, Kevin Bowyer, Lawrence Hall, and Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16(2):321–357, 2002.
- [35] Rama Chellappa, Amit Roy-Chowdhury, and Amit Kale. Human identification using gait and face. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, Minnesota, USA, June 2007.
- [36] David Chen, Sam Tsai, Vijay Chandrasekhara, Gabriel Takacs, Ramakrishna Vedantham, Radek Grzeszczuk, and Bernd Girod. Residual enhanced visual vector as a compact signature for mobile visual search. *Signal Processing*, 93(8):2316–2327, 2013.
- [37] Kuan-Wen Chen, Chih-Chuan Lai, Pei-Jyun Lee, Chu-Song Chen, and Yi-Ping Hung. Adaptive learning for target tracking and true linking discovering across multiple non-overlapping cameras. *IEEE Transactions on Multimedia*, 13(4):625–638, 2011.
- [38] Dong Seon Cheng, Marco Cristani, Michele Stoppa, Loris Bazzani, and Vittorio Murino. Custom pictorial structures for re-identification. In *Proc. of British Machine Vision Conference*, Dundee, Scotland, August 2011.
- [39] Yongmei Cheng, Wen tian Zhou, Yi Wang, Chun hui Zhao, and Shao wu Zhang. Multi-camera-based object handoff using decision-level fusion. In *Proc. of IEEE Int. Congress on Image and Signal Processing*, Tianjin, China, October 2009.
- [40] Scott Cohen. Background estimation as a labeling problem. In *Proc. of IEEE Int. Conf. on Computer Vision*, Beijing, China, October 2005.
- [41] Dung-Nghi Truong Cong, Louahdi Khoudour, Catherine Achard, Cyril Meurie, and Olivier Lezoray. People re-identification by spectral classification of silhouettes. *Signal Processing*, 90(8):2362–2374, 2010.

- [42] Dajana Conte, Pasquale Foggia, Gennaro Percannella, and Mario Vento. A multiview appearance model for people re-identification. In *Proc. of IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*, Klagenfurt, Austria, August 2011.
- [43] Etienne Corvee, Slawomir Bak, and Francois Bremond. People detection and re-identification for multi surveillance cameras. In *Proc. of Int. Conf. on Computer Vision Theory and Applications*, Rome, Italy, February 2012.
- [44] Thomas Cover and Joy Thomas. *Elements of Information Theory*. Wiley, UK, 1991.
- [45] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, San Diego, USA, June 2005.
- [46] Van Dang and Bruce Croft. Feature selection for document ranking using best first search and coordinate ascent. In *Proc. of ACM workshop in Conf. on Special Interest Group on Information Retrieval*, Geneva, Switzerland, July 2010.
- [47] Antitza Dantcheva and Jean-Luc Dugelay. Frontal-to-side face re-identification based on hair, skin and clothes patches. In *Proc. of IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*, Klagenfurt, Austria, August 2011.
- [48] Abir Das, Anirban Chakraborty, and Amit Roy-Chowdhury. Consistent re-identification in a camera network. In *Proc. of Europ. Conf. on Computer Vision*, Zurich, Switzerland, September 2014.
- [49] Jason Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit Dhillon. Information-theoretic metric learning. In *Proc. of Int. Conf. on Machine Learning*, Oregon, USA, June 2007.
- [50] Icaro Oliveira de Oliveira and Jose Luiz de Souza Pio. People re-identification in a camera network. In *Proc. of IEEE Int. Conf. on Dependable, Autonomic and Secure Computing*, Chengdu, China, September 2009.
- [51] Simon Denman, Clinton Fookes, Alina Bialkowski, and Sridha Sridharan. Soft-biometrics: Unconstrained authentication in a surveillance environment. In *Proc. of IEEE Int. Conf. on Digital Image Computing: Techniques and Applications*, Melbourne, Australia, December 2009.

- [52] Anthony Dick and Michael Brooks. A stochastic approach to tracking objects across multiple cameras. In *Australian Conf. on Artificial Intelligence*, Cairns, Australia, December 2004.
- [53] Mert Dikmen, Emre Akbas, Thomas Huang, and Narendra Ahuja. Pedestrian recognition with a learned metric. In *Proc. of Asian Conf. on Computer Vision*, Queenstown, Newzealand, November 2010.
- [54] Gianfranco Doretto, Thomas Sebastian, Peter Tu, and Jens Rittscher. Appearance-based person reidentification in camera networks: problem overview and current approaches. *Journal of Ambient Intelligence and Humanized Computing*, 2(2):127–151, 2011.
- [55] Richard Duda, Peter Hart, and David Stork. *Pattern Classification*. John Wiley & Sons, USA, 2001.
- [56] Rala Ebied. Feature extraction using pca and kernel-pca for face recognition. In *Proc. of Int. Conf. on Informatics and Systems*, Cairo, Egypt, May 2012.
- [57] Markus Enzweiler and Dariu M. Gavrila. Monocular pedestrian detection: Survey and experiments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(12):2179 – 2195, 2009.
- [58] Lukas Esterle, Peter Lewis, Xin Yao, and Bernhard Rinner. Socio-economic vision graph generation and handover in distributed smart camera networks. *ACM Transactions on Sensor Networks*, 10(2):1–24, 2014.
- [59] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, San Francisco, USA, June 2010.
- [60] Pedro Felzenszwalb. Representation and detection of deformable shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(2):208–220, 2005.
- [61] Marin Ferecatu and Hichem Sahbi. Multi-view object matching and tracking using canonical correlation analysis. In *Proc. of IEEE Int. Conf. on Image Processing*, Cairo, Egypt, November 2009.
- [62] Xiubo Geng, Tie yan Liu, Tao Qin, and Hang Li. Feature selection for ranking. In *Proc. of ACM Conf. on Special Interest Group on Information Retrieval*, Amsterdam, Netherlands, July 2007.
- [63] James Gentle. *Computational Statistics*. Springer, UK, 2009.

- [64] Niloofar Gheissari, Thomas Sebastian, Peter Tu, Jens Rittscher, and Richard Hartley. Person reidentification using spatiotemporal appearance. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, New York, USA, June 2006.
- [65] Andrew Gilbert and Richard Bowden. Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity. In *Proc. of Europ. Conf. on Computer Vision*, Graz, Austria, May 2006.
- [66] Andrew Gilbert and Richard Bowden. Incremental, scalable tracking of objects inter camera. *Computer Vision and Image Understanding*, 111(1):43–58, 2008.
- [67] Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen Change Loy. *Person Reidentification*. Springer, UK, 2014.
- [68] Shaogang Gong, Tao Xiang, and Somboon Hongeng. Learning human pose in crowd. In *Proc. of ACM Int. Conf. on Multimedia*, Firenze, Italy, October 2010.
- [69] Rafael C. Gonzalez and Richard E. Woods. *Digital image processing*. Pearson Prentice Hall, Pearson Education, Inc., 2008.
- [70] Douglas Gray, Shane Brennan, and Hai Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. of IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance*, Rio de Janeiro, Brazil, September 2007.
- [71] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Proc. of Europ. Conf. on Computer Vision*, Marseille, France, October 2008.
- [72] Prithwijiit Guha, Amitabha Mukerjee, and Venkatesh Subramanian. Formulation, detection and application of occlusion states (oc-7) in the context of multiple object tracking. In *Proc. of IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*, Klagenfurt, Austria, August 2011.
- [73] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. Is that you? metric learning approaches for face identification. In *Proc. of IEEE Int. Conf. on Computer Vision*, Kyoto, Japan, September 2009.
- [74] Esin Guldogan and Moncef Gabbouj. Feature selection for content-based image retrieval. *Signal, Image and Video Processing*, 2(3):241–250, 2008.

- [75] Omar Hamdoun, Fabien Moutarde, Bogdan Stanculescu, and Bruno Steux. Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In *Proc. of ACM/IEEE Int. Conf. on Distributed Smart Cameras*, California, USA, September 2008.
- [76] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, UK, 2006.
- [77] Martin Hirzer, Csaba Beleznai, Martin Kostinger, Peter Roth, and Horst Bischof. Dense appearance modeling and efficient learning of camera transitions for person re-identification. In *Proc. of IEEE Int. Conf. on Image Processing*, Florida, USA, October 2012.
- [78] Martin Hirzer, Csaba Beleznai, Peter Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Proc. of Scandinavian Conf. on Image Analysis*, Ystad, Sweden, May 2011.
- [79] Martin Hirzer, Peter Roth, and Horst Bischof. Person re-identification by efficient impostor-based metric learning. In *Proc. of IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*, Beijing, China, September 2012.
- [80] Weiming Hu, Min Hu, Xue Zhou, Tieniu Tan, Jianguang Lou, and Steve Maybank. Principal axis-based correspondence between multiple cameras for people tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):663–671, 2006.
- [81] Guichun Hua, Min Zhang, Yiqun Liu, Shaoping Ma, and Liyun Ru. Hierarchical feature selection for ranking. In *Proc. of Int. Conf. on World wide web*, New York, USA, April 2010.
- [82] Qinghua Huang, Dacheng Tao, Xuelong Li, Lianwen Jin, and Gang Wei. Exploiting local coherent patterns for unsupervised feature ranking. *IEEE Trans. on Systems, Man, and Cybernetics, Part B*, 41(6):1471–1482, 2011.
- [83] David Huffman. A method for the construction of minimum-redundancy codes. *Proceedings of the Institute of Radio Engineers*, 40(9):1098–1101, September 1952.
- [84] iLIDS. Home office multiple camera tracking scenario definition, 2008.
- [85] Adrian Ilie and Greg Welch. Ensuring color consistency across multiple cameras. In *Proc. of IEEE Int. Conf. on Computer Vision*, Beijing, China, September 2005.
- [86] Yohei Ishii, Hitoshi Hongo, Kazuhiko Yamamoto, and Yoshinori Niwa. Face and head

- detection for a real-time surveillance system. In *Proc. of IEEE Int. Conf. on Pattern Recognition*, Cambridge, UK, August 2004.
- [87] Ramesh Jain and Hans Helmut Nagel. On the analysis of accumulative difference pictures from image sequences of real world scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2):206–214, 1979.
- [88] Omar Javed, Zeeshan Rasheed, Khurram Shafique, and Mubarak Shah. Tracking across multiple cameras with disjoint views. In *Proc. of IEEE Int. Conf. on Computer Vision*, Nice, France, October 2003.
- [89] Omar Javed, Khurram Shafique, Zeeshan Rasheed, and Mubarak Shah. Modeling inter-camera space-time and appearance relationships for tracking across non-overlapping views. *Computer Vision and Image Understanding*, 109(2):146–162, 2008.
- [90] Omar Javed, Khurram Shafique, and Mubarak Shah. Appearance modeling for tracking in multiple non-overlapping cameras. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, San Diego, USA, June 2005.
- [91] Omar Javed and Mubarak Shah. *Automated Multi-Camera Surveillance: Algorithms and Practice*. Springer, UK, 2008.
- [92] Kideog Jeong and Christopher Jaynes. Object matching in disjoint cameras using a colour transfer approach. *Springer Journal of Machine Vision and Applications*, 19(5):88–96, 2008.
- [93] Kai Jungling and Michael Arens. Local feature based person reidentification in infrared image sequences. In *Proc. of IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*, Boston, USA, August 2010.
- [94] Kai Jungling and Michael Arens. View-invariant person re-identification with an implicit shape model. In *Proc. of IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*, Klagenfurt, Austria, August 2011.
- [95] Vera Kettner and Ramin Zabih. Bayesian multi-camera surveillance. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, Ft. Collins, USA, June 1999.
- [96] Sohaib Khan and Mubarak Shah. Consistent labeling of tracked objects in multiple cameras with overlapping fields of view. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10):1355–1360, 2003.



- [97] Ron Kohavi and George John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1):273–324, 1997.
- [98] Martin Kstinger, Martin Hirzer, Paul Wohlhart, Peter Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, Providence, Rhode Island, June 2012.
- [99] Harold William Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1):83–97, 1955.
- [100] Cheng-Hao Kuo, Chang Huang, and Ram Nevatia. Inter-camera association of multi-target tracks by on-line learned appearance affinity models. In *Proc. of Europ. Conf. on Computer Vision*, Crete, Greece, September 2010.
- [101] Igor Kviatkovsky, Amit Adam, and Ehud Rivlin. Color invariants for person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1622–1634, 2013.
- [102] Leslie Lamport. Time, clocks, and the ordering of events in a distributed system. *Communications ACM*, 21(7):558–565, 1978.
- [103] Ryan Layne, Timothy Hospedales, and Shaogang Gong. Towards person identification and re-identification with attributes. In *Proc. of Europ. Conf. on Computer Vision*, Florence, Italy, October 2012.
- [104] Ryan Layne, Timothy Hospedales, and Shaogang Gong. Domain transfer for person re-identification. In *Proc. of ACM/IEEE International Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Stream*, Barcelona, Spain, October 2013.
- [105] Min Li, Zhaoxiang Zhang, Kaiqi Huang, and Tieniu Tan. Rapid and robust human detection and tracking based on omega-shape features. In *Proc. of IEEE Int. Conf. on Image Processing*, Cairo, Egypt, November 2009.
- [106] Wei Li and Xiaogang Wang. Locally aligned feature transforms across views. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, Portland, USA, June 2013.
- [107] Guoyun Lian, Jian Huang Lai, and Yang Gao. People consistent labeling between uncalibrated cameras without planar ground assumption. In *Proc. of IEEE Int. Conf. on Image Processing*, Hong Kong, China, September 2010.

- [108] Guoyun Lian, Jian-Huang Lai, Ching Suen, and Pei Chen. Matching of tracked pedestrians across disjoint camera views using ci-dlbp. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(7):1087–1099, 2012.
- [109] Jianning Liang, Su Yang, and Adam Winstanley. Invariant optimal feature selection: A distance discriminant and feature ranking based solution. *Pattern Recognition*, 41(5):1429–1439, 2008.
- [110] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, Boston, USA, June 2015.
- [111] Chunxiao Liu, Shaogang Gong, and Chen Change Loy. On-the-fly feature importance mining for person re-identification. *Pattern Recognition*, 47(4):1602–1615, 2014.
- [112] Chunxiao Liu, Shaogang Gong, Chen Change Loy, and Xinggang Lin. Person re-identification: What features are important? In *Proc. of Europ. Conf. on Computer Vision, International Workshop on Re-Identification*, Firenze, Italy, October 2012.
- [113] Stuart Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.
- [114] Chen Change Loy, Tao Xiang, and Shaogang Gong. Multi-camera activity correlation analysis. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, Miami, USA, June 2009.
- [115] Lianyang Ma, Xiaokang Yang, and Dacheng Tao. Person re-identification over camera networks using multi-task distance metric learning. *IEEE Transactions on Image Processing*, 23(8):3656–3670, 2014.
- [116] Christopher Madden, Eric Dahai Cheng, and Massimo Piccardi. Tracking people across disjoint camera views by an illumination tolerant appearance representation. *Springer Journal of Machine Vision and Applications*, 18(3):233–247, 2007.
- [117] Francisco Madrigal and Jean-Bernard Hayet. Multiple view, multiple target tracking with principal axis-based data association. In *Proc. of IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*, Klagenfurt, Austria, August 2011.
- [118] Emilio Maggio and Andrea Cavallaro. *Video Tracking: Theory and Practice*. Wiley and Sons, USA, 2010.

- [119] Dimitrios Makris and Tim Ellis. Automatic learning of an activity-based semantic scene model. In *Proc. of IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*, Miami, USA, July 2003.
- [120] Dimitrios Makris, Tim Ellis, and James Black. Bridging the gaps between cameras. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, Washington, USA, June 2004.
- [121] Niki Martinel and Christian Micheloni. Re-identify people in wide area camera network. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition Workshop*, Providence, USA, June 2012.
- [122] Niki Martinel, Christian Micheloni, and Claudio Piciarelli. Distributed signature fusion for person re-identification. In *Proc. of ACM/IEEE Int. Conf. on Distributed Smart Cameras*, Hong Kong, China, Oct 2012.
- [123] Riccardo Mazzon and Andrea Cavallaro. Multi-camera tracking using a multi-goal social force model. *Neurocomputing*, 100(0):41–50, 2013.
- [124] David Mills. Internet time synchronization: the network time protocol. *IEEE Transactions on Communications*, 39:1482–1493, 1991.
- [125] Sushmita Mitra, Partha Pratim Kundu, and Witold Pedrycz. Feature selection using structural similarity. *Information Science*, 198(0):48–61, 2012.
- [126] Adam Nilski. Evaluating multiple camera tracking systems - the i-lids 5th scenario. In *Proc. of IEEE Int. Carnahan Conf. on Security Technology*, Prague, Czech Republic, October 2008.
- [127] Flavio Padua, Rodrigo Carceroni, Geraldo Santos, and Kiriakos Kutulakos. Linear sequence-to-sequence alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(2):304–320, 2010.
- [128] Feng Pan, Tim Converse, David Ahn, Franco Salvetti, and Gianluca Donato. Feature selection for ranking using boosted trees. In *Proc. of ACM Int. Conf. on Information and Knowledge Management*, New York, USA, November 2009.
- [129] Unsang Park, Anil Jain, Itaru Kitahara, Kiyoshi Kogure, and Norihiro Hagita. Vise: Visual search engine using multiple networked cameras. In *Proc. of IEEE Int. Conf. on Pattern Recognition*, Hong Kong, China 2006.

- [130] Sateesh Pedagadi, James Orwell, Sergio Velastin, and Boghos Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, Columbus, Ohio, June 2013.
- [131] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.
- [132] Fatih Porikli. Inter-camera color calibration using cross-correlation model function. In *Proc. of IEEE Int. Conf. on Image Processing*, Barcelona, Spain, September 2003.
- [133] Saurabh Prasad and Lori Mann Bruce. Limitations of principal components analysis for hyperspectral target recognition. *IEEE Geoscience and Remote Sensing Letters*, 5(4):625–629, 2008.
- [134] Bryan Prosser, Shaogang Gong, and Tao Xiang. Multi-camera Matching under Illumination Change Over Time. In *Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications*, Marseille, France, October 2008.
- [135] Bryan Prosser, Shaogang Gong, and Tao Xiang. Multi-camera matching using bi-directional cumulative brightness transfer functions. In *Proc. of British Machine Vision Conference*, Leeds, UK, September 2008.
- [136] Bryan Prosser, Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Person re-identification by support vector ranking. In *Proc. of British Machine Vision Conference*, Aberystwyth, UK, August 2010.
- [137] Ali Rahimi, Brian Dunagan, and Trevor Darrell. Simultaneous calibration and tracking with a network of non-overlapping sensors. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, Washington DC, USA, July 2004.
- [138] Bernhard Rinner and Wayne Wolf. A bright future for distributed smart cameras. *Proceedings of the IEEE*, 96(10):1562–1564, 2008.
- [139] Bernhard Rinner and Wayne Wolf. An introduction to distributed smart cameras. *Proceedings of the IEEE*, 96(10):1565–1575, 2008.
- [140] Marko Robnik-Sikonja and Igor Kononenko. Theoretical and empirical analysis of relieff and rrelieff. *Machine Learning*, 53(2):23–69, 2003.

- [141] Fred Rothganger, Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. 3d object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints. *International Journal of Computer Vision*, 66(3):231–259, 2006.
- [142] David Mark Russell and Shaogang Gong. Segmenting highly textured nonstationary background. In *Proc. of British Machine Vision Conference*, Warwick, UK, September 2007.
- [143] Mohammad Ali Saghafi, Aini Hussain, Halimah Badioze Zaman, and Mohamad Hanif Saad. Review of person re-identification techniques. *IET Computer Vision*, 8(6):455–474, 2014.
- [144] David Salomon. *Data Compression: The Complete Reference*. Springer, UK, 2007.
- [145] Riccardo Satta, Giorgio Fumera, and Fabio Roli. Fast person re-identification based on dissimilarity representations. *Pattern Recognition Letters*, 33(14):1838–1848, 2012.
- [146] William Robson Schwartz and Larry Davis. Learning discriminative appearance-based models using partial least squares. In *Proc. of IEEE Int. Brazilian Symp. on Computer Graphics and Image Processing*, Rio De Janiero, Brasil, October 2009.
- [147] Gideon Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.
- [148] Jianbo Shi and Carlo Tomasi. Performance evaluation of tracking and surveillance. In *PETS2009*, <http://www.cvg.cs.rdg.ac.uk/slides/pets.html>, 2000-09.
- [149] Alan Smeaton, Paul Over, and Wessel Kraaij. Evaluation campaigns and trecvid. In *Proc. of ACM International Workshop on Multimedia Information Retrieval*, Santa Barbara, USA, October 2006.
- [150] Bi Song, Ting-Yueh Jeng, Elliot Staudt, and Amit Roy-Chowdhury. A stochastic graph evolution framework for robust multi-target tracking. In *Proc. of Europ. Conf. on Computer Vision*, Heraklion, Greece, September 2010.
- [151] Fengxi Song, Zhongwei Guo, and Dayong Mei. Feature selection using principal component analysis. In *Proc. of IEEE Int. Conf. on System Science, Engineering Design and Manufacturing Informatization*, Yichang, China, Nov 2010.
- [152] Stanislava Soro and Wendi Heinzelman. A survey of visual sensor networks. *Advances in Multimedia*, 9(9):1–21, 2009.

- [153] Masashi Sugiyama. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *Journal of Machine Learning Research*, 8:1027–1061, 2007.
- [154] Vildana Sulic, Janez Pers, Matej Kristan, and Stanislav Kovaci. Efficient feature distribution for object matching in visual-sensor networks. *IEEE Trans. on Circuits and Systems for Video Technology*, 21(7):903–916, 2011.
- [155] Xin Sun, Yanheng Liu, Jin Li, Jianqi Zhu, Huiling Chen, and Xuejie Liu. Feature evaluation and selection with cooperative game theory. *Pattern Recognition*, 45(8):2992–3002, 2012.
- [156] Murtaza Taj and Andrea Cavallaro. Distributed and decentralized multicamera tracking. *IEEE Signal Processing Magazine*, 28(3):46–58, 2011.
- [157] Dapeng Tao, Lianwen Jin, Yongfei Wang, Yuan Yuan, and Xuelong Li. Person re-identification by regularized smoothing kiss metric learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(10):1675–1685, 2013.
- [158] Luis Teixeira and Luis Corte-Real. Video object matching across multiple independent views using local descriptors and adaptive learning. *Pattern Recognition Letters*, 320(2):157–167, 2009.
- [159] Kinh Tieu, Gerald Dalley, and Eric Grimson Grimson. Inference of non-overlapping camera network topology by measuring statistical dependence. In *Proc. of IEEE Int. Conf. on Computer Vision*, Beijing, China, October 2005.
- [160] Oncel Tuzel, Fatih Porikli, and Peter Meer. Region covariance: A fast descriptor for detection and classification. In *Proc. of Europ. Conf. on Computer Vision*, Graz, Austria, 2006.
- [161] Roberto Vezzani, Davide Baltieri, and Rita Cucchiara. Path nodes integration of standalone particle filters for people tracking on distributed surveillance systems. In *Proc. of IEEE Int. Conf. on Image Analysis and Processing*, Vietri sul Mare, Italy, September 2009.
- [162] Roberto Vezzani, Davide Baltieri, and Rita Cucchiara. People re-identification in surveillance and forensics: a survey. *ACM Computing Surveys*, 46(2):1–37, 2013.
- [163] Lei Wang. Feature selection with kernel class separability. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(9):1534–1546, 2008.

- [164] Simi Wang, Michal Lewandowski, James Annesley, and James Orwel. Re-identification of pedestrians with variable occlusion and scale. In *Proc. of IEEE Int. Conf. on Computer Vision workshops*, Barcelona, Spain, November 2011.
- [165] Xiaogang Wang, Gianfranco Doretto, Thomas Sebastian, Jens Rittscher, and Peter Tu. Shape and appearance context modeling. In *Proc. of IEEE Int. Conf. on Computer Vision*, Rio de Janeiro, Brasil, October 2007.
- [166] Yimin Wang, Ruimin Hu, Chao Liang, Chunjie Zhang, and Qingming Leng. Camera compensation using a feature projection matrix for person reidentification. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(8):1350–1361, 2014.
- [167] Michael Weber, Martin Bauml, and Rainer Stiefelbogen. Part-based clothing segmentation for person retrieval. In *Proc. of IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*, Klagenfurt, Austria, August 2011.
- [168] Hua-Liang Wei and Stephen Billings. Feature subset selection and ranking for data dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):162–166, 2007.
- [169] Kilian Weinberger and Lawrence Saul. Distance metric learning for large margin nearest neighbour classification. *Journal of Machine Learning Research*, 10(2):207–244, 2009.
- [170] Terry Welch. A technique for high-performance data compression. *IEEE Computer*, 17(6):8–19, 1984.
- [171] Brett Williams, Ted Brown, and Andrys Onsman. Exploratory factor analysis: A five-step guide for novices. *Australasian Journal of Paramedicine*, 8(3):1–13, 2010.
- [172] Huazhong Xu, Pei Lv, and Lei Meng. A people counting system based on head-shoulder detection and tracking in surveillance video. In *Proc. of Int. Conf. on Computer Design and Applications*, Qinhuangdao, China, June 2010.
- [173] Rui Xu and II Wunsch, D. Survey of clustering algorithms. *Neural Networks, IEEE Transactions on*, 16(3):645–678, 2005.
- [174] Neil Yager and Ted Dunstone. The biometric menagerie. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(2):220–230, 2010.
- [175] Tingxu Yan and Yuexian Hou. An unsupervised feature selection method based on de-

- gree of feature cooperation. In *Proc. of IEEE Int. Conf. on Fuzzy Systems and Knowledge Discovery*, Shanghai, China, July 2011.
- [176] Ming Yang, Fengjun Lv, Wei Xu, and Yihong Gong. Detection driven adaptive multi-cue integration for multiple human tracking. In *Proc. of IEEE Int. Conf. on Computer Vision*, Kyoto, Japan, September 2009.
- [177] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Associating groups of people. In *Proc. of British Machine Vision Conference*, London, United Kingdom, September 2009.
- [178] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Person re-identification by probabilistic relative distance comparison. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, Colorado Springs, USA, June 2011.
- [179] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Re-identification by relative distance comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):653–668, 2013.