Advances in Computer Vision and Pattern Recognition

Shaogang Gong
Marco Cristani
Shuicheng Yan
Chen Change Loy *Editors*

Person Re-Identification

Springer

Advances in Computer Vision and Pattern Recognition

For further volumes:
http://www.springer.com/series/4205

Shaogang Gong · Marco Cristani
Shuicheng Yan · Chen Change Loy
Editors

Person Re-Identification

Springer

Editors
Shaogang Gong
Queen Mary University
London
UK

Marco Cristani
University of Verona
Verona
Italy

Shuicheng Yan
National University of Singapore
Singapore

Chen Change Loy
The Chinese University of Hong Kong
Shatin
Hong Kong SAR

Series editors
Sameer Singh
Rail Vision Europe Ltd.
Castle Donington
Leicestershire, UK

Sing Bing Kang
Interactive Visual Media Group
Microsoft Research
Redmond, WA, USA

ISSN 2191-6586 ISSN 2191-6594 (electronic)
Advances in Computer Vision and Pattern Recognition
ISBN 978-1-4471-6295-7 ISBN 978-1-4471-6296-4 (eBook)
DOI 10.1007/978-1-4471-6296-4
Springer London Heidelberg New York Dordrecht

Library of Congress Control Number: 2013957125

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Person re-identification is the problem of recognising and associating a person at different physical locations over time after the person had been previously observed visually elsewhere. Solving the re-identification problem has gained a rapid increase in attention in both academic research communities and industrial laboratories in recent years. The problem has many manifestations from different application domains. For instance, the problem is known as “re-acquisition” when the aim is to associate a target (person) when it is temporarily occluded during the tracking in a single camera view. On the other hand, in domotics applications or personalised healthcare environments, the primary aim is to retain the identity of a person whilst one is moving about in a private home of distributed spaces, e.g. crossing multiple rooms. Re-identification can provide a useful tool for validating the identity of impaired or elderly people in a seamless way without the need for more invasive biometric verification procedures, e.g. controlled face or fingerprint recognition. Moreover, in a human–robot interaction scenario, solving the re-identification problem can be considered as “non-cooperative target recognition”, where the identity of the interlocutor is maintained, allowing the robot to be continuously aware of the surrounding people. In larger distributed spaces such as airport terminals and shopping malls, re-identification is mostly considered as the task of “object association” in a distributed multi-camera network, where the goal is to keep track of an individual across different cameras with non-overlapping field of views. For instance, in a multi-camera surveillance system, re-identification is needed to trace the inter-camera whereabouts of individuals of interest (a watch-list), or simply to understand how people move in complex environments such as an airport and a train station for better crowd traffic management and crowding control. In a retail environment, re-identification can provide useful information for improving customer service and shopping space management. In a more general setting of online shopping, re-identification of visual objects of different categories, e.g. clothing, can help in tagging automatically huge volumes of visual samples of consumer goods in Internet image indexing, search and retrieval.

Solving the person re-identification problem poses a considerable challenge that requires visually detecting and recognising a person (subject) at different space time locations observed under substantially different, and often unknown, viewing conditions without subject collaboration. Early published work on re-identification can date back a decade ago to 2003, but most contemporary techniques have been developed since 2008, and in particular in the last 2–3 years. In the past 5 years, there has been a tremendous increase in computer vision research on solving the re-identification problem, evident from a large number of academic papers published in all the major conferences (ICCV, CVPR, ECCV, BMVC, ICIP) and journals (TPAMI, IJCV, Pattern Recognition). This trend will increase further in the coming years, given that many open problems remain unsolved.

Inspired by the First International Workshop on Re-Identification held at Florence in Italy in October 2012, this book is a collection of invited chapters from some of the world′s leading researchers working on solving the re-identification problem. It aims to provide a comprehensive and in-depth presentation of recent progress and the current state-of-the-art approaches to solving some of the fundamental challenges in person re-identification, benefiting from wider research in the computer vision, pattern recognition and machine learning communities, and drawing insights from video analytics system design considerations for engineering practical solutions. Due to its diverse nature, the development of person re-identification methods by visual matching has been reported in a wide range of fields, from multimedia to robotics, from domotics to visual surveillance, but all with an underlying computer vision theme. Re-identification exploits extensively many core computer vision techniques that aim at extracting and representing an individual's visual appearance in a scene, e.g. pedestrian detection and tracking, and object representation; and machine learning techniques for discriminative matching, e.g. distance metric learning and transfer learning. Moreover, solving the person re-identification problem can benefit from exploiting heterogeneous information by learning more effective semantic attributes, exploiting spatio-temporal statistics, estimating feature transformation across different cameras, taking into account soft-biometric cues (e.g. height, gender) and considering contextual cues (e.g. baggage, other people nearby).

This book is the first dedicated treatment on the subject of *Person Re-Identification* that aims to address a highly focused problem with a strong multidisciplinary appeal to practitioners in both fundamental research and practical applications. In the context of video content analysis, visual surveillance and human recognition, there are a number of other books published recently that aim to address a wider range of topics, e.g. *Video Analytics for Business Intelligence,* by Caifeng Shan, Fatih Porikli, Tao Xiang and Shaogang Gong (2012); *Visual Analysis of Behaviour: From Pixels to Semantics,* by Shaogang Gong and Tao Xiang (2011); and *Visual Analysis of Humans: Looking at People*, by Thomas Moeslund, Adrian Hilton, Volker Kruger and Leonid Sigal (2011). In contrast to those other books, this book provides a more in-depth analysis and a more comprehensive presentation of techniques required specifically for solving the problem of person re-identification. Despite aiming to address a highly focused problem, the techniques presented in this book, e.g. feature representation, attribute learning, ranking, active learning and transfer learning, are highly applicable to other more general problems in computer vision, pattern recognition and machine learning. Therefore, the book should also be of considerable interest to a wider audience.

We anticipate that this book will be of special interest to academics, postgraduates and industrial researchers specialised in computer vision and machine learning, database (including internet) image retrieval, big data mining and search engines. It should also be of interest to commercial developers and managers keen to exploit this emerging technology for a host of applications including security and surveillance, personalised healthcare, commercial information profiling, business intelligence gathering, smart city, public space infrastructure management, consumer electronics and retails. Finally, this book will also be of use to postgraduate students of computer science, engineering, applied mathematics and statistics, cognitive and social studies.

London, October 2013	Shaogang Gong
Verona	Marco Cristani
Singapore	Shuicheng Yan
Hong Kong	Chen Change Loy

Acknowledgments

The preparation of this book has required the dedication of many people. First of all, we thank all the contributing authors for their extraordinary effort and dedication in preparing the book chapters within a very tight time frame. Second, we express our gratitude to all the reviewers. Their critical and constructive feedback helped in improving the quality of the book. Finally, we thank Simon Rees and Wayne Wheeler at Springer for their support throughout the preparation of this book. The book was typeset using LaTeX.

This book was inspired by the First International Workshop on Re-Identification (Re-Id 2012), in conjunction with the European Conference on Computer Vision, held at Florence in Italy in October 2012. To that end, we thank the workshop programme committee and the authors who made the workshop a huge success. We also thank the workshop industrial sponsors Bosch, KAI Square, Vision Semantics and Embedded Vision Systems who sponsored the Best Paper Award prize and made the workshop a more rewarding experience.

Contents

Part I Features and Representations

Part II Matching and Distance Metric

Part III Evaluation and Application

Contributors

Tamar Avraham Technion Israel Institute of Technology, Haifa, Israel, e-mail: tammya@cs.technion.ac.il

Andrew D. Bagdanov University of Florence, Florence, Italy, e-mail: bagdanov@dsi.unifi.it

Sławomir Bąk INRIA, Sophia Antipolis, France, e-mail: slawomir.bak@inria.fr

Alberto Basso University of Padua, Padua, Italy, e-mail: bassoal1@dei.unipd.it

Loris Bazzani Istituto Italiano di Tecnologia, Genova, Italy, e-mail: loris.bazzani@iit.it

Csaba Beleznai Austrian Institute of Technology, Vienna, Austria, e-mail: Csaba.Beleznai@ait.ac.at

Alberto Del Bimbo University of Florence, Florence, Italy, e-mail: delbimbo@dsi.unifi.it

Horst Bischof Graz University of Technology, Graz, Austria, e-mail: bischof@icg.tugraz.at

François Brémond INRIA, Sophia Antipolis, France, e-mail: francois.bremond@inria.fr

Rama Chellappa University of Maryland, College Park, USA, e-mail: rama@umiacs.umd.edu

Dong Seon Cheng Hankuk University of Foreign Studies, Seoul, Korea, e-mail: cheng_ds@hufs.ac.kr

Marco Cristani University of Verona, Verona, Italy, e-mail: marco.cristani@univr.it

Rita Cucchiara University of Modena and Reggio Emilia, Modena, Italy, e-mail: rita.cucchiara@unimore.it

Henry Detmold Snap Network Surveillance, Adelaide, Australia, e-mail: henry@snapcontrol.com

Anthony Dick University of Adelaide, Adelaide, Australia, e-mail: anthony.dick@adelaide.edu.au

François Fleuret IDIAP, Martigny, Switzerland, e-mail: francois.fleuret@idiap.ch

Andrea Fossati ETH Zurich, Zurich, Switzerland, e-mail: fossati@vision.ee.ethz.ch

Pascal Fua EPFL, Lausanne, Switzerland, e-mail: pascal.fua@epfl.ch

Giorgio Fumera University of Cagliari, Cagliari, Italy, e-mail: fumera@diee.unica.it

Shaogang Gong Queen Mary University of London, London, UK, e-mail: sgg@eecs.qmul.ac.uk

Martin Hirzer Graz University of Technology, Graz, Austria, e-mail: hirzer@icg.tugraz.at

Timothy M. Hospedales Queen Mary University of London, London, UK, e-mail: tmh@eecs.qmul.ac.uk

Frédéric Jurie University of Caen Basse-Normandie, Caen, France, e-mail: frederic.jurie@unicaen.fr

Svebor Karaman University of Florence, Florence, Italy, e-mail: svebor.karaman@unifi.it

Martin Köstinger Graz University of Technology, Graz, Austria, e-mail: koestinger@icg.tugraz.at

Ryan Layne Queen Mary University of London, London, UK, e-mail: rlayne@eecs.qmul.ac.uk

Annan Li National University of Singapore, Singapore, Singapore, e-mail: elela@nus.edu.sg

Xinggang Lin Tsinghua University, Beijing, China, e-mail: xglin@mail.tsinghua.edu.cn

Michael Lindenbaum Technion Israel Institute of Technology, Haifa, Israel, e-mail: mic@cs.technion.ac.il

Giuseppe Lisanti University of Florence, Florence, Italy, e-mail: lisanti@dsi.unifi.it

Chunxiao Liu Tsinghua University, Beijing, China, e-mail: lcx08@mails.tsinghua.edu.cn

Luoqi Liu National University of Singapore, Singapore, Singapore, e-mail: liuluoqi@nus.edu.sg

Chen Change Loy The Chinese University of Hong Kong, Shatin, Hong Kong, e-mail: ccloy@ie.cuhk.edu.hk

Bingpeng Ma University of Chinese Academy of Sciences, Beijing, China, e-mail: bpma@ucas.ac.cn

Emanuele Menegatti University of Padua, Padua, Italy, e-mail: emg@dei.unipd.it

Matteo Munaro University of Padua, Padua, Italy, e-mail: munaro@dei.unipd.it

Vittorio Murino Istituto Italiano di Tecnologia, Genova, Italy, e-mail: vittorio.murino@iit.it

Jie Ni University of Maryland, College Park, USA, e-mail: jni@umiacs.umd.edu

Federico Pala University of Cagliari, Cagliari, Italy, e-mail: federico.pala@diee.unica.it

Qiang Qiu Duke University, Durham, USA, e-mail: qiang.qiu@duke.edu

Yogesh Raja Vision Semantics Ltd, London, UK, e-mail: yraja@visionsemantics.com

Fabio Roli University of Cagliari, Cagliari, Italy, e-mail: roli@diee.unica.it

Peter M. Roth Graz University of Technology, Graz, Austria, e-mail: pmroth@icg.tugraz.at

Riccardo Satta European Commission JRC Institute for the Protection and Security of the Citizen, Ispra, Italy, e-mail: Riccardo.Satta@jrc.ec.europa.eu

Horesh Ben Shitrit EPFL, Lausanne, Switzerland, e-mail: horesh.benshitrit@epfl.ch

Yu Su University of Caen Basse-Normandie, Caen, France, e-mail: yu.su@unicaen.fr

Luc Van Gool ETH Zurich, Zurich, Switzerland, e-mail: vangool@vision.ee.ethz.ch

Anton van den Hengel University of Adelaide, Adelaide, Australia, e-mail: anton.vandenhengel@adelaide.edu.au

Roberto Vezzani University of Modena and Reggio Emilia, Modena, Italy, e-mail: roberto.vezzani@unimore.it

Xiaogang Wang The Chinese University of Hong Kong, Shatin, Hong Kong, e-mail: xgwang@ee.cuhk.edu.hk

Tao Xiang Queen Mary University of London, London, UK, e-mail: txiang@eecs.qmul.ac.uk

Shuicheng Yan National University of Singapore, Singapore, Singapore, e-mail: eleyans@nus.edu.sg

Rui Zhao The Chinese University of Hong Kong, Shatin, Hong Kong, e-mail: rzhao@ee.cuhk.edu.hk

Wei-Shi Zheng Sun Yat-sen University, Guangzhou, China, e-mail: wszheng@ieee.org

Chapter 1
The Re-identification Challenge

Shaogang Gong, Marco Cristani, Chen Change Loy and Timothy M. Hospedales

Abstract For making sense of the vast quantity of visual data generated by the rapid expansion of large-scale distributed multi-camera systems, automated person re-identification is essential. However, it poses a significant challenge to computer vision systems. Fundamentally, person re-identification requires to solve two difficult problems of '*finding needles in haystacks*' and '*connecting the dots*' by identifying instances and associating the whereabouts of targeted people travelling across large distributed space–time locations in often crowded environments. This capability would enable the discovery of, and reasoning about, individual-specific long-term structured activities and behaviours. Whilst solving the person re-identification problem is inherently challenging, it also promises enormous potential for a wide range of practical applications, ranging from security and surveillance to retail and health care. As a result, the field has drawn growing and wide interest from academic researchers and industrial developers. This chapter introduces the re-identification problem, highlights the difficulties in building person re-identification systems, and presents an overview of recent progress and the state-of-the-art approaches to solving some of the fundamental challenges in person re-identification, benefiting from research in computer vision, pattern recognition and machine learning, and drawing insights from video analytics system design considerations for engineering practical solutions. It also provides an introduction of the contributing chapters of this book.

S. Gong (✉) · T. M. Hospedales
Queen Mary University of London, London, UK
e-mail: sgg@eecs.qmul.ac.uk

M. Cristani
University of Verona and Istituto Italiano di Tecnologia, Verona, Italy
e-mail: marco.cristani@univr.it

C. C. Loy
The Chinese University of Hong Kong, Shatin, Hong Kong
e-mail: ccloy@ie.cuhk.edu.hk

T. M. Hospedales
e-mail: tmh@eecs.qmul.ac.uk

S. Gong et al. (eds.), *Person Re-Identification*,
Advances in Computer Vision and Pattern Recognition,
DOI: 10.1007/978-1-4471-6296-4_1, © Springer-Verlag London 2014

The chapter ends by posing some open questions for the re-identification challenge arising from emerging and future applications.

1.1 Introduction

A fundamental task for a distributed multi-camera surveillance system is to associate people across camera views at different locations and time. This is known as the person re-identification (re-id) problem, and it underpins many crucial applications such as long-term multi-camera tracking and forensic search. More specifically, re-identification of an individual or a group of people collectively is the task of visually matching a single person or a group in diverse scenes, obtained from different cameras distributed over non-overlapping scenes (physical locations) of potentially substantial distances and time differences. In particular, for surveillance applications performed over space and time, an individual disappearing from one view would need to be matched in one or more other views at different physical locations over a period of time, and be differentiated from numerous visually similar but different candidates in those views. Potentially, each view may be taken from a different angle and distance, featuring different static and dynamic backgrounds under different lighting conditions, degrees of occlusion and other view-specific variables. A re-identification computer system aims to automatically match and track individuals either retrospectively or on-the-fly when they move across different locations.

Relying on human operator manual re-identification in large camera networks is prohibitively costly and inaccurate. Operators are often assigned more cameras than they can feasibly monitor simultaneously, and even within a single camera, manual matching is vulnerable to inevitable attentional gaps [1]. Moreover, baseline human performance is determined by the individual operator's experience amongst other factors. It is difficult to transfer this expertise directly between operators, and it is difficult to obtain consistent performance due to operator bias [2]. As public space camera networks have grown quickly in recent years, it is becoming increasingly clear that manual re-identification is not scalable. There is therefore a growing interest within the computer vision community in developing automated re-identification solutions.

In a crowded and uncontrolled environment observed by cameras from an unknown distance, person re-identification relying upon conventional biometrics such as face recognition is neither feasible nor reliable due to insufficiently constrained conditions and insufficient image detail for extracting robust biometrics. Instead, visual features based on the appearance of people, determined by their clothing and objects carried or associated with them, can be exploited more reliably for re-identification. However, visual appearance is intrinsically weak for matching people. For instance, most people in public spaces wear dark clothes in winter, so most colour pixels are not informative about identity in a unique way. To further compound the problem, a person's appearance can change significantly between different camera views if large changes occur in view angle, lighting, background clutter and occlusion. This results in different people often appearing more alike than the same person across different camera views. That is, intra-class variability can be, and is often, significantly larger than inter-class variability when camera view changes are involved. Current research efforts for solving the re-identification problem have primarily focused on two aspects:

1. Developing feature representations which are discriminative for identity, yet invariant to view angle and lighting [3–5];
2. Developing machine learning methods to discriminatively optimise parameters of a re-identification model [6]; and with some studies further attempting to bridge the gap by learning an effective class of features from data [7, 8].

Nevertheless, achieving automated re-identification remains a significant challenge due to the inherent limitation that most visual features generated from people's visual appearance are either insufficiently discriminative for cross-view matching, especially with low resolution images, or insufficiently robust to viewing condition changes, and under extreme circumstances, totally unreliable if clothing is changed substantially.

Sustained research on addressing the re-identification challenge benefits other computer vision domains beyond visual surveillance. For instance, feature descriptor design in re-identification can be exploited to enhance tracking [9] and identification of people (e.g. players in sport videos) from medium to far distance; the metric learning and ranking approaches developed for re-identification can be adapted for face verification and content-based image analysis in general. Research efforts in re-identification also contribute to the development of various machine learning topics, e.g. similarity and distance metric learning, ranking and preference learning, sparsity and feature selection, and transfer learning.

This chapter is organised as follows. We introduce the typical processing steps of re-id in Sect. 1.2. In Sect. 1.3, we highlight the challenges commonly encountered in formulating a person re-identification framework. In particular, we discuss challenges related to feature construction, model design, evaluation and system implementation. In Sect. 1.4, we review the most recent developments in person re-identification, introduce the contributing chapters of this book and place them in context. Finally in Sect. 1.5, we discuss a few possible new directions and open questions to be solved in order to meet the re-identification challenge in emerging and future real-world applications.

1.2 Re-identification Pipeline

Human investigators tasked with the forensic analysis of video from multi-camera CCTV networks face many challenges, including data overload from large numbers of cameras, limited attention span leading to important events and targets being missed, a lack of contextual knowledge indicating what to look for, and limited ability or inability to utilise complementary non-visual sources of knowledge to assist the search process. Consequently, there is a distinct need for a technology to alleviate the burden placed on limited human resources and augment human capabilities.

An automated re-identification mechanism takes as input either tracks or bounding-boxes containing segmented images of individual persons, as generated by a localised tracking or detection process of a visual surveillance system. To automatically match people at different locations over time captured by different camera views, a re-identification process typically takes the following steps:

1. Extracting imagery features that are more reliable, robust and concise than raw pixel data;
2. Constructing a descriptor or representation, e.g. a histogram of features, capable of both describing and discriminating individuals; and
3. Matching specified probe images or tracks against a gallery of persons in another camera view by measuring the similarity between the images, or using some model-based matching procedure. A training stage to optimise the matching parameters may or may not be required depending on the matching strategy.

Such processing steps raise certain demands on algorithm and system design. This has led to both the development of new and the exploitation of existing computer vision techniques for addressing the problems of feature representation, model matching and inference in context.

Representation: Contemporary approaches to re-identification typically exploit low-level features such as colour [10], texture, spatial structure [5] or combinations thereof [4, 11, 12]. This is because these features can be relatively easily and reliably measured, and provide a reasonable level of inter-person discrimination together with inter-camera invariance. Such features are further encoded into fixed-length person descriptors, e.g. in the form of histograms [4], covariances [13] or fisher vectors [14].

Matching: Once a suitable representation has been obtained, nearest-neighbour [5] or model-based matching algorithms such as support-vector ranking [4] may be used for re-identification. In each case, a distance metric (e.g. Euclidean or Bhattacharyya) must be chosen to measure the similarity between two samples. Model-based matching approaches [15, 16] and nearest-neighbor distance metrics [6, 17] can both be discriminatively optimised to maximise re-identification performance given annotated training data of person images. Bridging these two stages, some studies [7, 8, 18] have also attempted to learn discriminative low-level features directly from data.

Context: Other complementary aspects of the re-identification problem have also been pursued to improve performance, such as improving robustness by combining multiple frames worth of features along a trajectory tracklet [9, 12], set-based analysis [19, 20], considering external context such as groups of persons [21], and learning the topology of camera networks [22, 23] in order to reduce the matching search space and hence reduce false-positives.

1.2.1 A Taxonomy of Methods

Different approaches (as illustrated in different chapters of this book) use slightly different taxonomies in categorising existing person re-identification methods.

In general, when only an image pair is matched, the method is considered as a *single-shot recognition* method. If matching is conducted between two sets of images, e.g. frames obtained from two separate trajectories, the method is known as a *multi-shot recognition* approach. An approach is categorised as a *supervised* method if prior to application, and it exploits labelled samples for tuning model parameters such as distance metrics, feature weight or decision boundaries. Otherwise a method is regarded as an *unsupervised* approach if it concerns the extraction of robust visual features and does not rely on training data. Blurring these boundaries somewhat are methods which do learn from training data prior to deployment, but do not rely on annotation for these data.

1.3 The Challenge

1.3.1 Feature Representation

Designing suitable feature representation for person re-identification is a critical and challenging problem. Ideally, the features extracted should be robust to changes in illumination, viewpoint, background clutter, occlusion and image quality/resolution. In the context of re-id, however, it is unclear whether there exists universally important and salient features that can be applied readily to different camera views and for all individuals. The discriminative power, reliability and computability of features are largely governed by the camera-pair viewing conditions and unique appearance characteristics of different persons captured in the given views. Moreover, the difficulty in obtaining an aligned bounding box, and accurately segmenting a person from cluttered background makes extracting pure and reliable features depicting the person of interest even harder.

1.3.2 Model and System Design

There are a variety of challenges that arise during model and system design:

1. *Inter- and Intra-class variations*: A fundamental challenge in constructing a re-id model is to overcome the inter-class confusion, i.e. different persons can look alike across camera views; and intra-class variation, i.e. the same individual may look different when observed under different camera views. Such variations between camera view pairs are in general complex and multi-modal, and therefore are necessarily non-trivial for a model to learn.
2. *Small sample size*: In general a re-id module may be required to match single probe images to single gallery images. This means from a conventional classification perspective, there is likely to be insufficient data to learn a good model of each person's intra-class variability. 'One-shot' learning may be required under

which only a single pair of examples is available for model learning. For this reason, many frameworks treat re-id as a pairwise binary classification (same vs. different) problem [4, 16] instead of a conventional multi-class classification problem.

3. *Data labelling requirement*: For exploiting a supervised learning strategy to train a good model robust to cross-camera view variations, persons from each view annotated with identity or binary labels depicting same versus different are required. Consequently, models which can be learned with less training data are preferred since for a large camera network, collecting extensive labelled data from every camera would be prohibitively expensive.
4. *Generalisation capability*: This is the flip side of training data scalability. Once trained for a specific pair of cameras, most models do not generalise well to another pair of cameras with different viewing conditions [24]. In general, one seeks for a model with good generalisation ability that can be trained once and then applied to a variety of different camera configurations from different locations. This would sidestep the issue of training data scalability.
5. *Scalability*: Given a topologically complex and large camera network, the search space for person matching can be extremely large with numerous potential of candidates to be discriminated. Thus test-time (probe-time) scalability is crucial, as well as real-time low latency implementation for processing numerous input video streams, and returning query results promptly for on-the-fly real-time response.
6. *Long-term re-identification*: The longer the time and space separation between views is, the greater the chance will be that people may appear with some changes of clothes or carried objects in different camera views. Ideally a re-identification system should have some robustness to such changes.

1.3.3 Data and Evaluation

Many standard benchmark datasets reflect a ‘closed-world’ scenario, e.g. exactly two camera views with exactly one instance of each person per camera and 1:1 exact identity correspondence between the cameras. This is in contrast to a more realistic ‘open-world’ scenario, where persons in each camera may be only partially overlapping and the number of cameras, spatial size of the environment and number of people may be unknown and at a significantly larger scale. Thus the search space is of unknown size and contains a potentially unlimited number of candidate matches for a target. Re-identification of targets in such open environments can potentially scale to arbitrary levels, covering huge spatial areas spanning not just different buildings but different cities, countries or even continents, leading to an overwhelming quantity of ‘big data’.

There are a variety of metrics that are useful for quantifying the effectiveness of a re-identification system. The two most common metrics are ‘Rank-1 accuracy’, and the ‘CMC curve’. Rank-1 accuracy refers to the conventional notion of classification accuracy: the percentage of probe images which are perfectly matched to their corresponding gallery image. High Rank-1 accuracy is notoriously hard to obtain on challenging re-id problems. More realistically a model is expected to report a ranked list of matches which the operator can inspect manually to confirm the true match. The question is how high true matches typically appear on the ranked list. The CMC (Cumulative Match Characteristic) curve summarises this: the chance of the true match appearing in the top 1, 2, . . . , N of the ranked list (the first point on the CMC curve being Rank-1 accuracy). Other metrics which can be derived from the CMC curve include the scalar area under the curve, and expected rank (on average how far down the list is the true match). Which of these two metrics is the most relevant arguably depends on the specific application scenario: Whether a (probably low in absolute terms) chance of perfect match or a good average ranking is preferred. This dichotomy raises the further interesting question of which evaluation criterion is the relevant one to optimise when designing discriminatively trained re-identification models.

1.4 Perspectives and Progress

1.4.1 On Feature Representation

Seeking Robust Features

A large number of feature types have been proposed for re-identification, e.g. colour, textures, edges, shape, global features, regional features, and patch-based features. In order to cope with sparsity of data and the challenging view conditions, most person re-identification methods benefit from integrating several types of features with complementary nature [4–6, 9, 11, 12, 25–29]. Often, each type of visual feature is represented by a bag-of-words scheme in the form of a histogram. Feature histograms are then concatenated with some weighting between different types of features in accordance to their perceived importance, i.e. based on some empirical or assumed discriminative power of certain type of features in distinguishing visual appearance of individuals. Spatial information about the layout of these features is also an important cue. However, there is a tradeoff between more granular spatial decomposition providing a more detailed cue and increasing risk of mis-alignment between regions in image pairs, and thus brittleness of the match. To integrate spatial information into the feature representation, images are typically partitioned into different segments or regions, from which features are extracted. Existing partitioning schemes include horizontal stripes [4, 6, 18, 29], triangulated graphs [30], concentric rings [21], and localised patches [8, 13]. Chapters 2, 3 and 4 introduce some examples of robust feature representations for re-identification, such as fisher vectors and covariance descriptors. Chapters 5 and 6 take a different view of learning mid-level semantic attribute features reflecting a low-dimensional human-interpretable description of each person's appearance. Chapter 17 provides a detailed analysis and comparison of the different feature types used in re-identification.

Exploiting Shape and Structural Constraints

Re-identification requires first detecting a person prior to feature extraction. Performance of existing pedestrian detection techniques is still far from accurate for the re-identification purpose. Without a tight detection bounding box, the features extracted are likely to be affected by background clutter. Many approaches start with attempting to segment the pixels of a person in the bounding box (foreground) from background also included in the bounding box. This increases the purity of extracted features by eliminating contamination by background information.

If different body parts can be detected with pose estimation (parts configuration rather than 3D orientation) and human parsing systems, the symmetry and shape of a person can be exploited to extract more robust and relevant imagery features from different body parts. In particular, natural objects reveal symmetry in some form and background clutter rarely exhibits a coherent and symmetric pattern. One can exploit these symmetric and asymmetric principles to segregate meaningful body parts as the foreground, while discard distracting background clutter. Chapter 3 presents a robust symmetry-based descriptor for modelling the human appearance, which localises perceptually relevant body parts driven by asymmetry and/or symmetry principles. Specifically, the descriptor imposes higher weights to features located near to the vertical symmetry axis than those that are far from it. This permits higher preference to internal body foreground, rather than peripheral background portions in the image. The descriptor, when enriched with chromatic and texture information, shows exceptional robustness to low resolution, pose, viewpoint and illumination variations.

Another way of reducing the influence of background clutter is by decomposing a full pedestrian image into articulated body parts, e.g. head, torso, arms and legs. In this way, one wishes to focus selectively on similarities between the appearance of body parts whilst filtering out as much of the background pixels in proximity to the foreground as possible. Naturally, a part-based re-identification representation exhibits better robustness to partial (self) occlusion and changes in local appearances. Chapters 6 and 7 describe methods for representing the pedestrian body parts as 'Pictorial Structures'. Chapter 7 further demonstrates an approach to obtaining robust signatures from the segmented parts not only for 'single-shot' but also 'multi-shot' recognition.

Beyond 2D Appearance Features

Re-identification methods based on entirely 2D visual appearance-based features would fail when individuals change their clothing completely. To address this problem, one can attempt to measure soft biometric cues that are less sensitive to clothing appearance, such as the height of a person, the length of his arms and legs and the ratios between different body parts. However, soft biometrics are exceptionally difficult to measure reliably in typical impoverished surveillance video at ‘stand off’ distances and unconstrained viewing angles. Chapter 8 describes an approach to recover skeleton lengths and global body shape from calibrated 3D depth images obtained from depth-sensing cameras. It shows that using such non-2D appearance features as a form of soft biometrics promises more robust re-identification for long-term video surveillance.

Exploiting Local Contextual Constraints

In crowded public spaces such as transport hubs, achieving accurate pedestrian detection is hard, let alone extracting robust features for re-identification purpose. The problem is further compounded by the fact that many people are wearing clothing with similar colour and style, increasing the ambiguity and uncertainty in the matching process. Where possible, one aims to seek more holistic contextual constraints in addition to localised visual appearance of isolated (segmented) individuals.

In public scenes people often walk in groups, either with people they know or strangers. The availability of more and richer visual content in a group of people over space and time could provide vital contextual constraints for more accurate matching of individuals within the group. Chapter 9 goes beyond the conventional individual person re-identification by casting the re-identification problem in the context of associating groups of people in proximity over different camera views [21]. It aims to address the problem of associating groups of people over large space and time gaps. Solving the group association problem is challenging in that a group of people can be highly non-rigid with changing relative position of people within the group, as well as individuals being subject to severe self-occlusions.

Not All Are Equal: Salient Feature Selection

Two questions arise: (1) Are all features equal? (2) Does the usefulness of a feature (type) universally hold? Unfortunately, not all features are equally important or useful for re-identification. Some features are more discriminative for identity, whilst others more tolerant or invariant to camera view changes. It is important to determine both the circumstances and the extent of the usefulness of each feature. This is considered as the problem of feature weighting or feature selection. Existing re-identification techniques [4, 6, 11, 31] mostly assume implicitly a feature weighting or selection mechanism that is *global*, i.e. a set of generic weights on feature types invariant to a population. That is, to assume a single weight vector or distance metric (e.g. mahalanobis distance metric) that is globally optimal for all people. For instance, one often assumes colour is the most important (intuitively so) and universally a good feature for matching all individuals. Besides heuristic or empirical tuning, such weightings can be learned through boosting [11], ranking [4], or distance metric learning [6] (see Sect. Learning Distance Metric).

Humans often rely on *salient* features for distinguishing one from the other. Such feature saliency is valuable for person re-identification but is often too subtle to be captured when computing generic feature weights using existing techniques. Chapter 10 considers an alternative perspective that different appearance features are more important or salient than others in describing each particular individual and distinguishing him/her from other people. Specifically, it provides empirical evidence to demonstrate that some re-identification advantages can be gained from unsupervised feature importance mining guided by a person's appearance attribute classification. Chapter 17 considers a similar concept in designing a patch-based re-identification system, which aims to discover salient patches of each individual in an unsupervised manner in order to achieve more robust re-identification [8].

Exploiting Semantic Attributes

When performing person re-identification, human experts rely upon matching appearance or functional attributes that are discrete and unambiguous in interpretation, such as hairstyle, shoe-type or clothing-style [32]. This is in contrast to the continuous and more ambiguous 'bottom-up' imagery features used by contemporary computer vision based re-identification approaches, such as colour and texture [3–5]. This 'semantic attribute' centric representation is similar to a description provided verbally to a human operator, e.g. by an eyewitness.

Attribute representations may start with the same low-level feature representation that conventional re-identification models use. However, they use these to generate a low-dimensional attribute description of an individual. In contrast to standard unsupervised dimensionality reduction methods such as Principal Component Analysis (PCA), attribute learning focuses on representing persons by projecting them onto a basis set defined by axes of appearance which are semantically meaningful to humans.

Semantic attribute representations have various benefits: (1) In re-identification, a single pair of images may be available for each target. This exhibits the challenging case of 'one-shot' learning. Attributes can be more powerful than low-level features [33–35], as pre-trained attribute classifiers learn implicitly the variance in appearance of each particular attribute and invariances to the appearance of that attribute across cameras. (2) Attributes can be used synergistically in conjunction with raw data for greater effectiveness [7, 35]. (3) Attributes are a suitable representation for direct human interaction, therefore allowing searches to be specified, initialised or constrained using human-labelled attribute-profiles [33, 34, 36], i.e. enabling forensic person search. Chapter 5 defines 21 binary attributes regarding clothing-style, hairstyle, carried objects and gender to be learned with Support Vector Machines (SVMs). It evaluates the theoretical discriminative potential of the attributes, how reliably they can be detected in practice, how their weighting can be discriminatively learned and how they can be used in synergy with low-level features to re-identify accurately. Finally, it is shown that attributes are also useful for zero-shot identification, i.e. replacing the probe image with a specified attribute semantic description without visual probe. Chapter 6 embeds middle-level cloth attributes via a latent SVM framework for more robust person re-identification. The pairwise potentials in the latent SVM allow attribute correlation to be considered. Chapter 10 takes a different approach to discover a set of prototypes in an unsupervised manner. Each prototype reveals a mixture set of attributes to describe a specific population of people with similar appearance characteristics. This alleviates the labelling effort for training attribute classifiers.

1.4.2 On Model Learning

Learning Feature Transforms

If camera pair correspondences are known, one can learn a feature transfer function for modelling camera-dependent photometric or geometric transformations. In particular, a photometric function captures the changes of colour distribution of objects transiting from one camera view to another. The changes are mainly caused by different lighting and viewing conditions. Geometric transfer functions can also be learned from the correspondences of interest points. Following the work of Porikli [37], a number of studies have proposed different ways for estimating the Brightness Transfer Function (BTF) [4, 38–42]. The BTF can be learned either separately on different colour channels, or taking into account the dependencies between channels [41]. Some BTFs are defined on each individual, whilst other studies learn a cumulative function on the full available training set [4]. A detailed review of different BTF approaches can be found in Chap. 11. Most BTF approaches assume the availability of perfect foreground segments, from which robust colour features can be extracted. This assumption is often invalid in real-world scenarios. Chapter 11 relaxes this assumption through performing automatic feature selection with the aim to discard background clutter irrelevant to re-identification. It further demonstrates an approach to estimate a robust transfer function given only limited training pairs from two camera views.

In many cases the transfer functions between camera view pairs are complex and multi-modal. Specifically, the cross-views transfer functions can be different under the influence of multiple factors such as lighting, poses, camera calibration parameters and the background of a scene. Therefore, it is necessary to capture these different configurations during the learning stage. Chapter 17 provides a solution to this problem and demonstrates that the learned model is capable of generalising better to a novel view pair.

Learning Distance Metric

A popular alternative to colour transformation learning is distance metric learning. The idea of distance metric learning is to search for the optimal metric under which instances belonging to the same person are more similar, and instances belonging to different people are more different. It can be considered as a data-driven feature importance mining technique [18] to suppress cross-view variations.

Existing distance metric learning methods for re-identification include Large Margin Nearest Neighbour (LMNN) [43], Information Theoretic Metric Learning (ITML) [44], Logistic Discriminant Metric Learning (LDML) [45], KISSME [46], RankSVM [4], and Probabilistic Relative Distance Comparison (PRDC) [6]. Chapter 8 provides an introduction to using RankSVM for re-identification. In particular, it details how the re-identification task can be converted from a matching problem into a pairwise binary classification problem (correct match vs. incorrect match), and aims to find a linear function to weigh the absolute difference of samples via optimisation given pairwise relevance constraints.

In contrast to RankSVM which solely learns an independent weight for each feature, full Mahalanobis matrix metric learners optimise a full distance matrix, which is potentially significantly more powerful. Early metric learning methods [43, 44] are relatively slow and data hungry. More recently, re-identification research has driven the development of faster and lighter methods [46, 47]. Chapter 12 presents a metric learner for single-shot person re-identification and provides extensive comparisons on some of the widely used metric learning approaches. It has been shown that in general, metric learning is capable of boosting the re-identification performance without complicated and handcrafted feature representations. All the aforementioned methods learn a single metric space for matching. Chapter 17 suggests that different groups of people may be better distinguished by different types of features (a similar concept is also presented in Chap. 10). It proposes a candidate-set-specific metric for more discriminative matching given a specific group with small number of subjects.

Reduce the Need for Exhaustive Data Labelling

A major weakness of pairwise metric learning and other discriminative methods is the construction of a training set. This process requires manually annotating pairs of individuals across each camera pair. Such a requirement is reasonable for training and testing splits on controlled benchmark datasets, but limits their scalability to more realistic open-world problems, where there may be very many pairs of cameras, making this 'calibration' requirement impossible or prohibitively expensive. One possible solution has been presented in [48], where a per-patch representation of the human body is adopted, and each patch of the images of the original training dataset has been sampled many times in order to simulate diverse illumination conditions. Alternatively, other techniques have also been proposed [29, 49] that aim to exploit the structure of unlabelled samples in a semi-supervised multi-feature learning framework given very sparse labelled samples. Chapter 13 attempts to resolve this problem by dictionary-based domain adaptation, focusing on face re-identification. In particular, it assumes that the source domain (early location) has plenty of labelled data (subjects with known identities), whilst the target domain (different location) has limited labelled images. The approach learns a domain invariant sparse representation as a shared dictionary for cross-domain (cross-camera) re-identification. In this way, the quantity of pairwise correspondence annotations may be reduced.

Another perspective on this data scalability problem is that of transfer learning. Ideally one wishes to construct a re-identification system between a pair of cameras with minimal calibration/training annotation. To achieve this, re-identification models learned from an initial set of annotated camera-pairs should be able to be exploited and/or adapted to a new target camera pair (possibly located at a different site) without exhaustive annotation in the new camera pair. Adapting and transferring re-id models is a challenging problem which despite some initial work [24, 50] is still an open problem.

Re-identification as an Inference Problem

In many cases one would like to infer the identity of past and unlabelled observations on the basis of very few labelled examples of each person. In practice, the number of labelled images available is significantly fewer than the quantity of images for which one wants to identify. Chapter 14 introduces formally the problem of identity inference as a generalisation of the person re-identification problem. Identity inference addresses the situation of using few labelled images to label many unknown images without explicit knowledge that groups of images represent the same individual. The standard single- and multi-shot recognition problem commonly known in the literature can then be regarded as special cases of this formulation. This chapter discusses how such an identity inference task can be effectively solved through using a CRF (Conditional Random Field) model. Chapter 15 discusses a different facet of the re-identification problem. Instead of matching people across different camera views, the chapter explores identify inference within the same camera view. This problem is essentially a multi-object tracking problem, of which the aim is to mitigate the identity switching problem with the use of appearance cues. The study formulates a minimum-cost maximum-flow linear program to achieve robust multi target tracking.

1.4.3 From Closed- to Open-World Re-identification

Limitations of Existing Datasets

Much effort has been expended on developing methods for automatic person re-identification, with particular attention devoted to the problems of learning discriminative features and formulating robust discriminative distance metrics. Nevertheless, existing work is generally conditioned towards maximising ranking performance on small, carefully constructed *closed-world* benchmark datasets largely unrepresentative of the scale and complexity of more realistic *open-world* scenarios.

To bring re-identification from closed- to open-world deployment required by real-world applications, it is important to first understand the characteristics and limitations of existing benchmark datasets. Chapter 16 provides a comprehensive list of established re-identification benchmark datasets with highlights on their specific challenges and limitations. The chapter also discusses evaluation metrics such as Cumulative Match Characteristic (CMC) curve, which are commonly adopted by re-identification benchmarking methods. Chapter 17 provides an overview of various person re-identification systems and their evaluation on closed-world benchmark datasets. In addition, the chapter highlights a number of general limitations inherent to current re-identification databases, e.g. non-realistic assumption of perfectly aligned images, and limited number of camera views and test images for evaluation.

Exploiting Environmental Contextual Knowledge

Person re-identification cannot be achieved ultimately by matching imagery information alone. In particular, given a large camera network, the search space for re-identification can be enormous, leading to huge number of false matches. To reduce the very large number of possible candidates for matching, it is essential to discover and model the knowledge about inter-camera relationships as environmental contextual constraints to assist re-identification over different camera views.

The problem of inferring the spatial and temporal relationships among cameras is often known as *camera topology inference* [22, 23, 51–53], which involves the estimation of camera transition probabilities, i.e. (1) how likely people detected in one view are to appear in other views; and (2) an inter-camera transition time distribution, i.e. how much travel time is needed to cross a blind area [54]. State-of-the-art methods infer topology through searching for consistent spatiotemporal relationships from population activity patterns (rather than individual whereabouts) across views. For instance, methods presented in [51, 52] accumulate a large set of cross-camera entrance and exit events to establish a transition time distribution. Anton van den Hengel et al. [53] accumulate occupancy statistics in different regions of an overlapping camera network for scalable topology mapping. Loy et al. [22, 23] present a tracking-free method to infer camera transition probability and the associated time delay through correlating activity patterns in segmented regions across non-overlapping camera views over time. Chapter 19 describes a scalable approach based on [53] to automatically derive overlap topology for camera networks and evaluate its use for large-scale re-identification. Chapter 20 presents a re-identification prototype system that employs the global space–time profiling method proposed in [22] for real-world re-identification in disjoint cameras with non-overlapping fields of views.

Improving Post-Rank Search Efficiency

In open-world re-identification one may need to deal with an arbitrarily large number of individuals in multiple camera views during the query stage. After the ranking process, a ranked list of possibly hundreds of likely match images are returned by an appearance-based matching method. The final judgement is left to a human operator, who needs to inspect the list and manually localise the correct match against the query (probe) image. Existing re-identification methods generally assume the ranked list is good enough for decision making. In reality, such a ranking list is far from good and necessarily suboptimal, due to (1) visual ambiguities and disparities, and (2) lack of sufficient labelled pairs of training samples to cover diverse appearance variations from unknown changes in viewing conditions. Often, an operator needs to scroll down hundreds of images to find the true re-identification. For viable open-world re-identification, this *post-rank searching* problem needs be resolved.

Zheng et al. [19] takes a *set-based verification* perspective. More precisely, the study re-defines the re-identification problem as a verification problem of a small set of target people (which they call a watch list) from a large group of irrelevant individuals. The post-rank search thus becomes more realistic and relevant, as one only needs to verify a query against a watch-list, rather than matching the query against everyone in the scene exhaustively. Liu et al. [55] further present a *man-in-the-loop* method to make the post-rank search much more efficient. Specifically, they propose a manifold-based re-ranking method that allows a user to quickly refine their search by either 'one-shot' or a couple of sparse negative selections. Their study shows that the method allows correct re-identification converges three times faster than ordinary exhaustive search.

Chapter 18 proposes an attribute-centric alternative to improve target search by using textual description such as 'white upper garment and blue trousers'. Such a complex description can be conveniently obtained through combining a set of 'atomic' or basic attribute descriptions using Boolean operators. The resulting description is subsequently matched against the attribute profile of every image in the gallery to locate the target. Chapter 5 also explores a similar idea, which they call as 'zero-shot' re-identification. In a more practical sense, rather than using textual description solely for target search, Chap. 20 exploits the description to complement the ranking of candidate matches. In particular, a user may select multiple attributes describing the target to re-rank the initial list so as to promote target with similar attributes to a higher rank, leading to much faster target search in the rank list.

System Design and Implementation Considerations

To date, very little work has focused on addressing the practical question of how to best leverage the current state-of-the-art in re-identification techniques whilst tolerating their limitations in *engineering* practical systems that are *scalable* to typical real-world operational scenarios. Chapter 20 describes design rationale and implementational considerations of building a practical re-identification system that scales to arbitrarily large, busy, and visually complex spaces. The chapter defines three scalability requirements, i.e. associativity, capacity and accessibility. Associativity underpins the system's capability of accurate target extraction from a large search space. Several computer vision techniques such as tracklet association and global