Human Re-identification Through a Video Camera Network

Slawomir Bak

UNIVERSITE DE NICE-SOPHIA ANTIPOLIS

ECOLE DOCTORALE STIC

SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

THESE

pour l'obtention du grade de

Docteur en Sciences

de l'Université de Nice - Sophia Antipolis

Mention : INFORMATIQUE

presentée et soutenue par

Sławomir BĄK

Human Re-identification Through a Video Camera Network

Thèse dirigée par François BRÉMOND
et co-dirigée par Jan WĘGLARZ

soutenue le 5/7/2012

Jury:

Bernard	MERIALDO	Pr., Eurecom, France	Président
Mubarak	SHAH	Pr., University of Central Florida, USA	Rapporteur
Frédéric	JURIE	Pr., Université de Caen, France	Rapporteur
Cordelia	SCHMID	DR, INRIA Grenoble, France	Examinatrice
François	BRÉMOND	DR, INRIA Sophia Antipolis, France	Directeur de thèse
Jan	WĘGLARZ	Pr., Poznan University of Technology, Poland	Co-directeur de thèse

UNIVERSITY OF NICE - SOPHIA ANTIPOLIS

DOCTORAL SCHOOL STIC

SCIENCES ET TECHNOLOGIES DE L'INFORMATION
ET DE LA COMMUNICATION

THESIS

to obtain the degree of

Doctor of Philosophy

at the University of Nice - Sophia Antipolis
Specialty : Computer Vision

by

Sławomir Bąk

Human Re-identification Through a Video Camera Network

Supervisor: **François Brémond**
Co-supervisor: **Jan Węglarz**

INRIA Sophia Antipolis, PULSAR Team

July 5, 2012

Jury

M.	Bernard Merialdo	-	Eurecom	*President*
M.	Mubarak Shah	-	University of Central Florida	*Reviewer*
M.	Frédéric Jurie	-	University of Caen	*Reviewer*
Mme.	Cordelia Schmid	-	INRIA Grenoble	*Examiner*
M.	François Brémond	-	INRIA Sophia Antipolis	*Supervisor*
M.	Jan Węglarz	-	Poznan University of Technology	*Co-supervisor*

Dedicated to
my parents, Urszula and Krzysztof,
my brother Jarek
and my sister Kasia who is waiting for us in the afterlife...

Abstract

This thesis targets the appearance-based re-identification of humans in images and videos. Human re-identification is defined as a requirement to determine whether a given individual has already appeared over a network of cameras. This problem is particularly hard by significant appearance changes across different camera views, where variations in viewing angle, illumination and object pose, make the problem challenging.

We focus on developing robust appearance models that are able to match human appearances registered in disjoint camera views. As encoding of image regions is fundamental for appearance matching, we study different kinds of image descriptors. These different descriptors imply different strategies for appearance matching, bringing different models for the human appearance representation. By applying machine learning techniques, we generate descriptive and discriminative models, which enhance distinctive characteristics of extracted features, improving re-identification accuracy.

This thesis makes the following contributions. We propose six techniques for human re-identification. The first two belong to *single-shot* approaches, in which a single image is sufficient to extract a robust human signature. These approaches divide the human body into the predefined body parts and then extract image features. This allows to establish the corresponding body parts, while comparing signatures. The remaining four methods address the re-identification problem using signatures computed from multiple images (*multiple-shot* case). We propose two techniques which learn online the human appearance model using a boosting scheme. The boosting approaches improve recognition accuracy at the expense of time consumption. The last two approaches either assume the predefined model, or learn offline a model, to meet time requirements. We find that covariance feature is in general the best descriptor for matching appearances across disjoint camera views. As a distance operator of this descriptor is computationally intensive, we also propose a new GPU-based implementation which significantly speeds up computations. Our experiments suggest that *mean Riemannian covariance* computed from multiple images improves state of the art performance of human re-identification techniques. Finally, we extract two new image sets of individuals for evaluating the *multiple-shot* scenario.

Résumé

Ce manuscrit de thèse à pour sujet la ré-identification de personne basée sur leur apparence à partir d'images et de vidéos. La ré-identification de personne consiste à déterminer si un individu donné est déjà apparu sur un réseau de caméras. Ce problème est particulièrement difficile car l'apparence change significativement entre les différentes vues de caméra, où les variations de points de vue, d'illumination et de position de l'objet, rendent le problème difficile.

Nous nous concentrons sur le développement de modèles d'apparence robustes qui sont en mesure de faire correspondre les apparences humaines enregistrées dans des vues de caméra disjointes. Comme la représentation de régions d'image est fondamentale pour la mise en correspondance d'apparence, nous étudions différents types de descripteurs d'images. Ces différents descripteurs impliquent des stratégies différentes pour la mise en correspondance d'apparence, impliquant des modèles différents pour la représentation des apparences de personne. En appliquant des techniques d'apprentissage automatique, nous générons des modèles descriptifs et discriminatoires, qui améliorent la distinction des caractéristiques extraites, améliorant ainsi la précision de la ré-identification.

Cette thèse à les contributions suivantes. Nous proposons six techniques de ré-identification humaine. Les deux premières appartiennent aux approches *single-shot*, dans lesquelles une seule image est suffisante pour extraire une signature fiable de personne. Ces approches divisent le corps humain en différentes parties de corps prédéfinies, puis extraient les caractéristiques de l'image. Cela permet de mettre en correspondance les différentes parties du corps en comparant les signatures. Les quatre autres méthodes abordent le problème de ré-identification à l'aide de signatures calculées à partir de plusieurs images (*multiple-shot*). Nous proposons deux techniques qui apprennent en ligne le modèle d'apparence humaine en utilisant un schéma de boosting. Les approches de boosting améliorent la précision de la reconnaissance, au détriment du temps de calcul. Les deux dernières approches assument un modèle prédéfini, ou un apprentissage hors ligne des modèles, pour réduire le temps de calcul. Nous constatons que le descripteur de covariance est en général le meilleur descripteur pour la mise en correspondance des apparences dans des vues de caméras disjointes. Comme l'opérateur de distance de ce descripteur nécessite un calcul intensif, nous proposons également une nouvelle implémentation utilisant le GPU qui accélère considérablement les temps de calcul. Nos expériences suggèrent que la *moyenne Riemannienne des covariances* calculée à partir de plusieurs images améliore les performances par rapport aux techniques de ré-identification de personne de l'état de l'art. Enfin, nous proposons deux nouvelles bases d'images d'individus pour évaluer le scénario *multiple-shot*.

Acknowledgements

First, I would like to express my gratitude to my thesis supervisors, François Brémond and Jan Węglarz. Thanks to them, I was able to collaborate between two research units in two different countries while doing one PhD. Big thanks to François, who has been a great mentor, a guide, and a friend. I thank him for offering me the opportunity of doing a PhD on such an interesting and challenging topic. I also appreciate that he gave me a lot of freedom to apply my own ideas.

Thanks to my thesis reviewers, Mubarak Shah and Frédéric Jurie, for their very pertinent advices and remarks. I would like to thank Cordelia Schmid, my thesis examiner, for her interest in my work and Bernard Merialdo, the thesis committee president, for making his time available for me.

I would like to thank Monique Thonnat, for teaching me skills to express and to formalize the scientific ideas, especially at the abstract level. I am grateful to Krzysztof Kurowski, for his scientific support and help in solving administrative issues. The collaboration between two research units would be impossible without his commitment.

Thanks to Bernard Boulay, Etienne Corvee, Marek Błażewicz and Krystyna Napierała, for their technical and scientific support. They were always helping me in switching between research units, providing the required hardware architecture and a remote access to data, no matter where I was.

I would like to thank all my colleagues from the PULSAR team for making INRIA a fun and friendly place to work. Our collaboration, fruitful discussions and common nights while meeting deadlines, gave me moments I will always keep in mind. Special thanks to Etienne, Julien, Guido, Bernard, Sonia, Phu, Ikhlef, Guillaume, Rim, Ratnesh and Piotr for our friendship. Thanks to Sorana and Ezequiel, who despite being only a moment in our team, brought a lot of warmness and friendship.

I will never forget my good friends Mauri, Tamara, Ezequiel, Marek and Miłosz for joint expeditions discovering the beauty of the Maritime Alps. Big thanks to one of my best friends, Andrzej who made me addicted to the French Riviera and provided unconditional support through both the highs and lows of my time. I have to admit that I would never decide to travel every 6 months to Poland without having there such friends as Piątuś, Olek, Andrzej, Julia, Ewelina, Tuptuś, Michał, Buśko, Wojtek, Danka, Miłosz, Asia, Lukas, Sylwia, Filip, Ada, Antek, Iza, Błażej, Alicja and Mariusz. Thank you for your support and understanding.

My final words go to my family. I would like to thank my parents, my brother and the whole family, for their support and love. Special thanks to my brother who has encouraged me almost every day, while exchanging research and life ideas.

Contents

Nomenclature

List of Symbols

$\mathfrak{s}$	signature
S	similarity between two signatures
C	covariance matrix
μ	Mean Riemannian Covariance
α	classifier's weight
I	image
$\mathfrak{I}$	integral image
$\mathcal{M}$	manifold

Acronyms

DCD	Dominant Color Descriptor
SCR	Spatial Covariance Regions
MRC	Mean Riemannian Covariance
RCP	Reliable Covariance Patches
LCP	Learned Covariance Patches
MRCG	Mean Riemannian Covariance Grid
COSMATI	COrrelation-based Selection of covariance MATrIces

Chapter 1

Introduction

"*Computing is not about computers any more. It is about living.*"
(Nicholas Negroponte)

Computers, tablets and smartphones have taken possession of our lives. It is hard to even imagine a world without them. They extend our possibilities to communicate, to share information, and to entertain. Computers can perform repetitive, data intensive and computational tasks, much more efficiently and more accurately than humans.

However, there is a lot of high-level tasks that we humans perform automatically, subconsciously, and with so much ease that we do not realize how sophisticated processes have place in our brains. Video analysis is one of many tasks, in which computers cannot take full advantage of information given by camera sensors. Such tasks as object recognition, object detection or object tracking are straightforward for humans but extremely difficult for computers.

In the result, the main goal of studies in computer vision is to improve computer abilities to perform more intelligent tasks such as visual analysis and interpretation of images or videos. One of the primary tasks is the recognition of humans using information extracted from camera. Human recognition is of particular interest not only for people behavior analysis in large area networks but also in security applications for people tracking. The capability of recognizing humans would allow a video analysis system to associate an identity with a desired person and her/his occurrence in place and time. Determining whether a given person of interest has already been observed over a network of cameras is referred to as *human re-identification.*

This chapter introduces the problem of human re-identification in images and videos. Section 1.1 gives a formal definition of the problem and presents the key application of human re-identification. We discuss the involved challenges in section 1.2, illustrating the main difficulties. We conclude with a list of major contributions in section 1.3. The structure of the dissertation is presented in section 1.4.

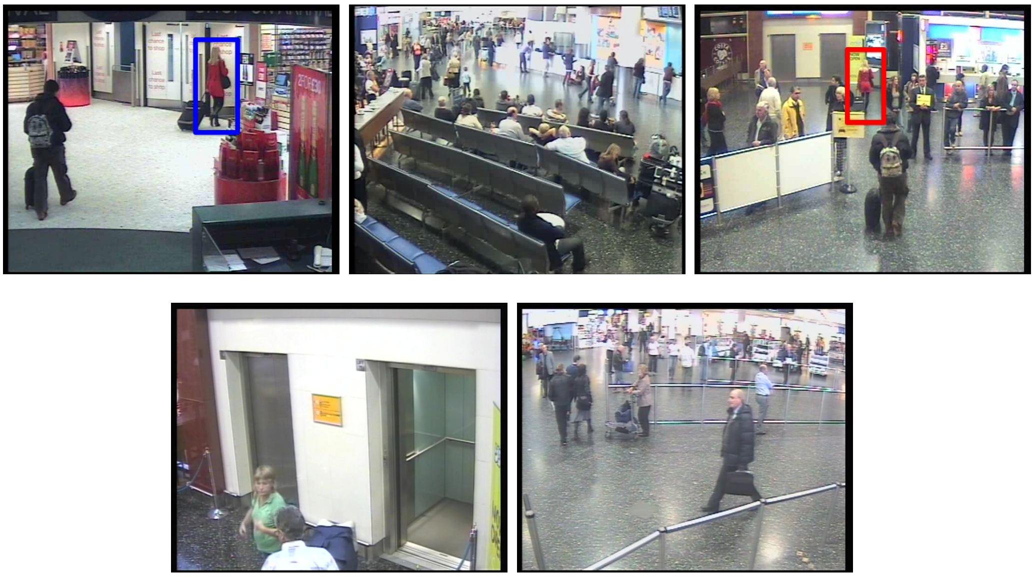


Figure 1.1: Human re-identification in a network of 5 CCTV cameras: the system should be able to associate all appearances of the same person with a single *identity* across disjoint camera views (*e.g.* the lady in red dress appears in two cameras). This video footage is distributed by the UK's government as the image Library for Intelligent Detection Systems (i-LIDS).

1.1 Motivation and problem statement

Recently, cameras spread out across various domains that range from personal computers, video games, home surveillance applications, to large camera networks which facility access to sports venue, monitored environments, such as airports, metro stations or car parks. A natural consequence of such situation is a need for an automated extraction of high-level semantic information from extremely large volumes of recorded video data.

In many surveillance systems, detection and tracking of moving objects constitute the main problem. The number of targets and occlusions produce ambiguities which introduce a requirement for reacquiring objects which have been lost during tracking. However, the ultimate goal of any surveillance system is not to track and reacquire targets, but to understand a scene and to establish an *identity* of the desired object.

Human re-identification can be defined as a determination whether a given person of interest has already been observed over a network of cameras (see figure 1.1). This issue is also known as the *person re-identification* problem. Person re-identification can be considered at different levels depending on information cues, which are currently available in the system. Biometrics such as iris, face or gait can be used to recognize identities. However, in most video surveillance scenarios such detailed

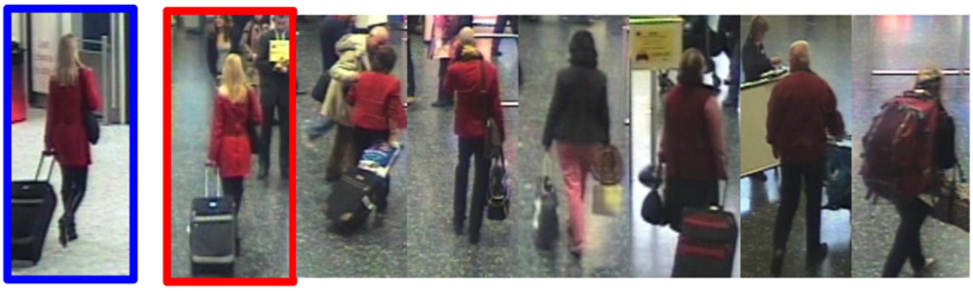


Figure 1.2: The results of the query. The first image on the left is the query image. The true match is on the first position in the list.

information might not be available, *e.g.* due to sensor scarce resolution or low frame rate. Therefore a robust modeling of a global appearance of an individual is necessary to re-identify a given person of interest. These identification techniques rely on clothing assuming that individuals wear the same clothes between different sightings. These approaches are referred to as *appearance-based re-identification*, which is the main topic of this dissertation.

It is worth noting that in these appearance-based approaches, the re-identification term stands for matching similar appearances over a network of cameras. These techniques cannot be so accurate as iris recognition or face recognition because the clothing is not distinctive enough. However, a robust appearance model can provide an effective interface to the operator to search the camera network for a person of interest. The search results are presented to the operator using the list of the most similar appearances to the appearance of interest (see figure 1.2).

Recognizing humans in a network of cameras and providing an efficient interface to the operator is the key application of human re-identification. However, matching appearances of the same object between different cameras is of primary importance not only in video analysis systems but also in video broadcasting applications (*e.g.* live broadcasting of sports events), while a set of cameras may need to track in parallel the object of interest (*e.g.* player tracking).

Human re-identification problem

We lay the problem as the following. We generate a human signature for each person detected and tracked in a video surveillance system. Let us denote a signature as $\mathfrak{s}_i^c$, where i encodes the person's identity and c denotes the camera. The task is to find for each signature its corresponding signature in another camera. It is realized by querying the database of signatures $\mathfrak{s}_j^{c'}$, where $c \neq c'$ with signature of interest $\mathfrak{s}_i^c$. The results of the query is the list of the most similar signatures ordered by increasing dissimilarity (figure 1.2). The position in the list of the true match is called the rank score. The main objective is to maximize recognition accuracy by

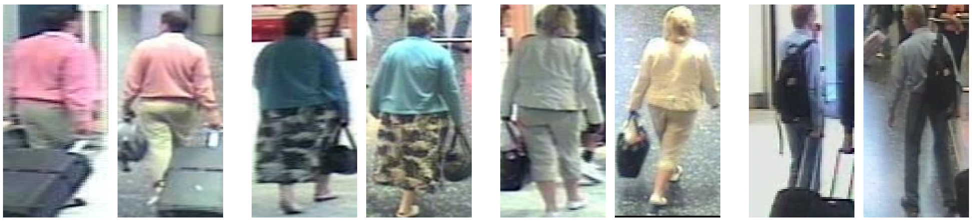


Figure 1.3: Color changes: each pair of images shows the same person captured from different cameras in different environments. These images highlight color dissimilarities of the same object acquired under different conditions.

minimizing the rank of the true match (see section 7.1 for details on evaluation metrics).

1.2 Research challenges

Appearance matching of the same person registered in disjoint camera views is one of the most challenging issues in every video analysis system. The underlying challenge of the human re-identification problem arises from significant appearance changes, caused by variations in viewing angle, illumination and object pose. Different color responses, different camera viewpoints and different camera parameters impede establishing correspondence between parts of the human body, while matching appearances.

Color changes

The same object acquired by different cameras shows color dissimilarities (see figure 1.3). Even identical cameras which have the same optical properties and work under the same lighting conditions may not match in their color responses. Besides, inter-camera variations in lighting conditions and differences in illumination significantly change the appearance of a person.

Pose changes

Moreover, the pose variations due to camera and viewpoint change as well as the articulation of the human body lead to significant differences in appearances of the same individual observed from different cameras (see figure 1.4). The first pair shows upright and bending poses. In this case, if we would have a good matching algorithm most of features from the first image could be found in the second image (only the spatial position of features has changed). However, in the rest of the examples, the first image contains visual features which do not appear in the second image.

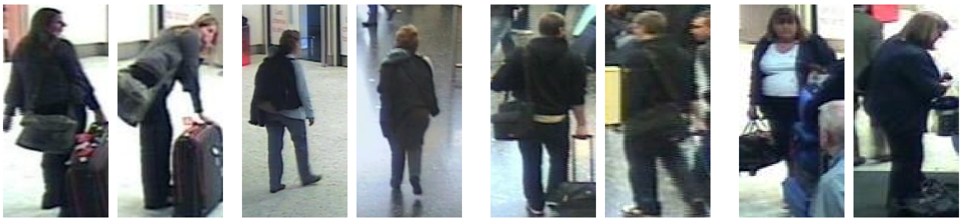


Figure 1.4: Pose changes: each pair of images presents how the appearance of the person can change due to pose changes. Note that in some cases there are visual features which are present in one camera, while they are not visible in the second one (the last pair: in the first image we can notice white t-shirt, which is not visible in the second image).

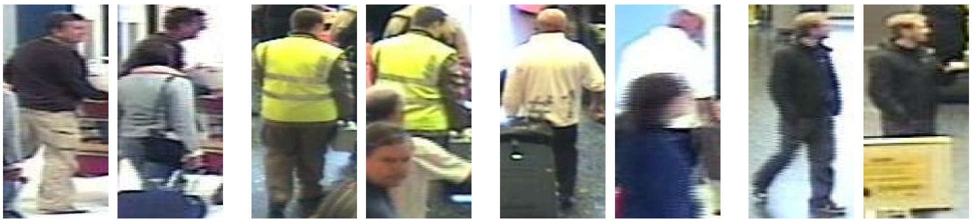


Figure 1.5: Occlusions: examples extracted from a real video surveillance system.

Finding corresponding body parts is referred to as the *correspondence problem.* This step is of particular interest as it has a significant influence on recognition accuracy, while matching appearances.

Occlusions

Further, self-occlusions (caused by body parts) and occlusions caused by other people or objects of the scene make the appearance matching task extremely difficult (see figure 1.5).

Detections

In order to recognize individuals in a real video surveillance system, we need to detect human silhouettes automatically. This brings a new challenge which comes from the preprocessing step and is not directly related to the re-identification problem but has a large influence on recognition accuracy.

Human detection is supposed to detect image regions where a human is present. Unfortunately, existing algorithms very often return noisy human detections (detected bounding boxes are not accurately centred around the people, only part of

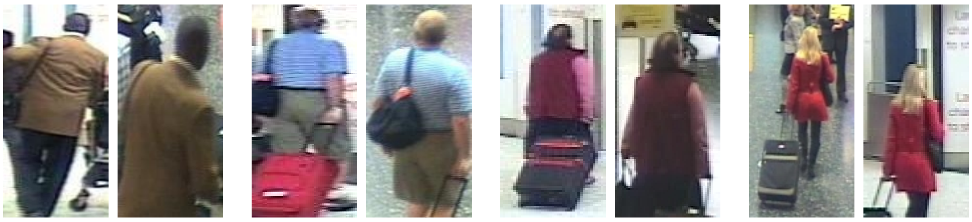


Figure 1.6: Noisy human detection results: right image in a pair shows noisy detection. Note that the size of appearance differs significantly when camera changes.

the people are detected). In most video analysis systems, tracking of objects is used to select the most informative appearance (called the *key* frame). However, the selection of the best candidate is still an open issue.

Noisy detections make the re-identification problem more challenging, especially when the full body is detected in the first camera and is not detected in the second one (*e.g.* only upper body is detected, see figure 1.6). Similarly as in pose variations, it can often happen that features which describe an appearance extracted from the first camera are not visible in an appearance extracted from the second camera. Moreover, in this case scale and resolution changes appear (person in the first image is much smaller than in the second).

Discriminative features

Inspired by *human memory* [Bjork 1996] and in particular - *recognition memory*, we can infer that not all cues provided by images should be used in the same manner. Recognition memory is the ability to recognize previously encountered events, objects, or people, involving processes of discriminative learning. Learning in humans is performed by stimuli which are probably an input into sensory memory even when a person tries to ignore them at the time they are presented. In the result, we, humans learn continuously to deal with the factors of associability and discriminability, while computers do not possess such capability. Thus, it raises numerous questions how to teach computers to match (associate) human appearances and how to focus on distinctive cues (features), providing within-group discriminations.

1.3 Main contributions

This thesis studies the question of feature sets for human re-identification in non-overlapping camera views. We employ different kinds of image descriptors to examine their invariance to camera changes and we investigate different strategies to enhance discriminative properties of these feature descriptors. We find *covariance* *matrix* to be the best descriptor for matching appearances registered in disjoint camera views. The thesis makes the following contributions:

- We formulate the appearance matching problem as the task of learning a model that selects the most descriptive features for a specific class of objects (*e.g.* humans). The main idea of the proposed approach is that different regions of the object appearance ought to be matched using different strategies to obtain a distinctive representation. This work has been submitted for publication [Bak 2012a] and is detailed in chapter 6.

- We offer a novel kind of feature, *i.e.* the *mean Riemannian covariance grid* (MRCG). Our idea is to combine efficiently the *mean Riemannian covariance* descriptor with a spatial information carried out by a dense grid structure. The matching accuracy is improved by our efficient discriminative method. This work was published in [Bak 2011b] and further developed in [Bak 2012b, Corvee 2012]. Detailed description can be found in chapter 6.

- We investigate discriminative approaches by boosting human appearances in *one-against-all* learning scheme. We employ Haar-wavelets and covariance descriptors as the features for describing image regions. In these approaches we enhance distinctive characteristics of a specific appearance by using information from appearances of other individuals. Boosting approaches were published in [Bak 2010a, Bak 2011a] and are discussed in detail in chapter 5.

- We develop two approaches which extract human appearance from a single image using body-part detection methods. Localization of body parts is based either on an asymmetry of the human body or on learned characteristics using *histogram of oriented gradients* (HOG). These approaches were published in [Bak 2010a, Bak 2010b] and are the main focus of chapter 4.

- We design a new GPU-based re-identification framework, offering a new implementation of the covariance distance operator. This implementation significantly speeds up computation with comparison to the optimal CPU implementation of Jacobi algorithm [Jacobi 1846]. The speed-up grows with a number of matrices, reaching its maximum (66) from about 1500 matrices. This work was published in [Bak 2011c] and is described in chapter 8.

- We extract two new image sets of individuals from i-LIDS data to study more carefully advantages of using multiple images in generating human signature. These two publicly available datasets fully satisfy requirements of *multiple-shot* person re-identification. These datasets were introduced in [Bak 2011a]. We present specifications of these datasets in chapter 7.

- We perform an extensive evaluation of our descriptors on various datasets. Performance comparison to state of the art approaches and detailed analysis can be found in chapter 7.

1.4 Thesis structure

Chapter 1 introduces the human re-identification problem describing its key application. The overall goals together with the research challenges and our main contributions are summarized. The remaining chapters are organised as follows:

- **Chapter 2** introduces the reader to state of the art methods for human re-identification, focusing particularly on *appearance-based* approaches. The first part describes re-identification by biometrics, showing constraints of these approaches due to sensor scarce resolution or low frame-rate. The remaining part focuses on models for appearance matching, considering *feature-oriented* and *learning* techniques. It also presents our motivation for contributing to both, *feature-oriented* and *learning* approaches.

- **Chapter 3** describes an overview of the re-identification framework. A short discussion on different appearance models for detection and tracking is presented as these steps are common requirements in most video analysis systems. We also indicate steps in the processing chain, which are the main focus of this dissertation.

- **Chapter 4** presents two *single-shot* approaches for human re-identification. These proposed techniques extract human characteristics using a single image. Both approaches belong to *feature-oriented* methods which use *a priori* information on structure of the human body parts. The chapter examines two kinds of image descriptors: (1) dominant color descriptor; and (2) covariance matrix. We investigate their invariance to appearance changes in disjoint camera views.

- **Chapter 5** contains two techniques for generating a human signature by *boosting* - a meta machine learning algorithm. Both approaches employ multiple images of many humans, using *one-against-all* learning scheme to extract a discriminative representation of the human appearance. Distinctive characteristics of a specific individual are enhanced by confronting its appearance with the appearances of other individuals.

- **Chapter 6** proposes two approaches based on mean Riemannian covariance matrices. The first method avoids time consuming *boosting* algorithm by defining appearance representation as a dense grid of features and by using a simple discriminative method for improving signature matching accuracy. The second technique consists in offline learning which learns the best features to match appearances between two camera views. We propose to formulate the appearance matching problem as the task of learning a model that selects the most descriptive features for a specific class of objects. Learning is performed in a covariance metric space using an entropy-driven criterion.

- **Chapter 7** presents a detailed performance analysis of proposed descriptors for re-identification and shows a performance comparison of these approaches with state of the art techniques. We describe evaluation metrics and publicly available datasets. New sets of individuals from i-LIDS data are proposed for evaluating *single-shot* and *multiple-shot* scenarios. We evaluate our approaches on publicly available datasets with varying difficulties, investigating descriptor limitations.

- **Chapter 8** describes an extension of our re-identification framework to large scale systems. This chapter focuses on a distance operator between signatures to perform an efficient signature retrieval, especially once the real-time response is expected from the system. Our best descriptors are based on covariance matrix which distance operator requires solving the *generalized eigenvalue problem.* This problem is computationally intensive and must be repeated constantly in the re-identification system during the browsing signatures of interest. In the result, we perform detailed analysis of the algorithm complexity and we explore for possibilities of parallelization. We find that some parts of the distance operator algorithm can be easily parallelized. Consequently, we take advantage of new high performance architectures to obtain the required high efficiency. We propose a new GPU-based implementation for solving the *generalized eigenvalues problem.* In the result, we significantly accelerate the distance computation, reaching 66 speedup in comparison with the CPU implementation.

- **Chapter 9** summaries our approaches and our key results, providing a discussion of the advantages and limitations of the work. It also suggests directions for future research in this area, presenting short-term and long-term perspectives.

Chapter 2

State of the art

"*All men by nature desire knowledge.*" (Aristotle)

In this chapter, the state of the art methods for human re-identification are presented. Identification techniques can be considered on different levels depending on information cues which are currently available in a video surveillance system. We take into account only approaches which use camera to acquire information cues. These approaches can be categorized into the two main families:

- biometrics,
- global appearance.

Biometrics belongs to the most effective approaches as they use biological characteristics to verify an identity. Iris (section 2.1.1), face (section 2.1.2) or gait (section 2.1.3) has proven to own one of the most discriminative information which can be used for recognizing individuals.

Unfortunately, in most video surveillance scenarios such detailed information might not be available due to sensor scarce resolution, low frame rate or difficult background subtraction (crowded environments, *e.g.* airports, metro stations). Therefore only a robust modeling of an individual's appearance (section 2.2) can provide cues for re-identifying a given person of interest. These identification techniques rely on clothing assuming that individuals wear the same clothes between different sightings. In state of the art these recognition methods are referred to as *appearance-based re-identification*, which is also the main topic of this dissertation. For the sake of clarity, before focusing on appearance-based methods, we present a brief overview of biometrics.

2.1 Human re-identification by biometrics

Biometrics is a term which marries two roots: *bio*, indicating a connection to living organisms; and *metric*, indicating a measurement. In general, biometrics refers to the statistical study of biological phenomena, where a measurement of biological characteristics is used to verify an individual's identity.

Biometrics consists of metrics which take into account physical as well as behavioural

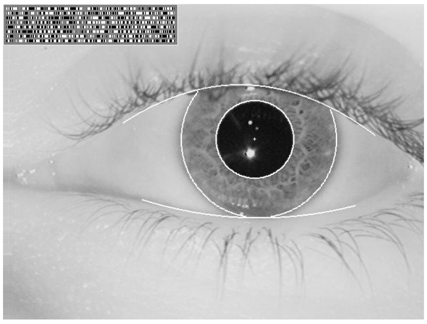


Figure 2.1: Example of an iris pattern, imaged in near infrared at a distance of about 35 cm. The outline overlay shows the results of iris, pupil, and eyelid localization steps, and the bit stream in the corner depicts the computed IrisCode [Daugman 2002].

characteristics. Physical characteristics refer to such techniques as *fingerprint*, *face recognition*, *DNA*, *iris recognition*, *retina scan* and *scent*. Behavioural characteristics are related to methods which examine how an individual behaves *e.g. keystroke/typing dynamics*, *handwritten signature*, *gait* and *voice*.

In this section we consider only these biometrics which can be acquired covertly by camera sensors. Thus, at least theoretically, biometrics recognition can be performed without subject's knowledge. The feasibility of this type of recognition receives increasing attention and is of particular interest for forensic and security purposes, such as the pursuit of criminals and terrorists.

We describe the following biometrics:

- iris - a physical characteristic of an individual's eye (section 2.1.1),
- face - a physical characteristic of an individual's face (section 2.1.2),
- gait - a behavioural characteristic of an individual's motion (section 2.1.3).

2.1.1 Iris recognition

Iris recognition is one of the most accurate techniques for identification of human beings which has ever been developed. Moreover, iris recognition supports non-invasive and covert data acquisition which makes this biometrics very desirable for security purposes. Although small (11 mm) and sometimes problematic to image, the iris has the great mathematical advantage that its pattern variability among different persons is enormous. In addition, as a planar object its image is relatively insensitive to illumination changes. Further, variations in viewing angle effect only affine transformations. Also the non-affine pattern distortion caused by pupillary dilation is readily reversible. The striated mesh of elastic pectinate ligament creates the predominant texture under visible light, whereas in the *near infrared* (NIR, 700-900 nm)

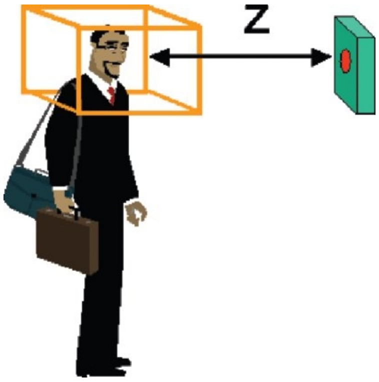


Figure 2.2: Illustration of standoff (Z), capture volume, and residence time. The recognition device is at the far right of the illustration. The capture volume is represented as a box around the subject's head [Matey 2008].

wavelengths, at close acquisition distance, deeper and somewhat more slowly modulated stromal features dominate the iris pattern. This pattern can be represented by IrisCode (see figure 2.1) which is extracted using bandpass filtering operations to take advantage of multiscale information [Daugman 2002, Wildes 1997].

NIR wavelengths reduce specular reflection from the convex cornea causing that even darkly pigmented irises reveal rich and complex features. However, the NIR wavelength is particularly risky because the eye does not instinctively respond with its natural mechanism (aversion, blinking and pupil contraction). Thus, there are also studies performing in visible light [Proenca 2010], to extract iris. Though, the use of visible light can decrease the quality of the captured data increasing the challenges in recognition.

Unfortunately, iris recognition systems impose significant constraints on the subject. The key constraints in current systems are standoff distance, capture volume, residence time, subject motion, subject gaze direction, and ambient environment [Matey 2008].

As we can see in figure 2.2, standoff is the distance between the subject and the iris image acquisition device. In some cases, the illumination components are located separately from the camera sensor. Then, we have to take into account both an illumination standoff and a camera standoff. The volume in which we expect sufficient quality of an iris image is called the capture volume. The capture volume is shown as a box around the subject's head in figure 2.2.

Residence time corresponds to the length of time in which the subject must remain within the capture volume to acquire the iris image. Moreover, in some systems there is a requirement forbidding movement of the subject (the subject should remain stationary) to minimize motion blur. Even if the system allows subject motion, there are limits on the subject velocity.

As the most of iris recognition systems assume that iris/sclera boundaries are circular (the subject should look approximately straight into the camera) the recognition performance decrease rapidly when gaze direction of the subject deviates from the optic axis of the camera by more than ≈ 15 degrees.

Finally, the ambient environment is disturbed by ambient light. The ambient light can affect the iris acquisition in ways that are unpredictable. There are two primary ways to reduce the effect of ambient light on the image: (1) temporal domain strobing/gating and (2) wavelength domain filtering. The former way of reducing the effect of ambient light is to shutter the camera with a gate that is shorter than the frame time and to strobe the controlled illumination synchronously with the gate. The latter way can be realized by usage of narrowband sources of illumination such as LEDs or lasers. In this case the bandpass optical filters at the camera can be used to block all wavelengths other than the ones of interest.

Even if iris recognition is one of the most accurate techniques for identification of human beings, above mentioned constraints show difficulties which appear during acquisition process. Another biometrics which can be used in less constrained environments is *face recognition*. The accuracy of face recognition is minor to iris recognition, as the face is a changeable social organ displaying a variety of expressions, being at the same time active $3D$ object whose image varies with viewing angle, pose, illumination, and age.

2.1.2 Face recognition

Face recognition, as one of the most important topics in computer vision, has already been investigated from at least three decades. The earliest successful approach of face recognition is the *Eigenface* method [Kirby 1990, Sirovich 1987]. This technique is based on linear projection of an image space to a low dimensional feature space. The *Eigenface* method uses principal components analysis (PCA) for dimensionality reduction which yields projection directions to maximize the total scatter across all classes, *i.e.*, across all images of all faces. Finally, each face can be considered as a combination of vectors (called *eigenfaces*) which correspond to a set of eigenvectors extracted from a training dataset. This eigenvectors (see figure 2.3) represent variations which can appear in the face appearance. This approach has been improved by applying Fisher's linear discriminant analysis [Fisher 1936] to work in the presence of lighting variations [Belhumeur 1997]. These approaches perform remarkably well only if pose variations are not significant. In [Bauml 2010, Lee 2003], we find approaches which struggle with pose variations and show promising results. However, even if face recognition approaches have no such strong constraints as iris, they still demand high resolution images to extract significant features from the face. Essentially, such features as eyes, nose, mount, has to be visible. Otherwise face pose alignment is difficult (alignment is necessary to perform matching between face appearances). In case when such features are not accessible (camera low res-

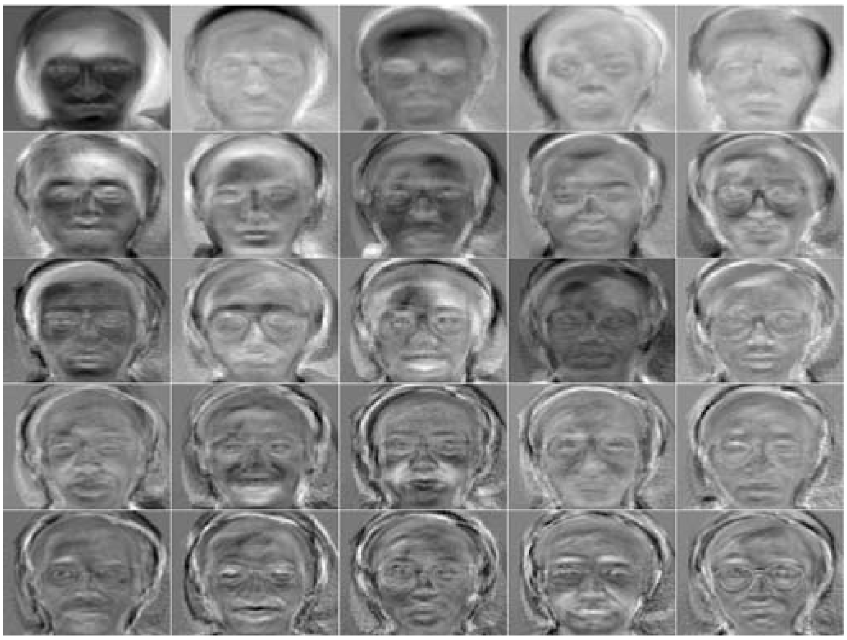


Figure 2.3: An example of the first 25 eigenfaces which represent variations of faces in a dataset.

olution, large distance from the object) gait recognition can be used to extract an identity.

2.1.3 Gait recognition

Gait recognition addresses the problem of human identification at a distance. It aims to discriminate individuals by the way they walk. Each person seems to have a distinctive and idiosyncratic way of walking. Human ambulation consists of synchronized integrated movements of hundreds of muscles and joints. Although these basic patterns of bipedal locomotion are similar for humans, gaits vary from one person to another in certain details such as their relative timing and magnitudes. The attractiveness of gait as a biometrics arises from the fact that is non-intrusive, can be measured from a great distance and can perform in a low resolution.

Gait approaches are mostly based on a silhouette estimated by background subtraction (see figure 2.4) and perform recognition by temporal correlation of silhouettes. Thus, background subtraction (extraction of a silhouette) is the main phase in gait recognition. After background subtraction, the detected silhouette can be represented by a distance signal [Wang 2003] obtained by a normalized distance of boundary pixels on the contour from the reference point (*e.g.* the centroid can be chosen as the reference point). The outer contour, unwrapped in counter-clockwise way, represents the original $2D$ silhouette shape in the $1D$ space (see figure 2.5).

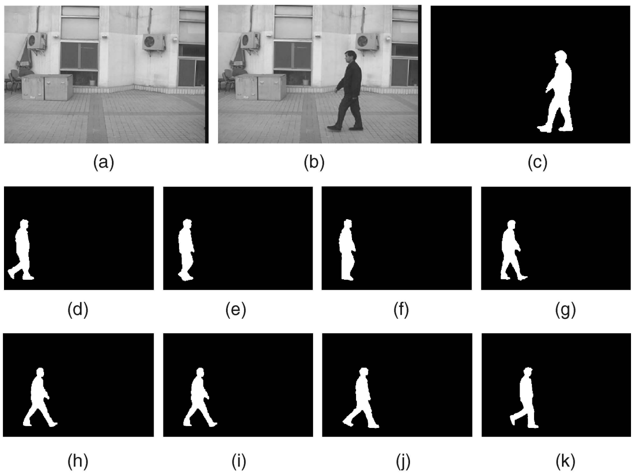


Figure 2.4: Examples of moving silhouette extraction and tracking [Wang 2003]: (a) a background image; (b) an original image; (c) an extracted silhouette from (b), and (d)-(k) temporal changes of moving silhouettes in a gait pattern.

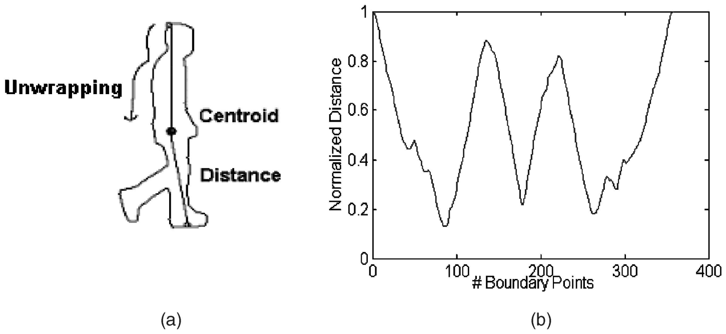


Figure 2.5: Silhouette representation [Wang 2003]: (a) illustration of boundary extraction and counter-clockwise unwrapping and (b) the normalized distance signal consisting of all distances between the centroid and the pixels on the boundary.

Compared with other widely used biometric features such as iris or face, gait recognition is relatively new area for computer vision researchers and still meets a lot of constraints. Most of the gait perception approaches assume a side-view of the person to extract distinctive motion templates. Moreover, silhouette extraction (background modelling) is still an open issue. Apparently, there are number of limitations in the use of gait in uncontrolled environments, *e.g.* outdoors under various lighting conditions [Chellappa 2007].

Unfortunately, in most video surveillance scenarios due to difficult background subtraction (crowded environments, *e.g.* airports, metro stations), gait still can not be used for identification. Therefore, a robust modelling of a global appearance of an individual is necessary to identify a given person of interest. In these identification techniques (named *appearance-based approaches*) clothing is the most reliable

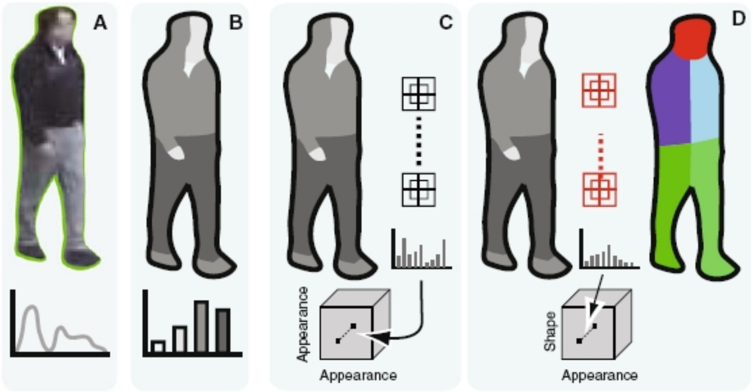


Figure 2.6: Appearance models [Doretto 2011]: (A) histogram representation (*e.g.* intensity value); (B) more sophisticated histograms, based on vector-quantization according to an appearance dictionary; (C) aggregation strategies take into account spatial correlations; (D) part based methods divide the body into parts and the description of corresponding parts determines the matching.

information on an identity of an individual (there is an assumption that individuals wear the same clothes between different sightings). Obviously, the appearance model can not be such distinctive as biometrics but can provide an effective interface to the operator, which could be able to search the camera network for a person of interest.

2.2 Human re-identification by appearance model

Recently, the person re-identification problem became one of the most important tasks in video surveillance applications. A natural consequence of an invention of robust human detection algorithms is to extend approaches for recognition purposes.

As human re-identification concerns a large set of individuals acquired from different cameras, it is necessary to provide a *distinctive* and *invariant* to camera changes signature. It has to be based on *discriminative* features to allow browsing the most similar signatures over a network of cameras. It can be achieved by signature matching which has to handle difference in illumination, pose and camera parameters. Note that, inter-camera variations in lighting conditions, difference in illumination, different camera parameters, changes in object orientation and object pose make this task extremely difficult. Besides, occlusions (caused by other people or objects of the scene) and self-occlusions (caused by body parts) form the re-identification problem as one of the hardest tasks in the video surveillance applications.

In *appearance-based re-identification* approaches, a human appearance is represented by a set of features extracted from a human image. The main issue appears in aggregating these features into discriminative representation which should yield correct matching of the same individual seen from different cameras. The simplest approaches do not take into account a spatial correlation between image features, applying a histogram representation. More sophisticated techniques embed a spatial information inside the feature vector to cope with the *correspondence problem*. Finally, the human body part detectors can be used to extract features from corresponding body parts to perform correct matching (see figure 2.6).

Further, person re-identification approaches concentrate either on learning aspects or on feature modeling. Learning approaches use training data of different individuals to search for strategies that combine given features maximizing inter-class variation whilst minimizing intra-class variation. These approaches focus either on *metric learning* for matching appearances regardless of the representation choice [Dikmen 2010, Zheng 2011], or on *discriminative methods* which enhance discriminative features of a specific individual [Lin 2008, Schwartz 2009]. Instead, *feature-oriented* approaches concentrate on designing an invariant feature, which should handle viewpoint and camera changes [Bazzani 2010, Farenzena 2010].

Appearance-based approaches rely on modeling a human signature using tracking and detection results. As the tracking can provide multiple images of the same individual, classification of appearance-based techniques distinguishes two main groups:

- *single-shot* approaches,
- *multiple-shot* approaches.

The former class exploits appearance information using only a singe image of an individual, while the latter employs multiple images of the same person as training data to obtain a reliable and distinctive representation.

2.2.1 Single-shot approaches

Single-shot approaches extract a human signature using only a single image. It is worth noting, that feature-oriented single-shot approaches exploit the least information among all appearance-based re-identification techniques. For each individual, they use a single image, whose features are independently matched against thousands of candidates. In contrary, learning approaches concentrate either on learning a metric regardless the representation choice (usually training phase is offline), or on enhancing discriminative features presented in an image (usually discriminative method is online).

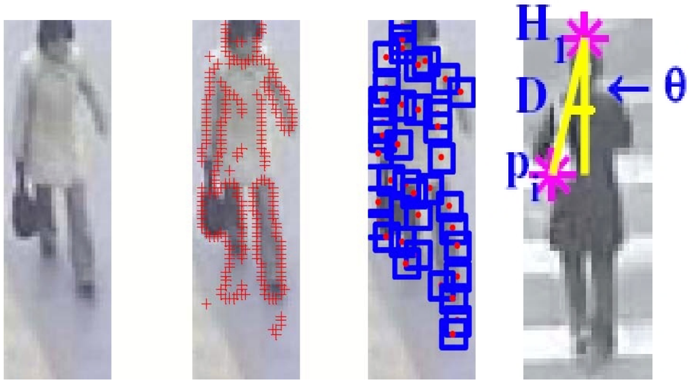
2.2.1.1 Feature-oriented approaches

Feature-oriented (FO) approaches concentrate on feature representation which should be invariant to pose and camera changes. These approaches usually assume *a priori* an appearance model, focusing on designing novel features for capturing the most distinctive aspects of an individual.

In [Park 2006] clothing color histograms taken over shirt and pants regions together with the approximated height of a person are used as primitive features to discriminate appearances. Head, shirt and pants regions are obtained by dividing a detected blob into three parts from the top to the bottom (at 1/5th and 3/5th of the blob's height). Only shirt and pants regions are taken into account to extract color appearance of a person. Dividing a blob into two parts helps to keep spatial correlation of feature distributions. Histograms are computed in HSV color space, based mostly on hue component (the authors assumed that hue is the most invariant component to illumination changes). Histograms with 10 bins (red, brown, yellow, green, blue, violet, pink, white, black and gray) are constructed from every pixel in the object segmentation. Then, final color is decided as the bin with the largest count.

Similarly, in [Gallagher 2008] clothing segmentation together with facial features is employed to recognize individuals. The clothing segmentation is obtained using graph cuts and clothing mask. The global clothing mask is learned automatically based on mutual information extracted from the set of images with the same qualitative appearance. Face is detected using [Viola 2003] approach. Then, each face is projected onto a set of Fisherfaces [Belhumeur 1997] to obtain 37-dimensional vector representing its appearance. Clothing regions are represented by 5-dimensional feature vectors. The first three elements in the vector are obtained by a linear transformation of RGB color values. The remaining two features describe texture as the responses to a horizontal and vertical edge detector. Finally, the appearance is represented by the segmented image (the set of histograms over each of the 5 features) or by visual world representation. This work shows that clothing can provide a significant improvement in face recognition accuracy for large image collections.

The human signature represented by color patches extracted from an appearance along edges is proposed in [Cai 2008]. The edges are detected by Canny edge detection algorithm [Canny 1986]. Then, square regions are selected around edges as shown in figure 2.7(a). The appearance of each region is characterized by dominant colors and their frequencies of occurrence in this region. The similarity of patches together with their geometric constraints are encoded in matching function. The geometric constraints are defined as a spatial correlation between a reference point and candidate points on the edges. The reference point is chosen as the top of the head. The spatial correlation of the candidate point is represented by its distance D from the reference point (the top of the head) and θ angle (see figure 2.7(b)). The distance between the head point and candidate point is normalized by the height of



(a) (b)

Figure 2.7: Human appearance [Cai 2008]: (a) From left: original image, results of Canny edge detection, region signatures on the edges; (b) geometric constraints.

the silhouette.

Similarly, in [Yu 2007] path-length features of the pixels inside the silhouette of a person are used to construct an appearance model. Path-length, the length of the shortest path from a distinguished point (in this work also the top of the head is chosen) to a point constrained to lie entirely within the body, captures the structural information of the appearance. The top of the head is relatively stable to the movement. Compared with the centroid of the silhouette, it can discriminate the features of the upper body and the lower body because they have different path-lengths and do not produce mixed distributions. Moreover, the top of the head is less sensitive to noise than the foot point, which can be hard to detect due to shadows. Path-length together with color of a pixel stands for the final feature extracted from the appearance. Given all the pixels of a snapshot of a person, the distribution of the feature vector is estimated. Then the similarity between the model distributions is measured by the Kullback-Leibler distance [Kullback 1978]. The authors also propose to match video sequences (*multiple-shot* case), where the simple threshold-based algorithm is applied to select key frames. Given video sequence with detected human appearance, the first frame is selected as the first key frame. Then, the following key frames are selected using the Kullback-Leibler distance and a threshold which specify if the new key frame should be selected. The final matching of video sequences is obtained by using the median of the closest distances between the key frames.

In [Kang 2004], a translation invariant model is proposed. The appearance is extracted using the smallest circle containing the blob (results of the segmentation algorithm). This circle is uniformly sampled into a set of control points, from which a set of concentric circles of various radii are used for defining bins of the appearance model. Inside each bin, a Gaussian color model is computed for modelling the color properties of the overlapping pixels of the detected blob (see figure 2.8). The final

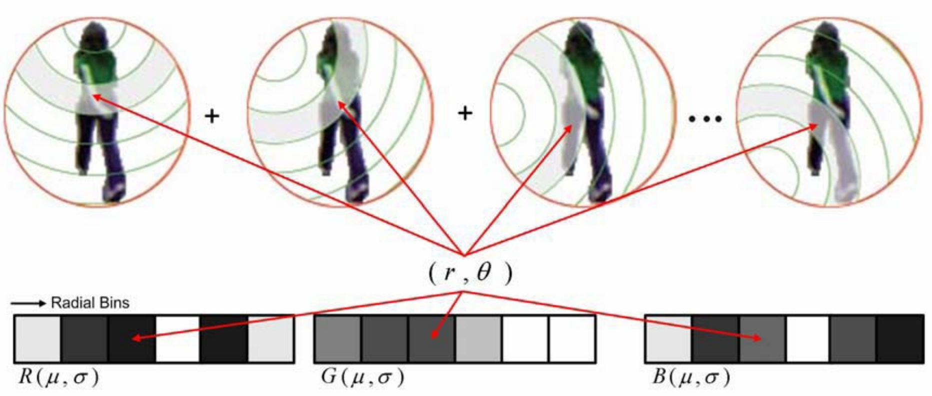


Figure 2.8: Computation of the color based appearance model [Kang 2004].

appearance model is represented by the normalized combination of the distributions obtained from each control point. Such representation guarantee rotation invariance as a rotation of the blob in the 2D image is equivalent to a permutation of the control points.

Shape and appearance context model is proposed in [Wang 2007]. A pedestrian image is segmented into regions and their color spatial information is registered into a co-occurrence matrix. A region appearance is represented by *histogram of oriented gradients* (HOG) in the *Log-RGB* color space [Funt 2002]. The gradient of the *Log-RGB* space works similar to the homomorphic filtering and makes the descriptor robust to illumination changes. The spatial correlation is handled by L-shaped regions of the image. Parts identification is done by modified shape context algorithm [Belongie 2002], which uses a shape dictionary learnt *a priori*. The context of the appearance and shape is handled by using occurrence/co-occurrence function which describes probability distributions and their correlations over the image region. This method works well if the system considers only a frontal viewpoint.

2.2.1.2 Learning approaches

Learning approaches use training data to search for strategies that improve appearance matching accuracy. These approaches concentrate either on *metric learning* (ML) regardless of the representation choice, or on *discriminative methods* (DM) which enhance discriminative features of an individual (*e.g.* by applying *one-against-all* learning scheme).

Usually learning a metric involves offline phase, in which training data is given as positive pairs (two images with the same person registered by different cameras) and negative pairs (two images with different person registered by different cameras) of images.

In contrary, applying *one-against-all* scheme, the discriminative approaches enhance

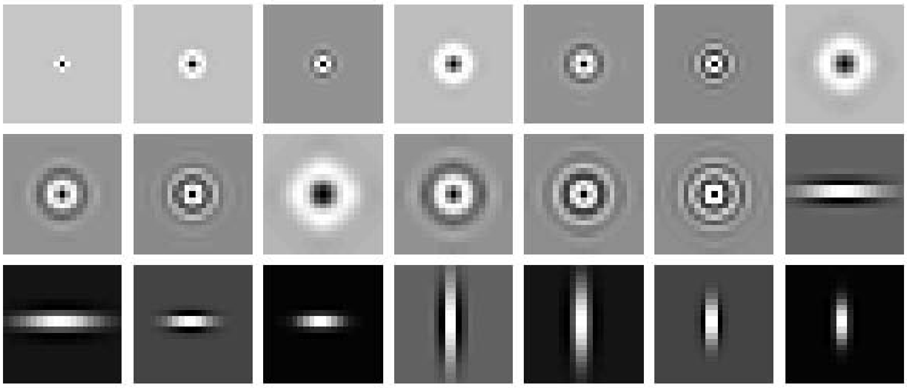


Figure 2.9: The filters used in the appearance model [Gray 2008] to describe texture features: rotationally symmetric Schmid filters and horizontal and vertical Gabor filters .

distinctive characteristics of a specific individual by using information from appearances of other individuals. In discriminative methods, learning is performed online, which makes these approaches difficult to apply in real-time systems.

Metric learning (ML)

In [Gray 2008], the ensemble of localized features (ELF) has been proposed. Instead of designing a specific feature for characterizing people appearance, a machine learning algorithm constructs a model that provides maximum discriminability by filtering a set of simple features. This simple features are extracted from a set of *strips* which span the entire horizontal dimension of the image. Each feature consists of three elements: a feature channel, a region and a histogram bin. A feature channel is any single channel transformation of the original image. In this work, different color channels (RGB/YCbCr/HSV) are used as well as texture channels. Texture channels are obtained by the result of convolution with a texture filter and the luminance channel. The authors use two families of texture filters: Schmid [Schmid 2001] and Gabor [Fogel 1989] (see figure 2.9). The similarity function between two appearances is learnt using AdaBoost algorithm. The model is extracted from pairs of pedestrian images. Positive sample is represented by a pair of images of the same person and negative sample is represented by a pair of images with different individuals. Thus, if there are N individuals imaged from two different cameras (one image from each camera), then there are N positive training examples and $N \times (N - 1)$ negative training examples. Classifiers are based on the absolute differences between two instances of the same feature. Finally, from the feature space (which is the product of the number of possible channels, regions and bins), the AdaBoost algorithm builds a model of the human appearance used for person re-identification.

The person re-identification problem is reformulated as a ranking problem in [Prosser 2010]. The authors present extensive evaluation of learning approaches and show that a ranking relevance based model can improve the reliability and accuracy in person re-identification under challenging viewing conditions. The presented idea is very similar to ELF [Gray 2008]: machine learning extracts a general representation of a human appearance. Also features used for region appearance representation are the same as in [Gray 2008]. The whole appearance is represented by six equally sized horizontal strips in order to roughly capture the head, upper and lower torso and upper and lower legs. The significant difference between these two approaches comes from a learning scheme. The novelty of this work is that the person re-identification problem is presented as a ranking problem. Here, the ranking algorithms learn a subspace where the potential true match is given by highest ranking rather than any direct distance measure. This work presents an extensive evaluation of different ranking algorithms such as RankBoost [Freund 2003] or RankSVM [Joachims 2002]. The proposed approach shows a significant improvement of SVM-based ranking approaches over Boosting-based ranking models where the weak rankers are constructed from individual features.

Distance learning is also the main topic in [Zheng 2011]. A probabilistic model maximizes the probability of true match pair having a smaller distance than that of a wrong match pair. This approach focuses on maximizing matching accuracy regardless of the representation choice. However, in order to benefit from different and complementary information captured by different features, the authors start with a mixture of color and texture histogram features similar to those used in [Gray 2008] and let the probabilistic model automatically discover an optimal feature distance.

[Dikmen 2010] also use a metric learning framework to obtain a robust metric for pedestrian recognition. The authors propose a novel cost function similar to the *large margin nearest neighbor* (LMNN [Weinberger 2009]), introducing the rejection condition (*i.e.* classifier returns no matches if all neighbors are beyond a certain distance). This method is referred to as LMNN-R. Further, this technique shows that color correction improves the matching accuracy when Euclidean distance is used to compare images (*i.e.* no learning). It is not the case for learned metrics of LMNN and of LMNN-R, which suggests that a learned metric is more robust in handling illumination changes using original images than using corrected images. The authors propose letting the learning algorithm to handle the color corrections issues.

Discriminative methods (DM)

Enhancement of discriminative power of each individual signature with respect to the others is the main issue in [Lin 2008]. Here, the appearance is encoded by 4-dimensional vectors, containing 3 color components and the height coordinate.

Only the heigh coordinate instead of using 2D spatial coordinates is used to handle viewpoint and pose variations. Features are aggregated using the probability density function (PDF). The distance between two appearances is established using pairwise dissimilarity profiles which are learned beforehand. The nearest neighbour classification is adapted to perform re-identification.

Similarly, in [Schwartz 2009], a rich set of feature descriptors based on color, textures and edges is used to reduce the amount of ambiguity among human class. Features are extracted from overlapping blocks constructing a high-dimensional feature vector. The high-dimensional signature is transformed into a low-dimensional discriminant latent space using a statistical tool called Partial Least Squares (PLS) in *one-against-all* scheme (the discriminative appearance of person is learned using information about the appearances of other persons). For the *one-against-all* scheme, PLS gives higher weights to features located in regions containing discriminative characteristics.

However, discriminative approaches such as [Lin 2008, Schwartz 2009] are often accused of non-scalability. In these approaches an extensive learning phase is necessary to extract discriminative signatures every time when a new person is added to the set of existing signatures. This makes these approaches very difficult to apply in real scenario where in every new minute new people appear.

2.2.1.3 Group context approaches

Approaches which potentially open a new direction, use information from a context environment to improve recognition accuracy. Visual information coming from surrounding people have been used in [Zheng 2009] to reduce ambiguity in person identification. This method shows that group association between two or more people can give valuable information about identity of an individual. It is shown that person re-identification method is improved by utilising associated group of people as visual context. Here, the appearance of group is represented by visual words. SIFT [Lowe 2004] features for each RGB channel are extracted and concatenated for each pixel representation. Then, this concatenated vectors are quantized into n clusters by *k-means* and a code book of n visual words is built. Finally, to represent the spatial distribution of visual words in an image, a *center rectangular ring ratio-occurrence descriptor* (CRRRO) is proposed. CRRRO aims to describe the ratio information of visual words within and between different rectangular ring regions. It is assumed that the distribution of constituent patches of each person in each ring is likely to be more stable against changes in relative positions between the two people over different viewpoints or scaling (see figure 2.10).

Similarly, in [Cai 2010] group context information handled by covariance descriptor [Tuzel 2006] improves the performance of person re-identification. It is shown that contextual cues coming from group of people around a person of interest can

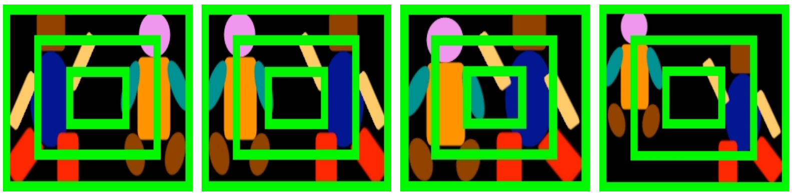


Figure 2.10: Rectangular ring regions used for group association [Zheng 2009]

significantly improve matching accuracy. In this paper, the appearance of a person as well as the appearance of a group have been represented using covariance descriptor which is able to capture statistical properties of features inside an image region.

2.2.1.4 Space-time approaches

In the state of the art approaches we can find that not only visual features can be used to associate appearances between disjoint camera views. In [Javed 2003] a Bayesian framework is proposed to fuse a camera topology and a simple color histogram representation to perform person re-identification. Camera topology is learned using space-time cues including inter-camera time intervals, location of exits/entrances between cameras and directions of movement. The tracking system begins with some prior knowledge gained from an initial training phase. The learned parameters are continuously updated to keep up with the changing motion and appearance patterns throughout the life-time of the system. Given the Bayesian framework, we are able to compute probability of an object entering a certain camera at a certain time using the location, time and velocity of its exit from other cameras. The change in appearance of a person between certain cameras is modelled using a Gaussian distribution of distances between their appearances. As the appearance is represented by color histograms the modified Bhattacharyya distance is applied as a metric to distinguish different appearances. Nevertheless, the assumption that all the colors change similarly is not always true. In order to handle this problem in [Javed 2005] is shown that the inter-camera brightness transfer function (BTF) lies in a low-dimensional subspace, and can be learnt using a set of corresponding calibration objects. However, finding pixel to pixel correspondences from appearances of the same person in two different cameras is not possible due to pose changes and self-occlusion. Hence, normalized histograms of person brightness values were used to compute BTF. The low-dimensional subspace is learnt using probabilistic Principal Component Analysis [Tipping 1999] on BTFs obtained from the training data and used for the appearance matching.

2.2.2 Multiple-shot approaches

Multiple-shot approaches employ multiple images of the same person to obtain a more informative signature. These approaches can also be divided into two classes: feature-oriented approaches and learning approaches. Having several images of the same person enables to apply different kinds of statistics on feature distribution to obtain a reliable signature even in the case of feature-oriented approaches. Moreover, using multiple images per individual, we can apply tools like SVM [Cortes 1995] or Fisher discriminant analysis [Fisher 1936] for separating a specific individual from the rest of people in a given feature space. In multiple-shot approaches, we can also find dimensionality reduction methods which are used to speed up processing and minimize influence of noise for recognition task.

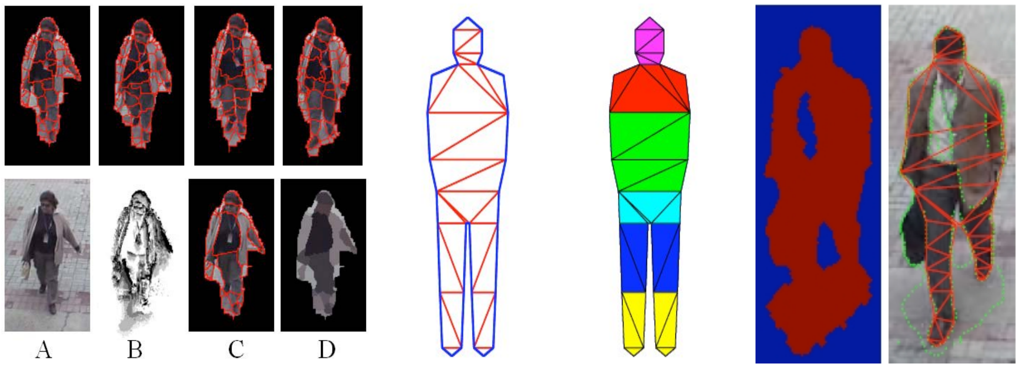
2.2.2.1 Feature-oriented approaches

In the multiple-shot case, the ability of acquiring more than one image for generating signature, allows to apply different kinds of statistics for extracting reliable features. Such techniques as clustering, Bayesian or even PageRank, were investigated to obtain the most reliable and distinctive features.

In [Madden 2007b] an illumination-tolerant histogram represents an appearance of a person. An extended histogram equalization (named *controlled equalization*) is applied to a foreground image to obtain color representation invariant to illumination changes. Then, *k-means* color clustering algorithm extracts the major colors combined into the histogram. The final appearance representation is computed over N frames by incrementally augmenting color clusters. Similarity between histograms is determined by a modified *Kolmogorov* distance.

[Gheissari 2006] proposes the spatiotemporal graph which uses multiple images to group spatiotemporally similar regions. The spatiotemporal segmentation is applied to reject edge information which is temporally unstable. Only edges which are interior to the foreground are considered. Then, a triangulated person model is used to handle a correspondence between different body parts. The person model is represented by a decomposable triangulated graph as a method for model fitting to people. A dynamic programming algorithm is used to fit the model to the image of the person (see figure 2.11). Image regions are compared using color and structural information. The color information is represented by normalized histograms based on hue and saturation. The structural appearance is captured using edge detector (orientation as well as the ratios between RGB color components of the two regions on either side of the edge determine structural qualities).

In [Huang 2009], the appearance of a person is segmented into three parts with height ratios of 2:4:4, and the features are extracted from the lower two parts by ignoring the head (top) part (see figure 2.12 (c)). Then, a color sampling strategy is applied to obtain a tree structure containing medians of colors. In this structure



(a) Spatiotemporal Segmentation (b) Triangulated Model (c) Model fitting

Figure 2.11: Spatiotemporal appearance [Gheissari 2006]. (a) Upper row: segmentation for single frames. Lower row: A) original image, B) frequency image, C) final segmentation after graph partitioning, D) median image for final segmentation; (b) left: an example of a decomposable triangulated graph used as a person model. The blue edges correspond to the boundary of the person while the red edges are interior edges. Right: partitioning of the person into body parts. (c) Left: foreground mask. Right: fitting results.

child histograms are computed by separation of a parent histogram using its median value. The feature vector is obtained by merging median vectors acquired from RGB channels. Finally, a Bayesian-based tracker combines a set of features using a multivariate normal distribution.

In [Hamdoun 2008] re-identification is performed using optimized implementation of SURF [Bay 2008] interest points collected during short video sequences. A signature is built for each detected and tracked person. The accumulated interest point descriptors represent the final signature. The interest points are stored in KD-tree to speed-up the query processing time. The association of the models is obtained by a voting approach: every interest point extracted from the query is compared to all models points stored in the KD-tree, and a vote is added for each model containing the nearest descriptor. Finally the re-identification is performed with the highest vote for the model.

It is worth of noting that interest point descriptors like SIFT [Lowe 2004] or SURF, which are extensively used in computer vision, have not found so much interest in re-identification community. [Gheissari 2006] shows that interest-point-based approach performs significantly worse than color- and region-based descriptor. Interest point descriptors are good at handling serious blurring and handling image rotation but unfortunately they are poor at handling significant changes in viewpoint and illumination.

In [Farenzena 2010], the authors propose to select salient parts of the body by adopting perceptual principles of symmetry and asymmetry. Two horizontal axes

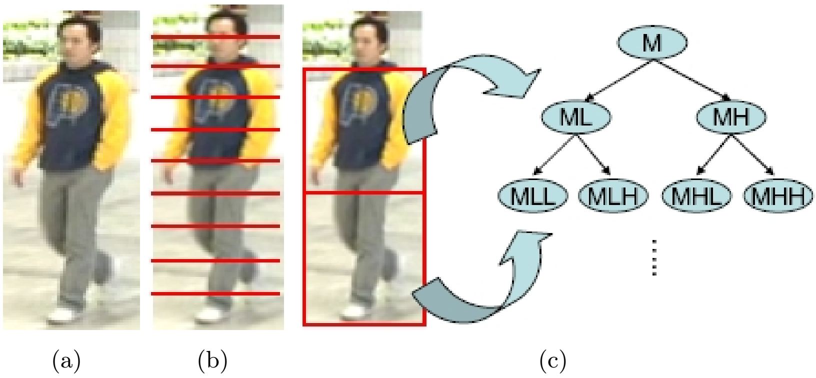


Figure 2.12: Segmented parts of the human body: (a) original image; (b) 10 horizontal strips [Bird 2005]; (c) The tree structure extracted from upper and lower body [Huang 2009].

of asymmetry are used to isolate three main body regions, usually corresponding to head, torso and legs. On the last two, a vertical axis of appearance symmetry is estimated (see figure 2.13). Then, the appearance of each body region is represented by combination of three features: (1) chromatic content (HSV histogram); (2) *maximally stable color regions* (MSCR) [Forssén 2007] and (3) *recurrent highly structured patches* (RHSP). The extracted features are weighted by the distance with respect to the vertical axis to minimize effects of pose variations. The authors name this approach *symmetry-driven accumulation of local features* (SDALF).

Recurrent patches were also proposed in [Bazzani 2010]. The authors use epitome analysis to extract highly informative patches from a set of images. The epitome of an image is its miniature containing the essence of the textural and shape properties of the image [Jojic 2003]. The human appearance is represented by a combination of three features: (1) HSV histogram; (2) local epitome; (3) global epitome. The last two analyze the presence of recurrent patterns in the appearance. In order to discard redundant information from the set of images, for each individual the appearance model is built on images which has been selected using unsupervised Gaussian clustering based on HSV histograms. The matching distance between appearances is represented by a function which combines distances on particular features using different weights.

[Oreifej 2010] performs re-identification in aerial images. Here, the PageRank algorithm is applied to extract the most informative regions to distinguish individuals. PageRank originally was invented to grade websites based on a random walk algorithm which not only gives higher scores to the websites that have more incoming links but also to the pages that are referred to from prominent webpages. Therefore, in a graph of connected webpages, the most informative pages are associated with higher ranks. Here, the same idea is used to seek the most informative image regions from the set of all images. Thus, the multiple images are integrated using

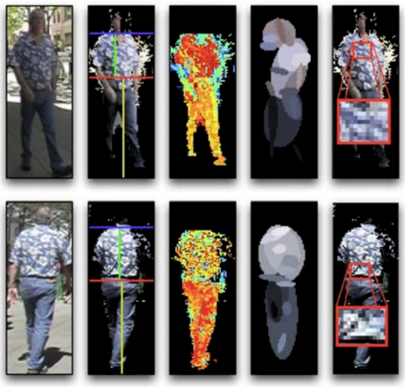
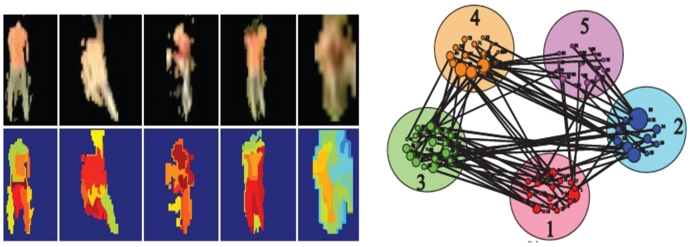


Figure 2.13: Sketch of SDALF approach [Farenzena 2010]: From left in columns: 1) two instances of the same person; 2) x- and y-axes of asymmetry and symmetry, respectively; 3) weighted histogram back-projection (brighter pixels mean a more important color), 4) Maximally Stable Color Regions; 5) Recurrent Highly Structured Patches.



(a) Blobs and their weights (b) Graph for five images

Figure 2.14: The PageRank Weighting [Oreifej 2010]: (a) The first row shows blobs extracted from five images. The second row shows the region weights assigned by PageRank. (b) Outer circles represent images. Inner circles represent regions, where the circle size corresponds to the region weight.

undirected graph and then PageRank assigns higher weights to prominent regions and degrades noisy regions by assigning them lower weights (see figure 2.14).

In [Cheng 2011], the authors show that features are not so important as precise body part detection, looking for part-to-part correspondences. Body parts are detected using *pictorial structures* (PS) [Andriluka 2009], which search for the appearance of body parts by using densely sampled shape context descriptors and discriminatively trained AdaBoost classifiers. In this work a body is composed by 6 parts (chest, head, thighs and legs as in figure 2.15 (a)). After fitting PS, the appearance of each body part is described by HSV histogram and MSCR [Forssén 2007]. Using multiple images for each subject, each body part occurs more times, thus PS fitting can be improved by exploring variety of detections. The authors proposed iterative process (called Custom PS - CPS), which alternates between estimating the body

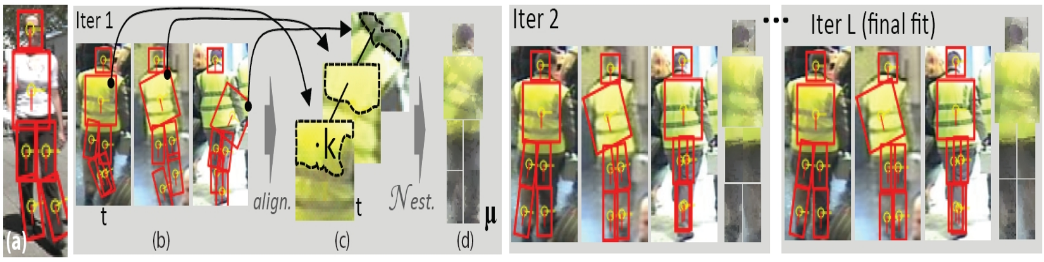


Figure 2.15: CPS method: iterative searching for body part localization [Cheng 2011].

configuration and updating the appearance model. Processing N images (see figure 2.15 (b)), a Gaussian distribution is estimated for all pixels, taking into account their spatial neighbors with similar color (figure 2.15 (c)). PS fitting together with these Gaussian scores form the final body part localization.

2.2.2.2 Learning approaches

In multiple-shot approaches, the underlying assumption is that the knowledge extracted from a training dataset of different individuals could generalize to unseen examples. These approaches employ very often dimensionality reduction methods, Fisher discriminants or SVM classifiers to discriminate between different individuals. We can also find re-identification methods which combine generative and discriminative models. These techniques focus on both aspects - design of descriptive features and enhancement of their discriminative properties [Hu 2008, Hirzer 2011].

Discriminative methods (DM)

In [Hu 2008] a generative and discriminative models are combined together with online learning strategy to perform re-identification. The human appearance is represented by histograms computed over upper body, lower body and full body segment. The histograms are composed of color features, autocorrelograms [Huang 1997] and a bag of features based on SIFT [Lowe 2004] descriptor. Discriminative and generative models are handled by Naive Bayes classifier. For each individual, both the generative models and the discriminative models make decisions about its identity. This classification is performed independently. Then, the final decision is based on the decision fusion module which works in an adaptive weighted voting scheme. The classifier's weights are adjusted accordingly to the correct/wrong decision. True labels of the appearances are obtained by tracking a person and by so called *identity mutual exclusiveness*: if multiple individuals exist in the same scene and if we are sure about ones identity (*e.g.* because of tracking), without loss of generality, we can definitively be sure that other appearances must not be our known identity. This property allows to acquire negative samples which can be used to update discriminative models.

In [Nakajima 2003], a human appearance is represented by a holistic approach based on histograms of different features. The authors propose to use histograms based on color, normalized color, shape features and local shape patterns. All these representations are used to learn multi-class SVMs to estimate pose and to recognize individuals. Each appearance is separated from the rest by one SVM classifier or by a multi-class SVM based on the tree/graph structure of pairwise SVMs. The best performance is obtained using two dimensional normalized color histograms, where $d_1 = R/(R+G+B)$ and $d_2 = G/(R+G+B)$ build two dimensional space (every dimension is represented by 32 bins).

In [Truong Cong 2009], a color feature is considered as the main cue for the identity management. A graph-based approach for a non-linear dimensionality reduction is applied to extract the most informative color representation. Dimensionality reduction is often used for processing of the data, such as classification, visualization and compression. A representation of the data in lower dimension speeds up processing and minimizes influence of noise for classification task. In this work a comparison of color normalization techniques together with different histogram representations is presented. Three color normalization techniques of the RGB color space are compared:

- Gray normalization obtained by dividing the pixel value by the average of the image (or the segmented image - corresponding to the moving object). For each channel we have:

  $$I_k^* = \frac{I_k}{mean(I_k)} \tag{2.1}$$

  where I_k is the color value of the channel k.

- Histogram equalization [Hordley 2005] based on the assumption that the rank ordering of sensor responses is preserved across a change in imaging illuminations.

- Affine normalization, defined as

  $$I_k^* = \frac{I_k - mean(I_k)}{std(I_k)} \tag{2.2}$$

Finally, human appearances are represented by a set of points in a low-dimensional space. Every point is obtained by projecting histogram representation into the new coordinate system computed by the dimensionality reduction method. Since each person is represented by a set of images (sets of points), the dissimilarity between signatures is computed using the distance between centroids of these sets corresponding to signatures of interest. The best performance is obtained using the histogram equalization technique.

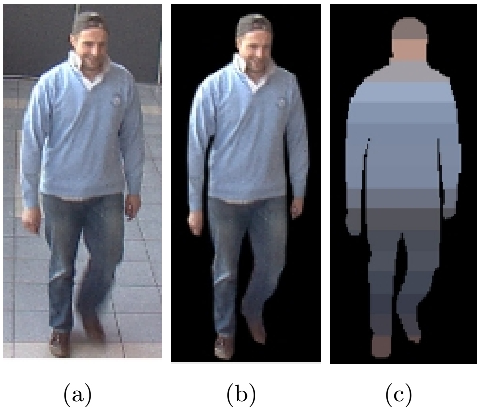


Figure 2.16: Color-position histogram [Truong Cong 2010]: (a) original image; (b) localization of the silhouette; (c) color distribution in the silhouette.

In [Truong Cong 2010], an extension of [Truong Cong 2009] is proposed. A new descriptor for static images called the *color-position histogram* is introduced (see figure 2.16). The silhouette is vertically divided into n equal parts and the mean color is computed to characterize each part. Compared to the classical color histogram, it leads to better results (thanks to the spatial information) and uses less memory. Moreover, a new dissimilarity measure between two sets of points in low-dimensional space is presented. The authors use the optimal margin and the miss-classification error obtained by SVM to compute distance between signatures which can not be separated by a linear model (note that previously, in [Truong Cong 2009], dissimilarity did not take into account separability of points; distance was computed using only position of centroids).

In [Bird 2005] the re-identification is performed to detect loitering individuals in public transportation areas. If a person is present for a long time, it is suspected as a drug dealer. The pedestrian appearance is segmented into 10 equally spaced horizontal strips (see figure 2.12 (b)). Then, the median HSL colour of the foreground pixels of each of these ten strips is used as the feature vector defining a signature. The lightness component is scaled by 0.5 to minimize variance to lighting variations. Then, Fisher linear discriminant analysis is applied to enhance the differences between signatures of the different individuals in the Fisher space.

Bag of features based on SIFT [Lowe 2004] descriptor together with online learning is proposed in [Teixeira 2009] to improve matching accuracy. SIFT vectors are quantized into visual words using the vocabulary tree to define the final space of descriptor. In this work a large vocabulary is a key factor as spatial information of the geometric layout of visual words is not taken into account.

In [Hirzer 2011], every individual is represented by two models: descriptive and discriminative. The discriminative model is learned using the descriptive model as an assistance. The descriptive model is represented by $N = 7$ horizontal stripes and

	single-shot	multiple-shot
feature-oriented	[Park 2006, Gallagher 2008, Cai 2008, Yu 2007, Kang 2004, Wang 2007]	[Madden 2007b, Gheissari 2006, Huang 2009, Hamdoun 2008, Farenzena 2010, Bazzani 2010, Oreifej 2010, Cheng 2011]
learning	[Gray 2008, Prosser 2010, Zheng 2011, Dikmen 2010, Lin 2008, Schwartz 2009]	[Hu 2008, Nakajima 2003, Truong Cong 2009, Truong Cong 2010, Bird 2005, Teixeira 2009, Hirzer 2011]

Table 2.1: Taxonomy of appearance-based re-identification methods.

within each stripe the covariance descriptor [Tuzel 2006] is computed. Employing covariance descriptor, the authors generate an initial ranked list of person images. Images from the end of this list are used as negative samples to learn a discriminative model. Learning is based on boosting, where weak classifiers are built using Haar like features and covariance descriptors. Boosting selects discriminative features, while producing the final ranking list.

2.3 Summary

In this chapter, the previous work on the human re-identification problem has been presented. We show that human re-identification can be considered on different levels depending on information cues which are available in a video analysis system. This dissertation focuses on appearance-based approaches, which are the least accurate among discussed techniques but they are also the least constraint concerning an acquisition environment. Our objective is to perform re-identification on low-resolution data (in most video surveillance scenarios, iris, face or gait characteristics are difficult to acquire). Thus, the main cue for recognition task is obtained by a global appearance, in which clothing provides the most reliable information about an identity.

Global appearance can be modeled by features which has been obtained from only a single image (*single-shot* approaches) or from multiple images (*multiple-shot* approaches). These features should compose invariant representation to human appearance changes across different camera views. The appearance representation can

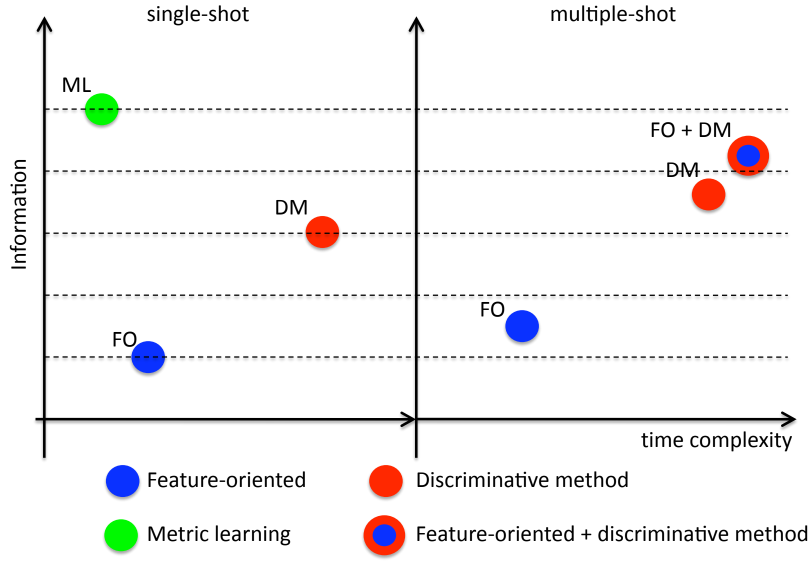


Figure 2.17: State of the art appearance-based re-identification techniques as a relationship between information provided for computing signature and time complexity.

be extracted using *feature-oriented* procedures or by *learning* approaches (see table 2.1).

Appearance-based re-identification techniques can also be considered in the context of information used for generating a signature representation and time complexity (see figure 2.17). We classify *feature-oriented* approaches that use only a single image as techniques which use the least information among all appearance-based re-identification techniques. These approaches are based only on expertise of a designer and a single image. We rank *metric learning* class higher than *discriminative approaches* (more information is needed to generate a signature), because for learning a metric we need knowledge given by images of the same human registered in different cameras, while for discriminative approaches training data can be obtain automatically by applying spatiotemporal constraints. Concerning time complexity of *metric learning* approaches, we did not take into account offline learning phase, while plotting the graph.

In this thesis, we propose several methods, investigating different levels of information. We propose *single-shot* as well as *multiple-shot* approaches, exploring different discriminative techniques. However, before explaining details of the proposed techniques, in the next chapter we present an overview of the framework for human re-identification.

CHAPTER 3

Human re-identification framework overview

"*A journey of a thousand miles must begin with a single step.*" (Lao Tzu)

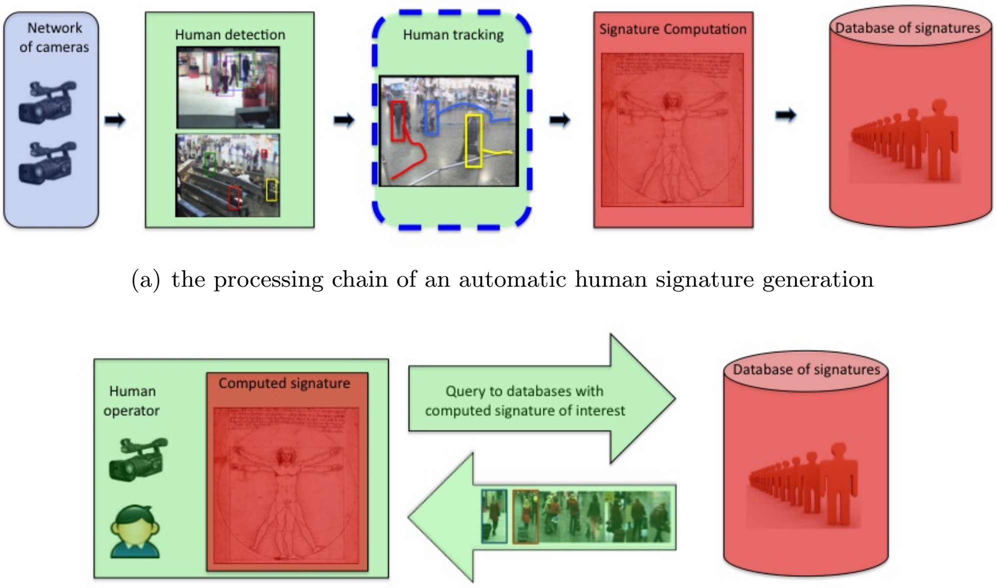
A typical way of extracting an appearance model in an automatic surveillance system is by first detecting an object in an image, and then by tracking it using different strategies. In this chapter, we describe a general framework for a human re-identification system, indicating the steps in the processing chain, which are the main focus of this dissertation. We also present a short discussion on different appearance models for detection and tracking, as these steps are essential to automatically extract a human appearance from a video content.

3.1 Human re-identification framework

This section describes a general framework for a human re-identification system. In figure 3.1, we illustrate the main components which are necessary for a human operator to use a re-identification system. Usually, a video surveillance system consists of number of cameras which are distributed over an area of interest (*e.g.* airport, metro stations, car parks, etc.). The data from such network of cameras can be processed in parallel or by a central unit depending on the hardware architecture of the system.

In every video analysis system, the first task is to determine whether the moving regions in an image belong to a predefined object class (*e.g.* a human) in the scene. This task is called *object detection* and it plays a very important role. The results of this step have a significant influence on the next stages in the processing chain. In this thesis we are only interested in objects which are humans.

Human detection (section 3.2) is the first module in our re-identification framework. Then, *human tracking* can be applied (section 3.3) to obtain a set of image regions corresponding to the tracked object. In a *single-shot* case, tracking is optional as only a single image is used to create a human signature. However, in most video analysis systems, even if the whole trajectory is not used for signature computation, the object is usually tracked to obtain the most informative appearance (called *key*



(a) the processing chain of an automatic human signature generation

(b) an interaction of a human operator with a database of signatures

Figure 3.1: The re-identification framework. We highlight the main interest of dissertation by red color.

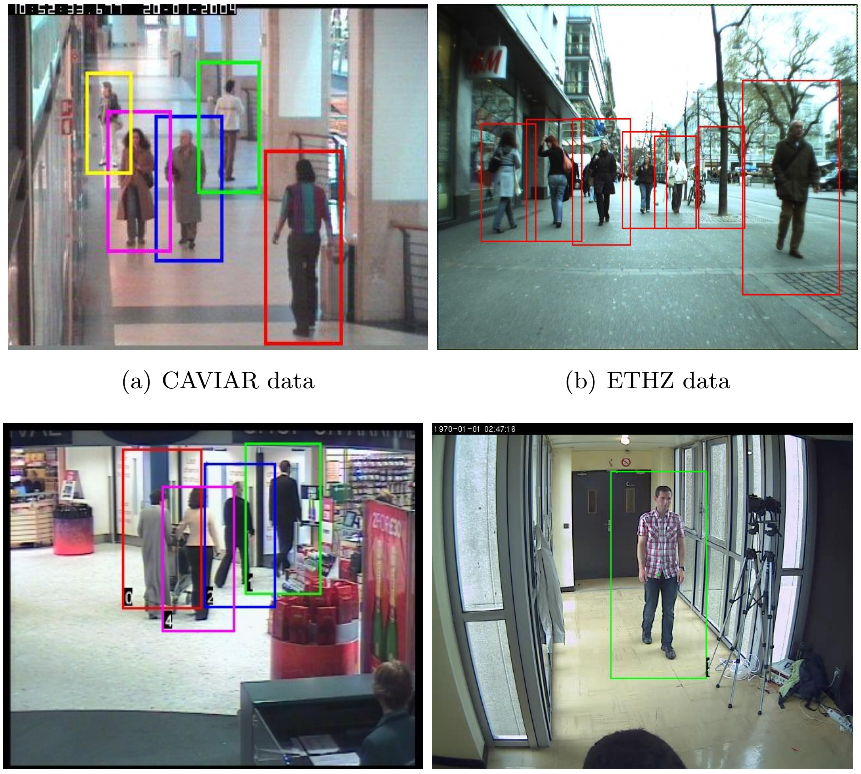
frame). The key frames are supposed to contain all useful information of a video sequence.

After detection and tracking of a person, the system computes a human signature. Computed signatures are stored in a database. This provides an effective interface for a human operator to search the most similar signatures to a signature of interest (figure 3.1(b)). In this dissertation, we mainly focus on the signature computation step (chapter 4, chapter 5 and chapter 6) and also we propose a solution for efficient storage of signatures in a database (chapter 8).

Human detection and *human tracking* are fundamental for computing human signatures. As recognition accuracy of re-identification algorithms is dependent on these two steps, in the next two sections we shortly overview state of the art methods for detecting and tracking humans.

3.2 Human detection

Detecting different categories of objects in an image or in a video content is one of the fundamental tasks in computer vision. Two main steps can be distinguished in a typical object detection algorithm. The first task is *feature extraction*, in which the most informative object descriptors regarding the detection process are obtained from the visual content. The second task is *classification*, in which the obtained



(a) CAVIAR data

(b) ETHZ data

(c) i-LIDS data

(d) TSP data

Figure 3.2: Examples of human detection results. 2D bounding rectangles indicate image regions which have been classified as positive samples.

object descriptors are used to classify the object of interest.

Human detection is considered among the hardest examples of object detection problems. The articulated structure and variable appearance of the human body, combined with illumination and pose variations, contribute to the complexity of the problem. Most of the leading approaches in human detection are discriminative methods, which use machine learning techniques such as *support vector machines* (SVMs) [Cortes 1995] or *boosting* [Freund 1997], to differentiate a human appearance from the rest of the world. These methods have become increasingly popular since they can cope with high-dimensional spaces and they are able to select relevant descriptors among a large set. The feature extraction methods for human detection mostly focus on *dense representations*, where descriptors are obtained inside a detection window. The image is scanned densely at different scales and a learned classifier of a human model is evaluated. The exemplary results of a human detector performed on different datasets are illustrated in figure 3.2.

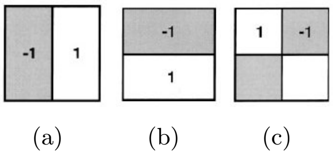


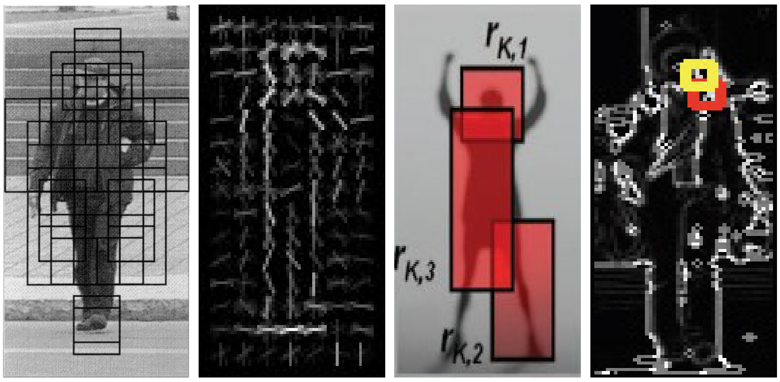
Figure 3.3: The three types of 2-dimensional non-standard Haar wavelets: (a) vertical; (b) horizontal; (c) diagonal.

Human detection approaches

One of the first robust representations of human class is based on **Haar wavelet** transform [Papageorgiou 2000]. Images are mapped from the space of pixels to an over-complete dictionary of Haar wavelet features. Haar wavelet based descriptors encode intensity differences between two adjacent regions, providing a rich description of the pattern (see figure 3.3). Such patterns are scaled and moved over a detection window to generate the over-complete dictionary of Haar wavelet features. This coding of local intensity differences at several scales provides a flexible and expressive representation that can characterize a human class (see figure 3.4(a)). Instead of using the over-complete dictionary, a small set of important features can be chosen via a greedy selection method using AdaBoost [Viola 2001].

Integral image (a new image representation introduced by [Viola 2001], see section 4.2.3) together with cascade structure of boosting, allow to process images extremely fast with a high detection rate. Haar wavelet features have become increasingly popular due to this efficient computation. Further, this work was extended in [Viola 2005] to extract Haar wavelets from video content (space-time differences in video) to integrate intensity information with motion information. Patterns of motion and appearance were extracted to build a robust model of walking humans. The results have shown that combination of both sources can outperform the original Haar wavelet descriptor.

In [Dalal 2005], an effective human detector is described. SVM classifier together with a densely sampled *histogram of oriented gradient* (**HOG**) inside the detection window has improved the state of the art performance. Even though, this approach is computationally more complex than Haar wavelets, it has obtained significantly outperforming results. The method is based on evaluating well-normalized local histograms of image gradient orientations in a dense overlapping grid. The main idea is to characterize an object appearance and a shape by the distribution of local intensity gradients or edge directions. The image window is divided into small spatial regions (called *cells*). For each cell, 1-D histogram of gradient directions or edge orientations is accumulated. Histograms extracted from all cells are merged into the feature vector which stands for the appearance representation. These vectors are used to learn SVM classifier to create a human detector. The descriptors weighted by the positive SVM weights are illustrated in figure 3.4(b).



(a) Haar (b) HOG (c) Covariance (d) Granules

Figure 3.4: Human detection approaches:(a) Haar wavelets [Papageorgiou 2000]; (b) HOG descriptor weighted by the positive SVM weights [Dalal 2005]; (c) three possible covariance descriptor sub-windows [Tuzel 2008]; (d) granules comparing gradient orientations [Duan 2009].

Robust human detection algorithm using covariance matrices as object descriptors is presented in [Tuzel 2008]. Here, a dense model of **covariance features** extracted inside a detection window is proposed. Covariance descriptor is computed from sub-windows with different sizes sampled from different locations (see figure 3.4(c)). Then, a boosting mechanism selects the best regions characterizing a human silhouette. Unfortunately, it is not trivial to build a classifier working in the space of covariance matrices (see section 5.2.1). Using covariance matrices, we also influence significantly computational complexity. The classification step involves the eigenvalue decomposition to compute the logarithm of a symmetric matrix which requires $O(d^3)$ arithmetic operations, where d is the size of covariance matrix (*e.g.* the processing of a 320 × 240 image requires 3 seconds).

Recently, [Duan 2009] proposed a novel approach to boost a set of *associated pairing comparison features* (APCFs) called **granules**, which are represented by square window patches (see figure 3.4(d)). The main idea is to take advantage of comparing features from different locations to search meaningful correlations. The Granular space increases enormously when granules with different sizes are compared in different locations and at different scales. Even if we use simple features to characterize a square patch [Huang 2010], the size of granular space is so tremendous that conventional exhaustive search methods used for Haar-wavelets, HOG, or covariance features are inapplicable to select discriminative granular features. Thus, to alleviate this issue, heuristic approaches have to be applied to speed up the feature selection process. In spite of significant computational complexity during the learning phase, the classification step is not so time consuming (detection can be almost as fast as using Haar wavelets) and the accuracy outperforms existing approaches.

In our re-identification framework, we have decided to use HOG-based detector. The

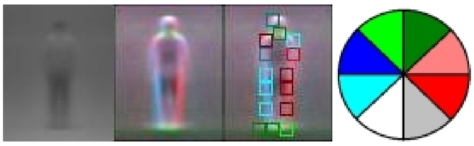


Figure 3.5: Mean human image with corresponding edge magnitudes and the 15 most dominant cells. From the left: the first image shows the mean human image calculated over all positive samples in the database; the second image shows the corresponding mean edge magnitude response; the third image shows this later image superposed with the 15 most dominant cells of size 8 × 8 pixels. The cell bounding boxes are drawn with a color set by their most dominant edge orientation with the scheme defined in the last image.

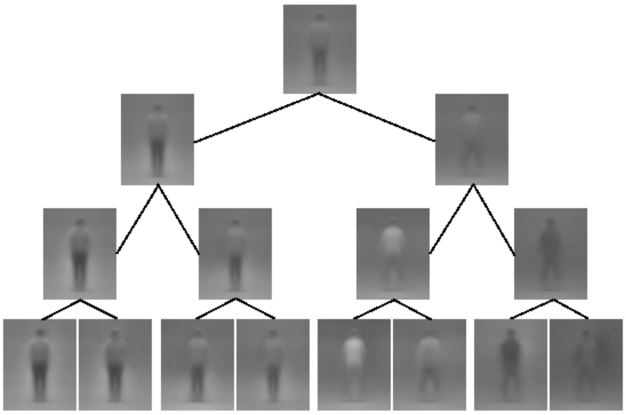


Figure 3.6: The tree of human appearances.

choice of HOG features is a compromise between time complexity and descriptive power. Our HOG technique is adapted from a face detection method [Corvee 2009] to detect human silhouette. The detection algorithm extracts *histograms of oriented gradients*, using Sobel convolution kernel, in a multi-resolution framework. The technique was originally designed to detect faces using the assumption that facial features remain approximatively at the same location. However, locations of human silhouette features do not remain constant through out the time with varying poses (*e.g.* knees are constantly changing position when walking; a shoulder changes position from walking to slightly bending when pushing a trolley). Hence, we have modified the algorithm to detect humans using cells located at specific locations around the human silhouette as shown in figure 3.5. These most dominant cells are the cells having the closest HOG vector to the mean HOG vector calculated over the vectors (of the corresponding cell) from a human database.

We integrate these dominant cells using *decision tree learning* [Corvee 2010] based on hierarchical classification. Figure 3.6 shows an example of tree, where each image is the average of the training samples of corresponding people with similar features. Finer level of granularity of human posture is reached as we go down the

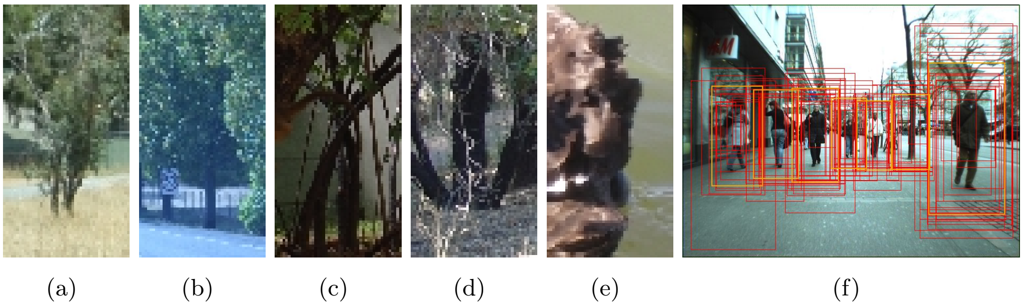


Figure 3.7: Human detection issues: (a)-(e) false positives, appearances which are very similar to a human silhouette; (f) numerous candidates of detected humans.

tree of features. The first node (root) corresponds to the most dominant cell among all possible cells accumulated by using the whole training dataset Two sub nodes are initiated by splitting the database in two according to the distribution of cell orientation error. We do not go into details concerning learning as it is not the main topic of this thesis. The interested reader is referred to [Corvee 2010].

Human detection issues

As we have already mentioned, the articulated structure of the human body, illumination changes and pose variations make the human detection problem one of the most challenging examples of object detection. In real scenario, a human detection algorithm has to handle partial occlusions, shadows and scale changes. Moreover, often classifier fails on the appearance which is very similar to a human silhouette (see figure 3.7 (a)-(e)). Finally, classifier can provide a lot of detections (rectangles around the desired object, see figure 3.7 (f)) from which the best candidate has to be chosen. In our detection framework, we select the best candidate using confidence level returned by classifier. Having a set of overlapping detections, we compute weighted average using confidences as weights of rectangles, producing the final detection (yellow rectangles in figure 3.7 (f)). The selection of the best candidate is significant as it has an influence on the next step in the processing chain: *human tracking*.

3.3 Human tracking

Tracking of objects in complex real world scenes involves dealing with multi targets, complicated occlusions, and cluttered or moving backgrounds. In complex and dense scenarios, many objects may have similar appearance. Occlusions, such as object self-occlusion, occlusion between multiple objects, and occlusion by other static scene objects happen very often. As significant improvement in human detection algorithms has been achieved, detection-based tracking methods gain more and more attentions. Especially for complex environments, where detection-based approaches are more flexible and robust (automatic initialization, no issues with camera movements).

Unfortunately, even if the object detectors can perform in such conditions relatively well, the tracking of particular objects still remains challenging. Detection performance is usually a trade off between the detection rate and the false alarm rate. False alarms, missed detections and inaccurate responses of the detector (detected bounding boxes are not accurately centred around the desired object, only part of the object is detected) happen frequently in the detection procedure which provides misleading information to tracking algorithms. Detection-based human tracking must overcome these failures of the human detector. Moreover, the difficulties caused by occlusions and similar appearance among multiple humans do not facilitate the task.

Human tracking approaches

In [Xing 2009], a detection-based tracking algorithm is presented. This approach aims at overcoming the limitations of human detector to track multiple objects through occlusions. The tracking performs in two-stage mode (local and global). In the local stage, the tracking is improved by: (1) a human partition method which focuses on the upper human body at three different levels (head-shoulder, head-torso, full body); (2) a particle filter which is used to deal with partial object occlusions for generating reliable tracklets. In the global stage, the detection and tracking results are collected from a temporal sliding window to deal with ambiguity caused by full object occlusion. The tracklets generated in the local stage, guided by the potential tracklets generated within the temporal sliding window are associated by *Hungarian* algorithm [Kuhn 1955]. In order to calculate the association cost matrix, the likelihood between two tracklets is extracted using appearance, shape and motion attributes of the tracklet. The appearance is represented by color histogram, the shape is described by the object height and the motion model is characterized by its velocity. The idea of two-stage approach allows to seek both, the local optimum trajectory for each object and the global optimum trajectories for all the tracked objects.

Occlusion can also be treated as a scenario in which an object leaves and re-enters the scene. In this case [Snidaro 2008] approach can be applied. The method is based on the on-line boosting framework which learns a specific model for each individual. The model is used to locate a target in the next frame and then the observed positive sample is used to update the model by on-line learning. The negative samples are obtained using randomly selected background regions. The specific model for each instance of the class allows at the same time robust tracking and recognition of the particular instance as it leaves and re-enters the scene. The

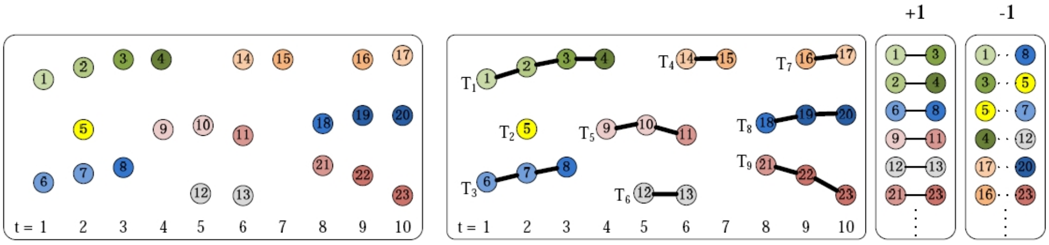


Figure 3.8: The overview of the process of obtaining on-line training samples. From the left: the detection responses in t frames; the result of reliable tracklets based on spatiotemporal correlations; positive training samples; negative training samples, [Kuo 2010].

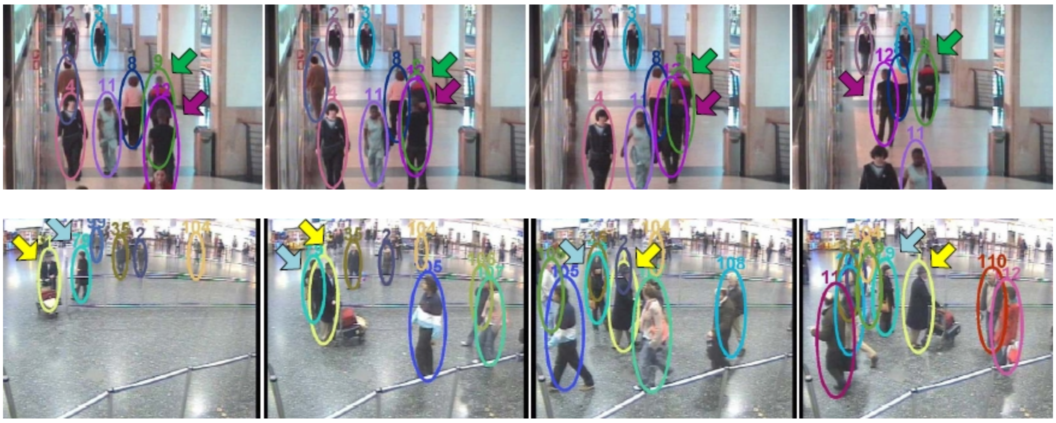


Figure 3.9: Example of tracking results on CAVIAR (top row) and i-LIDS (bottom row) dataset [Kuo 2010].

human model is represented by Haar wavelet features and Local Binary Patterns (LBP) [Ojala 2002]. Haar features encode the generic shape of the object and LBP features capture texture details. Both classifiers are employed, exploiting the on-line boosting approach. The Haar-based classifier is mainly used to detect a generic target class of objects, while the boosted LBP features are used to discriminate the texture of the single instance. The boosting uses just background images as the negative examples, which means that both classifiers are mostly focused on discriminating between the target and the background.

On-line learning of discriminative appearance model for robust multi-target tracking is also the main topic in [Kuo 2010]. The proposed human model is designed to distinguish different targets from each other, rather than from the background. The general model of human class is learned using training samples collected on-line from tracklets within a time sliding window. Positive samples are represented by pairs of detection responses within the same tracklet (tracklets are collected using spatiotemporal relationships). Negative samples are collected by extracting pairs of detection responses from tracklets which do not belong to the same target (see figure 3.8). The similarity between two appearances (pair of detection responses) is computed using color histograms, covariance matrices and HOG descriptors extracted from image patches at different locations and different scales. The distances between these patches are used by Adaboost algorithm to learn a strong classifier to discriminate different appearances. Examples of tracking results are illustrated in figure 3.9.

3.4 Conclusion

This chapter has described a general framework for human re-identification. We have presented the processing chain of an automatic video analysis system, discussing the main steps for extracting a human appearance from a video content. We have also highlighted components in the processing chain, which are the main focus in this dissertation.

Further, we have presented different models for human detection and tracking, as these tasks are fundamental for computing automatically a human signature. We have also showed that these tasks are related to each other and they have to cope with the articulated structure of the human body, complicated occlusions, illumination changes and pose variations. All these issues make human detection, human tracking and human re-identification extremely challenging tasks in a video analysis system.

In the following chapters we focus on extracting a human appearance from cropped images provided by detection and tracking components. We assume that each image is a region of interest centered on the human body. The next chapter presents *single-shot* approaches which extract human characteristics using only a single image.

Chapter 4

Single-shot human re-identification

"*All models are wrong; some models are useful*" (George Box)

Appearance-based re-identification techniques that focus on associating pairs of images, each containing one instance of an individual, are named *single-shot* approaches. These methods extract human characteristics from a single image to perform human re-identification. It is worth noting that these approaches can also be applied to *multi-shot* case, by extending the signature matching to *many-to-many* scenario. In this chapter, we propose two *single-shot* approaches: DCD and SCR. The former method is based on *dominant color descriptor* (DCD) extracted from the upper and the lower body parts detected by using an asymmetry-driven approach. While the latter employs *covariance matrix descriptor* together with a new spatial pyramid scheme to characterize the human body parts detected by *histogram of oriented gradients* (HOG). This method is referred to as *spatial covariance regions* (SCR).

4.1 DCD signature

In this section we present a human appearance model based on *dominant color descriptor* (DCD). Figure 4.1 illustrates phases of our approach. As an input we take cropped image obtained by a human detector. Then, a *background subtraction* technique (section 4.1.1) is applied to extract image regions corresponding to an object of interest (*foreground*). Afterwards, we normalize color (section 4.1.2) to obtain color invariant representation. Finally, we extract our DCD signature as a human appearance descriptor (section 4.1.4).

4.1.1 Background subtraction

Background subtraction is the most fundamental image processing operation in video analysis application. In video surveillance, the camera mostly focuses on objects of interest (foreground objects), like people, vehicles, abandoned luggages. So far, many different methods have been proposed to isolate objects from the rest

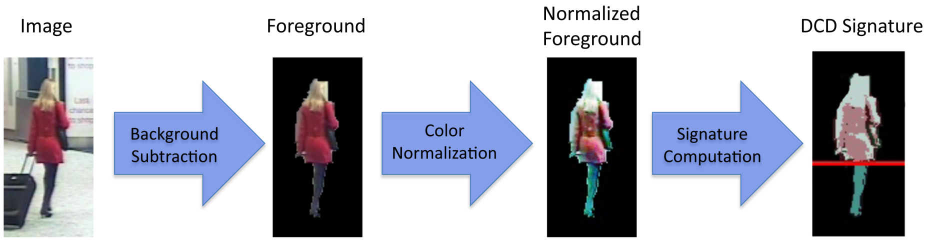


Figure 4.1: Extraction of DCD signature.

of the image. When we work on video content the very simplest background subtraction method is to subtract the current frame from the previously learned model of the background. This model may correspond to a reference image (background image given before processing) or can be extracted by learning methods, such as mixture of Gaussians, kernel density estimation or bag of words. These models might also be based on motion detection (frame differencing) so that, when an object moves an algorithm is able to identify both, object regions and background regions, and update a model respectively.

Unfortunately, aforementioned approaches do not perform very well in crowded environments, such as airports or metro stations, where an object appearance can be altered by shadows and significant illumination changes. Besides, in many contexts, we wish to be able to extract a foreground without having a separate background image (*e.g.* foreground/background segmentation in still images). In this particular case, we can apply segmentation tools which use either texture (color) information or edge information to separate foreground and background regions. In [Boykov 2001] an approach based on optimization by graph-cut has been developed which successfully combines both types of information. This approach has been extended by GrabCut algorithm [Rother 2004].

GrabCut is based on foundation of graph cut, which is an optimization technique using a graph theory for solving energy minimization problems. Energy minimization problems can be reduced to instances of the maximum flow problem in a graph, which corresponds to finding a minimal cut of the graph.

The graph cut algorithm formulates foreground extraction as a binary labeling problem, which assigns to each node in a graph, *i.e.* pixel, a unique label (foreground or background). The solution is obtained by minimizing a energy (cost) function

$$E(A) = \lambda \sum_{p \in P} R(A_p) + \sum_{\{p,q\} \in N} B(A_p, A_q), \tag{4.1}$$

where $A = (A_1, \ldots, A_p, \ldots, A_{|P|})$ is a binary vector whose components A_p specify label assignments to pixels p in P. Each A_p can be either foreground or background. Coefficient λ encodes a relative importance of the region properties term $R(A_p)$ *vs.*

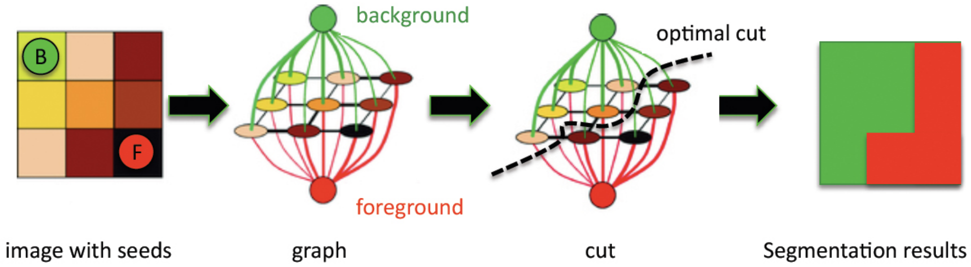
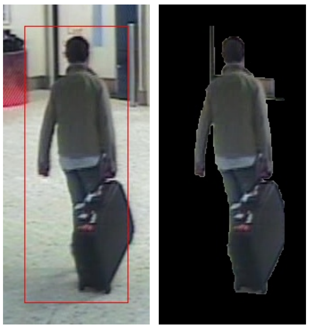


Figure 4.2: Graph cut: a simple $2D$ segmentation example. The seeds B (background pixel) and F (foreground pixel) are necessary for initialization step. These cues can be provided by a user or by automatic foreground detectors. The cost of each edge is reflected by the edge's thickness. The edges between pixels corresponds to $B(A_p, A_q)$ term, and the edges connected to foreground and to background are defined by $R(A_p)$.

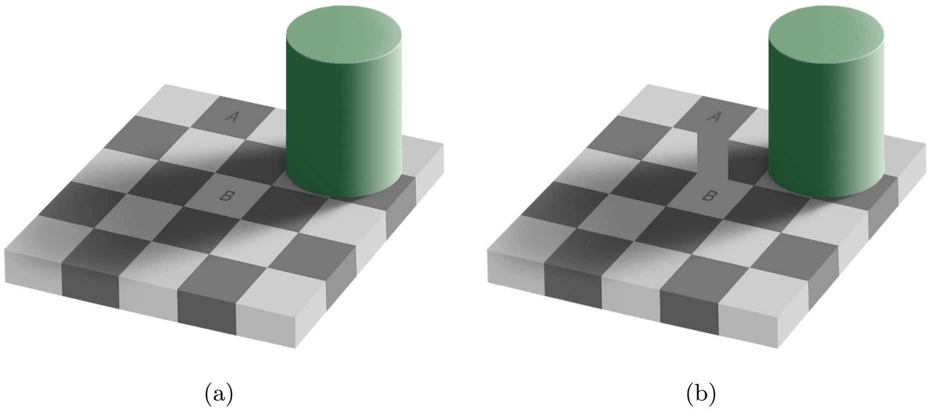


(a) input (b) output

Figure 4.3: GrabCut: the red rectangle can be given either by a user or by a human detector.

the boundary properties term $B(A_p, A_q)$. $R(A_p)$ is a likelihood that pixel p belongs to foreground or background. The likelihood is estimated as the color similarity of the pixel's color to the color distribution of the areas marked by a user as foreground or background. $B(A_p, A_q)$ is a penalty term when adjacent nodes are assigned to different labels. The more similar the colors of the two nodes are, the larger $B(A_p, A_q)$, and thus the less likely the edge is on the object boundary. Figure 4.2 illustrates a graph cut based segmentation. The user has to initialize segmentation by indicating some certain pixels of foreground and background regions.

GrabCut, optimized version of graph cut, allows to initialize segmentation without specifying foreground regions. GrabCut is based on iterative minimization algorithm, which deals with missing foreground pixels by putting provisional labels on some pixels (in the foreground) which can subsequently be retraced. Only background regions, which are specified by the user, are taken to be firm - guaranteed not to be retraced later. In GrabCut, background regions are determined by the



(a)

(b)

Figure 4.4: Optical illusion: (a) square A appears darker than square B, when in fact they are both exactly the same color; (b) prove: a rectangle of the same color has been drawn connecting square A and B. In this example, humans perceive the squares as having different reflectance, and interpret the colors as different.

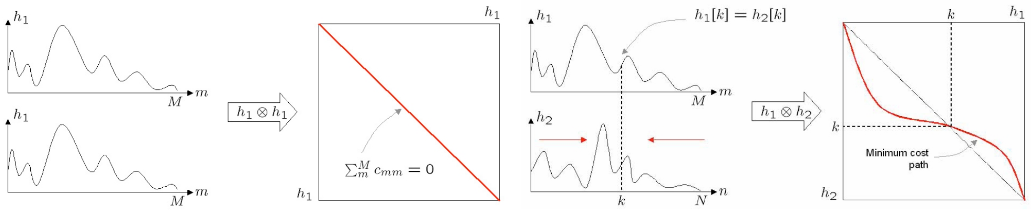
user as a strip of pixels around the outside of the marked rectangle (marked in red in figure 4.3).

In our case, we can provide such rectangle using the output of a human detection algorithm. Applying GrabCut, we extract foreground regions of a human silhouette. Then, we represent a human appearance by *color* feature.

4.1.2 Color feature

Color is the visual perceptual property related to a phenomenon that takes place in a human eye. Human perception of color derives from the spectrum of light interacting with cone cells in the retina eye and depends on spectral sensitivity of these cells. We associate color with objects, materials and light sources, based on their physical properties such as light absorption, reflection, or emission spectrum. Consequently, light conditions and physical characteristics of object, have a significant influence on color perception. Moreover, color perception is relative to the context in which it is viewed and depends on subjective color experience (see figure 4.4).

Concerning acquisition devices like cameras, characteristics of color sensors are often far from characteristics of receptors in a human eye. Additionally, even identical cameras which have the same optical properties and are working under the same lighting conditions may not match in their color responses (slight differences in spectral sensitivity). Different physical characteristics of cameras, different lighting conditions and different environments cause significant color dissimilarities in appearances of the same object acquired by different cameras. This issue is one of the most challenging problems using color as a feature while two appearances from different cameras are compared. Fortunately, we can reduce color dissimilarities by



(a) (b)

Figure 4.5: BTF [Porikli 2003]: (a) minimum cost path for the same histograms; (b) minimum cost path for the different histograms.

applying either *color calibration* [Javed 2005, Porikli 2003] or *color normalization* [Buchsbaum 1980, Gevers 2004, Madden 2007a, Hordley 2005] techniques.

Color calibration

In order to minimize a significant differences in illumination of the same object acquired by different cameras, *color calibration* techniques search a *brightness transfer function* (BTF) that maps color distributions from one camera to another. One possible solution to this problem is proposed by [Porikli 2003]. The method is based on correlation matrix analysis and dynamic programming. The authors assume that images of the same object are available for different cameras. Then, the correlation matrix is computed from a pair of 1-D color histograms. BTF is obtained from this matrix by finding a minimum cost path during bins mapping. Correlation matrix C between two histograms $h_1[m]$ and $h_2[n]$, where $m = 1 \ldots M$ and $n = 1 \ldots N$ correspond to bins, is defined as:

$$C_{M \times N} = h_1 \otimes h_2 = \begin{bmatrix} c_{1,1} & c_{1,2} & \cdots & c_{1,N} \\ c_{2,1} & c_{2,2} & \cdots & c_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ c_{M,1} & c_{M,2} & \cdots & c_{M,N} \end{bmatrix} \tag{4.2}$$

where each element $c_{m,n}$ is the distance between the corresponding histogram bins. The minimum cost path from the $c_{1,1}$ to $c_{M,N}$ is found by dynamic programming. The sum of the matrix elements on the path gives the minimum score among all possible routes (note that this mapping may not be one-to one). Finally, this path is converted to the inter-camera model function (see figure 4.5).

Unfortunately, this mapping is not unique and it varies from frame to frame depending on a large number of parameters that include illumination, scene geometry, exposure time, focal length, and aperture size of each camera. It is worth noting, that color dissimilarities already appear during tracking a person in a single camera. Thus, BTF function should also take into account spatial position of the appearance during color mapping. Additionally, as BTF maps color spaces from one camera to another, we need to compute such mapping for each pair of cameras in the network. In large camera networks, this operation is impracticable, therefore in our approaches we propose to use *color normalization* technique.

Color normalization

In contrast to calibration techniques, color normalization is focused on an invariant color description. Normalization approaches seek transformations of the image data into the illuminant and device independent representation. This topic has been investigated in the last few decades bringing many invariant color representations (Greyworld normalization [Buchsbaum 1980], kernel density estimation [Gevers 2004], controlled equalization [Madden 2007a]). In our approaches, a technique called *histogram equalization* [Hordley 2005] leads to the best performance.

Histogram equalization (HE) is based on the assumption that the rank ordering of sensor responses is preserved across a change in imaging conditions (lighting or device). Histogram equalisation is an image enhancement technique originally developed for a single channel image. The aim was to increase the overall contrast by brighting dark areas of the image, emphasizing details in those regions. Histogram equalization achieves this aim by stretching range of histogram to be as close as possible to a uniform histogram (see figure 4.6). Cumulative distribution function is used to spread out distribution values. The approach is based on the idea that amongst all possible histograms, a uniformly distributed histogram has maximum entropy [Gonzalez 2001]. Maximizing the entropy of a distribution we maximize its information and thus histogram equalization maximizes the information content of the output image. We apply the histogram equalization to each color channel (RGB) to maximize the entropy in each channel and to obtain the invariant image. Figure 4.6 illustrates the effect of applying the histogram equalization to: (1) images of the same individual captured by different cameras; (2) Gaussian distribution. These images highlight the fact that a change in illumination leads to a significant change in the colors captured by the camera. Normalized images are much more similar than the two original images.

4.1.3 Dominant color descriptor

Dominant color descriptor (DCD) was proposed by MPEG-7 for image retrieval [Deng 2001, Yang 2008]. The main idea of this descriptor is based on the observation that a small number of colors is usually sufficient to characterize the color information in an image region. DCD is defined as:

$$F = \{\{c_i, p_i\}, i = 1, \dots, N\}, \tag{4.3}$$

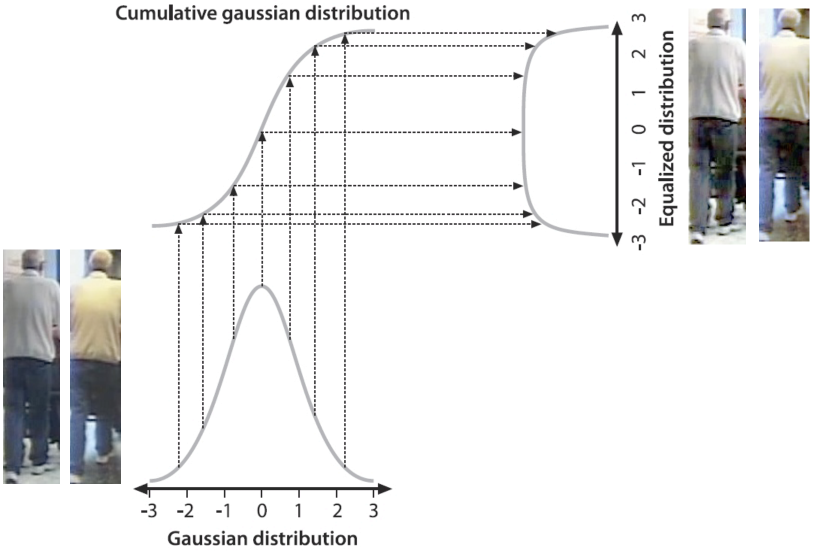


Figure 4.6: Histogram equalization: two images and Gaussian distribution equalized using the cumulative density function.

where N is the total number of dominant colors in the image, c_i is a 3-D color vector, p_i is its percentage, and $\sum_i p_i = 1$ (see figure 4.7). Given N, DCD divides color space of an image into N clusters. As DCD is quite sensitive to the number of clusters N, we evaluate this descriptor using different settings (see section 7.3.1).

Distance between two dominant color descriptors F_1 and F_2 is defined using the improved dissimilarity measure [Yang 2008]:

$$D^2(F_1, F_2) = 1 - \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} a_{ij} S_{ij}, \tag{4.4}$$

where $S_{ij} = [1 - |p_i - p_j|] * \min(p_i, p_j)$ and

$$a_{ij} = \begin{cases} 1 - d_{ij}/d_{max}, & d_{ij} \leq T_d \\ 0, & d_{ij} > T_d \end{cases} \tag{4.5}$$

where d_{ij} is the Euclidean distance between two colors. Threshold T_d is the maximum distance for two colors to be considered as similar ($i.e.$ threshold T_d is the maximum distance used to judge whether two color clusters are similar), and $d_{max} = \alpha T_d$. Let $\alpha = 1$, and consider the case when the Euclidean distance of two color clusters is slightly smaller than maximum distance. From equation (4.5), the similarity coefficient a_{ij} between the two color clusters will be very close to zero; therefore, it

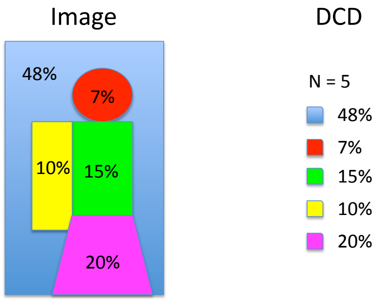


Figure 4.7: Example of an artificial image with the dominant colors and their percentage values.

cannot clearly distinguish between the colors exceeding the maximum distance. In order to properly reflect similarity coefficient between two color clusters, we set parameter $\alpha = 2$ and $T_d = 25$ (the same values of parameters were assumed in [Yang 2008]).

4.1.4 Asymmetry-driven human body separation

In our approach we use dominant colors to create a human signature. The main limitation of dominant color descriptor is that it does not handle a spatial relationship between regions in an image. We overcome this issue by dividing a person silhouette into the two main body parts: the upper body part and the lower body part. The body separation is obtained by our asymmetry-driven approach. A straightforward way would be to simply use fixed partitions based on the bounding box. However, as human detection does not always return accurately centred bounding box, we propose to use a search strategy to refine the separation line.

We assume that the human body can be often characterized by asymmetry appearance (in most cases colors corresponding to the upper body are different than colors corresponding to the lower body). As this assumption is not always true, we constrain the search of an asymmetry axis ψ to vertical extension $[m - \varepsilon, m + \varepsilon]$, where m corresponds to the horizontal axis which separates the extracted silhouette into the two equal areas defined on each side of the border (see figure 4.8).

The separation quality is evaluated using the dissimilarity measure between two dominant color descriptors extracted from regions obtained by this separation. We choose the axis which maximize this dissimilarity. Finally, the human signature is represented by two dominant color descriptors corresponding to the upper body part (U) and the lower body part (L). Defining the similarity between two human signatures $\mathfrak{s}_A$ and $\mathfrak{s}_B$, we compute a sum of dissimilarities obtained by comparing

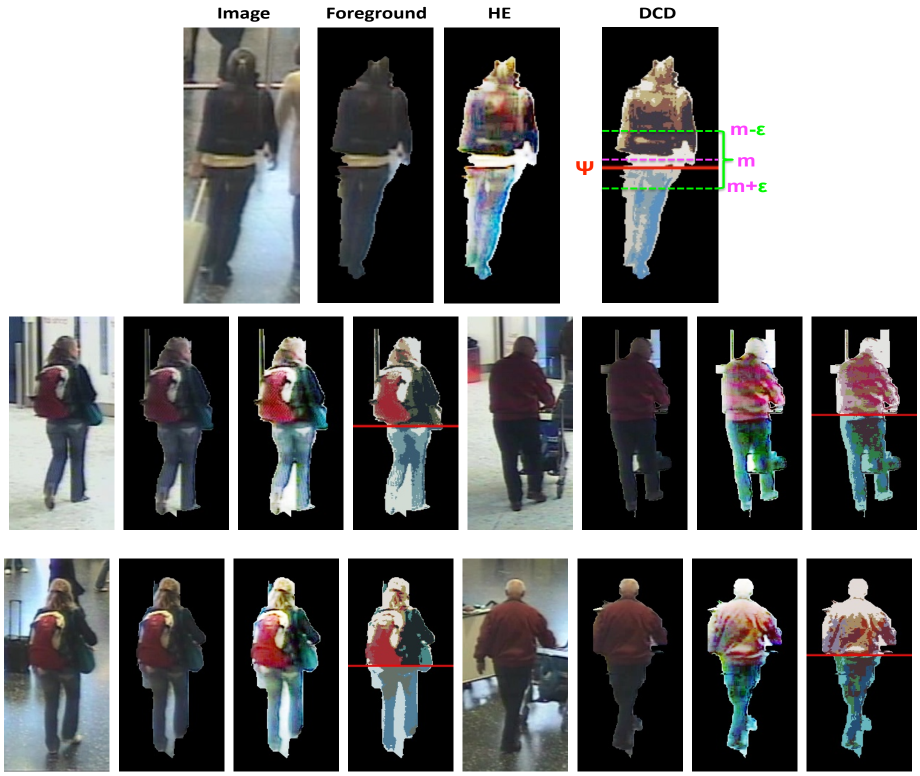


Figure 4.8: Asymmetry-driven DCD signature. The top row presents all phases of our appearance extraction method. The bottom rows illustrate examples of DCD-based signatures extracted from i-LIDS-40 images.

corresponding body parts

$$S(\mathfrak{s}_A, \mathfrak{s}_B) = \frac{1}{D^2(U_A, U_B) + D^2(L_A, L_B)}. \tag{4.6}$$

4.2 SCR signature

This section describes a human appearance model based on *spatial covariance regions* (SCR) extracted from human body parts. We present the sketch of our approach in figure 4.9. Having a cropped image, we detect human body parts using HOG-based detector (section 4.2.1). These body parts are used to address *the correspondence problem* between appearances registered in different cameras. Once the body parts are detected, the next step is to handle color dissimilarities caused by camera and illumination differences. Thus, we applied *histogram equalization* to obtain invariant color representation. Then, on such normalized image, the covariance regions (see section 4.2.2) of body parts are computed at different grid cell resolutions to

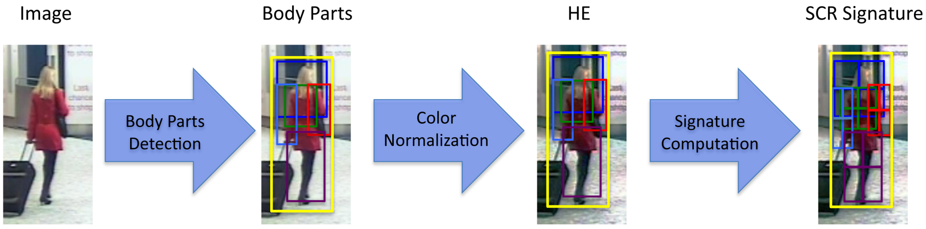


Figure 4.9: Extraction of SCR signature.

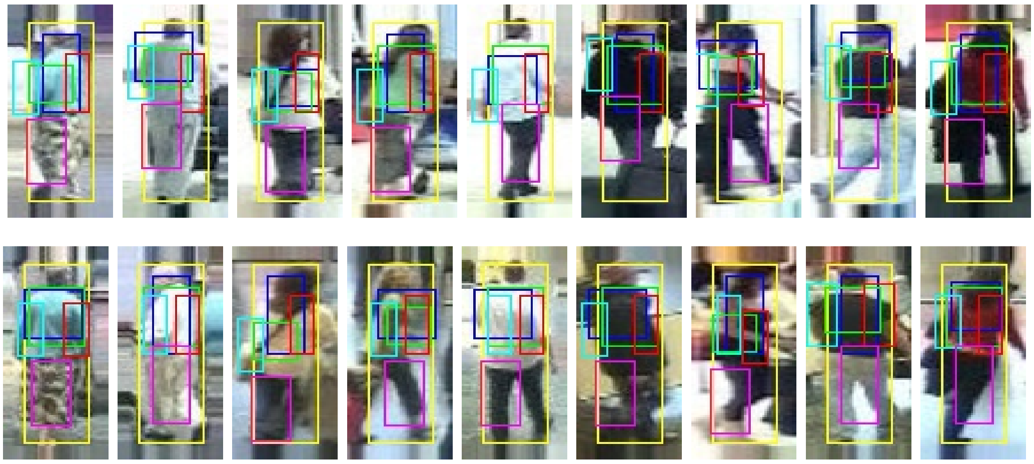


Figure 4.10: Illustration of human detection and body part detection results. Detections are indicated by 2D bounding boxes. Colors correspond to different body parts: the full body (yellow), the top (dark blue), the torso (green), legs (violet), the left arm (light blue) and the right arm (red). Every column shows two instances of the same person acquired by different camera.

generate a human signature. Furthermore, the dissimilarities between these regions corresponding to different images are combined using an idea derived from the spatial pyramid match kernels (section 4.2.4).

4.2.1 Body part detection

In order to detect body parts of a person we have applied a hierarchical tree of histogram of oriented gradients [Corvee 2010]. We have used the same HOG detector as described in section 3.2, where it was specialized for the full human body detection. In the case of body parts, we have trained independently classifiers on various areas of a person, corresponding to five body parts: top, torso, legs, left arm and right arm (see figure 4.10).

Having these five body part detectors we are able to localize corresponding body parts during matching appearances acquired from different cameras. We represent an appearance of the body part using *covariance matrix descriptor*.

4.2.2 Covariance matrix descriptor

In [Tuzel 2006] covariance of d-features has been proposed to characterize a region of interest. The descriptor encodes information of the variances of defined features inside the region, their correlations with each other and a spatial layout.

Covariance matrix can be computed from any type of image such as a one dimensional intensity image, three channel color image or even other types of images, *e.g.* infrared. Let I be an image and F be a d-dimensional feature image extracted from I

$$F(x, y) = \phi(I, x, y), \tag{4.7}$$

where function ϕ can be any mapping, such as color, intensity, gradients, filter responses, *etc.* For a given rectangular region $Reg \subset F$, let $\{f_k\}_{k=1...n}$ be the d-dimensional feature points inside Reg. Each feature point f_k is characterized by function ϕ. We represent region Reg by the $d \times d$ covariance matrix of the feature points

$$C_{Reg} = \frac{1}{n-1} \sum_{k=1}^{n} (f_k - \mu)(f_k - \mu)^T, \tag{4.8}$$

where μ is the mean of the points.

Such defined a positive definite and symmetric matrix can be seen as a tensor. The main problem is that such defined tensor space is a manifold that is not a vector space with the usual additive structure (do not lie on Euclidean space). Hence, many usual operations, such as mean or distance, need a special treatment. Therefore, our covariance manifold is specified as Riemannian to determine a powerful framework using tools from differential geometry [Pennec 2006].

We use the distance definition proposed by [Förstner 1999] to compute the dissimilarity between two covariance matrices C_i and C_j

$$\rho(C_i, C_j) = \sqrt{\sum_{k=1}^{d} \ln^2 \lambda_k(C_i, C_j)}, \tag{4.9}$$

where $\lambda_k(C_i, C_j)_{k=1...d}$ are the generalized eigenvalues of C_i and C_j, determined by

$$\lambda_k C_i x_k - C_j x_k = 0, \qquad k = 1 \dots d \tag{4.10}$$

and $x_k \neq 0$ are the generalized eigenvectors.

4.2.3 Integral image and fast covariance computation

Integral image is an intermediate image representation proposed in [Viola 2001]. This representation allows to compute the sum of features inside any rectangular region in constant time ($O(1)$), significantly speeding up the feature extraction process.

The integral image $\mathfrak{I}$ of the image I contains at location x, y the sum of the pixels above and to the left of x, y, inclusive:

$$\mathfrak{I}(x, y) = \sum_{x' \leq x, y' \leq y} I(x', y'). \tag{4.11}$$

We can also use this representation for fast calculation of covariance features [Tuzel 2006]. Let us write (i, j)-th element of the covariance matrix defined in equation (4.8) as

$$C_{Reg}(i, j) = \frac{1}{n-1} \sum_{k=1}^{n} (f_k(i) - \mu(i))(f_k(j) - \mu(j)). \tag{4.12}$$

Expanding the mean and rearranging the terms we can write

$$C_{Reg}(i, j) = \frac{1}{n-1} \left[\sum_{k=1}^{n} f_k(i) f_k(j) - \frac{1}{n} \sum_{k=1}^{n} f_k(i) \sum_{k=1}^{n} f_k(j) \right]. \tag{4.13}$$

Using *integral images* we are able to compute the sum terms in constant time, thus computation of the covariance matrix in region Reg takes also constant time. However, to perform such computation, we have to pre-compute the sum of each feature dimension $f(i)_{i=1 \ldots n}$, as well as the sum of the multiplication of any two feature dimensions, $f(i)f(j)_{i,j=1 \ldots n}$. Hence, we need to construct d integral images for each feature dimension $f(i)$ and $\frac{d^2+d}{2}$ integral images for each multiplication. Factor 2 comes from commutative property of multiplication operation ($f_k(i) f_k(j) = f_k(j) f_k(i)$). In total we need

$$d + \frac{d^2 + d}{2} = \frac{d(d+3)}{2} \tag{4.14}$$

integral images to perform fast covariance computation. As integral image can be computed in one pass over the original image [Viola 2001], the time and memory complexity of constructing the integral images is $O(d^2 WH)$.

4.2.4 Spatial pyramid matching

This section describes how the human signature is computed and how the dissimilarity between two signatures is obtained. First, we introduce the notion of *pyramid match kernel*. Then, the signature levels are described. Finally, the dissimilarity functions between two signatures at each level are defined.

Pyramid match kernel

The original formulation of pyramid matching has been proposed in [Grauman 2005]. The pyramid matching allows for precise matching of two collections of features in a high dimensional appearance space. Unfortunately, it discards all spatial information. Hence, in [Lazebnik 2006] an orthogonal approach (pyramid matching in the two-dimennsional image space) has been proposed.

Let us assume that X and Y are two sets of vectors in a d-dimensional feature space. The pyramid matching finds an approximate correspondence between these two sets. We construct a sequence of grids at resolutions $0, \dots, L$. The grid at level l has 2^l cells along each dimension, for a total of $D = 2^{dl}$ cells.

Informally, pyramid matching works by placing a sequence of increasingly coarser grids over the feature space and taking a weighted sum of the number of matches that occur at each level of resolution. Intuitively, we penalize matches found in larger cells because they involve increasingly dissimilar features. Matches found at finer resolutions are weighted more highly than matches found at coarser resolutions.

The pyramid match kernel is defined as

$$\kappa^L(X,Y) = \mathcal{I}^L + \sum_{l=0}^{L-1} \frac{1}{2^{L-l}}(\mathcal{I}^l - \mathcal{I}^{l+1}) \tag{4.15}$$

$$= \frac{1}{2^L}\mathcal{I}^0 + \sum_{l=1}^{L} \frac{1}{2^{L-l+1}}\mathcal{I}^l, \tag{4.16}$$

where $\mathcal{I}^l$ is the matching function at level l (in original formulation the matching function was represented by histogram intersection).

Spatial matching scheme

In our approach the matching function is based on a similarity defined using the covariance matrix distance. In figure 4.11 grid levels are presented. Level 0 corresponds to the full body. Level 1 consist of detected body parts: top, torso, left arm, right arm and legs. Finally, level 2 is described by grid cells inside detected body parts. The human signature is represented by the set of covariance matrices computed at all levels.

The matching function at level 0 is defined as

$$\mathcal{I}^0 = \frac{1}{\rho(C_i, C_j)}, \tag{4.17}$$

where C_i and C_j are covariance matrices computed on regions corresponding to the full body of individual i and individual j, respectively.

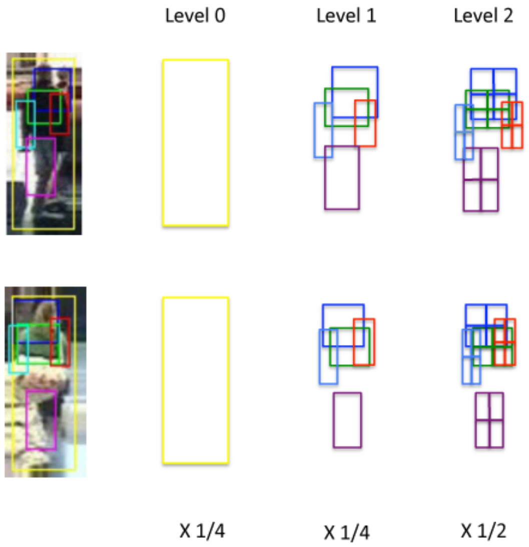


Figure 4.11: Example of constructing a three-level pyramid. Level 0 corresponds to the full body. Level 1 and level 2 correspond to the rest of detected body parts and grids inside body parts, respectively. Finally, we weight each level according to equation (4.16).

Concerning following levels, we define the matching function as

$$\mathcal{I}^l = \frac{1}{\mathcal{D}^l}, \tag{4.18}$$

where $\mathcal{D}^l$ is a dissimilarity function at level l between two sets of covariance matrices $\mathfrak{C}^i$ and $\mathfrak{C}^j$ computed inside regions of interest. Let us assume that $\mathcal{M}_z$ is the set of z largest ρ distances between corresponding covariance matrices. Then, the dissimilarity functions is defined as

$$\mathcal{D}^l = \frac{\sum_{k=1}^{n(l)} \rho(C_k^i, C_k^j) - \sum_{m \in \mathcal{M}_z} \rho_m}{n(l) - z}, \tag{4.19}$$

where i and j correspond to individuals and $n(l)$ is the number of compared covariance matrices. The introduction of $\mathcal{M}_z$ increases robustness towards outliers coming from possible occlusions. In experimental evaluation we set $z = \frac{n(l)}{2}$.

Finally, the similarity between two signatures extracted from images I_i and I_j, is defined as

$$S(I_i, I_j) = \frac{1}{\kappa^L(I_i, I_j)}. \tag{4.20}$$

4.3 Conclusion

This chapter introduced two *single-shot* approaches for re-identificaiton: DCD signature and SCR signature.

The former method extracts a human silhouette from a still image using a graph cut based approach. Applying *histogram equalization*, we obtain color representation invariant to camera changes. The foreground region is divided into the upper and the lower body parts, by using an asymmetry-driven approach. Both body parts are then represented by dominant color descriptor.

The latter method employs the human body part detector for solving *the correspondence problem* during matching two appearances from different cameras. Again, we apply *histogram equalization* to handle color dissimilarities caused by camera and illumination differences. Appearances of body parts are represented by covariance descriptor to enhance invariant representation. Such extracted and normalized human appearances are matched using the spatial pyramid matching.

Both aforementioned techniques belong to *feature-oriented* methods, which usually assume *a priori* appearance model and focus on designing novel features for capturing the most distinguishing aspects of an individual. The presented approaches are dependent on body part detection accuracy where inaccurate detections can significantly deteriorate recognition performance.

In the next chapter we propose more general techniques which learn an appearance model using training data of multiple people. Applying machine learning techniques, we can obtain an appearance representation, which is not constraint to the fixed and predefined model but it is chosen during the training process. Further, using a discriminative learning, our models focus only on discriminative features of an individual providing higher recognition accuracy.

Chapter 5

Human re-identification by boosting

"*The only source of knowledge is experience.*" (Albert Einstein)

We, humans learn continuously to deal with the factors of associability and discriminability, for recognizing people. Inspired by *human memory*, the re-identification system should learn how to associate human appearances and how to focus on distinctive cues (features) of a specific individual, providing within-group discriminations. Defining machine learning, we can say that: *"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E"* [Mitchell 1997]. Following the idea of learning to distinguish people, this chapter offers two discriminative approaches. The proposed techniques generate a human signature by employing multiple images of many humans. Both methods use *one-against-all* learning scheme to extract a discriminative representation of the human appearance. It means that distinctive characteristics of a specific individual are enhanced by using information from images of other individuals. Learning is performed by *boosting* - a meta machine learning algorithm. We employ *Haar-wavelets* (section 5.1) and *covariance descriptors* (section 5.2) as the features for describing image regions.

5.1 Binary classification in Haar-like feature space

This section presents the human re-identification approach based on Haar-like features. Figure 5.3 shows the key steps of our technique. This approach belongs to the *multiple-shot* case, where multiple images of the same person are used to compute a signature. Given a set of images corresponding to the same individual obtained by tracking (section 3.3), we extract *foreground* regions using background subtraction technique (section 4.1.1). These foreground regions serve as positive samples in *one-against-all* learning scheme. Negative samples are formed by the rest of humans acquired in the same camera. By using Haar-like features (section 5.1.1) and *AdaBoost* (section 5.1.2), we learn a signature for a specific individual. The result signature is represented by a strong classifier given by *AdaBoost*. Similarity

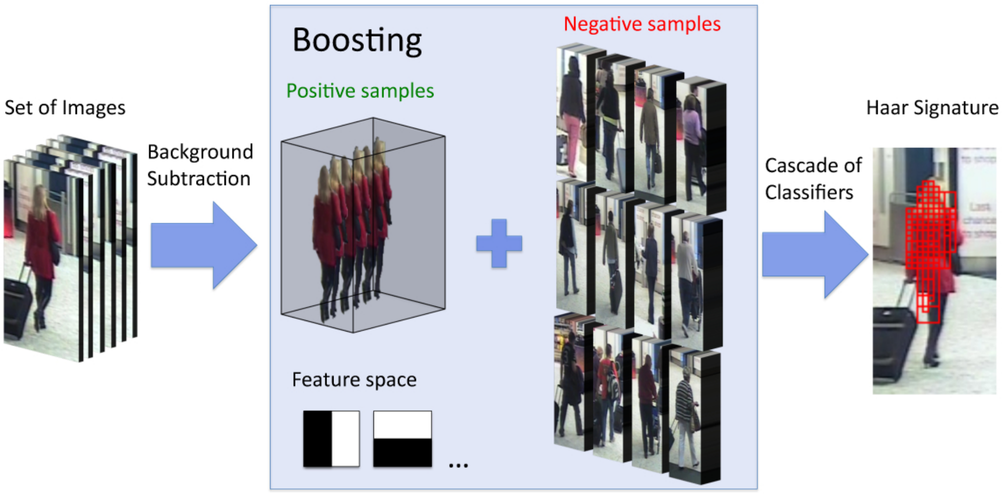


Figure 5.1: Extraction of Haar signature.

measures for comparing signatures are investigated in section 5.1.3.

5.1.1 Haar-like features

Haar-like features inspired by [Papageorgiou 1998] compose rectangular patterns based on the weighted sums of pixels. We use an extended set of Haar-like features [Lienhart 2002], which significantly enrich the basic set proposed by Viola and Jones [Viola 2001]. Figure 5.2 illustrates feature prototypes, which are scaled and shifted independently in vertical and horizontal direction in order to generate an over-complete set of features. This set of features is called as over-complete because it has much more elements than a basis window size, *e.g.* given the base resolution of 20×40 pixels, the tremendous number of Haar-like features $(435, 750)$ has to be considered.

Let us define a Haar-like feature as

$$f_{\mathfrak{R}} = \sum_{r \in \mathfrak{R}} \omega_r s(r), \tag{5.1}$$

where weights $\omega_r \in \{-1, +1\}$ and $s(r)$ is the pixel intensity sum of rectangular region r, $\mathfrak{R}$ is the set of rectangular regions describing the feature prototype. This Haar-like feature encodes intensity difference between adjacent regions, providing a flexible and expressive representation that is able to describe the pattern. Instead of using the over-complete dictionary, a learning algorithm can select a small set of important features to characterize an object class.

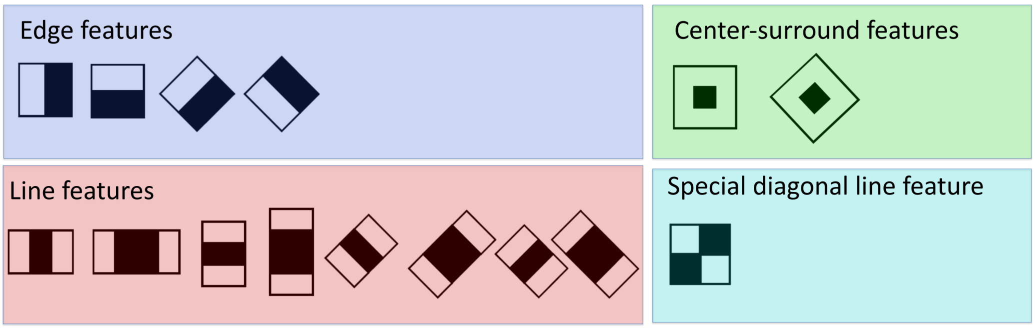


Figure 5.2: Feature prototypes of simple Haar-like and center-surround features. Dark areas have negative weights while bright areas have positive weights [Lienhart 2002].

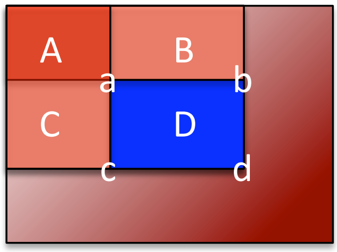


Figure 5.3: Calculation of the pixel sum of rectangle D. This can be done with four array references: a, b, c, d are the values of the integral image at given positions. Thus, $a = s(A)$, $b = s(A) + s(B)$, $c = s(A) + s(C)$, $d = s(A) + s(B) + s(C) + s(D)$. The sum within D can be computed as $s(D) = d + a - b - c$.

5.1.1.1 Fast feature computation

In section 4.2.3, we have already introduced an intermediate image representation called *integral image*. This representation allows to compute the sum of features inside any rectangular region in constant time ($O(1)$).

The integral image $\mathfrak{I}$ of the image I contains at location x, y the sum of the pixels above and to the left of x, y, inclusive:

$$\mathfrak{I}(x, y) = \sum_{x' \leq x, y' \leq y} I(x', y'). \tag{5.2}$$

By using the *integral image*, we can compute any rectangular sum in four array references (see figure 5.3). The difference between two rectangular sums can be computed in eight references. Since the *edge features* involve adjacent rectangular sums they can be computed in six array references, eight in the case of the *line features*, and nine for the *special diagonal line features* (see figure 5.2).

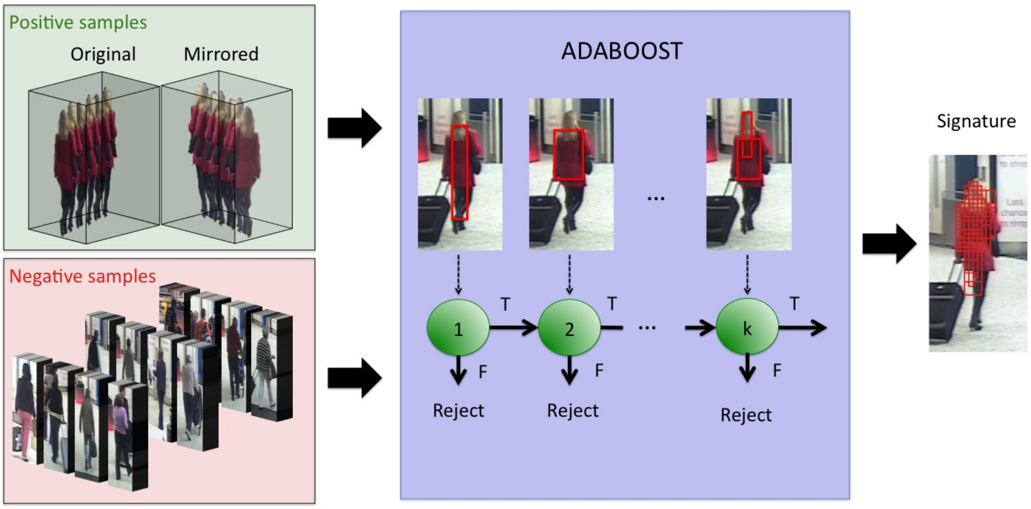


Figure 5.4: Haar signature: cascade of k strong classifiers. At each stage a new strong classifier is trained to eliminate additional negatives, requiring additional computation.

5.1.2 Learning a signature

We learn a signature of a specific individual using *one-against-all* learning scheme. Figure 5.4 presents the sketch of the learning phase.

5.1.2.1 Training data

Given a set of images corresponding to the same individual tracked in a video sequence, we generate positive samples for a learning algorithm. Positive samples are obtained by cropped images of the tracked individual and their mirror counterparts. Mirror images bring a larger set of positive samples and enable to create pose-invariant signature. Negative samples are formed by cropped images corresponding to the rest of humans tracked in the same camera.

5.1.2.2 Haar-like feature set reduction

We scale every image into a fixed size of 20×40 pixels. Given this base resolution, the tremendous number of Haar-like features needs to be considered $(435, 750)$. Even though Haar-like features can be calculated efficiently using an *integral image*, this number of features makes the learning phase time consuming. By employing Haar-like features only for foreground regions, we speed up significantly the learning process. We filter out meaningless Haar-like features and decrease the features set

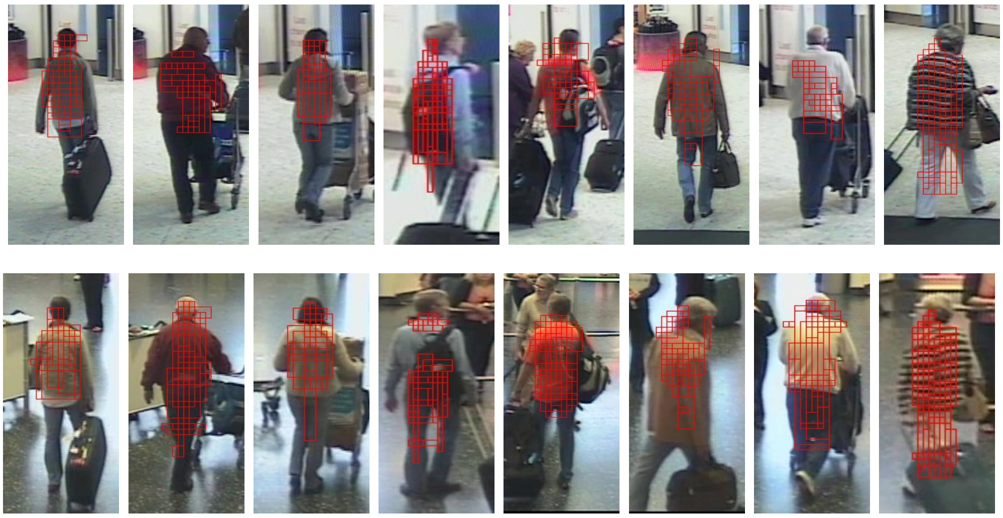


Figure 5.5: Illustration of Haar-based signatures.

from 435, 750 to around 20, 000 features depending on the size of foreground region. This huge decrease of the feature set is obtained by ignoring all patterns which intersect the background area, and at the same time by considering only features which are inside the foreground area.

5.1.2.3 AdaBoost

Given a feature set and a training set of positive and negative samples, any number of machine learning approaches could be used to learn a signature. Similarly to [Viola 2001], we develop AdaBoost to select a small set of relevant features and to train the signature, as this technique has successfully been applied to human detection.

AdaBoost refers to *adaptive boosting*, which is a machine learning algorithm, introduced by [Freund 1995]. In its original form, the AdaBoost is used to boost the classification performance of a simple learning algorithm (often called *weak classifier*). AdaBoost maintains a discrete distribution (set of weights) over the training examples, and selects a weak classifier via the weak learning algorithm at each iteration. AdaBoost is adaptive in the sense that training examples that are misclassified by the weak classifier at the current iteration, receive higher weights at the following iteration. Thus, the subsequent classifiers which are built, focus on instances misclassified by previous classifiers. The accuracy of *weak classifiers* have to be only better than random guessing (*e.g.* in the case of binary classification, an error has to be less than 0.5), to allow the AdaBoost to improve the accuracy, by producing a final linear combination of weak classifiers (often called *strong classifier*). This strong classifier works similar to *majority voting strategy*, where the final classification decision depends on a weighted majority hypothesis in which the weight of

- Given example images $(x_1, y_1), \dots, (x_n, y_n)$, where $y_i \in Y = \{-1, +1\}$, for negative and positive examples respectively.
- Initialize weights $w_{1,i} = \frac{1}{2m}$ and $w_{1,j} = \frac{1}{2l}$ for $y_i = -1$ and $y_j = +1$, where m and l are the number of negatives and positives respectively.
- For $t = 1, \dots, T_s$:
  1. Normalize the weights,

     $$w_{t,i} = \frac{w_{t,i}}{\sum_{j=1}^{n} w_{t,j}}, \tag{5.3}$$

     so that w_t is a probability distribution.
  2. Train each weak classifier h_j using distribution w_t.
  3. Calculate the error of h_j: $\varepsilon_j = Pr_{i \sim w_t}[h_j(x_i) \neq y_i]$.
  4. Choose the classifier h_t with the lowest error ε_t.
  5. Set $\alpha_t = \frac{1}{2} \ln \left(\frac{1-\varepsilon_t}{\varepsilon_t} \right)$
  6. Update the weights

     $$w_{t+1,i} = w_{t,i} \exp(-\alpha_t y_i h_t(x_i)) \tag{5.4}$$

- The final strong classifier is

  $$\mathcal{H}(x) = \operatorname{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x) \right). \tag{5.5}$$

Figure 5.6: The adaptive boosting algorithm (AdaBoost).

each weak classifier is a function of its accuracy.

Figure 5.6 presents the boosting algorithm. We adapt this boosting scheme to our signature learning, by defining a weak classifier as a threshold function based on a Haar-wavelet response. Term T_s stands for *stop criterion*, by which AdaBoost stops adding weak classifiers to the final strong classifier (figure 5.5 visualizes learned Haar-based signatures).

5.1.2.4 Weak classifier

Our *weak classifier* $h_j(x)$ consists of feature f_j, threshold θ_j and parity $\mathfrak{p}_j$ indicating the direction of the inequality sign

$$h_j(x) = \begin{cases} +1, & \text{if } \ \mathfrak{p}_j f_j(x) < \mathfrak{p}_j \theta_j \\ -1, & \text{otherwise} \end{cases} \tag{5.6}$$

Feature f_j corresponds to a single Haar-like feature. In contrary to Tieu and Viola [Tieu 2004], we do not use Gaussian models to compute the threshold of the weak classifier. In our approach the threshold computation is based on the concept of information entropy used as a heuristic to produce the smallest set of features. This idea is originated from Quinlan [Quinlan 1986]. The threshold separates the training data, while maximizing the *information gain* (*mutual information*).

5.1.2.5 Cascade of strong classifiers

Similarly to [Viola 2001], we develop a cascade of classifiers generated by a boosting scheme. Combining increasingly more complex classifiers into a cascade structure, we reduce radically computation time. The key idea is that smaller, and therefore more efficient, boosted classifiers can be constructed which disregard many of the negative individuals while accepting almost all positive instances (*i.e.* the stop criterion T_s of a boosted classifier can be adjusted so that the false negative rate is close to zero). Simpler classifiers are used to reject the majority of individuals before more complex classifiers are called upon to achieve low false positive rates.

Cascade is a degenerated decision tree (the list), in which a positive result from one classifier triggers the evaluation of the following one (see figure 5.4). A negative outcome at any strong classifier leads to the immediate negative classification of an image.

Stages in the cascade are constructed by training classifiers using AdaBoost and by adjusting the stop criterion T_s to minimize false negatives. Subsequent strong classifiers are trained using those samples which pass through all the previous stages. As a result, the successive classifiers face a more difficult task than the preceding classifiers. Consequently, it appears that the deeper classifier, the more complex it is. The strong classifiers are added to the cascade until the overall target for false positive and false negative rate is met (stopping criterion T_c). The final cascade of strong classifiers stands for a human signature.

5.1.2.6 Stop criteria

Having a boosting classifier, we always have to define term T_s, which stands for stop criterion of adding weak classifiers to the strong classifier. Moreover as we use the cascade of strong classifiers, we also have to provide stopping criterion T_c of adding new stages in the cascade. Both criteria are based on classification accuracy obtained on training data.

Given training data $X = \{(x_1, y_1), \dots, (x_n, y_n)$, we split it into two parts: $X' = \{(x_1, y_1), \dots, (x_{2n/3}, y_{2n/3})\}$ and $X'' = \{(x_{2n/3+1}, y_{2n/3+1}), \dots, (x_n, y_n)\}$. X' is used to learn classifiers and X'' is used to stop boosting algorithm. Testing our classifier on these dataset, we can obtain TP, TN,FP, FN, which correspond to *true positive*, *true negative*, *false positive* and *false negative* samples. We build Haar signature (classifier) by applying the following two stopping criteria:

- T_s: the boosting stops adding new weak classifiers to the strong classifiers if

$$(TP''_t < TP''_{t-1}) \wedge (TN''_t < TN''_{t-1}), \tag{5.7}$$

where t corresponds to the iteration step in the boosting algorithm (classifiers loose performance) or if:

$$(TP''/|X''_{y_i=1}| > p) \wedge (TN''/|X''_{y_i=-1}| > 0.5), \tag{5.8}$$

classifiers eliminates more than 50% of negatives and classifies correctly $p \times 100\%$ of positive samples (in our experiments we set $p \in \{0.9, 0.95, 0.99\}$).

- T_c: we stop adding new stages in the cascade if

$$(TP_t < TP_{t-1}) \wedge (TN_t < TN_{t-1}), \tag{5.9}$$

which means that classifiers loose performance. We also finish training when the number of stages is greater than k.

Different values of parameters p and k are evaluated in experiments.

It is worth noting that if we assume $p = 0.99$ and $k = 20$, while eliminating 50% negatives at each stage, we can expect false positives rate about $0.5^{20} \approx 9.5 * 10^{-7}$ and a true positives rate about $0.99^{20} \approx 0.8$.

5.1.3 Haar similarity

Given a human signature we can find out whether a person of interest appeared in a video sequence. This can be determined by applying the given cascade of strong classifiers (*signature*) to all frames of the video sequence. In this case, the signature can be used as a detector of a specific individual. Unfortunately, such solution is not efficient. Each time when we search for a person of interest, we would need to process all frames from a video sequence.

As we have already mentioned in section 3.1, we can optimize re-identification by applying pre-processing steps. The first step consists in detecting general class of objects (*i.e.* humans). Then, cropped images with detected humans are used to build signatures. These signatures should be stored in a database to provide the most efficient way for browsing the video content restricted to detected objects. Having a database of signatures we need to be able to compare different signatures. As Haar-based signature is represented by a classifier, there is not straightforward way to compute similarity between two signatures. Consequently, we propose two ways for comparing Haar-based signatures.

5.1.3.1 Volume-based similarity

Let us scale every cropped image into a fixed size of $W \times H$ pixels (in experiments we assume 20×40 window). This gives us $n = W \times H$ ($20 \times 40 = 800$) dimensional

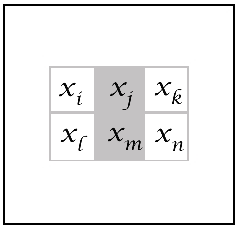


Figure 5.7: Illustration of a Haar-like feature. The sum of the white pixels is subtracted from the sum of the grey pixels.

space. If we assume that a pixel value $x \in \Psi$ and $|\Psi| = \tau$ (in general a range of intensity is $\tau = 256$), then the size of our image space is τ^n.

According to equations (5.1) and (5.6), the weak classifier corresponding to the Haar-like feature (line feature) illustrated in figure 5.7 can be expressed as inequality

$$\mathfrak{p}_z(\omega_a(x_i + x_l) + \omega_b(x_j + x_m) + \omega_c(x_k + x_n)) < \mathfrak{p}_z\theta_z, \tag{5.10}$$

where parity $\mathfrak{p}_z$ indicates the direction of the inequality sign, coefficients ω_a, ω_b and ω_c are chosen arbitrarily, θ_z is the threshold and x represents a pixel value. In general, each Haar-like classifier can be written in the similar way. Therefore a strong classifier as well as a cascade of them can be represented by a set of linear inequalities like equation (5.10). Each such equation can be seen as a hyperplane, which separates the image space for the images which meet inequality and the images which do not. As the Haar-based signatures consists of a set of strong classifiers, which are composed of a set of weak classifiers, the final signature can be seen as a set of weak classifiers. The set of weak classifiers gives us the set of inequalities, which can be seen as the set of hyperplanes. In addition, $2n$ default hyperplanes: $x = \min(\Psi)$ and $x = \max(\Psi)$ for each of the n dimensions are given (constraints of the image space). This hyperplanes together with the hyperplanes given by the weak classifiers, cut a hypercube in the image space. Therefore, let us assume that each signature is such hypercube.

Let $\mathcal{V}_{\mathfrak{s}_A}$ be the volume of the hypercube generated by weak classifiers of signature $\mathfrak{s}_A$. We define similarity S between two signatures $\mathfrak{s}_A$ and $\mathfrak{s}_B$ as

$$S(\mathfrak{s}_A, \mathfrak{s}_B) = \frac{\mathcal{V}_{\mathfrak{s}_A\mathfrak{s}_B}}{\min(\mathcal{V}_{\mathfrak{s}_A}, \mathcal{V}_{\mathfrak{s}_B})}, \tag{5.11}$$

where $\mathcal{V}_{\mathfrak{s}_A\mathfrak{s}_A}$ is the volume of the hypercube produced by merging weak classifiers of signature $\mathfrak{s}_A$ with weak classifiers of signature $\mathfrak{s}_B$.

Computing the volume of hypercube, we need to store 2^n vertices. In our approach we use 20×40 pixel sub-window which leads to $n = 800$ dimensions, bringing an

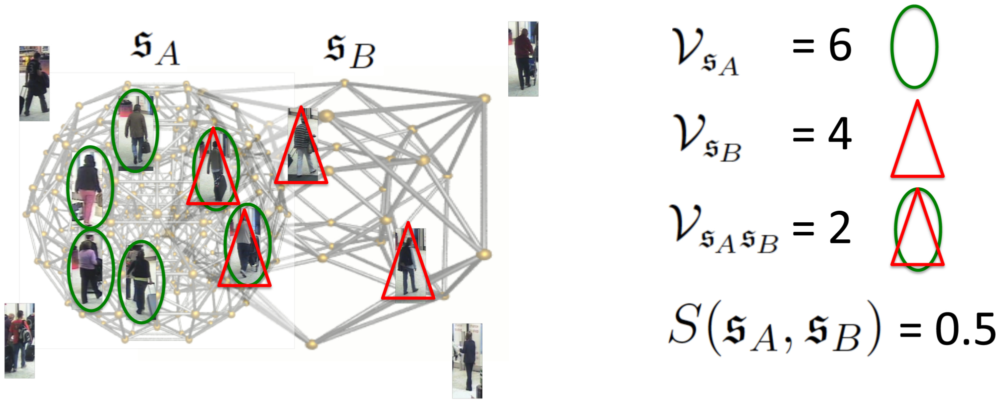

$\mathfrak{s}_A$
$\mathfrak{s}_B$
$\mathcal{V}_{\mathfrak{s}_A}$ = 6
$\mathcal{V}_{\mathfrak{s}_B}$ = 4
$\mathcal{V}_{\mathfrak{s}_A\mathfrak{s}_B}$ = 2
$S(\mathfrak{s}_A, \mathfrak{s}_B)$ = 0.5

Figure 5.8: The volume-based similarity computation: the signatures (hypercubes) are illustrated as the meshes; the sparse space is given by images. Images which satisfy weak classifiers of $\mathfrak{s}_A$ and weak classifiers of $\mathfrak{s}_A$ are highlighted using ellipses and triangles, respectively.

unattainable memory and time requirement. For that reason our volume computation does not consider the whole space τ^n.

We compute the volume based only on a sparse space of images. This sparse space corresponds to all cropped images registered in a network of cameras. Thus, we define volume $\mathcal{V}_{\mathfrak{s}_A}$ as the number of images which satisfy inequalities given by weak classifiers of $\mathfrak{s}_A$, along all cropped images process by the system. Figure 5.8 illustrates volumes and the similarity computation.

5.1.3.2 Margin-based similarity

We propose a similarity which is based on a concept of *margin* [Schapire 1998, Rudin 2007]. The margin theory is related to the phenomenon that AdaBoost often does not seem to suffer from overfitting, in the sense that the test error does not go up even after a large number of iterations T. The margin of a boosted classifier on a particular example can be interpreted as a measure of the classifier's confidence on this particular example. The margin theory states that AdaBoost tends to increase the margins of the training examples, and in the result, the increase in the margins implies better generalization performance.

The margin of example (x_i, y_i) with respect to classifier $\mathcal{H}$ is defined as

$$\mathfrak{M}^i_{\mathcal{H}} = \frac{y_i \sum_{t=1}^{T} \alpha_t h_t(x_i)}{\sum_{t=1}^{T} \alpha_t}. \tag{5.12}$$

This number is between -1 and $+1$. It is positive if and only if $\mathcal{H}$ correctly classifies the example.

Let us represent a signature by the cascade of strong classifiers and the positive training samples on which this classifiers has been learned

$$\mathfrak{s}_A = \{(\mathcal{H}^A_z) : z = 1 \ldots k_A; (x_{A,i}, y_{A,i}) : i = 1 \ldots l_A, y_{A,i} = +1\}, \tag{5.13}$$

where k_A is the number of strong classifiers in the cascade and l_A is the number of positive training examples which has been classified correctly by the cascade during the training process.

Given two signatures $\mathfrak{s}_A$ and $\mathfrak{s}_B$, we define the similarity S as

$$S(\mathfrak{s}_A, \mathfrak{s}_B) = \mathbb{M}_{A,B} + \mathbb{M}_{B,A}, \tag{5.14}$$

where

$$\mathbb{M}_{A,B} = \frac{1}{l_B} \sum_{i=1}^{l_B} \min_z \left(\mathfrak{M}^{B,i}_{\mathcal{H}^A_z} \right) \ln(e + \frac{z}{k_A}). \tag{5.15}$$

The key insight of this formula is that we evaluate examples, which were used to learn signature $\mathfrak{s}_B$, using the strong classifiers of signature $\mathfrak{s}_A$. As the margins give us the confidence of classification, we use them to establish how similar are the signatures (classifiers). We compute the minimum margin over the strong classifiers to obtain the confidence of the whole cascade. By logarithm in equation (5.15), we put greater weight if the minimum margin is extracted at deeper stage in the cascade. Figure 5.9 shows the sketch of the margin-based similarity computation.

5.1.4 Discussion

The main disadvantage of Haar-based signatures is that we need to keep learning examples to compute the similarity between two signatures. Computing either the volume-based similarity or the margin based similarity, we always test our signatures on a set of images. Computing the volume-based similarity we need to execute signature classifiers on continuously growing sparse space of images. Using the margin-based similarity, it is necessary to test each signature from a database on images, which were used for computing a query signature. Both solutions are not computationally tractable, bringing slow search time, while browsing the database of signatures.

The necessity of keeping training examples comes from a design of Haar-like features, which can be seen as filters separating positive and negative samples. Generating signature, the boosting selects only a few features (from the over-complete set) per signature, making the distance computation more sophisticated (difficulties with comparing two different Haar-like patterns, which also can correspond to different spatial locations).

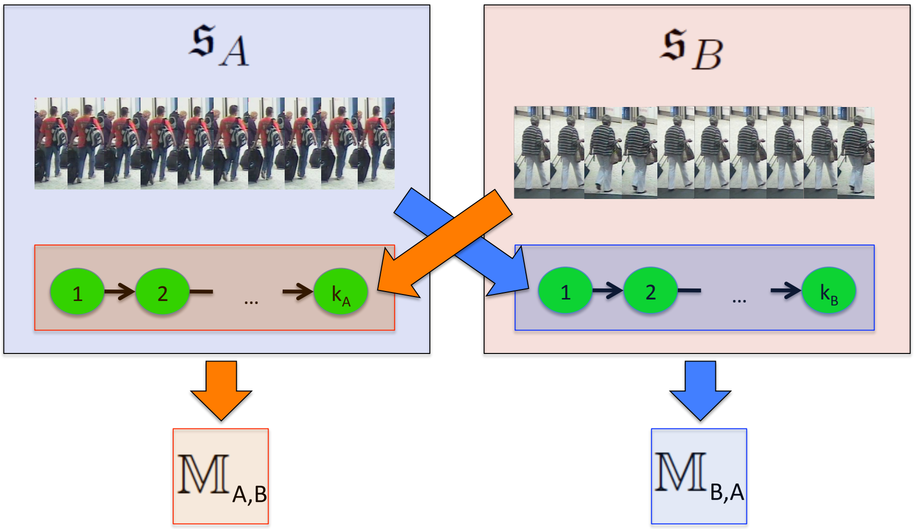


Figure 5.9: The margin-based similarity computation: each signature is represented by a cascade of strong classifiers and the positive training examples.

In the next section, we propose an approach which creates an appearance representation, which does not require to keep training data. This approach use boosting to select discriminative features, which are used for matching signatures.

5.2 Binary classification using Riemannian manifolds

In this section we present a human appearance model based on *Mean Riemannian Covariance* (MRC) (section 5.2.2) extracted from tracks of a particular individual. Figure 5.10 illustrates the key insight of the approach. Similarly to the Haar-based approach (section 5.1), this approach uses multiple images of the same person to generate a signature. The input of our approach is a set of cropped images corresponding to human detection and tracking results, normalized by the histogram equalization applied to each color channel (RGB) (see section 4.1.2). From such normalized images we extract MRC patches (section 5.2.3). We investigate the significance of MRC patches based on their reliability extracted during tracking (section 5.2.4) and their discriminative power obtained by a boosting scheme (section 5.2.5.3). These two alternative ways for feature selection has been evaluated to discuss the utility of a discriminative learning. The proposed selection method based only on reliability of patches is referred to as *Reliable Covariance Patches (RCP)* method and the method based on a boosting scheme is denoted as *Learned Covariance Patches (LCP)* method. However, before explaining details of the proposed MRC patches, we present a short introduction to Riemannian geometry, which provides tools for

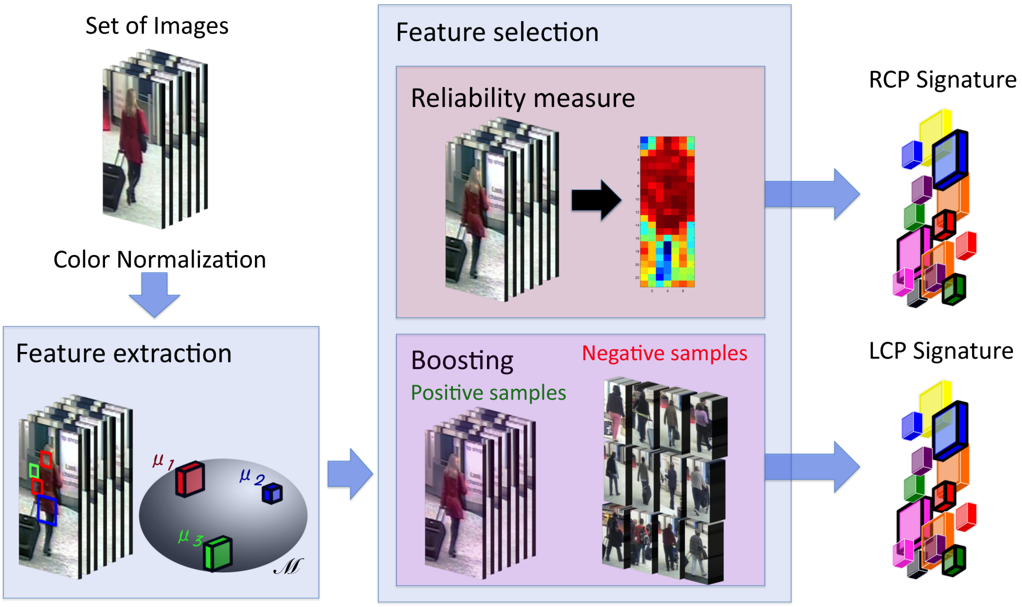


Figure 5.10: Extraction of the signature based on MRC patches.

using covariance matrix descriptor.

5.2.1 Riemannian geometry

Covariance matrix as a positive definite and symmetric matrix can be seen as a tensor. The main problem is that such defined tensor space is a manifold that is not a vector space with the usual additive structure. A manifold is a topological space which is locally similar to an Euclidean space. It means that every point on the m-dimensional manifold has a neighbourhood homeomorphic to an open subset of the m-dimensional space $\Re^m$. Performing operations on the manifold involves choosing a metric.

Specifying manifold as Riemannian gives us Riemannian metric. This automatically determines a powerful framework to work on the manifold by using tools from differential geometry [Pennec 2006]. Riemannian manifold $\mathcal{M}$ is a differentiable manifold in which each tangent space has an inner product which varies smoothly from point to point. Since covariance matrices can be represented as a connected Riemannian manifold, we apply operations such as *distance* and *mean* computation using this differential geometry.

Figure 5.11 shows an example of a two-dimensional manifold, a smooth surface living in $\Re^3$. Tangent space $T_x\mathcal{M}$ at x is the vector space that contains the tangent vectors to all 1-D curves on $\mathcal{M}$ passing through x. Riemannian metric on manifold

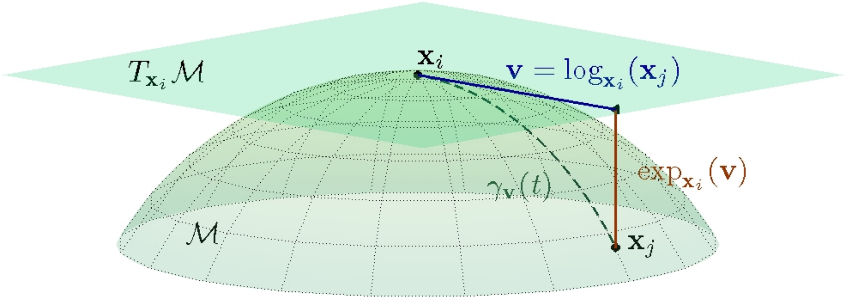

$T_{\mathbf{x}_i}\mathcal{M}$
$\mathbf{x}_i$
$\mathbf{v} = \log_{\mathbf{x}_i}(\mathbf{x}_j)$
$\exp_{\mathbf{x}_i}(\mathbf{v})$
$\gamma_{\mathbf{v}}(t)$
$\mathcal{M}$
$\mathbf{x}_j$

Figure 5.11: An example of a two-dimensional manifold. We show the tangent plane at x_i, together with the exponential and logarithm mappings related to x_i and x_j [Goh 2008b].

$\mathcal{M}$ associates to each point $x \in \mathcal{M}$, a differentiable varying inner product $\langle \cdot, \cdot \rangle_x$ on tangent space $T_x\mathcal{M}$ at x. This induces a norm of tangent vector $v \in T_x\mathcal{M}$ such that $\|v\|_x^2 = \langle v, v \rangle_x$. The minimum length curve over all possible smooth curves $(\gamma_v(t))$ on the manifold between x_i and x_j is called *geodesic*, and the length of this curve stands for geodesic distance $\rho(x_i, x_j)$.

Before defining geodesic distance, let us introduce the exponential and the logarithm functions, which take as an argument a square matrix. The exponential of any matrix can be defined as the series

$$\exp(W) = \sum_{k=0}^{\infty} \frac{W^k}{k!}. \tag{5.16}$$

In the case of symmetric matrices, we can apply simplification. Let $W = U\ D\ U^T$ be a diagonalization, where U is an orthogonal matrix, and $D = DIAG(d_i)$ is the diagonal matrix of the eigenvalues. We can write any power of W in the same way $W^k = U\ D^k\ U^T$. Thus

$$\exp(W) = U\ DIAG(\exp(d_i))\ U^T, \tag{5.17}$$

and similarly the logarithm is given by

$$\log(W) = U\ DIAG(\log(d_i))\ U^T. \tag{5.18}$$

According to a general property of Riemannian manifolds, geodesics realize a local diffeomorphism from the tangent space at a given point of the manifold to the manifold. It means that there is the mapping which associates to each tangent vector $v \in T_x\mathcal{M}$ a point of the manifold. This mapping is called the exponential map, because it corresponds to the usual exponential in some matrix groups.

The exponential and logarithmical mappings have the following expressions [Pennec 2006]:

$$\exp_\Sigma(W) = \Sigma^{\frac{1}{2}} \exp(\Sigma^{-\frac{1}{2}} W \Sigma^{-\frac{1}{2}}) \Sigma^{\frac{1}{2}}, \tag{5.19}$$

$$\log_{\Sigma}(W) = \Sigma^{\frac{1}{2}} \log(\Sigma^{-\frac{1}{2}} W \Sigma^{-\frac{1}{2}}) \Sigma^{\frac{1}{2}}, \tag{5.20}$$

where

$$\Sigma^{\frac{1}{2}} = \exp\left(\frac{1}{2}(\log(\Sigma))\right). \tag{5.21}$$

Given tangent vector $v \in T_x\mathcal{M}$, there exists a unique geodesic $\gamma_v(t)$ starting at x (see figure 5.11). The exponential map $\exp_x : T_x\mathcal{M} \rightarrow \mathcal{M}$ maps tangent vector v to the point on the manifold that is reached by this geodesic. The inverse mapping is given by logarithm map denoted by $\log_x : \mathcal{M} \rightarrow T_x\mathcal{M}$. For two points x_i and x_j on manifold $\mathcal{M}$, the tangent vector to the geodesic curve from x_i to x_j is defined as $v = \overrightarrow{x_i x_j} = \log_{x_i}(x_j)$, where the exponential map takes v to the point $x_j = \exp_{x_i}(\log_{x_i}(x_j))$. The Riemannian distance between x_i and x_j is defined as $\rho(x_i, x_j) = \| \log_{x_i}(x_j) \|_{x_i}$.

5.2.2 Mean Riemannian covariance (MRC)

Let $C_1, \ldots, C_N$ be a set of covariance matrices. The Karcher or Fréchet mean is the set of tensors minimizing the sum of squared distances. In the case of tensors, the manifold has a non-positive curvature, so there is a unique mean value μ:

$$\mu = arg \min_{C \in \mathcal{M}} \sum_{i=1}^{N} \rho^2(C, C_i), \tag{5.22}$$

where ρ is the covariance matrix distance (equation 4.9).

Since covariance matrices lay on a Riemannian manifold we use the intrinsic Newton gradient descent algorithm to compute the approximation mean covariance at step $t+1$:

$$\mu_{t+1} = exp_{\mu_t}\left[\frac{1}{N}\sum_{i=1}^{N} log_{\mu_t}(C_i)\right], \tag{5.23}$$

where exp_{μ_t} and log_{μ_t} are specific operators (mapping functions) uniquely defined on a Riemannian manifold (see section 5.2.1). This iterative gradient descent algorithm usually converges very fast (in experiments 5 iterations were sufficient, which is similar to [Pennec 2006]).

5.2.2.1 Mean Riemannian covariance *vs.* volume covariance

Covariance matrix could be directly computed from a video by merging feature vectors from many frames into a single content (similarly to 3D descriptors, *i.e.* 3D HOG). Then, this covariance could be seen as *mean covariance*, describing characteristics of the video. Unfortunately, such solution disturbs time dependencies (time order of features is lost). Further, context of the features might be lost and at the same time some features will not appear in the covariance.

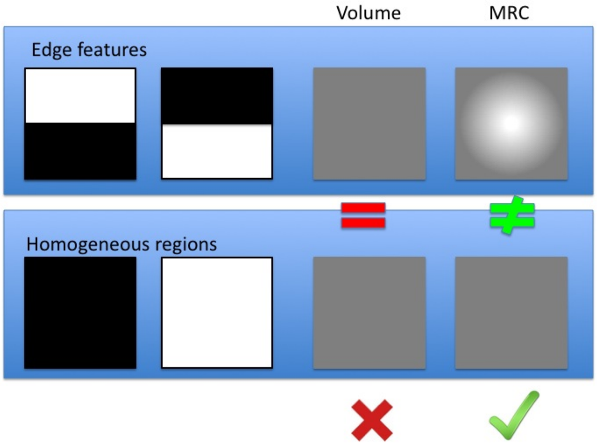


Figure 5.12: Difference between covariance computed directly from the video content (volume covariance) and MRC. Volume covariance looses information on edge features and can not distinguish two given cases - two edge features (first row) from two homogeneous regions (second row). MRC holds information on the edges, being able to differentiate both cases.

Figure 5.12 illustrates the case, in which edge features are lost during computation of the volume covariance. This is a consequence of loosing information that the feature appeared in specific time. Computing volume covariance, order of the feature appearances and their spatial correlations can be lost by merging feature distribution in time. MRC holds much more information than covariance computed directly from the volume.

There have already been proposed some other statistics on covariance functions which take into account spatial and temporal changes [Cressie 1999, Fuentes 2006]. However, these approaches assume that covariance function is a stationary (or weakly stationary) process in space and time. Unfortunately in our case we can not make such strong assumptions as visual features mostly do not meet this requirement in a video sequence (mean function of features is not constant over time). Moreover, noisy human detections and gaps in tracking prevent the covariance functions to be separable in space and time. As it is hard to apply these statistic methods to our covariance matrices, we have decided to use the mean covariance computed on a Riemannian manifold as a descriptor of a region in a video sequence.

5.2.3 MRC patches

In this section we define *Mean Riemannian Covariance* (MRC) patches and explain their merits. Once a human has been detected and color has been normalized, we scale every cropped image into a fixed size $W \times H$. MRC patch corresponds to a square region (of size $\frac{W}{4} \times \frac{W}{4}$ or $\frac{W}{2} \times \frac{W}{2}$). We assume that such MRC patches can

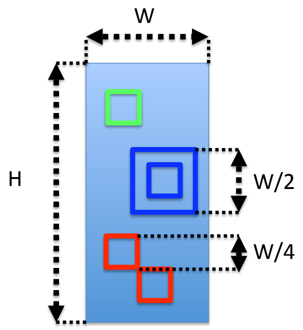


Figure 5.13: MRC Patches. Green patch corresponds to the single square region (the top). Blue patch is an example of combination (small - big) of two patches with different size (the middle). Red patch illustrates combination (small - small) of patches with the same size (the bottom).

form patch combinations (see figure 5.13). In our approach we only consider patch combinations built using maximally two MRC patches. The choice of the number of patches per patch combination and of the square regions with fixed size is a compromise between time complexity and descriptive power. The patch combination may consist of two patches with the same size as well as with different size. Different patterns of a human appearance are captured from a window $W \times H$. We extract patches of size $\frac{W}{4} \times \frac{W}{4}$ shifted horizontally and vertically by $\frac{W}{8}$ and patches of size $\frac{W}{2} \times \frac{W}{2}$ shifted horizontally and vertically by $\frac{W}{4}$, producing a dense representation. The position of MRC patch on the fixed size window and the spatial correlation between patches is essential to carry out discriminative power of rectangular patterns.

Let $C_1^p, \ldots, C_N^p$ be a set of covariance matrices extracted during tracking of N frames corresponding to image square regions at position of patch p. We define MRC patch as the mean covariance of these covariance matrices (see section 5.2.2) computed using a Riemannian space (see figure 5.14). The mean covariance matrix as an intrinsic average blends all extracted matrices. This mean covariance matrix not only keeps information on feature distribution but also carries out essential cues on temporal changes of the appearance related to the position of patch p. Moreover, as we normalize covariance matrix, rotation and illumination changes are absorbed. MRC patch (the mean covariance matrix together with its position) is fundamental in our approach. Further, patch combinations are able to catch different patterns and spatial correlations of MRC patches. Now we introduce a new reliability measure, which describes an invariance of a patch.

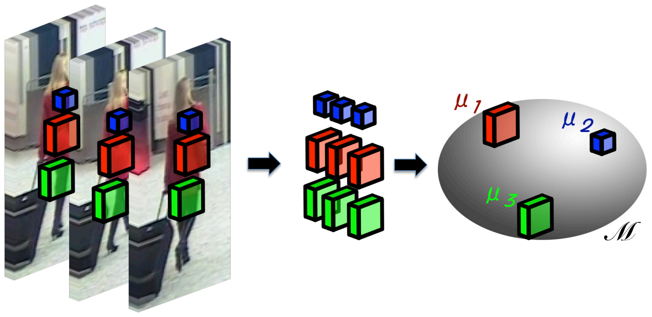


Figure 5.14: Computation of three MRC patches. Covariances gathered from tracking results are used to compute the mean covariance using Riemannian manifold space (depicted with the surface of the sphere). The mean covariance forms MRC patch.

5.2.4 Reliability-based MRC patch selection (RCP)

5.2.4.1 Patch reliability

Let $r_1^p, \ldots, r_N^p$ be a set of image square regions extracted during tracking of N frames at position of patch p. For each such region we extract covariance matrix C_i^p which encodes information of the variances of defined features f inside the region and their correlations.

Together with the mean covariance matrix we define the reliability measure $\mathfrak{R}$ which describes stability of the tracked image region

$$\mathfrak{R} = 1 - \frac{\sigma}{\max \sigma}, \quad \sigma = \frac{1}{N-1}\sum_{k=1}^{N}(\bar{f}_k - \bar{\Gamma})^2, \quad \bar{\Gamma} = \frac{1}{N}\sum_{k=1}^{N}\bar{f}_k \tag{5.24}$$

where $\bar{\Gamma}$ is the mean feature vector along N tracking regions and $\bar{f}_k$ is the mean feature vector corresponding to region k, $\max \sigma$ corresponds to the maximal value of σ for the specific individual. As we assume that only partial occlusions may occur, the reliability works similarly to background subtraction. Here, our idea is to remove most variable features because we assume that these features are the noisiest (containing background). Figure 5.16 (b) illustrates that the most unreliable features correspond to legs and background.

5.2.4.2 Patch selection

We select the most reliable patches with the highest reliability among others. If the patch combination is considered, the reliability is an average of reliabilities of both patches. This selection method allows us to create the human signature without any discriminative learning phase.

5.2.5 Boosted MRC patch selection (LCP)

Similarly to Haar-based approach (section 5.1), we learn a signature of a specific individual using *one-against-all* learning scheme. We again lay the re-identification problem as classification where we separate the specific individual class from the rest of humans. Assuming that discriminative features of an individual extracted from one camera correspond to discriminative features of the same individual extracted from another camera, we define similarity measure based on the output of the improved boosting algorithm.

In contrast to Haar-based signature, we do not structure our classifier into the cascade. Instead, we use an improved boosting algorithm to select discriminative features of an individual. The main advantage of using this algorithm is that it generates the confidence scores of selected classifiers. We use these scores as weights of selected features, while matching signatures.

5.2.5.1 Training data

Given a set of cropped images corresponding to the same individual tracked in a video sequence, we generate positive samples for a learning algorithm. Each signature has an associated set of relevant observations (tracking results of human j in N frames from camera c) represented by a set of images ${\mathcal{I}^{+}}^{c}_{j} = \{I^{c}_{j,1}, \dots I^{c}_{j,N}\}$. This set is used as positive samples in a learning phase. Negative samples are obtained from images extracted from the same camera during tracking the rest of humans: ${\mathcal{I}^{-}}^{c}_{j} = \{I^{c}_{i,k}\}_{i \neq j}$.

5.2.5.2 Improved boosting algorithm using confidence-rated predictions

Improved boosting algorithm is an extended boosting framework in which each weak classifier generates not only predicted classification, but also self-rated confidence score which estimates the confidence of its prediction [Schapire 1999]. We use this confidence value to define the similarity measure between two signatures (section 5.2.6).

Boosting based on confidence-rated predictions differs from the general AdaBoost in design of weak classifier. The weak classifier takes into account a detailed analysis of its performance on the training dataset. The output of hypotheses h_t is derived

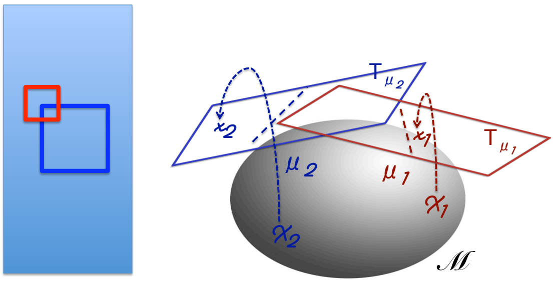


Figure 5.15: MRC Classifier. The manifold is depicted with the surface of the sphere, and the planes are the tangent spaces of the mean covariances. The samples $\mathfrak{X}_i$are projected to tangent space T_{μ_i}at means via log_μ operation.

with respect to the given distribution w_t over training samples, minimizing

$$Z = \sum_{i=1}^{n} w_t e^{-y_i h_t(x_i)}. \tag{5.25}$$

Let split equation (5.25) into a sum over negative and positive samples

$$Z = \sum_{i:y=+1} w_t e^{-h_t(x_i)} + \sum_{i:y=-1} w_t e^{h_t(x_i)} \tag{5.26}$$

$$= \sum_{\substack{i:y=+1\\ h_t(x_i)=y_i}} w_t e^{-\alpha_t} + \sum_{\substack{i:y=+1\\ h_t(x_i)\neq y_i}} w_t e^{-\beta_t} + \sum_{\substack{i:y=-1\\ h_t(x_i)\neq y_i}} w_t e^{\alpha_t} + \sum_{\substack{i:y=-1\\ h_t(x_i)=y_i}} w_t e^{\beta_t} \tag{5.27}$$

$$= W_+^+ e^{-\alpha_t} + W_+^- e^{-\beta_t} + W_-^+ e^{\alpha_t} + W_-^- e^{\beta_t}, \tag{5.28}$$

where W_+^+, W_-^+, W_+^-, W_-^- correspond to weighted sums of true positive, false positive, false negative and true negative samples, respectively. Using standard calculus, we see that this is minimized when

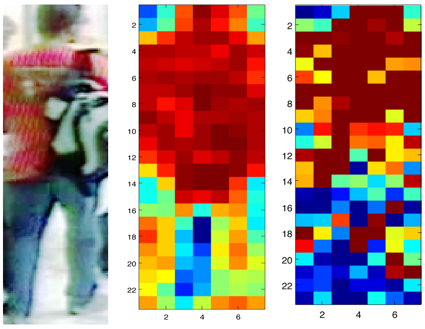
$$\alpha_t = \frac{1}{2} \ln\left(\frac{W_+^+}{W_-^+}\right) \quad \text{and} \quad \beta_t = \frac{1}{2} \ln\left(\frac{W_+^-}{W_-^-}\right). \tag{5.29}$$

These equations are used to define a weak classifier.

5.2.5.3 Weak classifier

We define a weak classifier as function h based on distance computation on Riemanian manifold space (see figure 5.15). Image I is classified by a weak classifier built on patch p using a threshold function defined as

$$h(p, I) = \begin{cases} \alpha = \frac{1}{2}\ln(\frac{W_+^+}{W_-^+}) & if \ \sum_j w_j \rho(\mu_j, C_j) \leq \theta_p \\ \beta = \frac{1}{2}\ln(\frac{W_+^-}{W_-^-}), & otherwise \end{cases} \tag{5.30}$$



(a) I (b) $\mathfrak{R}$ (c) α

Figure 5.16: Illustration of patch significance: (a) one of many frames obtained during tracking; (b) reliability map obtained by the first method; (c) confidence map obtained by boosting. Colours correspond to significance of patches (for clarity only $\frac{W}{4} \times \frac{W}{4}$ patches, shifted by $\frac{W}{8}$ pixels are illustrated, red indicates the highest significance, blue the lowest).

where W_+^+, W_-^+, W_+^-, W_-^- correspond to weighted sums of true positive, false positive, false negative and true negative samples, respectively. Threshold θ_p and weights w_j are obtained by minimizing error of weak classifier during learning; μ_j is the mean covariance of MRC patch classifier and covariance C_j corresponds to the image region of I.

Each MRC patch and MRC patch combination extracted during tracking forms a weak classifier. From the set of weak classifiers, boosting algorithm selects the ones which together form a strong classifier. Details of boosting algorithm using confidence-rated predictions can be found in [Schapire 1999].

We use the selected weak classifiers in MRC patch matching. In figure 5.16 we present reliability and confidence maps, based on reliability $\mathfrak{R}$ and confidence α, respectively.

5.2.5.4 Stop criterion

Using either reliability-based selection method or boosting, we select a set of MRC patches for generating a single signature. In both cases we constraint size of this set to predefined value z which sensitivity is evaluated in experiments (see section 7.4.2).

5.2.6 MRC patch matching

Given extracted human signatures, we introduce a way to effectively distinguish individuals. The human signature is represented by a set of relevant MRC patches,

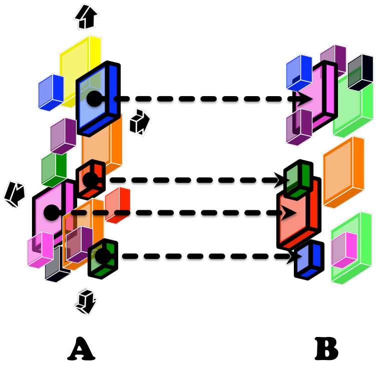


Figure 5.17: The similarity between two human signatures. Every signature is a set of MRC patches. Signature A is shifted left/right/up and down to find out the best corresponding patches in signature B (position of a patch determines matching). Connections in the figure represent corresponding patches. Some connections are suppressed for clarity.

extracted using one of the aforementioned selection methods. MRC patch consists of the mean covariance with its position. The position of the patch is essential to keep discriminative power of the human signature. In general, the matching of two signatures $\mathfrak{s}_A$ and $\mathfrak{s}_B$ is carried out by maximizing the similarity measure. We shift one signature over another one to reduce body alignment issues. When shifting signature (see figure 5.17) we preserve relative position between patches to avoid wasting of discriminative property of the patch position, while maximizing the similarity. In experiments we evaluate the influence of the shifting operation on matching accuracy. The similarity between two human signatures $\mathfrak{s}_A$ and $\mathfrak{s}_B$ is defined as

$$S(\mathfrak{s}_A, \mathfrak{s}_B) = \frac{1}{|K|} \sum_{i \in K} \frac{\gamma_{A_i} + \gamma_{B_i}}{\rho(\mu_{A_i}, \mu_{B_i})}, \tag{5.31}$$

where K stands for the set of corresponding patches i in signature $\mathfrak{s}_A$ and signature $\mathfrak{s}_B$; ρ is the covariance matrix distance; γ is a reliability ($\mathfrak{R}$) or a confidence (α) value of the corresponding patches depending on selection method. Here, we take advantages of the patch selection method which is able to choose the most informative patch combinations. The efficiency of patch combination is handled indirectly by γ coefficients.

5.2.7 Discussion

Discriminative approaches such as [Lin 2008, Schwartz 2009] are often accused of non-scalability. It is true that in these approaches (in our as well) an extensive learning phase is necessary to extract discriminative signatures. These approaches are difficult to apply to real scenarios where new people appear continuously. First, the learning phase prevents to generate signatures in real-time. Second, every time when a new signature is created we have to update all signatures in the database (*one-against-all* learning scheme). For example, in PLS [Schwartz 2009] there is a requirement to have all the gallery signatures beforehand, in order to estimate weights on the appearance model. When one pedestrian is added, weights must be recomputed which makes the approach not-suitable for video surveillance systems.

As a solution to scalability issues we can propose to extract a *reference dataset* which can be used as negative samples for learning a discriminative signature. This *reference dataset* can be chosen offline to be the most representative. Nevertheless, in this case, a time consumption issue still remains. As in the system, there is no constraints for a *real-time* signature generation, there already exist learning approaches which can operate in reasonable time. Moreover, recently, driven by the insatiable market demand for real-time, the programmable *Graphic Processor Unit* (GPU) has evolved into a highly parallel, multithreaded, manycore processor with tremendous computational power and very high memory bandwidth. These new hardware architectures such as NVIDIA [Lindholm 2008] are investigated to speed up the computation of pattern recognition problems. Currently, step by step, well known learning algorithms [Catanzaro 2008] or time consuming descriptors (see chapter 8) are ported to such specialized architectures. As image and media processing demand a lot of computation power, this direction seems to fit perfectly for more sophisticated approaches. We claim that the usage of high-performance computing can be a solution to make discriminative learning approaches more suitable for video-surveillance systems.

5.3 Conclusion

This chapter employed boosting scheme for extracting discriminative representation of a human appearance. We proposed two discriminative techniques to generate human signature using multiple images of many humans. Both methods apply *one-against-all* learning scheme to enhance distinctive characteristics of a specific appearance using information from appearances of other individuals.

The first method is based on Haar-like features. The main disadvantage of this technique is that we need to keep learning examples to compute the similarity between signatures. The necessity of keeping training examples comes from a design of Haar-like features, which can be seen as filters separating positive and negative samples.

The second technique uses covariance matrix as the feature for describing image region. We proposed to use *mean Riemannian covariance* (MRC) as the final descriptor which holds statistic information on temporal changes of the appearance. We offer two alternative ways for feature selection which performance is evaluated in section 7.4.2.

Although these discriminative approaches are inspired by *human memory* and they show promising results (see section 7.4), boosting approaches are computationally intensive which is unfavorable in practice. In the next chapter we propose less computationally expensive solutions which use information on the appearance of other individuals either by introducing a simple and efficient discriminative method or by applying an offline learning stage.

Chapter 6

Efficient human re-identification using Riemannian manifolds

"*Everything should be made as simple as possible, but not simpler.*"
(Albert Einstein)

Re-identification approaches based on discriminative learning are often computationally heavy and non-scalable. In this chapter, we propose two *multiple-shot* approaches, which are less computationally expensive than previously presented boosting-based methods, while enhancing discriminative and descriptive features. Although, we again employ *mean Riemannian covariance* as the feature for characterizing image regions, we reinforce MRC's properties using new appearance models. The first method assumes a predefined appearance model (section 6.1), while the second technique learns an appearance representation during an offline stage, guided by an entropy-driven criterion (section 6.2).

6.1 MRCG signature

This section describes a new appearance model combining information from multiple images to obtain highly discriminative human signature, called *mean Riemannian covariance grid* (MRCG). Figure 6.1 illustrates the sketch of our technique. The input of the approach is a set of cropped images corresponding to human detection and tracking results. We handle color dissimilarities caused by camera illumination difference by applying the histogram equalization to each color channel (RGB) (see section 4.1.2). The normalized image is scaled and divided into a grid structure of overlapping cells (section 6.1.1). Each *cell* is characterized by MRC descriptor and by its distinctiveness (the relevance of the feature) computed using discriminative method (section 6.1.2). The full grid of overlapping *cells* characterized by MRC-s and their distinctiveness, stands for the final appearance representation.

6.1.1 Mean Riemannian covariance grid (MRCG)

In this section we define the novel *mean Riemannian covariance grid* (MRCG) and explain its merits. The proposed human signature has been designed to deal with low

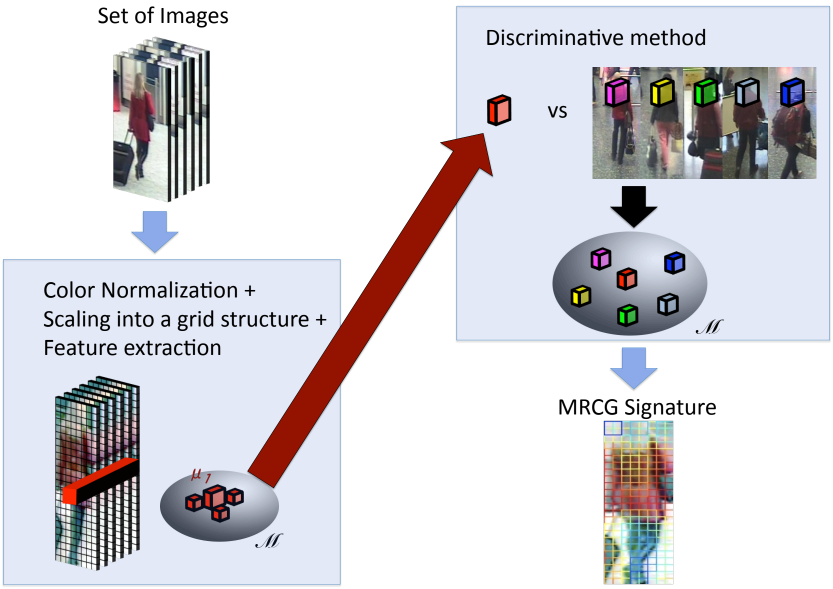


Figure 6.1: Extraction of MRCG signature.

resolution images and crowded environment where more specialized techniques (*e.g.* based on body part detectors) might fail. We combine dense descriptors philosophy [Dalal 2005] with the effectiveness of the MRC descriptor.

Once color has been normalized, we scale every human image into a fixed size of $W \times H$ pixels. Then, an image is divided into a dense grid structure with overlapping spatial square regions (*cells*). First, such dense representation makes the signature robust to partial occlusions. Second, as the grid structure, it contains relevant information on spatial correlations between MRC *cells*, which is essential to carry out discriminative power of the signature. Moreover, as we use covariance matrices to describe characteristic of the cells, this technique enables to fuse efficiently different types of features and their modalities.

MRC *cell* describes statistics of an image square region corresponding to the specific position in the grid structure. In contrary to MRC patch (section 5.2.3), MRC *cell* is computed from the fixed size square region (in experiments we set cells as 16×16 square regions) and we do not consider different MRC combinations. The appearance model is fixed as a grid structure.

Let $C_1^p, \dots, C_N^p$ be a set of covariance matrices extracted during tracking of N frames corresponding to image square regions at position of the cell p. We define the MRC as the mean covariance of these covariance matrices (see section 5.2.2) computed using Riemannian space (see figure 6.2). As an intrinsic average, the mean covariance matrix blends appearance information from multiple images. This

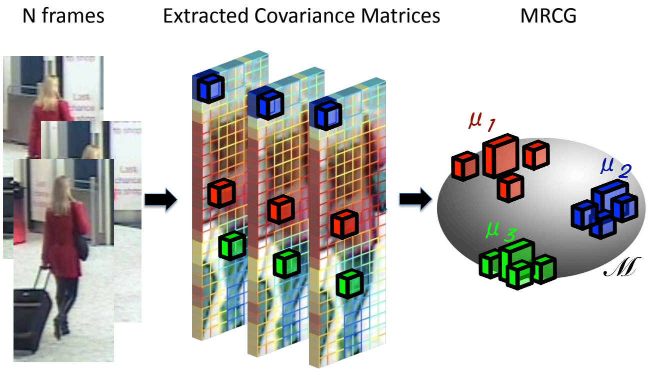


Figure 6.2: Computation of the MRCG. Covariances gathered from tracking results are used to compute the MRC using Riemannian manifold space (depicted with the surface of the sphere).

mean covariance matrix not only keeps information on feature distribution but also gives essential cues on temporal changes of the appearance related to the position of cell p. All MRC *cells* compose a full grid, named as *mean Riemannian covariance grid* (MRCG).

6.1.2 MRC discriminants

The goal of using discriminants, is to identify the relevance of MRC *cells.* We present an efficient way to enhance discriminative features, improving matching accuracy. Similar to boosting-based approach, this method allows us to highlight features which are distinctive for a particular individual. The main advantage of this method is its efficiency. Using simple statistics on Riemannian manifold we are able to enhance distinctive features, without applying any time consuming training process.

Given a set of signatures $\mathfrak{S}^c = \{\mathfrak{s}_i^c\}_{i=1}^n$ where s_i^c is signature i from camera c, each signature is represented by MRCG: $\mathfrak{s}_i^c = \{\mu_{i,1}^c, \mu_{i,2}^c, \ldots, \mu_{i,m}^c\}$, where $\mu_{i,j}^c$ stands for MRC *cell* and m is the number of *cells* in the grid. For each $\mu_{i,j}^c$ we compute the variance between the human signatures from camera c defined as

$$\sigma_{i,j}^c = \frac{1}{n-1} \sum_{k=1;k\neq i}^{n} \rho^2(\mu_{i,j}^c, \mu_{k,j}^c). \tag{6.1}$$

Hence for each human signature $\mathfrak{s}_i^c$ we obtain the vector of discriminants related to our MRC *cells*, $d_i^c = \{\sigma_{i,1}^c, \sigma_{i,2}^c, \ldots, \sigma_{i,m}^c\}$. This idea is similar to methods derived from text retrieval where a frequency of *terms* is used to weight relevance of a *word.* As we do not want to quantize covariance space, we use $\sigma_{i,j}^c$ of MRC *cell* to extract its relevance. The MRC is assumed to be more significant when its variance is larger in the class of humans. Here, it is a kind of *"killing two birds with one stone"*: (1) it is obvious that the most common patterns belong to the background (the variance is small); (2) the patterns which are far from the rest are at the same time the most discriminative (the variance is large).

We thought about normalizing the $\sigma^c_{i,j}$ by the variance *within the class* (similarly to Fisher's linear discriminants). However, the results have shown that such normalization does not improve matching accuracy. We think this is a consequence that the given number of frames per individual is not enough to obtain the reliable variance of MRC *within the class.*

6.1.2.1 Scalability

Discriminative approaches (*e.g.* [Lin 2008, Schwartz 2009]) are often accused of non-scalability. It is true that in these approaches an extensive learning phase is necessary to extract discriminative signatures. This makes these approaches very difficult to apply in real scenario where in every new minute new people appear.

Fortunately, our approach by using very simple discriminative method is able to perform in a real system. It is true that every time when a new signature is created we have to update all signatures in the database. However, for $10,000$ signatures, our update takes less than 30 seconds. Moreover, we do not expect more than such amount of signatures into database as the re-identification approaches are constraint to *one day period* (the strong assumption about the same clothes). Further, one alternative solution might be to use a *reference dataset* as it has already been mentioned in section 5.2.7.

6.1.3 Grid matching

Given the extracted human signatures, we introduce a way to effectively distinguish individuals. As already mentioned the human signatures consist of a set of MRC *cells* structured into a dense grid. In general the matching of two signatures $\mathfrak{s}_A$ and $\mathfrak{s}_B$ is carried out by maximizing the similarity measure. We shift one signature over another in x and y-direction to reduce body alignment issues (see figure 6.3). When shifting signature we preserve relative position between MRC cells to avoid wasting of discriminative property. In our experiments we evaluate the influence of shifting operation on matching accuracy. The similarity between two human signatures $\mathfrak{s}_A$ and $\mathfrak{s}_B$ is defined as

$$S(\mathfrak{s}_A, \mathfrak{s}_B) = \frac{1}{|K|} \sum_{i \in K} \frac{\sigma_{A,i} + \sigma_{B,i}}{\rho(\mu_{A,i}, \mu_{B,i})} \tag{6.2}$$

where K stands for the set of *cells* in signature $\mathfrak{s}_A$ which have corresponding *cells* in signature $\mathfrak{s}_B$; ρ is the covariance distance; $\sigma_{A,i}$ and $\sigma_{B,i}$ are the discriminants of the corresponding MRC-s.

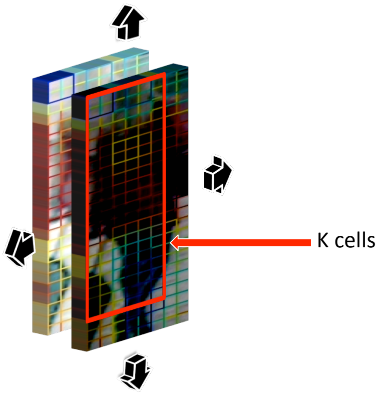


Figure 6.3: The similarity between two MRCG signatures. Every signature is a grid of MRC *cells*. One signature is shifted over another in x and y direction to reduce body alignment issues and maximize the similarity measure.

6.2 COSMATI signature

In this section, we propose to formulate the appearance matching problem as the task of learning a model that selects the most descriptive features for a specific class of objects. Even if we focus on human class, the proposed technique allows to learn a model for other classes of objects (*e.g.* animals, bikes, *etc.*). Our main idea is that different regions of the object appearance ought to be matched using different strategies to obtain a distinctive representation. We refer to this approach as *COrrelation-based Selection of covariance MATrIces* (COSMATI).

In contrast to the previous approach, we do not assume fixed appearance model, but we learn offline the best representation for a specific object class (*e.g.* class of humans). While learning a model (see figure 6.4), we use pairs of images registered in different cameras as training data. The relevant pairs (images of the same object captured in different cameras) are employed as positive samples, while irrelevant pairs (images of different objects captured in different cameras) correspond to negative samples. Learning is performed in a covariance metric space (section 6.2.2) using an entropy-driven criterion (section 6.2.2.2).

Having the learned model, we generate signatures using two operations (figure 6.5). The fist stage overcomes color dissimilarities caused by variations in lighting conditions. We again apply the *histogram equalization* [Hordley 2005] technique to the color channels (RGB) to maximize the entropy in each of these channels and to obtain camera-independent color representation. The second step is responsible for appearance extraction (section 6.2.3) using a general model learned for a specific class of objects (*e.g.* humans). The following sections describe our feature space and the learning, by which the appearance model for matching is generated.

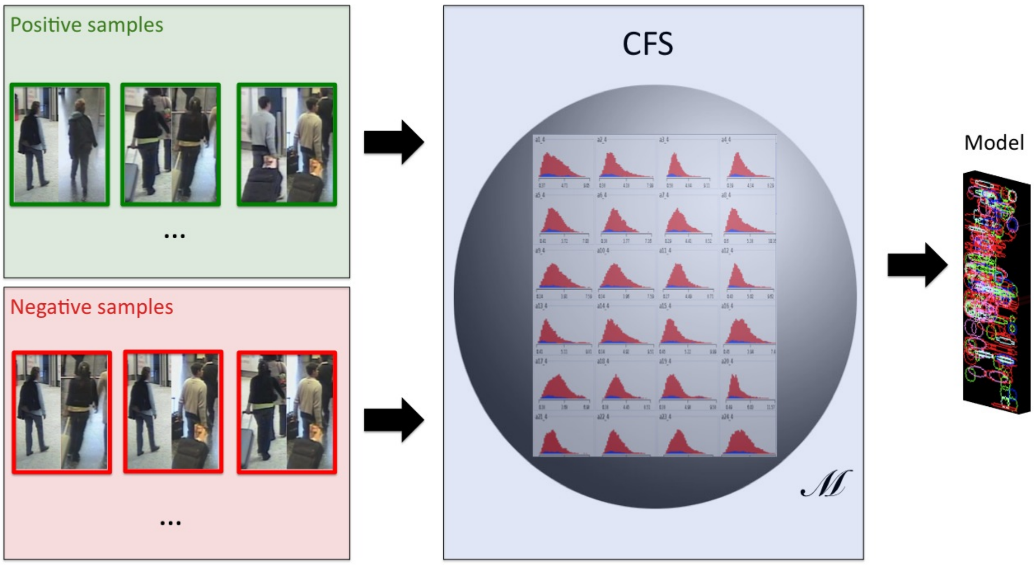


Figure 6.4: Learning a model: correlation-based feature selection (CFS) method selects the best subset of features for appearance matching.

6.2.1 Feature space

Our object appearance is again characterized using the *covariance descriptor* [Tuzel 2006]. We recall that this descriptor encodes information on feature variances inside an image region, their correlations with each other and their spatial layout. The performance of the covariance features is found to be superior to other methods, as rotation and illumination changes are absorbed by the covariance matrix.

In contrast to [Hirzer 2011, Tuzel 2006] and our previously presented approaches, this time we do not limit our covariance descriptor to a single feature vector. Instead of defining *a priori* feature vector, we use a machine learning technique to select covariances that provide the most descriptive representation of the appearance of an object.

Let $L = \{R, G, B, I, \nabla_I, \theta_I, \dots\}$ be a set of feature layers, in which each layer is a mapping such as color, intensity, gradients and filter responses (texture filters, *i.e.* Gabor, Laplacian or Gaussian). Instead of using covariance between all of these layers, which would be computationally expensive, we compute covariance matrices of a few relevant feature layers. These relevant layers are selected depending on the region of an object (see section 6.2.2).

In addition, let layer $\mathfrak{D}$ be a distance between the center of an object and the pixel location. Covariance of the distance layer $\mathfrak{D}$ and three other layers l ($l \in L$) form our descriptor, which is represented by a 4×4 covariance matrix. By using distance $\mathfrak{D}$

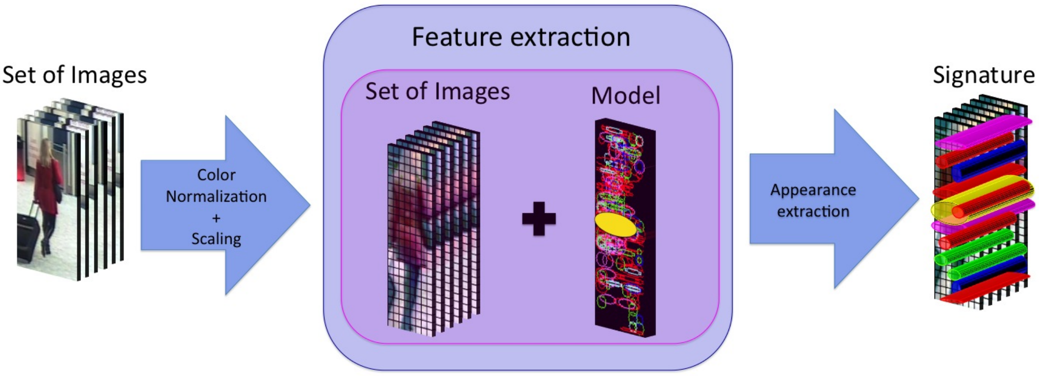


Figure 6.5: Extraction of signature using the previously learned model.

in every covariance, we keep a spatial layout of feature variances, which is rotation invariant. State of the art techniques very often use pixel location (x, y) instead of distance $\mathfrak{D}$, yielding better description of an image region. Conversely, among our detailed experimentation, using $\mathfrak{D}$ rather than (x, y), we did not decrease the recognition accuracy in the general case, while decreasing the number of features in covariance matrix. This discrepancy may be due to the fact that we hold spatial information twofold, (1) by $\mathfrak{D}$ in covariance matrix and (2) by location of a rectangular sub-region from which the covariance is extracted. We constraint our covariances to 4×4 matrices, ensuring computational efficiency. Also, bigger covariance matrices tend to include superfluous features which can clutter the appearance matching. 4×4 matrices provide sufficiently descriptive correlations while keeping low computational time needed for calculating generalized eigenvalues during distance computation.

Different combinations of three feature layers produce different types of covariance descriptor. By using different covariance descriptors, assigned to different locations in an object, we are able to select the most discriminative covariances according to their positions. The idea is to characterize different regions of an object by extracting different kinds of features (*e.g.* when comparing human appearances, edges coming from shapes of arms and legs are not discriminative enough in most cases as every instance possess similar features). Taking into account this phenomenon, we minimize redundancy in an appearance representation by an entropy-driven selection method.

Let us define a meta covariance feature space as

$$\mathfrak{C} = \{(P, cov(\mathfrak{D}, l_i, l_j, l_k)) : l_i, l_j, l_k \in L; P \in \mathbf{P}\}, \tag{6.3}$$

where $\mathbf{P}$ is a set of rectangular sub-regions of the object; cov is the covariance of features; l_i, l_j, l_k are color/intensity/filter layers. Figure 6.6 shows different feature layers and examples of three different types of covariance descriptor. The dimension

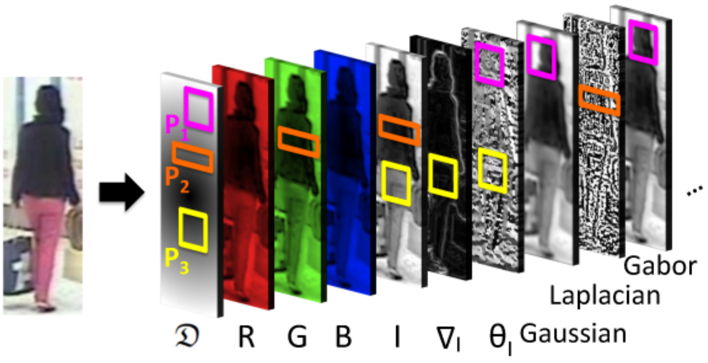


Figure 6.6: A meta covariance feature space. Example of three different covariance features. Every covariance is extracted from *distance* layer ($\mathfrak{D}$) and three channel functions (*e.g.* bottom covariance feature is extracted from region P_3 using layers: $\mathfrak{D}$, I-intensity, ∇_I-gradient magnitude and θ_I-gradient orientation).

n of our meta covariance feature space is the product of the number of possible rectangular regions and the number of different types of our covariance descriptor.

6.2.2 Learning by correlations in a covariance metric space

Let $\mathbf{a}_i^c = \{\mathbf{a}_{i,1}^c, \mathbf{a}_{i,2}^c, \dots \mathbf{a}_{i,m}^c\}$ be a set of relevant observations of an object i in camera c, where $\mathbf{a}_{ij}^c$ is a n-dimensional vector composed of different covariance features extracted from image j of object i using n-dimensional meta covariance feature space $\mathfrak{C}$. We define the distance vector between two samples $\mathbf{a}_{i,j}^c$ and $\mathbf{a}_{k,l}^{c'}$ as follows

$$\delta(\mathbf{a}_{i,j}^c, \mathbf{a}_{k,l}^{c'}) = \left[\rho(\mathbf{a}_{i,j}^c[z], \mathbf{a}_{k,l}^{c'}[z])\right]_{z=1\dots n}^T, \tag{6.4}$$

where ρ is a geodesic distance between covariance matrices [Förstner 1999], and $\mathbf{a}_{i,j}^c[z]$, $\mathbf{a}_{k,l}^{c'}[z]$ are the corresponding covariance matrices (the same region P and the same combination of layers). The index z is an iterator of $\mathfrak{C}$.

We cast the appearance matching problem into the following *distance learning* problem. Let δ^+ be the distance vectors computed using pairs of relevant samples (of the same people captured with different cameras, $i = k$, $c \neq c'$) and let δ^- be distance vectors computed between pairs of related irrelevant samples ($i \neq k$, $c \neq c'$). Pairwise elements δ^+ and δ^- are distance vectors, which stand for positive and negative samples, respectively. These distance vectors define *covariance metric space*. Given δ^+ and δ^- as training data, our task is to find a general model of appearance to maximize matching accuracy by selecting relevant covariances and thus defining a distance.

6.2.2.1 Riemannian geometry

Covariance descriptors as positive definite symmetric matrices lie on a manifold that is not a vector space (they do not lie on Euclidean space). We have already mentioned that specifying the covariance manifold as Riemannian we can apply differential geometry [Pennec 2006] to perform usual operations such as mean or distance. However, learning on a manifold space is a difficult and unsolved challenge. Most of methods (*e.g.* [Tuzel 2008] and also our boosting-based method) perform classification by regression over the mappings of the training data on a suitable tangent plane. Defining tangent plane over the Karcher mean of the positive training data points, we can preserve a local structure of the points. Unfortunately, the models extracted using means of the positive training data points tend to be overfitted. These models concentrate on tangent planes obtained from training data and do not have generalization properties. In section 5.2, we did not require generalized properties as we were learning the model for a particular individual (learning was used for selecting discriminative features). The present work addresses the problem of learning a general model for matching a specific class of objects (*e.g.* humans). Thus, we need a method which will not be based on planes created using fixed tangent planes dependent on particular humans. In [Hirzer 2011], authors avoid this problem by casting covariance matrices into *Sigma Points* that lie on approximated covariance space.

Neither using tangent planes over the Karcher means extracted from training data, nor casting covariances into *Sigma Points* satisfy our matching model. As we want to take full advantage of covariance manifold space, we propose to extract a general model for appearance matching by identifying the most salient features. Based on the hypothesis: *"A good feature subset is one that contains features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other"* [Hall 1999], we extract our appearance model using covariance features f_z ($f_z \in \mathfrak{C}$) selected by *correlation-based feature selection* technique.

6.2.2.2 Correlation-based feature selection (CFS)

Correlation-based feature selection (CFS) [Hall 1999] is a filter algorithm that ranks feature subsets according to a correlation-based evaluation function. This evaluation function favors feature subsets which contain features highly correlated with the class and uncorrelated with each other. In our *distance learning* problem, we define positive and negative class by δ^+ and δ^-, as relevant and irrelevant pairs of samples. Further, feature $f_z \in \mathfrak{C}$ is characterized by a distribution of the zth elements in distance vectors δ^+ and δ^-. The feature-class correlation and the feature-feature inter-correlation is measured using a symmetrical uncertainty model [Hall 1999]. As this model requires nominal valued features, we discretize f_z using the method of Fayyad and Irani [Fayyad 1993]. Let X be a nominal valued feature obtained by discretization of f_z.

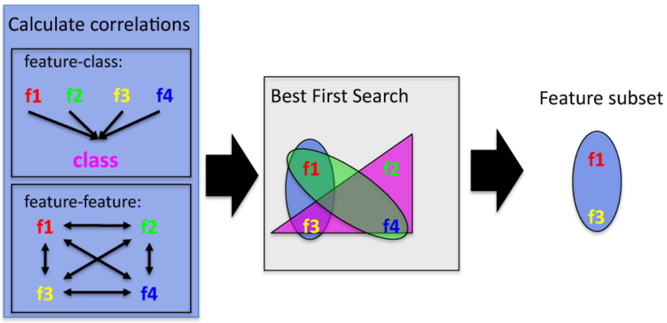


Figure 6.7: Correlation-based feature selection.

We assume that a probabilistic model of X can be formed by estimating the probabilities of the values $x \in X$ from the training data. The uncertainty can be measured by entropy defined as

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x). \tag{6.5}$$

A relationship between features X and Y can be given by

$$H(X \mid Y) = -\sum_{y \in Y} p(y) \sum_{x \in X} p(x \mid y) \log_2 p(x \mid y). \tag{6.6}$$

The amount by which the entropy of X decreases reflects additional information on X provided by Y and is called the *information gain* (*mutual information*)

$$\begin{aligned} Gain &= H(X) - H(X \mid Y) \qquad (6.7) \\ &= H(Y) - H(Y \mid X) \\ &= H(X) + H(Y) - H(X, Y). \end{aligned}$$

Even if the *information gain* is a symmetrical measure, it is biased in favor of features with more discrete values. Thus, the symmetrical uncertainty r_{XY} is used to overcome this problem

$$r_{XY} = 2 \times \left[\frac{Gain}{H(X) + H(Y)} \right]. \tag{6.8}$$

Having the correlation measure, subset of features $\mathfrak{S}$ is evaluated using function $\mathfrak{M}(\mathfrak{S})$ defined as

$$\mathfrak{M}(\mathfrak{S}) = \frac{k \, \overline{r_{cf}}}{\sqrt{k + k\,(k+1)\,\overline{r_{ff}}}}, \tag{6.9}$$

where k is the number of features in subset $\mathfrak{S}$, $\overline{r_{cf}}$ is the average feature-class correlation and $\overline{r_{ff}}$ is the average feature-feature inter-correlation

$$\overline{r_{cf}} = \frac{1}{k} \sum_{f_z \in \mathfrak{S}} r_{cf_z}, \qquad \overline{r_{ff}} = \frac{2}{k\,(k-1)} \sum_{\substack{f_i, f_j \in \mathfrak{S} \\ i<j}} r_{f_i f_j}, \tag{6.10}$$

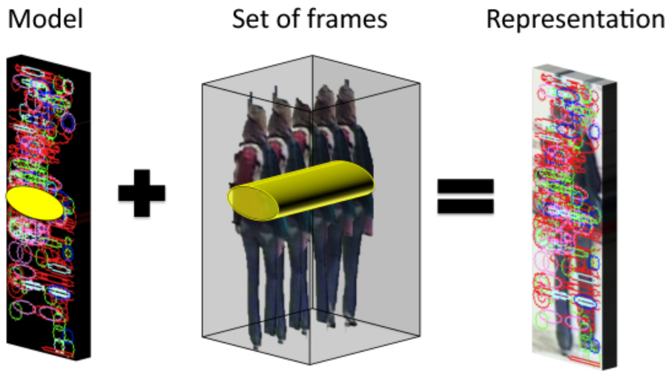


Figure 6.8: Extraction of the appearance using the model (the set of features selected by CFS). Different colors and shapes in the model refer to different kinds of covariance features.

where c is the class, or relevance feature, which is $+1$ on δ^+ and -1 on δ^-. The numerator in equation (6.9) indicates predictive ability of subset $\mathfrak{S}$ and the denominator stands for redundancy among the features.

Equation (6.9) is the core of CFS, which ranks feature subsets in the search space of all possible feature subsets. Since exhaustive enumeration of all possible feature subsets is prohibitive in most cases, a heuristic search strategy has to be applied. We have investigated different search strategies, among which *best first search* [Rich 1991] performs the best.

Best first search is an AI search strategy that allows backtracking along the search path. Our *best first* starts with no feature and progresses forward through the search space adding single features. The search terminates if T consecutive subsets show no improvement over the current best subset (we set $T = 5$ in experiments). By using this stopping criterion we prevent the best first search from exploring the entire feature subset search space. Figure 6.7 illustrates CFS method. Let Π be the output of CFS that is the feature subset of $\mathfrak{C}$. This feature subset Π forms a model that is used for appearance extraction and matching.

6.2.3 Appearance extraction

Having the general model Π for a specific object class (*e.g.* humans), we compute the appearance representation using a set of frames (see figure 6.8). Our method belongs to the group of *multiple-shot* approaches, where multiple images of the same person are used to extract a compact representation. This representation can be seen as a *signature* of the multiple instances.

Using our model Π, a more straightforward way to extract appearance would be to compute covariance matrices of a video volume directly for every $f_z \in \Pi$. However, using volume covariance we loose information on real feature distribution (time feature characteristics are merged - see section 5.2.2). Thus, we compute MRC using a Riemannian manifold using covariance features presented in a general model. The mean covariance matrix as an intrinsic average blends appearance information from multiple images. For every feature $f_z \in \Pi$ we compute the mean covariance matrix. The set of mean covariance matrices stands for an appearance representation of an object (*signature*).

6.2.4 Appearance matching

Let $\mathfrak{A}$ and $\mathfrak{B}$ be the object signatures. The signature consists of mean covariance matrices extracted using set Π. The similarity between two signatures $\mathfrak{A}$ and $\mathfrak{B}$ is defined as

$$S(\mathfrak{A}, \mathfrak{B}) = \frac{1}{|\Pi|} \sum_{i \in \Pi} \frac{1}{\rho(\mu_{\mathfrak{A},i}, \mu_{\mathfrak{B},i})}, \tag{6.11}$$

where ρ is a geodesic distances [Förstner 1999], $\mu_{\mathfrak{A},i}$ and $\mu_{\mathfrak{B},i}$ are mean covariance matrices extracted using covariance feature $i \in \Pi$. Using the average of similarities computed on feature set Π the appearance matching becomes robust to noise.

6.3 Conclusion

This chapter presented two efficient methods for human re-identification. These *mulitple-shot* approaches are less computationally expensive than boosting-based methods. We again employ *mean Riemannian covariance* as the feature for characterizing images regions. The first method (MRCG) is based on predefined appearance model (dense grid), while the second technique (COSMATI) learns an appearance representation during an offline stage.

MRCG is one alternative to the boosting scheme, giving an efficient way to take advantage of information from the appearance of other individuals. Nonetheless, as it was highlighted in section 5.2.7, we might need to extract a *reference dataset* to provide the scalable solution. Data selection needed for the discriminative analysis, a size of the reference dataset *etc.* raise numerous questions, producing new open issues.

COSMATI is designed to be faster than MRCG. It is based on small covariance matrices which are selected using an entropy-driven criterion. The disadvantage of this approach is that it requires the offline learning phase, which is based on training data from each camera to obtain the distinctive representation. We also underline that COSMATI can be combined with MRC discriminants, providing the descriptor which use the most information for generating human signature.

In section 6.2.2.1 we have discussed problem related to learning on a manifold, where model is usually created using tangent planes over the Karcher means of the positive training data which can lead to the overfitted classifier. In the result, we use CFS method to find the most descriptive features for a specific class of objects. However, we have to stress that such selection methods are usually used as pre-processing steps for machine learning algorithms. Although learning on a manifold is still an open issue, there are already approaches which explore a manifold space either by clustering methods [Goh 2008a] or by *control points* [Sen 2008]. Further studies should look for machine learning methods which could prosper in finding subspaces on a manifold determining the metric for appearance matching.

Chapter 7

Experimental results

"*It is the weight, not numbers of experiments that is to be regarded.*"
(Isaac Newton)

This chapter demonstrates an extensive evaluation of our techniques. Studying different descriptor parameters and analyzing recognition performance on various datasets, targeting different challenges, we show the main limitations of the proposed approaches. We firstly describe evaluation metrics for re-identification (section 7.1). Then, in section 7.2, we present publicly available state of the art datasets and introduce new data for exploring the *multiple-shot* case. We carry out detailed experiments of proposed methods, investigating their pros and cons in sections 7.3, 7.4 and 7.5. Finally, in section 7.6, the reader can find comparison to state of the art methods on various datasets.

7.1 Evaluation metrics for re-identification

The performance of appearance-based re-identification system can be difficult to quantify because re-identification performance is dependent upon the size of the dataset. Let us consider the performance of random guessing. On a dataset of size 50 pedestrians the performance is 2%. However, on a dataset of size 1000 pedestrians, the same algorithm performs 0.1%. If a human operator has an algorithm with a 10% re-identification rate, it would be useless on the first dataset, but very practical on the later. The problem is relatively easy for very small datasets. Nonetheless, it becomes challenging as the number of possible matches grows. Given a single human signature, the chance of choosing the correct match is inversely proportional to the number of considered signatures.

7.1.1 Cumulative matching characteristic (CMC) curve

In [Gray 2007], re-identification is considered as a ranking problem. In this framework a ranking is induced on the elements of the dataset and the probability that the correct match has a rank equal to or less than the rank score value is plotted over the size of the test set. This performance evaluation metric is known as the *cumulative matching characteristic (CMC) curve*. The CMC curve represents the

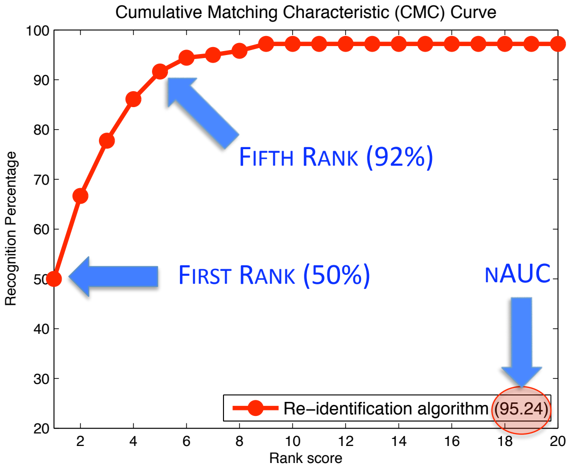


Figure 7.1: Example of CMC curve with computed nAUC value.

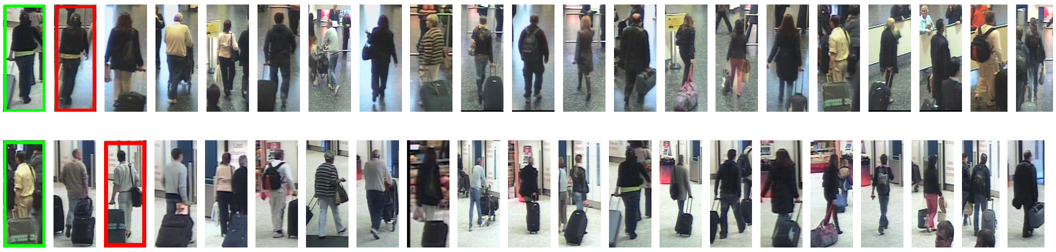


Figure 7.2: Example of the person re-identification on i-LIDS-MA. The left-most image is the probe image. The remaining images are the top 20 matched gallery images. The red boxes highlight the correct matches in the list.

expectation of finding the correct match in the top n matches. In figure 7.1, an example of the CMC curve is illustrated. The first rank describes the percentage of the queries in which the correct match is found on the first position in the list of the most similar signatures (see figure 7.2). The fifth rank represents the percentage of the queries in which the correct match is found in the first five positions in the list.

7.1.2 Normalized area under curve (nAUC)

We also employ a quantitative scalar appraisal of CMC curve by computing the normalized area under curve (nAUC) value (see figure 7.1).

7.1.3 Evaluation scheme

We generate a human signature for each person detected and tracked in a video analysis system. Let us denote a signature as $\mathfrak{s}_i^c$, where i encodes the person identity and c denotes the camera. The task is to find for each signature its corresponding signature in another camera. It is realized by querying the database of signatures $\mathfrak{s}_j^{c'}$, where $c \neq c'$ with signature of interest $\mathfrak{s}_i^c$. The results of the query is the list of the most similar signatures ordered by increasing dissimilarity (see figure 7.2). The position in the list of the true match is called the rank score. The performance is plotted into the CMC curve to present the re-identification accuracy. This evaluation scheme is analogous to a standard surveillance scenario where an operator queries a system with images of the same individual captured over a short period of time from a particular camera to find him/her in a network of cameras.

7.2 Datasets

In this section we present the most used state of the art datasets for person re-identification, highlighting their challenging aspects. We also report the best results that have been published so far on these data. Further, we introduce two new sets of individuals from i-LIDS data, focusing on evaluation of the *multiple-shot* case.

7.2.1 CAVIAR

This dataset comes from the EC Funded CAVIAR project[1]. Video clips were recorded from two different points of view in a shopping center in Lisbon. The first one shows a view of the corridor, while the second shows a frontal view of the scene. The resolution is half-resolution PAL standard (384 × 288 pixels, 25 frames per second). Small appearances and a low resolution in one of the two cameras make the data very challenging (see figure 7.3).

CAVIAR4REID

In [Cheng 2011], the authors used a ground truth provided by CAVIAR project to extract human appearances. Of the 72 different individuals identified (with images varying from 17 × 39 to 72 × 144), 50 are captured by both views and 22 from only one camera. For each pedestrian captured by both cameras, 10 images were selected by maximizing the variance with respect to resolution changes, light conditions, occlusions, and pose changes. In the result, this dataset is very challenging due to significant differences in human appearance registered in disjoint camera views. Table 7.1 summarizes current state-of-the-art results on the CAVIAR4REID

[1] CAVIAR webpage: *http://homepages.inf.ed.ac.uk/rbf/CAVIAR*

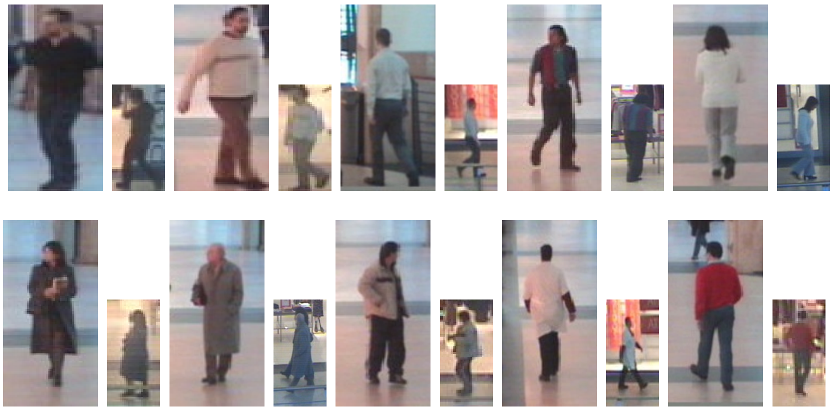


Figure 7.3: The sample images from CAVIAR dataset. Pairs of images highlight appearance changes due to different camera resolution and illumination changes.

dataset. It is worth noting, that even in the case of relatively small dataset (only 50 individuals), the recognition accuracy in the first rank is very low (less than 20%). This shows that appearance-based re-identification may be extremely difficult due to significant camera changes.

Reference	Method	Accuracy (1st - 5th rank)
[Cheng 2011]	PS - single-shot	9% - 31%
[Farenzena 2010]	SDALF - single-shot	7% - 25%
[Cheng 2011]	CPS - multiple-shot	17% - 47.8%
[Farenzena 2010]	SDALF - multiple-shot	9% - 39%

Table 7.1: State of the art results on CAVIAR4REID dataset. The first and the fifth ranks are reported.

7.2.2 ETHZ

ETHZ dataset was originally used for evaluation of human detection algorithms [Ess 2007]. The video sequences are captured from moving cameras with a resolution of 640×480 pixels, providing a range of variations in people appearances. In [Schwartz 2009] this dataset has been adjusted for re-identification purposes[2]. The modified dataset consists of three sequences:

- SEQ. #1 contains 83 pedestrians,
- SEQ. #2 contains 35 pedestrians,
- SEQ. #3 contains 28 pedestrians.

[2] ETHZ dataset for appearance-based modeling, William Robson Schwartz webpage.

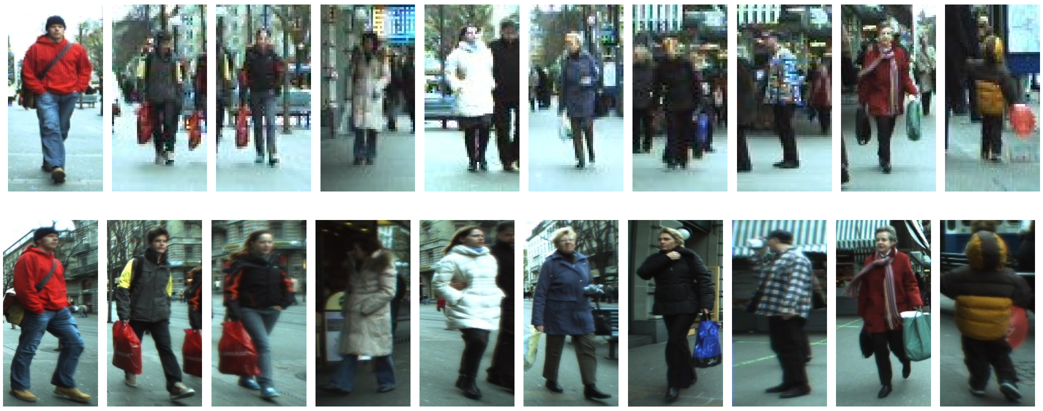


Figure 7.4: The sample images from ETHZ dataset (SEQ. #1). Top and bottom lines correspond to images from the beginning and from the end of the sequence, respectively.

Unfortunately, even if the video sequences are acquired from moving camera and the data contain changing appearances, all pedestrians are extracted from the same camera. In our belief, despite such challenging aspects as illumination changes and occlusions, the ETHZ dataset is not challenging enough to evaluate re-identification approaches. One of the most challenging issues in the re-identification problem is due to different camera settings, different color responses, different camera view points and different environments, which is not the case for this dataset. Thus, the performance accuracy obtained on this dataset is much higher than on the rest of considered datasets. However, as this dataset is popular for appearance-based model evaluation, we have also included the ETHZ in our testing database. Figure 7.4 illustrates a few examples of images from ETHZ dataset and table 7.2 presents state of the art results.

7.2.3 i-LIDS

The image library for intelligent detection systems (i-LIDS) is the Home Office benchmark for Video Analytics (VA) systems developed in partnership with the Centre for the Protection of National Infrastructure (CPNI). The i-LIDS video library contains different scenarios each representing real world CCTV footage (*e.g.* abandoned baggage, parked vehicle or multiple camera tracking scenario). The most valuable data for detection and re-identification problems can be found in the Multiple Camera Tracking Scenario (MCTS) dataset with five camera views. The MCTS dataset contains approximately 50 hours of footage. The following sets of individuals are extracted from the mentioned MCTS data.

Reference	Method	Sequence	Accuracy (1st - 5th rank)
[Cheng 2011]	CPS	SEQ. # 1	98% - 100%
		SEQ. # 2	95.5% - 100%
		SEQ. # 3	99% - 100%
[Farenzena 2010]	SDALF	SEQ. # 1	90% - 93%
		SEQ. # 2	91% - 98%
		SEQ. # 3	94% - 97%
[Bazzani 2010]	HPE	SEQ. # 1	85% - 93%
		SEQ. # 2	81% - 93%
		SEQ. # 3	87% - 96%
[Schwartz 2009]	PLS	SEQ. # 1	80% - 87%
		SEQ. # 2	75% - 84%
		SEQ. # 3	76.5% - 85%

Table 7.2: State of the art results on ETHZ dataset. The first and the fifth ranks are reported.

i-LIDS-119

This evaluation dataset has been extracted automatically by [Zheng 2009]. It contains 476 images with 119 individuals (see figure 7.5) registered by two different cameras. This dataset is very challenging since there are many occlusions and often only the top part of the person is visible. Table 7.3 presents state of the art performance on this dataset.

Reference	Method	Accuracy (1st - 5th rank)
[Cheng 2011]	PS - single-shot	30% - 51%
[Farenzena 2010]	SDALF - single-shot	28% - 48%
[Zheng 2009]	Group context - single-shot	24% - 44%
[Bazzani 2010]	HPE - multiple-shot	22% - 45%
[Cheng 2011]	CPS - multiple-shot	47% - 73%
[Farenzena 2010]	SDALF - multiple-shot	50% - 70%

Table 7.3: State of the art results on i-LIDS-119 dataset. The first and the fifth ranks are reported.

i-LIDS-119 has been prepared for *single-shot* case where the number of images per individual is not significant (only one image is necessary). Unfortunately, for *multiple-shot* case, this dataset does not fit very well because the number of images per individual is too low (in average 4). Therefore we have extracted two new sets of individuals from i-LIDS data (i-LIDS-MA and i-LIDS-AA) to investigate more precisely the advantage of using tracking results in building the human signatures. These datasets finally satisfy all requirements of *multiple-shot* person re-identification (at least 2 non-overlapping cameras and at least 10 images per individual extracted

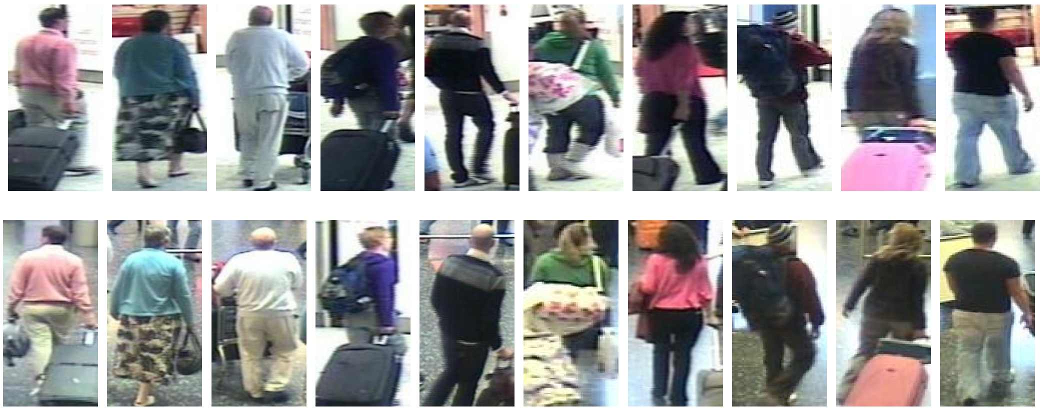


Figure 7.5: The sample images from i-LIDS-119 dataset. Top and bottom lines correspond to images from different cameras. Columns illustrate the same person.

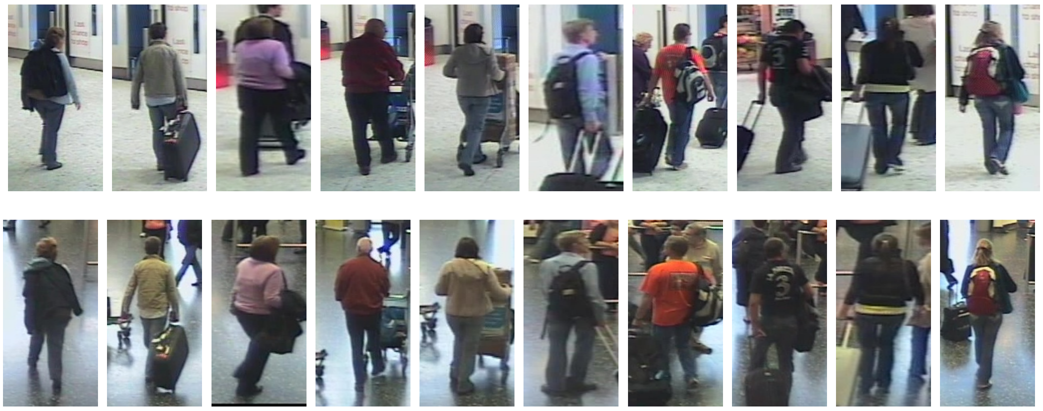


Figure 7.6: The sample images from i-LIDS-MA dataset. Top and bottom lines correspond to images from different cameras. Columns illustrate the same person.

from each camera).

i-LIDS-MA (Manually Annotated)

This dataset contains 40 individuals extracted from two cameras. For each individual 46 frames have been annotated manually from both cameras. Therefore we have $40 \times 2 \times 46 = 3680$ annotated images. Figure 7.6 illustrates sample images from i-LIDS-MA.

i-LIDS-AA (Automatically Annotated)

The manually annotated dataset (i-LIDS-MA) does not reflect real video surveillance scenario where humans are detected and tracked automatically. Consequently, we

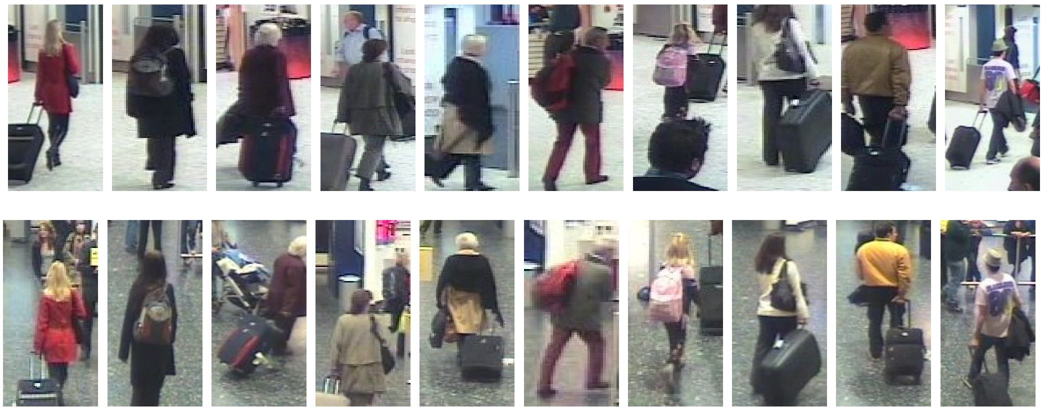


Figure 7.7: The sample images from i-LIDS-AA dataset. Top and bottom lines correspond to images from different cameras. Columns illustrate the same person.

have applied HOG-based human detector and tracker to obtain multiple images of individuals seen from both cameras. In this case, detection and tracking results are noisy which makes the dataset more challenging. This dataset contains 100 individuals. For each individual we have extracted automatically a different number of frames depending on tracking difficulties. In total, the dataset contains 10754 images. Figure 7.7 illustrates sample images from i-LIDS-AA.

Both datasets i-LIDS-MA and i-LIDS-AA have been extracted for studying more carefully advantages of using multiple images in generating human signature. We have already shared these datasets with at least 10 research laboratories.

7.2.4 TSP

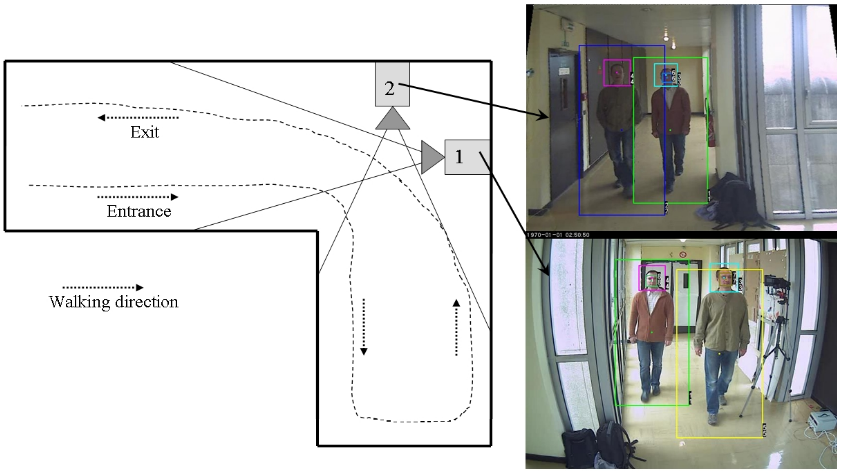


Figure 7.8: The acquisition framework of TSP data [Corvee 2012].

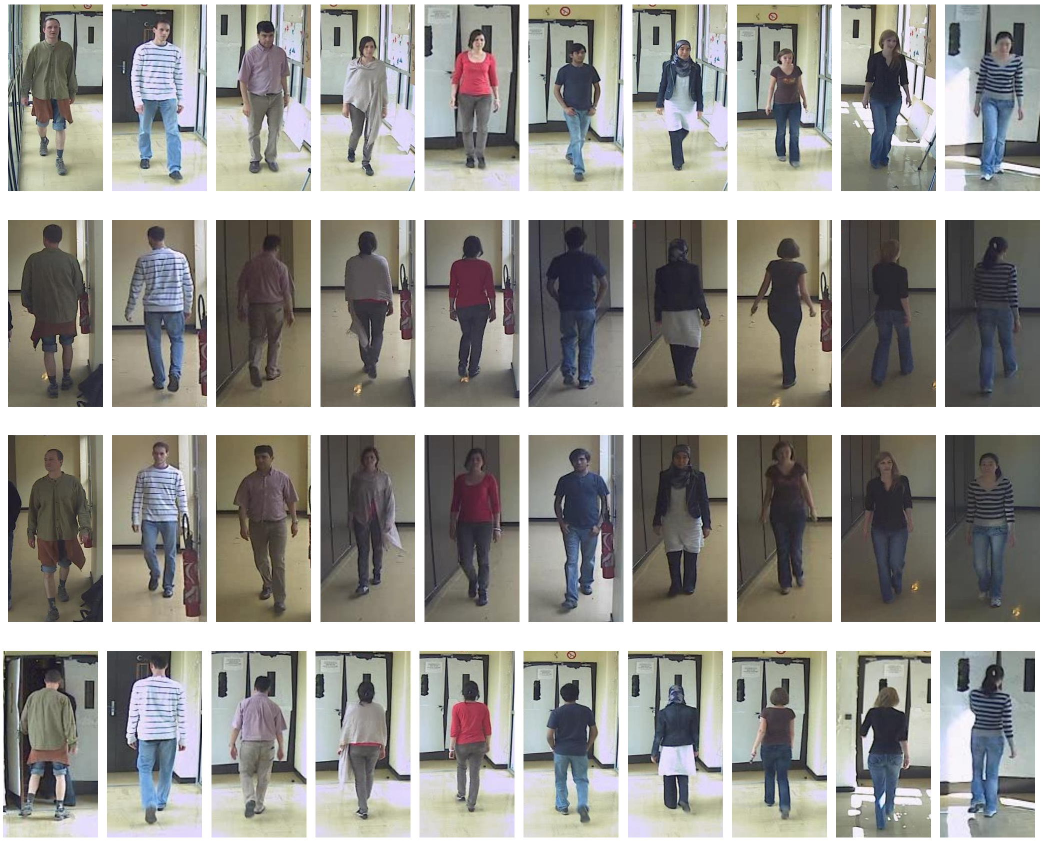


Figure 7.9: The sample images from TSP dataset: (1) the first row corresponds to the frontal view registered by camera 1, (2) the second and the third row correspond to the back view and the frontal view registered by camera 2, respectively, (2) the last row corresponds to the back view registered by camera 1.

TSP dataset was extracted in VIDEO-ID project[3] in TELECOM Sud Paris (TSP). This data was proposed for evaluation of human detection, face recognition and person re-identification algorithms [Corvee 2012]. People are detected and tracked in two cameras by employing LBP detector . This new database was created from two-camera network with non-overlapping views. Figure 7.8 illustrates an acquisition setup. The dataset contains 23 individuals which are changing clothing during acquisition process. In total we can distinguish 36 different subjects. Each subject is registered three times: (1) the frontal view while entering the scene registered by camera 1, (2) the back view and the frontal view registered while tracking a person in camera 2, (3) the back view while existing the scene registered by camera 1 (see figure 7.9). Concerning evaluation of person re-identification, the main challenging aspects of this dataset are significant illumination changes, inaccurate people detections and pose variations (frontal and back views).

[3] VIDEO-ID webpage: *http://www-sop.inria.fr/pulsar/projects/videoid/*

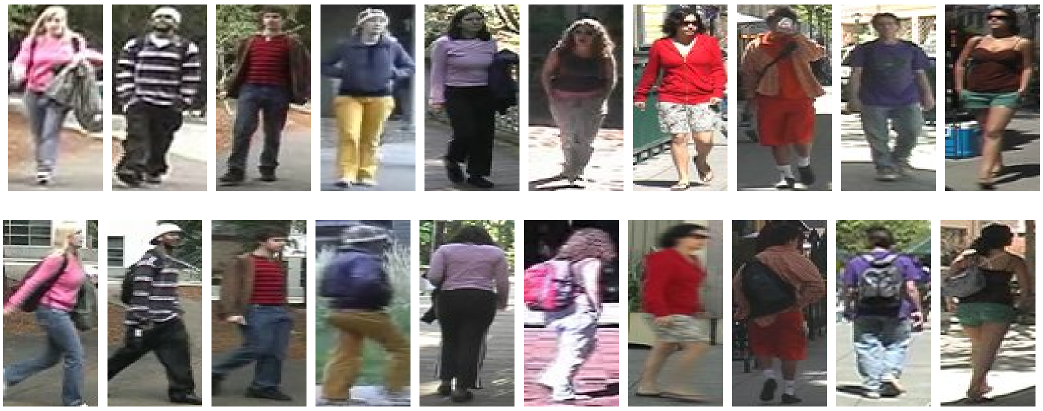


Figure 7.10: The sample images from VIPER dataset. Top and bottom lines correspond to images from different cameras. Columns illustrate the same person.

7.2.5 VIPER

VIPER [Gray 2007] dataset contains two views of 632 pedestrians. Each image is cropped and scaled to be 128 × 48 pixels. Images of the same pedestrian have been taken from different cameras, under different viewpoints, poses and lighting conditions. The primary motivation of [Gray 2007] was to propose a dataset which can be used for learning and evaluation of the viewpoint invariant approaches. Thus, the dataset contains pairs which viewpoint angle changes from 45 up to 180 degrees. The quality of the images varies. The video was compressed before processing and as a result, the images have spatially sub sampled chrominance channels, as well as some minor interlacing and compression artifacts. It is the most challenging dataset currently available for the *single-shot* human re-identification (see figure 7.10). Table 7.4 presents state of the art performance on 316 pairs randomly selected from VIPER dataset. Unfortunately, this dataset does not correspond to a video surveillance scenario, in which many images of the same person are available to generate a signature (*multiple-shot* case).

Reference	Method	Accuracy (1st - 5th rank)
[Cheng 2011]	PS	21% - 45%
[Zheng 2011]	PRDC	15% - 38%
[Hirzer 2011]	Des+Dis	19% - 39%
[Dikmen 2010]	LMNN-R	21% - 49%
[Farenzena 2010]	SDALF	20% - 39%
[Prosser 2010]	PRSVM	13% - 37%
[Gray 2008]	ELF	12% - 31%

Table 7.4: State of the art results on VIPER dataset. The first and the fifth ranks are reported.

7.3 Single-shot approaches

This section presents a detailed performance analysis of *dominant color descriptor* based (DCD) signature and of *spatial covariance region* based (SCR) signature. Both signatures are evaluated using i-LIDS-MA dataset (section 7.2.3). We study different steps of our methods, investigating the sensitivity of parameters and their impact on the recognition accuracy. The parameters can have a significant influence on performance so proper evaluation is necessary.

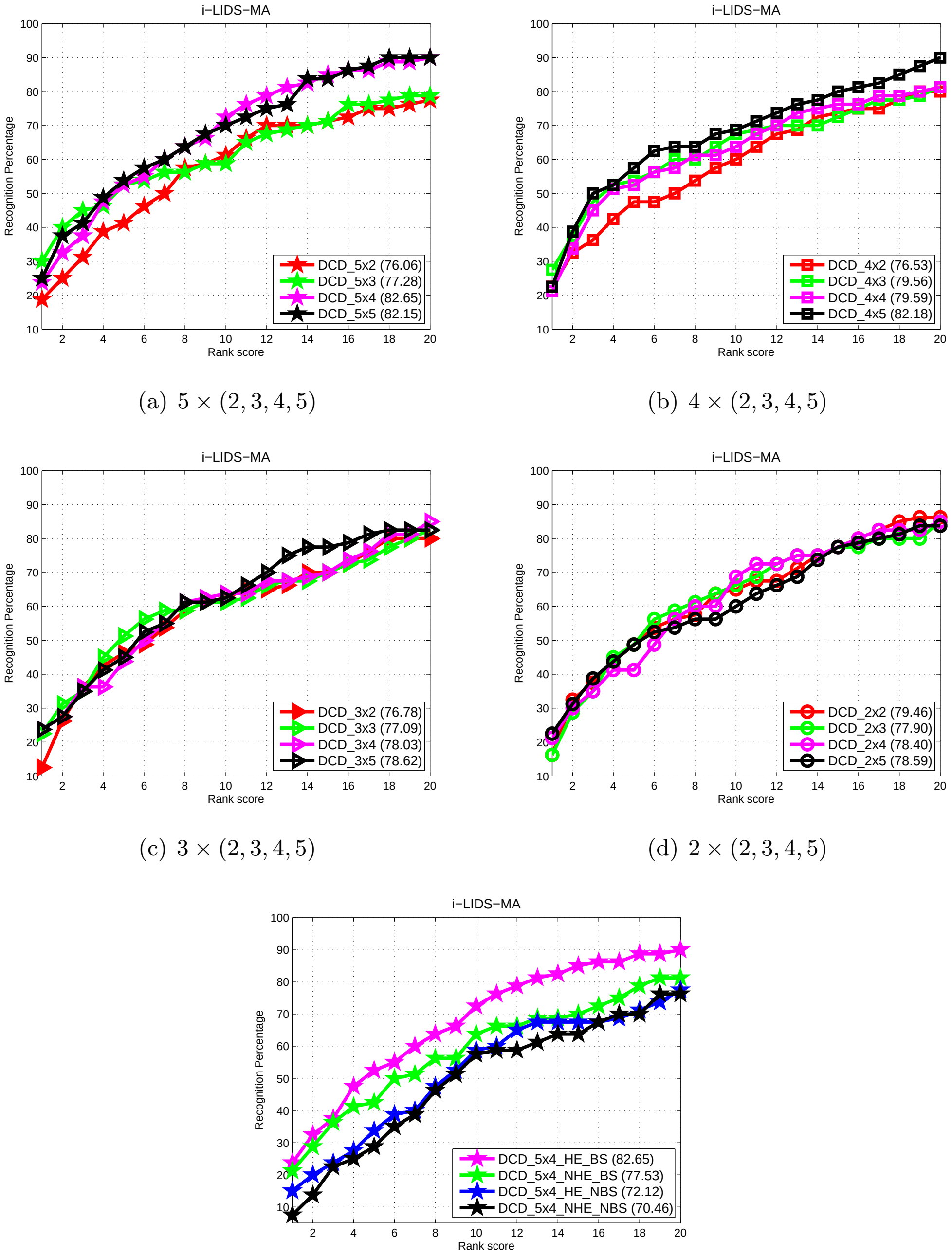
7.3.1 DCD signature

Similarly to [Yang 2008], we set parameter $\alpha = 2$ and $T_d = 25$ (see section 4.1.3). Colors are represented in the perceptually uniform CIE LUV color space. From every human image, foreground regions are extracted using *background subtraction* algorithm (section 4.1.1). Then, we normalize color by using *histogram equalization* (section 4.1.2). We investigate different numbers of dominant colors for the upper and the lower body representation. Then, the influence of *background subtraction* and *histogram equalization* is evaluated.

Number of dominant colors: Representing a human appearance, we have to provide the maximum number of dominant colors for each body part. Figure 7.11(a - d) illustrates different combinations for selecting the maximum number of dominant colors to represent the upper and the lower body part. Our results show that the best performance is obtained for the number of dominant colors around $4 - 5$. This result confirms the evaluation presented in [Yang 2008].

Background subtraction and histogram equalization: Figure 7.11(e) shows the impact of *background subtraction* and *histogram equalization* on recognition rate. We can notice that the *background subtraction* step is the crucial part, while extracting DCD signature. The results clearly show that background regions can significantly disturb our descriptor.

Histogram equalization also improves the recognition accuracy. However, we would expect more significant improvement after *histogram equalization* (especially in the first ranks), what is not confirmed by the evaluation. This discrepancy might be due to the fact, that *histogram equalization* disturbs an initial color space, deteriorating the clustering step. Histogram equalization is independently applied to each color channel, which in the result produce color artifacts slightly contaminating color space.



(a) $5 \times (2, 3, 4, 5)$

(b) $4 \times (2, 3, 4, 5)$

(c) $3 \times (2, 3, 4, 5)$

(d) $2 \times (2, 3, 4, 5)$

(e) $(HE, NHE) \times (BS, NBS)$

Figure 7.11: CMC curves obtained on i-LIDS-MA dataset using DCD signature: (a),(b),(c),(d) show different combinations of the maximum dominant color numbers for the upper and the lower body part (upper × lower); (e) the performance of DCD signature applied on images with and without *background subtraction* step (BS *vs*. NBS), and with and without *histogram equalization* (HE *vs*. NHE).

7.3.2 SCR signature

Feature vector: We evaluate our SCR signature using different feature combinations for covariance descriptor (section 4.2.2). We select the following five feature vectors:

$$F_1 = \left[x, y, R_{xy}, G_{xy}, B_{xy}, \nabla^R_{xy}, \theta^R_{xy}, \nabla^G_{xy}, \theta^G_{xy}, \nabla^B_{xy}, \theta^B_{xy}\right], \tag{7.1}$$

$$F_2 = \left[\mathcal{D}, R_{xy}, G_{xy}, B_{xy}, \nabla^R_{xy}, \theta^R_{xy}, \nabla^G_{xy}, \theta^G_{xy}, \nabla^B_{xy}, \theta^B_{xy}\right], \tag{7.2}$$

$$F_3 = \left[R_{xy}, G_{xy}, B_{xy}, \nabla^R_{xy}, \theta^R_{xy}, \nabla^G_{xy}, \theta^G_{xy}, \nabla^B_{xy}, \theta^B_{xy}\right], \tag{7.3}$$

$$F_4 = \left[x, y, I_{xy}, \nabla^I_{xy}, \theta^I_{xy}\right], \tag{7.4}$$

$$F_5 = \left[x, y, R_{xy}, G_{xy}, B_{xy}\right], \tag{7.5}$$

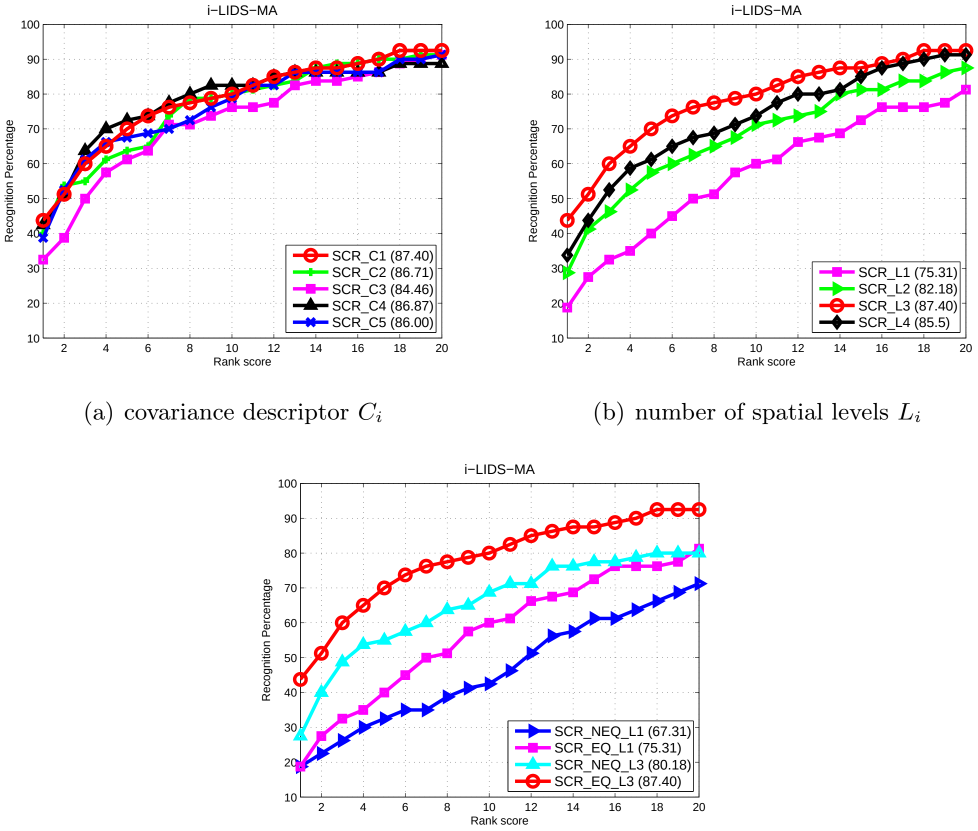
where x and y are pixel location, $\mathcal{D}$ is a distance ($L2$ norm) between the center of an object and the pixel location, $R_{xy}, G_{xy}, B_{xy}, I_{xy}$ are RGB and intensity channel values. ∇ and θ corresponds to gradient magnitude and orientation in each channel, respectively. While using feature vector F_i, we represent an image region by covariance matrix C_i.

Feature vectors are manually selected to minimize redundancy and maximize information of the covariance descriptor. Defining F_1 and F_2, we compare the usage of spatial features (x, y) *vs* $\mathcal{D}$. Then, F_3 illustrates the feature performance while not keeping spatial correlations inside the covariance descriptor. Finally, F_4 investigates the usage of only simple texture features (intensity + gradients) and F_5 represents only color statistics.

Figure 7.12(a) presents the performance of SCR using different types of covariance descriptor. It is apparent that the best performance is obtained by the feature vector which contains the largest number of features (F_1). Surprisingly, the recognition accuracy of feature vector F_4 is also very high. It means that the most discriminative information is given by texture features. The worst performance is obtained by feature vector F_3, in which there is no information on pixel location (no spatial correlations). It clearly shows that information on spatial correlation of features (texture) has significant impact on recognition accuracy.

Levels of spatial pyramid: In section 4.2.4, we have described the spatial pyramid matching, which compares two objects by computing the similarity on decreasingly finer cells. Matches found at smaller cells are weighted more highly than matches found at larger cells.

Investigating the impact of different number of levels, we carry out experiments using $L = 0, 1, 2, 3$, while applying the pyramid match kernel (see equation 4.16).



(a) covariance descriptor C_i

(b) number of spatial levels L_i

(c) histogram equalization

Figure 7.12: CMC curves obtained on i-LIDS-MA dataset using SCR signature: (a) performance comparison of different covariance descriptors; (b) analysis of the level's depth: $L1$ corresponds to the matching using only one level (level 0 refers to the full body, see figure 4.11), $L2$ corresponds to the matching using two levels (level 0 and level 1), *etc.*; (c) the performance of SCR signature with and without *histogram equalization* (HE *vs*. NHE).

Figure 7.12(b) illustrates the performance changes depending on the level's depth. The best recognition rate is achieved using a three-level pyramid ($L3$ refers to the pyramid matching at level 0,1 and 2). It appears that a four-level pyramid performs worse than a three-level pyramid. This is due to the fact that at level 3, image regions contain only few pixels, from which it is difficult to extract additional information.

Histogram equalization: Figure 7.12(c) shows the impact of *histogram equalization* step. We can notice the significant increase in performance, while applying both, the spatial pyramid matching and *histogram equalization*.

7.4 Boosting approaches

In this section, we evaluate our boosting approaches on i-LIDS-MA dataset (section 7.2.3). We study different stopping criteria while learning signatures and also investigate the matching performance w.r.t. the number of learning frames. We show that our methods based on MRC patches (RCP/LCP signature) significantly improve performance over already presented methods.

7.4.1 Haar signature

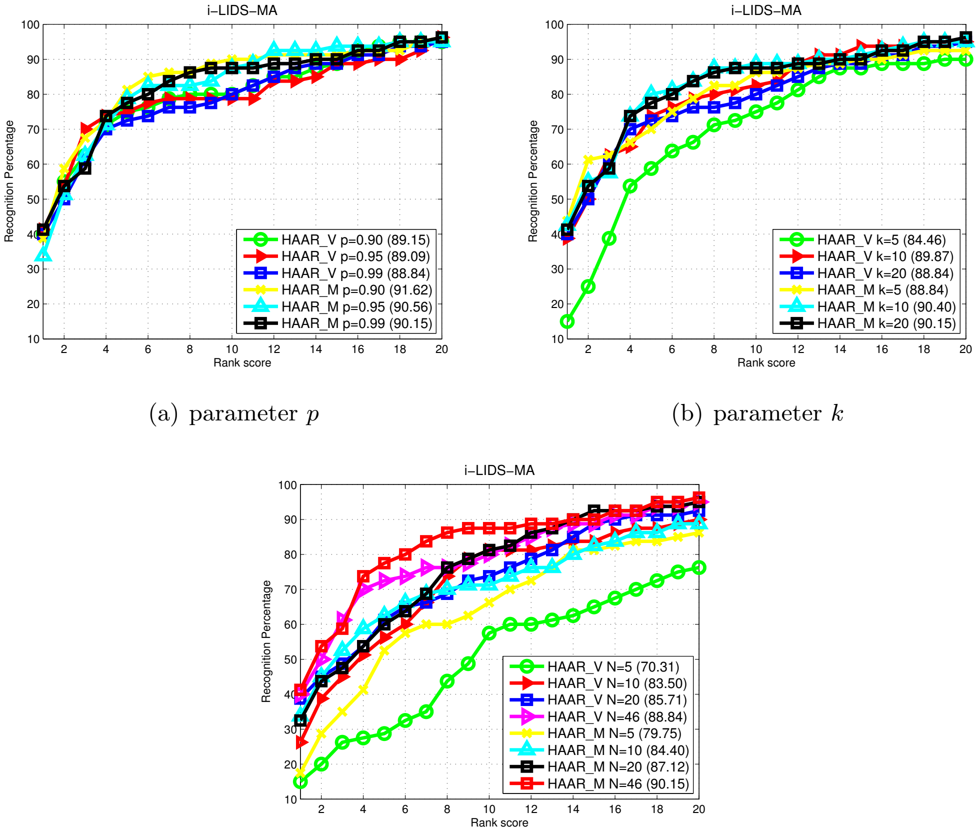
We present a performance of Haar-based signatures investigating different stopping criteria while employing two signature matching strategies: volume-based similarity (section 5.1.3.1) and margin-based similarity (section 5.1.3.2).

Stop criteria: We evaluate our Haar signature using different values of parameters p (the minimum rate of *true positives* on a boosting stage), and k (the maximum number of stages in a cascade of classifiers), which determine learning of the Haar classifier. In experiments we assume $p \in \{0.9, 0.95, 0.99\}$ and $k \in \{5, 10, 20\}$ (see section 5.1.2.6 for details).

Figure 7.13(a) illustrates the impact of parameter p. We could expect that the greater p, the higher probability of generating the overfitted classifier (especially in our case, where we compute similarity between signatures, testing the classifier on images from different cameras). However, the results show that the performance of our Haar signature does not decrease, while increasing p. This confirms the phenomenon that AdaBoost does not suffer from overfitting (the margin theory states that AdaBoost tends to increase the margins of the training examples, and in the result, the increase in the margins implies better generalization performance [Schapire 1998, Rudin 2007]).

We also analyze the size of the cascade k, while learning signatures. From figure 7.13(b), we can deduce that volume-based similarity depends more on parameter k than margin-based similarity. It appears that the margin-based similarity, which is based on the real-value classification (the margin) provides more useful information than the volume-based similarity, which is based on the binary classification. The results show that $k = 10$ gives relatively good performance.

Number of positive frame samples: We carry out experiments to show the evolution of the performance with the number of given frames (positive images) for training per individual. We learn signatures using $N \in \{5, 10, 20, 46\}$ positive samples. From figure 7.13(c) it is apparent that the larger number of frames, the better performance is achieved. It is worth noting that $N \approx 50$ is usually affordable in practice as it corresponds to only 2 seconds of a standard 25 frame rate camera.



(a) parameter p

(b) parameter k

(c) number of frames N

Figure 7.13: CMC curves obtained on i-LIDS-MA dataset using Haar signatures. HAAR_V and HAAR_M labels correspond to volume-based and margin-based similarity, respectively: (a) the minimum rate of *true positives* - parameter p; (b) the maximum number of stages in a cascade - parameter k; (c) number of frames N required for learning signatures.

Volume-based similarity *vs.* margin-based similarity: In all experiments margin-based similarity performed slightly better than volume-based similarity (figure 7.13). We again believe that this is due to the fact that margin-based similarity provides the confidence of classification (the margin), while classifying an image, which is not the case for volume-based similarity (the volume-based similarity is only based on binary response of a classifier).

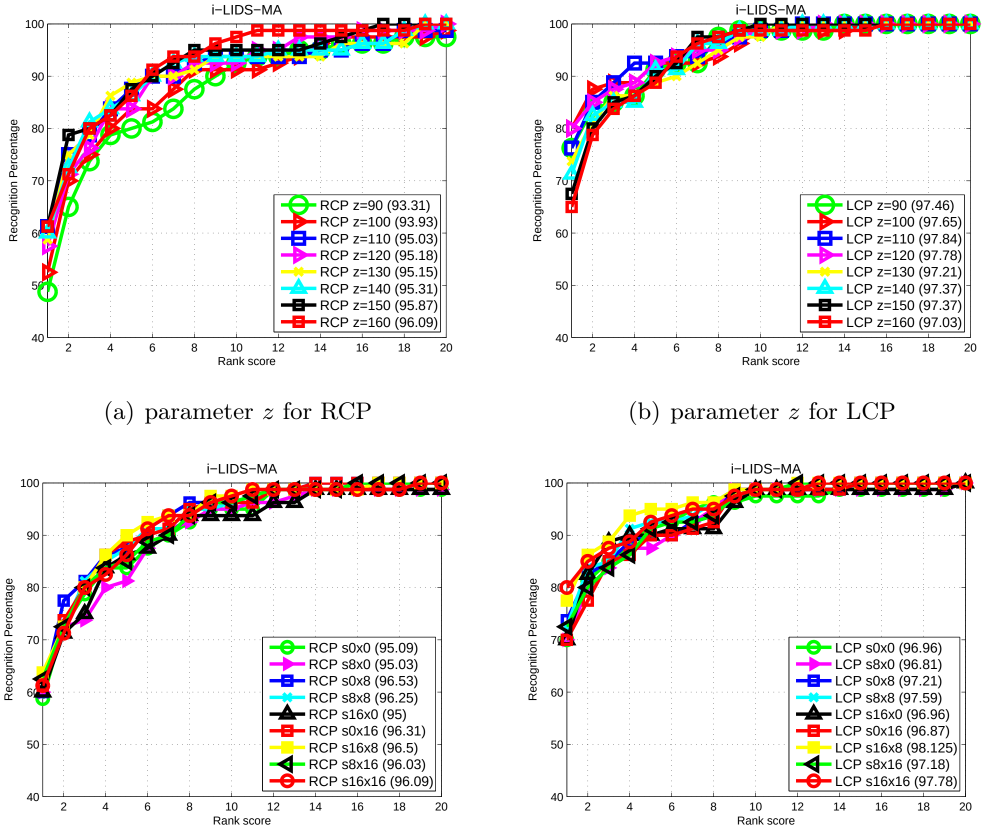
7.4.2 RCP/LCP signature

Experimental setup: Every human image is scaled into a fixed size of 64 × 192 pixels. We generate MRC patches of 16 × 16 pixels with 8 pixels step (it gives 161 patches). We generate MRC patches of 32 × 32 pixels with 16 pixels step (it gives

33 patches). Then, we generate combinations of MRC patches (small - small, small - big, big - big, we limit space of patches to those which are situated with the 16 pixels step). Finally, in total we have 3401 MRC weak classifiers.

Stop criterion: We evaluate our RCP and LCP signature using different values of parameter z, which determines the number of MRC patches used for generating signature. We set $z \in \{90, 100, \ldots, 160\}$. Analyzing the performance at figure 7.14(b), we should not expect any confirmation of boosting properties related to overfitting issues, as our similarity function is not directly based on the classification function. Parameter z determines the density of MRC patches. We can note that for RCP, the larger number of patches, the better performance is achieved, which is not the case for LCP method. LCP method significantly outperforms RCP (especially in the first ranks), which bears out advantages of using discriminative learning.

Signature matching: While matching, we shift one signature over another to reduce body alignment issues. Shifting is performed horizontally and vertically by fixed amount of pixels, maximizing the similarity measure (see section 5.2.6). From figure 7.14(d), we can deduce that shifting improves performance, especially in the first ranks.



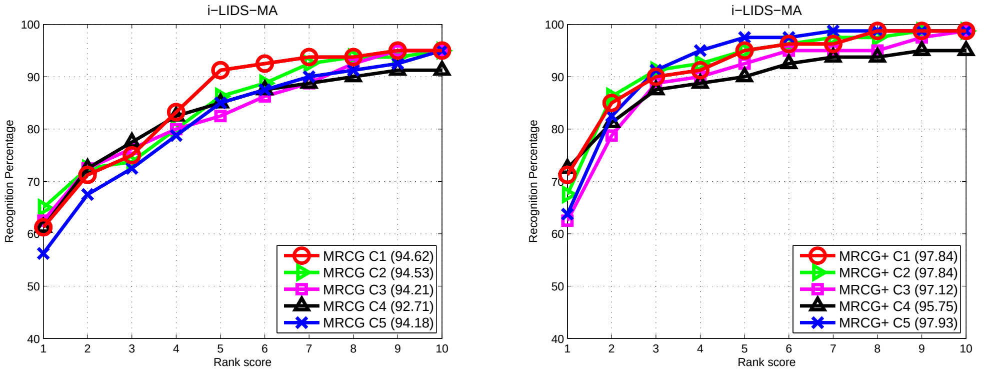
(a) parameter z for RCP

(b) parameter z for LCP

(c) matching for RCP

(d) matching for LCP

Figure 7.14: CMC curves obtained on i-LIDS-MA dataset using LCP/RCP signatures: (a,b) stop criterion; (c,d) signature matching (shifting parameters: $sW \times H$ corresponds to vertical and horizontal shift).



(a) covariance descriptor C_i for MRCG

(b) covariance descriptor C_i for MRCG+

Figure 7.15: CMC curves obtained on i-LIDS-MA dataset using MRCG signatures: performance analysis of different covariance descriptors: (a) without using discriminative method, (b) with using discriminative method.

7.5 Efficient re-identification

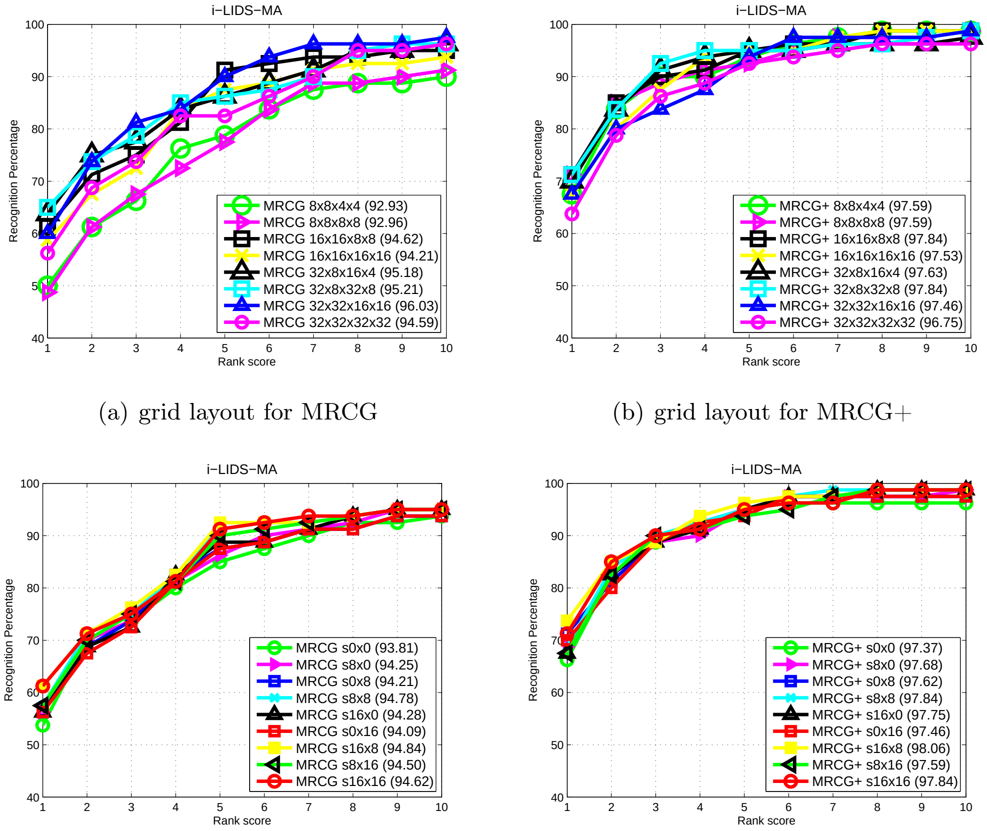
This section explores recognition performance of methods described in chapter 6. Again, i-LIDS-MA dataset (section 7.2.3) is employed to evaluate different settings of parameters.

7.5.1 MRCG signature

Experimental setup: Every human image is scaled into fixed size of 64 × 192 pixels (size of the grid). We extract MRC cells of $W \times H$ pixels, on a fixed grid of $w \times h$ pixel step ($w \times h$ corresponds to the overlap between each cell in the dense grid). In experiments we refer to Mean Riemannianc Covariance Grid without using discriminant method as MRCG and with using discriminant method as MRCG+.

Feature vector: We carry out experiments using different features inside the mean Riemannian covariance. Similarly as in section 7.3.2, we use the same five feature vectors (F_i), computing covariance matrices (C_i). From figure 7.15, it is clear that again the best performance is obtained by the feature vector which contains the largest number of features (F_1). We can also note that discriminative method significantly improves recognition rate, especially in the first ranks (figure 7.15(b)).

Grid layout: We perform experiments investigating different structures of the grid. Figure 7.16(a,b) shows performance of MRCG using different size of cells ($W \times H$) with different density step ($w \times h$). The dense grid slightly increases the recognition performance. We expect that a dense structure will have more relevant impact on



(a) grid layout for MRCG

(b) grid layout for MRCG+

(c) matching for MRCG

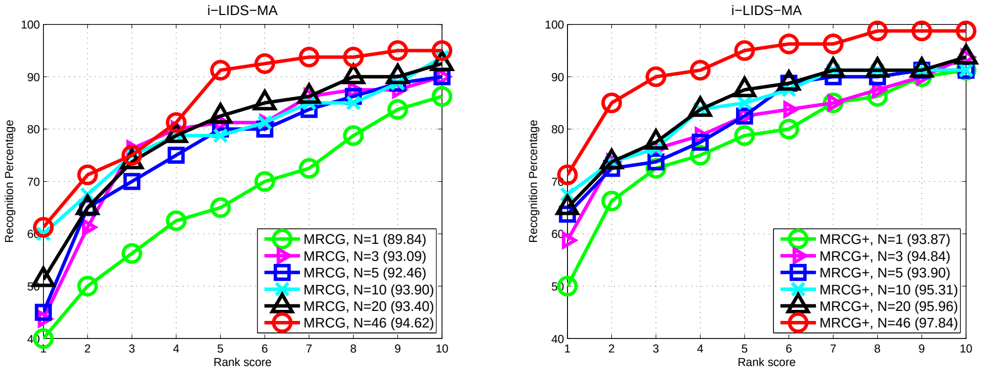
(d) matching for MRCG+

Figure 7.16: CMC curves obtained on i-LIDS-MA dataset using MRCG signatures: (a,b) performance analysis of grid layout ($W \times H \times w \times h$); (c,d) signature matching (shifting parameters: $sW \times H$ corresponds to vertical and horizontal shift).

recognition accuracy in the case of many occlusions.

Signature matching: While matching, we shift one signature over another to reduce body alignment issues. Shifting is performed horizontally and vertically by fixed amount of pixels, maximizing the similarity measure (see section 5.2.6). From figure 7.16(c,d), we can deduce that shifting slightly improves performance. We notice that improvement is more significant when human detections are less accurate (*e.g.* for i-LIDS-AA dataset).

Number of frames: We carry out experiments to show the evolution of the performance with the number of given frames per individual (figure 7.17). The results indicate that the larger number of frames, the better performance is achieved. It clearly shows that averaging covariance matrices on Riemannian manifold using multiple shots leads to a much better recognition accuracy.



(a) number of frames for MRCG

(b) number of frames for MRCG+

Figure 7.17: CMC curves obtained on i-LIDS-MA dataset using MRCG signatures: performance evolution w.r.t. the number of given frames.

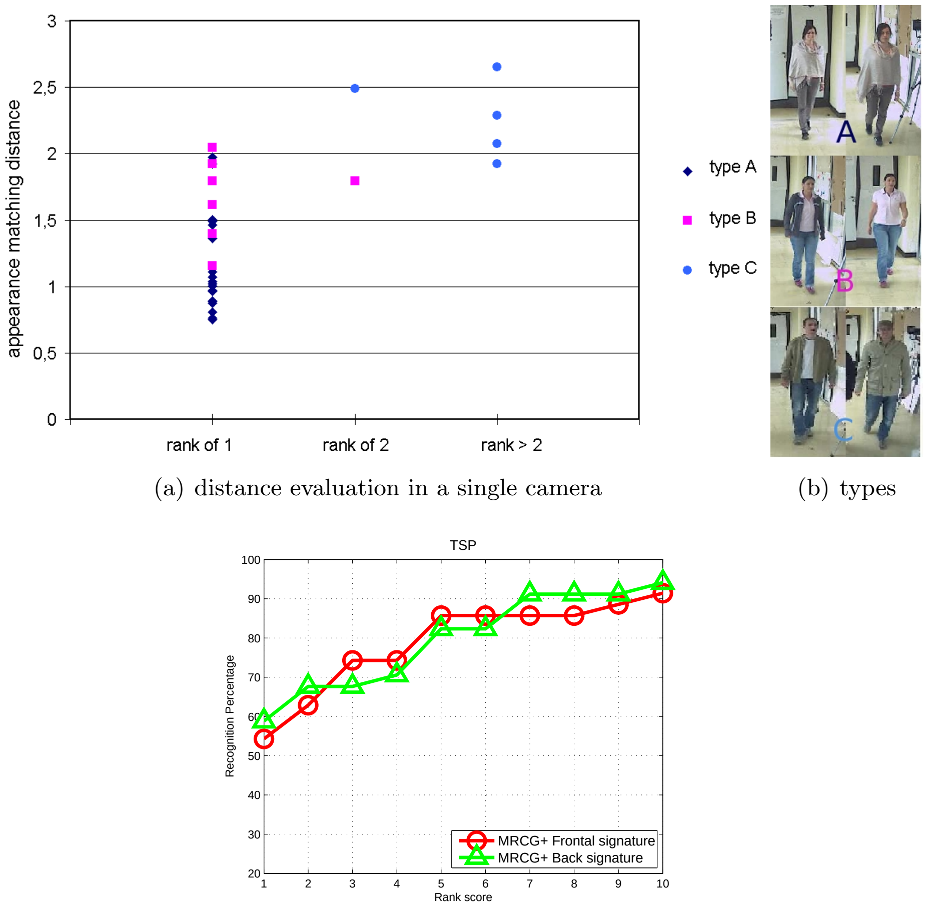
TSP

Using TSP dataset (section 7.2.4) we evaluate two common video surveillance scenarios: (1) person re-identification in a single camera and (2) multiple-appearance recognition employing two cameras. Person appearances are registered automatically by applying LBP-based detector [Corvee 2012].

In TSP dataset, we distinguish three types of appearance comparison (see figure 7.18(b)):

- type A - people wearing the same clothes in recording sessions,
- type B - people slightly change their appearances from one recording session to the other (*e.g.* someone unzipping his/her jacket or someone taking his/her scarf off),
- type C - people with great change of appearance (*e.g.* person adding a coat with a hat).

We analyze our MRCG descriptor comparing appearances registered in a single camera. The tracked people appearance signatures are recorded in a database during the first session. During the second session, the appearance signatures of all the tracked people are compared with the database. In figure 7.18(a), we plot the appearance matching distances between the same person appearances for type A, B and C. The results show that people type A and B are successfully retrieved with rank 1, except for 1 person with rank 2, when matching distances are below 2.1. People of type C (*i.e.* strong change of appearance) are not re-identified. However, a uncertainty zone exists for a matching distance between 1.75 and 2.1.



(a) distance evaluation in a single camera

(b) types

(c) re-identification in different cameras

Figure 7.18: Evaluation of MRCG descriptor using TSP data: (a) distance comparison for type A, B and C; (b) example images of types; (c) CMC curves for queering a database using signatures corresponding to frontal and back view, respectively.

We have also performed an evaluation using two cameras. As we have already mentioned (section 7.2.4), we have extracted three trajectories (1) the frontal view by camera 1, (2) the back view and the frontal view by camera 2 and (3) the back view by camera 1. We evaluate re-identification in the following way:

- mono appearance mode - frontal and back view of people tracked in camera 1 have their appearance signatures extracted separately,
- multiple appearance mode - both frontal and back view of people tracked in camera 2 have their appearance signatures extracted (we could differentiate two views based on the direction of the trajectory). Both signatures are merged into a single signature stored in a database (we create signature which contains two MRCG descriptors).

Every signature from camera 1 is used as a query to our database (multiple-appearance signatures extracted from camera 2). The CMC curves for back and frontal view signatures are presented in figure 7.18(c). The results show that despite automatic inaccurate people detection and the different lighting conditions of the two cameras, the system shows promising people appearance based re-identification.

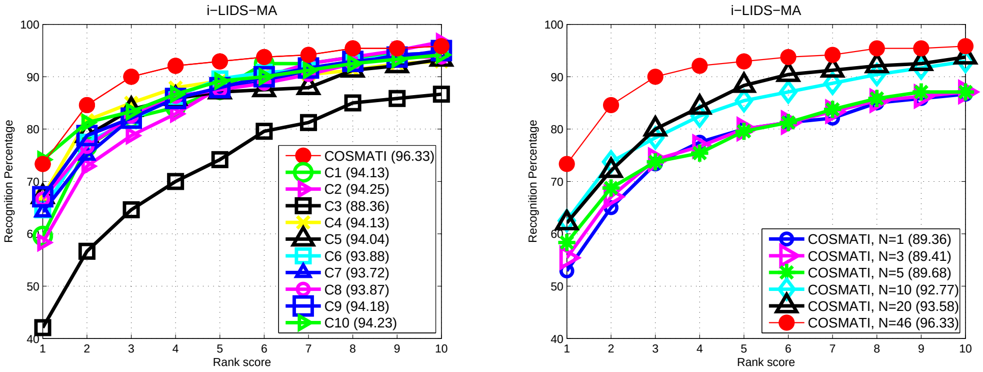
7.5.2 COSMATI signature

Feature space: We scale every human image into a fixed size window of 64×192 pixels. The set of rectangular sub-regions $\mathbf{P}$ is produced by shifting 32×8 and 16×16 pixel regions with 8 pixels step (up and down). It gives $|\mathbf{P}| = 281$ rectangular sub-regions. We set $L = \{(l, \nabla_l, \theta_l)_{l=I,R,G,B}, G_{i=1\ldots4}, \mathcal{N}, \mathfrak{L}\}$, where I, R, G, B refer to intensity, red, green and blue channel, respectively; ∇ is the gradient magnitude; θ corresponds to the gradient orientation; G_i are Gabor's filters with parameters $\gamma, \theta, \lambda, \sigma^2$ set to $(0.4, 0, 8, 2)$, $(0.4, \frac{\pi}{2}, 8, 2)$, $(0.8, \frac{\pi}{4}, 8, 2)$ and $(0.8, \frac{3\pi}{2}, 8, 2)$, respectively; $\mathcal{N}$ is a gaussian and $\mathfrak{L}$ is a laplacian filter. A learning process involving all possible combinations of three layers would not be computationally tractable (229296 covariances to consider in section 6.2.2.2). Thus instead, we experimented with different subsets of combinations and selected a reasonably efficient one. Among all possible combinations of the three layers, we choose 10 combinations ($C_{i=1\ldots10}$) to ensure inexpensive computation. We set C_i to (R, G, B), $(\nabla_R, \nabla_G, \nabla_B)$, $(\theta_R, \theta_G, \theta_B)$, (I, ∇_I, θ_I), (I, G_3, G_4), $(I, G_2, \mathfrak{L})$, $(I, G_2, \mathcal{N})$, $(I, G_1, \mathcal{N})$, $(I, G_1, \mathfrak{L})$, (I, G_1, G_2), respectively, separating color and texture features. Similar idea was already proposed in [Gray 2008]. Note that we add to every combination C_i layer $\mathfrak{D}$, thus generating our final 4×4 covariance descriptors. The dimension of our meta covariance feature space is $n = |\mathfrak{C}| = 10 \times |\mathbf{P}| = 2810$.

Learning and testing: Let us assume that we have $(h + q)$ individuals seen from two different cameras. For every individual, m images from each camera are given. We take q individuals for learning our model, while h individuals are used to setup the gallery set. We generate positive training examples by comparing m images of the same individual from one camera with m images from the second camera. Thus, we produce $|\delta^+| = q \times m^2$ positive samples. Pairs of images coming from different individuals stand for negative training examples, thus producing $|\delta^-| = q \times (q-1) \times m^2$.

For each individual we randomly select $m = 10$ images. Then, we randomly select $q = 10$ individuals to learn a model. The evaluation is performed on remaining $h = 30$ individuals. Every signature is used as a query to the gallery set of signatures from different camera. This procedure is repeated 10 times to obtain averaged CMC curves.

COSMATI *vs.* $\mathbf{C_i}$: We first evaluate the improvement in using different types of covariance descriptors for the appearance matching. We compare models based on a



(a) h=30 individuals

(b) h=30 individuals

Figure 7.19: Performance comparison: (a) with models based on a single type of covariance descriptor; (b) w.r.t. the number of given frames

single type of covariance with the model, which employs different kinds of covariance features. From figure 7.19(a) it is apparent that combination of different kinds of covariance features improves matching accuracy.

COSMATI w.r.t. the number of shots: We carry out experiments to show the evolution of the performance with the number of given frames per individual (figure 7.19(b)). The results indicate that the larger number of frames, the better performance is achieved. It clearly shows that averaging covariance matrices on Riemannian manifold using multiple shots leads to a much better recognition accuracy. We again highlight that $N \approx 50$ is usually affordable in practice as it corresponds to only 2 seconds of a standard 25 frame rate camera.

Computation complexity: In our experiments, for $q = 10$ and $m = 10$ we generate $|\delta^+| = 1000$ and $|\delta^-| = 9000$ training samples. Learning on 10.000 samples takes around 20 minutes on Intel quad-core 2.4GHz. The model in average is composed of 150 covariance features. The calculation of generalized eigenvalues of 4×4 covariance matrices (distance computation) takes $\sim 2\mu s$ without applying any hardware-depended optimization routines (*e.g.* LAPACK library can perform faster using *block operations* optimized for architecture).

7.6 Competitive evaluation

This section presents a comparative evaluation of our descriptors. We carry out experiments on various datasets, presenting the comparison to state of the art techniques. We discuss the results and explore the main limitations of our descriptors.

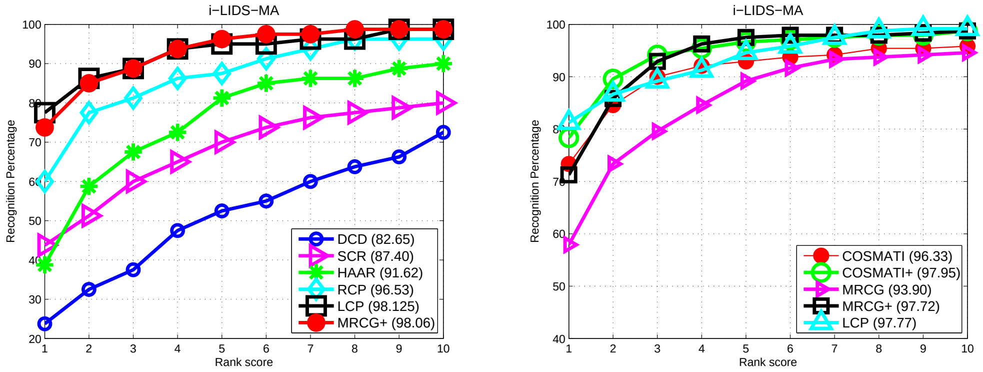
7.6.1 i-LIDS-MA

We have already used i-LIDS-MA dataset (section 7.2.3) to study different parameters of our methods in previous sections. Now, we select the best parameters per each method and present the comparison in figure 7.20(a,b). As COSMATI requires learning samples from both cameras, we also carry out experiments comparing COSMATI *vs*. MRCG and LCP, selecting $q = 10$ individuals for learning and $h = 30$ individuals for testing (the detailed evaluation procedure is described in section 7.5.2).

The best performance is achieved by COSMATI+. We remind that COSMATI is the fastest method among techniques presented in figure 7.20(b), as it uses small covariance matrices (4×4). The experiment bears out that a fusion of the strong covariance descriptor with an efficient selection method produces descriptive models for the appearance matching problem. In the result, in combination with our discriminative method (COSMATI+), we achieve the best recognition accuracy. The disadvantage of COSMATI is the offline learning phase which requires training data from each camera to obtain the distinctive representation. It is also worth noting that Haar-like features are not so discriminative as covariance descriptors. We have expected such results as Haar-like features are just based on simple rectangular intensity patterns, while covariance matrices keep information on feature distribution with their spatial layout.

Considering our *single-shot* approaches, SCR signature significantly outperforms DCD signature. We strongly believe that it is due to the fact that covariance matrix holds texture information which is not the case for dominant color descriptor.

In figure 7.21, we illustrate differences in performance of covariance-based, color-based and Haar-based signature. The covariance-based descriptor (MRCG) keeps texture information (query A) in contrast to DCD and HAAR signatures. However, its performance slightly decreases when image regions are homogeneous (small amount of edge features - query B, D). Although edges can also mislead the algorithm (query C), we find covariance feature as the best descriptor in general case for matching appearances across disjoint camera views. We also show that covariance-based descriptor can be outperformed by DCD (query B) or by Haar-like features (query C). Moreover, DCD can also outperform Haar-like features (query B) and *vice versa* (query C). It opens new directions in further studies *e.g.* looking for the



(a) h = 40 individuals

(b) h = 30 individuals

Figure 7.20: Performance comparison: '+' indicates the cases in which we have applied our discriminative method (section 6.1.2); (a) all methods except COSMATI; (b) COSMATI *vs.* MRCG and LCP.

combination of different features or selecting the best kind of feature depending on content of an image/video.

7.6.2 i-LIDS-AA

This dataset contains 100 individuals automatically detected and tracked in two cameras. Cropped images are noisy, which makes the dataset more challenging (*e.g.* detected bounding boxes are not accurately centered around the people, only a part of the people is detected due to occlusion).

The performance on this dataset is shown in figure 7.22. The results show again that descriptors which average covariances on Riemannian manifold outperform significantly *single-shot* approach. Unfortunately, achieved recognition rate is not very high in comparison with the results obtained on i-LIDS-MA. It shows one of the main limitations of the approaches - performance directly depends on human detection accuracy.

We also evaluate the proposed techniques using hierarchical matching (we extract human signatures on different resolutions and then during feature matching we maximize similarity through all extracted resolutions). Unfortunately, performance do not change significantly (1% - 2% improvement in the first rank has been noticed) with a great increase of time consumption. Moreover, using hierarchical matching (giving more flexibility to the appearance) we lose discriminative properties of the spatial correlation (*e.g.* we might match noise).

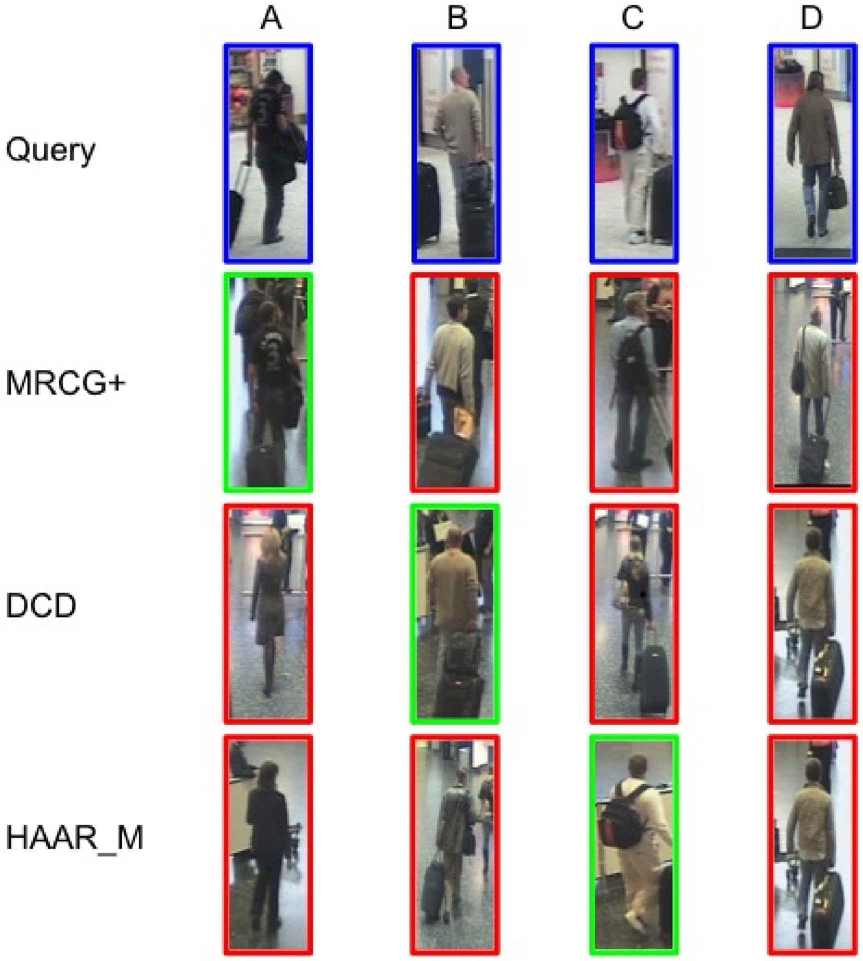


Figure 7.21: The query results of different algorithms: the first row corresponds to the query image and the remaining rows correspond to the first ranks returned by MRCG, DCD and HAAR_M algorithms, respectively. True matches are highlighted with green color.

7.6.3 i-LIDS-119

This dataset is extensively used in the literature for testing the person re-identification approaches. The dataset is very challenging since there are many occlusions and often only the top part of the person is visible. i-LIDS-119 has been prepared for *single-shot* case, thus it does not fit very well for evaluating *multiple-shot* case, because the number of images per individual is very low (in average 4). Moreover, for 22 individuals there are only 2 images given (one from each camera). Hence, in evaluation of *multiple-shot* signatures, we apply simple affine transformation on given images (coordination of transformation matrix are changed by 5% and rotation angle is in range of $[-6°; 6°]$) to obtain multiple images. As every transformation of an original image is allowed we claim that this solution is fair with the

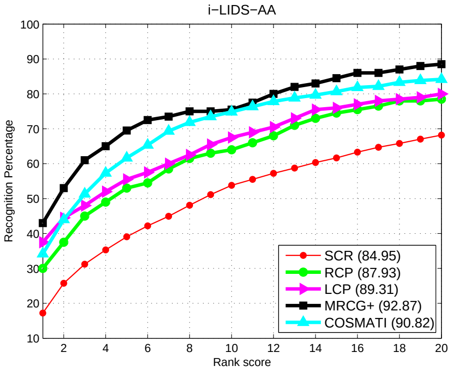


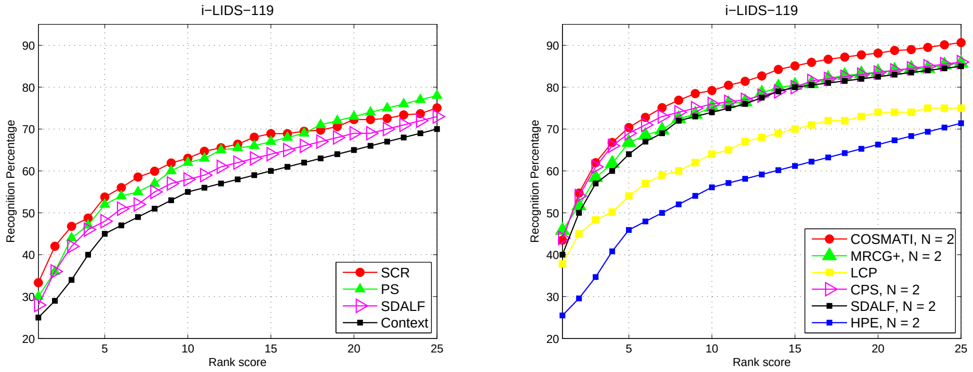
Figure 7.22: Performance comparison of covariance-based descriptors on i-LIDS-AA. Evaluation of COSMATI is performed using the models learned on i-LIDS-MA.

state of the art. In total, we use maximally $N = 2$ original images to create human signatures.

Single-shot approaches: We compare our SCR descriptor with current state of the art approaches: PS [Cheng 2011], SDALF [Farenzena 2010] and Group-Context [Zheng 2009]. Our descriptor outperforms all state of the art descriptors, achieving 32% in the first rank (figure 7.23). It is worth mentioning that the performance is not very high because the person images from the i-LIDS data are very challenging since they are captured from non-overlapping multiple camera views subject to significant occlusions, noisy body part detections and large variations in both view angle and illumination. PS achieves comparable results to our descriptor (both descriptors, SCR and PS are based on body part detections), while SDALF is slightly worse. It may be due to the fact that assumptions of SDALF descriptor (perceptual principles of symmetry and asymmetry of the human body) are not met in this dataset because of many occlusions.

Multiple-shot approaches: We carry out experiments using our covariance-based multiple-shot signatures. We compare with CPS [Cheng 2011], SDALF [Farenzena 2010] and HPE [Bazzani 2010]. In case of COSMATI, we have used models learned on i-LIDS-MA to evaluate our approach on the full dataset of 119 individuals. Our CMC curves are generated by averaged CMC over 10 trials. Results are presented in figure 7.23(b). COSMATI performs the best among all considered methods. We believe that it is due to the informative appearance representation obtained by CFS technique (section 6.2.2.2). It clearly shows that a combination of the strong covariance descriptor with the efficient selection method produces distinctive models for the appearance matching problem.

COSMATI *vs.* PRDC [Zheng 2011]: Additionally, we compare our approach with PRDC method. This method also requires offline learning, thus we can compare with state of the art approach, while learning models on i-LIDS-119. PRDC focuses on distance learning that can maximize matching accuracy regardless of the rep-



(a) single-shot

(b) multiple-shot

Figure 7.23: Comparison with state of the art approaches on i-LIDS-119: (a) with PS [Cheng 2011], SDALF [Farenzena 2010] and Context [Zheng 2009]; (b) with CPS [Cheng 2011], SDALF [Farenzena 2010] and HPE [Bazzani 2010].

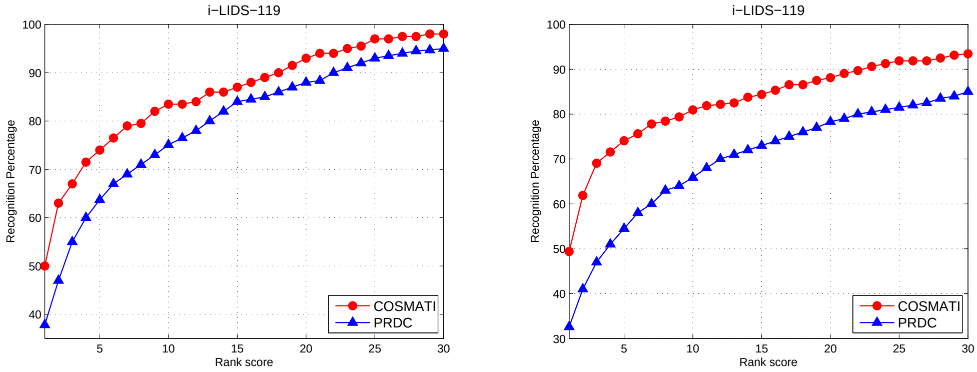
resentation choice. We reproduce the same experimental settings as [Zheng 2011]. All images of h randomly selected individuals are used to set up the test set. The remaining images form the training data. Each test set is composed of a gallery set and a probe set. Unlike [Zheng 2011], we use multiple images (maximally $N = 2$) to create the gallery and the probe set. The procedure is repeated 10 times to obtain reliable statistics. Our results (figure 7.24(a,b)) show clearly that COSMATI outperforms PRDC. The results indicate that using strong descriptors, we can significantly increase matching accuracy.

COSMATI - model analysis: The strength of our approach is the combination of many different kinds of features into one similarity function. Table 7.5 shows the percentage of different kinds of covariance features embedded in all our models, which were used during the evaluation. Unlike [Gray 2008], our method concentrates more on texture filters than on color features. It appears that the higher the resolution of images, the more frequent the usage of texture filters is. Figure 7.25 illustrates that models extracted on higher resolution employ more texture features.

$C_1(R,G,B)$	$C_2(\nabla_R,\nabla_G,\nabla_B)$	$C_3(\theta_R,\theta_G,\theta_B)$	$C_4(I,\nabla_I,\theta_I)$	$C_5(I,G_3,G_4)$
9.61 %	5.30%	4.62%	4.37%	5.47%
$C_6(I,G_2,\mathfrak{L})$	$C_7(I,G_2,\mathcal{N})$	$C_8(I,G_1,\mathcal{N})$	$C_9(I,G_1,\mathfrak{L})$	$C_{10}(I,G_1,G_2)$
12.70 %	14.29%	17.35%	12.12%	14.16 %

Table 7.5: A table showing the percentage of different covariance features embedded in models (choice of covariance layers C_i is detailed in section 7.5.2).

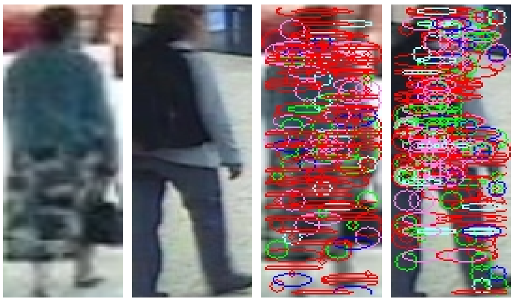
Background subtraction step: Using multiple images, a straightforward way



(a) h = 50 individuals

(b) h = 80 individuals

Figure 7.24: COSMATI *vs*. PRDC [Zheng 2011]: h is the size of the gallery set (larger h means smaller training set).

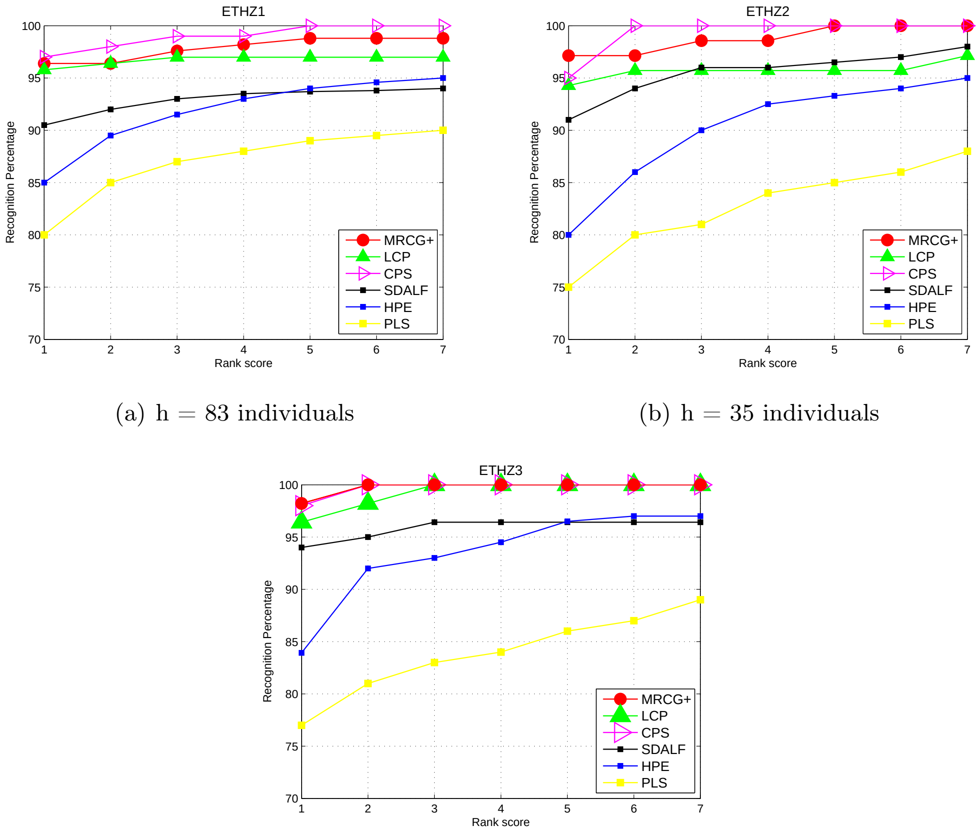


(a) (b) (c) (d)

Figure 7.25: Extracted models: (a,b) example image from i-LIDS-119 (lower resolution) and from i-LIDS-MA (higher resolution); (c,d) models extracted using images from i-LIDS-119 and i-LIDS-MA respectively; red indicates color features which turn out to be more prominent in (c).

for appearance extraction would be to use a *background subtraction* algorithm for obtaining foreground regions. Unfortunately in many real scenarios, consecutive frames are not available due to gaps in tracking results. Having a still image, we could apply the extended graph-cut approach: GrabCut [Rother 2004] (see section 4.1.1) to obtain a human silhouette. This approach can be driven by cues coming from detection results (a rectangle around the desired object). Surprisingly, GrabCut does not increase matching accuracy in our framework for all covariance-based techniques. The main reason of this result is that GrabCut often mis-segments significant parts of foreground regions. In our approaches we overcome the background issue either by learning a model which focuses on features corresponding to foreground (COSMATI), or by applying discriminative method (*e.g*. LCP by boosting and MRCG by discriminants).

7.6.4 ETHZ



(a) h = 83 individuals

(b) h = 35 individuals

(c) h = 28 individuals

Figure 7.26: CMC curves obtained on ETHZ dataset: (a) sequence SEQ. #1; (b) sequence SEQ. #2; (c) sequence SEQ. #3. We compare our methods with PLS [Schwartz 2009], HPE [Bazzani 2010], SDALF [Farenzena 2010] and CPS [Cheng 2011].

We analyze performance of our algorithms on ETHZ dataset. As mentioned previously, this dataset contains video sequences extracted only from a single camera. In the results, the performance accuracy obtained on this data is much higher than on the rest of considered datasets. The challenging aspects of ETHZ are illumination changes and occlusions.

Figure 7.26 illustrates performance comparison with state of the art techniques. Our approaches perform slightly worse than CPS [Cheng 2011], but significantly better than PLS [Schwartz 2009], HPE [Bazzani 2010] and SDALF [Farenzena 2010] . As CPS is based on densely sampled shape context descriptors - *pictorial structures (PS)*, which provide precise body part detections, we believe that this is the main reason of outperforming our descriptors.

The evaluation on this dataset bears out that re-identification techniques can easily be used for object re-acquisition, while tracking in a single camera (*e.g.* when object is lost due to occlusion). High recognition rate indicates that appearance matching in a single camera is much easier task than appearance matching across disjoint camera views.

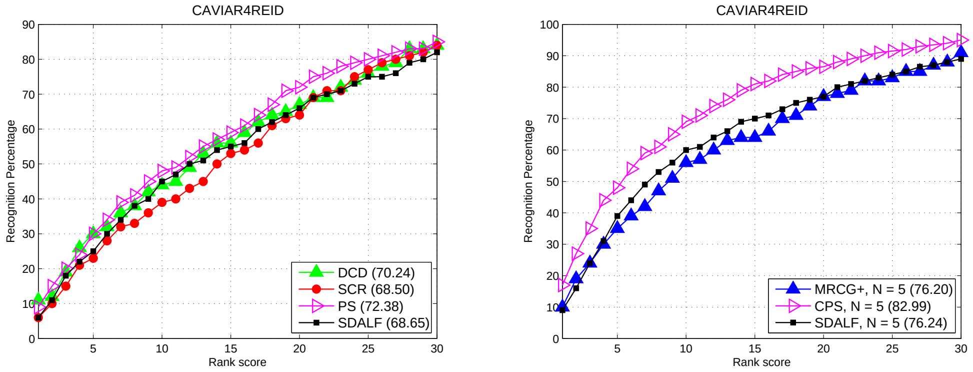
7.6.5 CAVIAR

The experiments are performed on CAVIAR4REID dataset. The main challenge of this data comes from significant illumination changes and extremely low resolution in one of the two cameras.

Figure 7.27 shows performance comparison of our descriptors with state of the art methods, considering both, *single-shot* and *multiple-shot* techniques. Surprisingly, our DCD signature performs slightly better than SCR and than SDALF in the *single-shot* mode. It appears that in low resolution, dominant color descriptor outperforms covariance matrix. In fact, covariance matrix holds information on feature distribution taking into account their spatial layout (texture). In very low resolution, we can expect that texture features do not appear. Until now, the resolution problem has been ignored as we assumed that texture information is often present in an image. This experiment raises several questions for appearance matching, including resolution-dependent re-identification (different descriptors for different resolutions) or looking for the optimal descriptor which performs the best in all resolutions. We also highlight a limitation of our covariance-based descriptors which perform relatively poor in case of very low resolution images. The best performance in *multiple-shot* case is achieved by CPS [Cheng 2011], which besides employing precise body part detector, it represents the appearance of each body part by HSV histogram and MSCR [Forssén 2007] descriptor. We also note that even in the case of relatively small dataset (only 50 individuals), the recognition accuracy in the first rank is very low for all state of the art techniques (less than 20%).

7.6.6 VIPER

Recently, we have also tested our *single-shot* covariance-based descriptor on VIPER dataset. VIPER data is dedicated for evaluation of viewpoint invariant approaches. Unfortunately, our SCR descriptor performs significantly worse than state of the art techniques. This discrepancy may be due to the fact that our body part detector performs very poorly on this data because of blurred images and numerous compression artifacts. Moreover, significant pose changes, different quality of images and also compression artifacts deteriorate covariance features. The best performance on this dataset is achieved by *metric learning* methods (LMNN-R [Dikmen 2010], PRDC [Zheng 2011]) and body-part based method (PS [Cheng 2011]). It is worth



(a) single-shot

(b) multiple-shot

Figure 7.27: Performance comparison on CAVIAR4REID dataset: (a) *single-shot* descriptors *vs*. PS [Cheng 2011] and SDALF [Farenzena 2010]; (b) *multiple-shot* descriptors *vs*. CPS [Cheng 2011] and SDALF [Farenzena 2010].

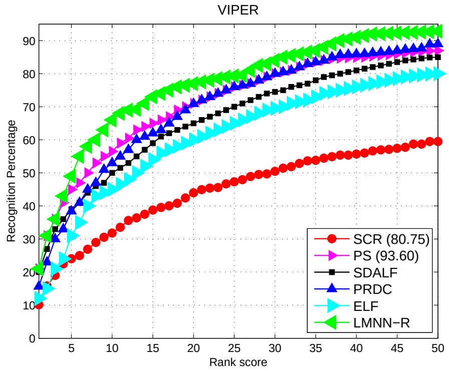
noting that all of these methods use color histograms as a feature representation. Figure 7.28 illustrates state of the art results on this dataset.

7.7 Conclusion

This chapter focuses on performance analysis of proposed techniques. We described metrics and publicly available datasets for evaluation of human re-identification. We demonstrated an extensive evaluation of our methods, investigating their pros and cons. Our techniques are explored in comparison with state of the art approaches on various datasets, showing competitive performance.

The best performance of our methods is achieved by COSMATI+ which in fact requires the most information, which must be provided for generating signatures (see figure 7.29(b)). However, we want to highlight that approaches which are based on training data requiring positive pairs (two images with the same person registered in different cameras), may have difficulties while employed in real systems. Annotations of training data from c cameras and training $\binom{c}{2} = \frac{c!}{2!(c-2)!}$ models, can be unaffordable in practice in case of large c.

Discriminative approaches (*e.g.* HAAR, LCP) also show promising performance, however they are often accused on non-scalability (extensive online learning phase is necessary to extract discriminative signatures every time when a new person is added to the database of signatures). Some solutions for this issue are discussed in section 5.2.7, suggesting the usage of new GPU architectures and a reference dataset. Moreover, our MRCG+ is one alternative which provides an efficient descriptor. Nonetheless, selection of training samples needed for discriminative



(a) p = 316

Figure 7.28: Performance comparison on VIPER dataset with: PS [Cheng 2011], SDALF [Farenzena 2010], PRDC [Zheng 2011], ELF [Gray 2008], LMNN-R [Dikmen 2010].

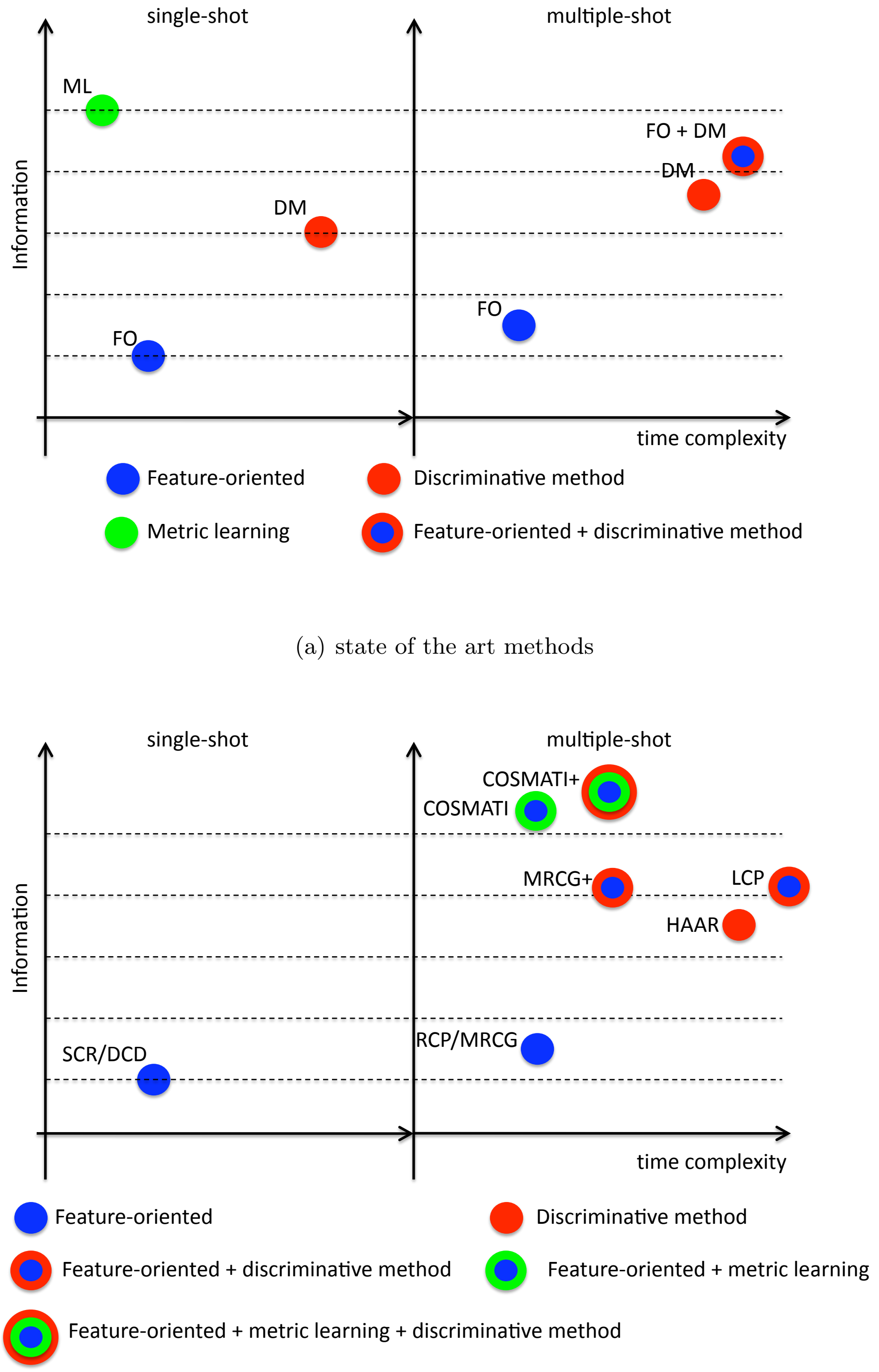
method, size of the reference dataset *etc.* raise numerous questions, producing new open issues.

Our feature-oriented descriptors which do not use any discriminative method show worse performance. Such conclusion was expected as these methods use less information to build signature representation. Moreover, both *single-shot* approaches (DCD and SCR) are dependent on body part detections, which can have significant impact on recognition accuracy.

All experiments confirm that signatures generated in the *multiple-shot* mode outperform the *single-shot* case. Using multiple images, we can provide more information, which can be managed in more reliable way.

We found that the mean Riemannian covariance in general case is the best descriptor for matching appearances across disjoint camera views. Nonetheless, performance of covariance matrix goes down while matching low resolution images (20×40). It appears that simple features can perform better than covariance descriptors in the case of low resolution images. One direction in further studies might be to use content-dependent signature generation. It means that the appearance representation could be dependent on cues available in an image/video (*e.g.* different kinds of features could be used depending on image resolution).

In figure 7.29, we illustrate taxonomy of state of the art approaches compared to our proposed techniques. As most of our approaches are based on *covariance descriptor*, which is computationally intensive, we classify these methods as partially *feature-oriented*. We can note that *metric learning* approaches which are based on strong descriptors may open new directions for future research in this field.



(a) state of the art methods

(b) techniques proposed in the dissertation

Figure 7.29: Appearance-based re-identification techniques as a relationship between information required for computing signature and time complexity of a method: (a) general classes of state of the art approaches; (b) our techniques: labels in the chart refer to acronyms of proposed techniques.

Chapter 8

Human re-identification on highly parallel GPU architecture

"*It's not the size of the dog in the fight, it's the size of the fight in the dog.*"
(Mark Twain)

This chapter discusses the application of re-identification techniques to the real-time video analysis systems. These systems demand for effective processing, *i.e.* they should perform low-cost, low-power and high-speed operations. Our best descriptors are based on covariance matrix which distance operator requires solving the *generalized eigenvalue problem*. This problem is computationally intensive and must be repeated constantly in the re-identification system during the browsing signatures of interest. In the result, we perform detailed analysis of the algorithm complexity and we explore for possibilities of parallelization. We find that some parts of the distance operator algorithm can be easily parallelized. Consequently, we take advantage of new high performance architectures to obtain the required high efficiency. We design a new GPU-based re-identification framework (section 8.1). GPU has evolved into a highly parallel computing architecture applied not only in $3D$ graphics but also for problems which can be expressed as data-parallel computations (section 8.1.1). We propose a new implementation of the distance operator for querying the example database of signatures stored on GPU (section 8.2). In the result, we significantly accelerate the distance computation, reaching 66 speedup in comparison with the CPU implementation (section 8.3).

8.1 GPU-based re-identification framework

In section 3.1 we have already presented the re-identification framework. Now, we explore the component which has not been investigated in the previous chapters: a database of signatures (see figure 8.1). Providing an effective interface for a human operator, we propose a solution which significantly speeds up response time of a database.

Given a query signature, the database should response with the list of the most similar signatures stored in the database. We assume that space-time models like [Javed 2007] should be applied before the searching process starts. The time con-

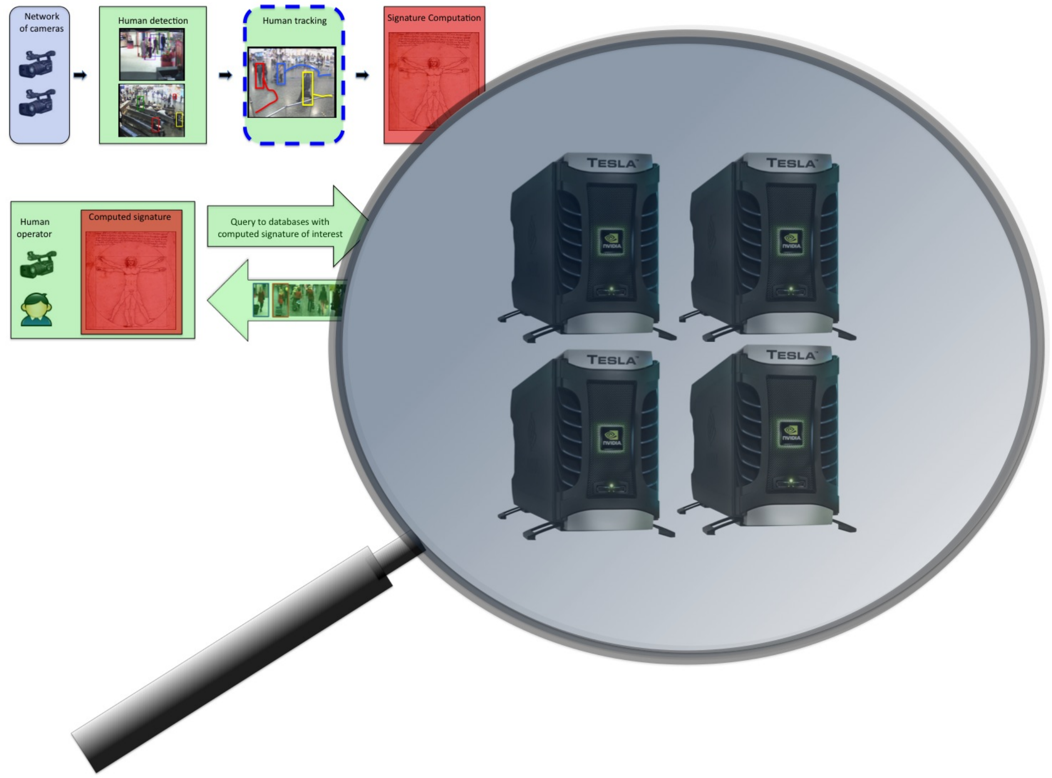


Figure 8.1: The GPU-based re-identification framework. The loupe highlights a GPU-based architecture of a database for storing human signatures.

straints extracted from the topology of cameras can significantly prune the candidate set to be matched. Further, the browsing of the most similar signatures can be optimized in the two following ways:

- by applying the searching strategies,
- by using a new highly parallel architectures.

The first solution refers to strategies which take advantage of a metric space. If signature similarity is a metric, then the searching procedure can be transposed to the *nearest neighbor search* (NNS) problem. As the *nearest neighbor search* problem arises in numerous fields of application, various solutions have already been proposed in the literature [Mico 1992, Yianilos 1993]. Further, there are numerous variants of the NNS problem, while the two most well-known are the *k-nearest neighbor search* and the ε-approximate nearest neighbor search.

The second solution takes advantages of novel architectures which enable the *high performance computing* (HPC). In November 2006, NVIDIA's Tesla architecture was introduced in the GeForce 8800 GPU. This modern $3D$ graphics processing unit has evolved from a fixed-function graphics pipeline to a programmable parallel processor with computing power exceeding multicore CPUs. Together with this architecture, NVIDIA enables high-performance parallel computing applications written in the C language to use the *compute unified device architecture* (CUDA) parallel programming model and development tools.

In the thesis, the most of the proposed appearance models do not meet metric requirements. In the result, we explore the second solution making the following contributions.

- We propose to store the database of signatures on a GPU unit (section 8.1.2).
- We offer a new GPU-based implementation of the distance operator between covariance matrices (section 8.2).

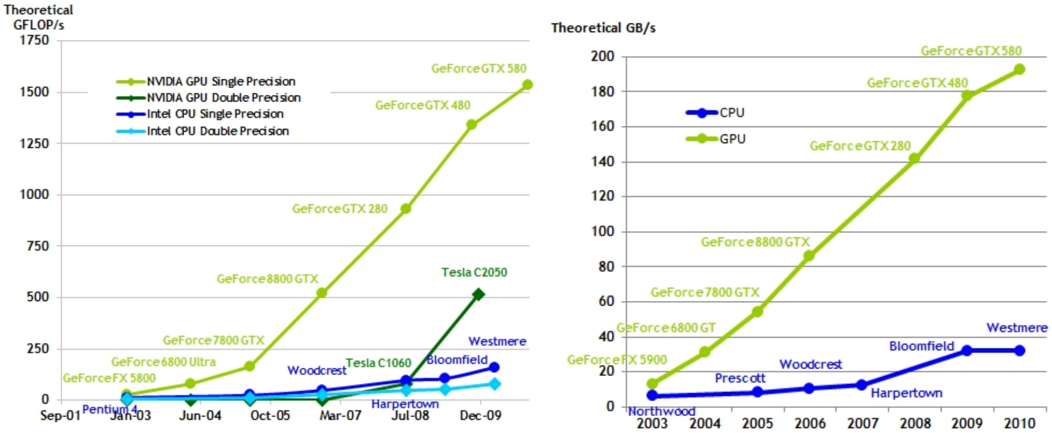
Concerning the former contribution, it is worth noting that we can easily store on GPU all of our appearance models except Haar-based signature. Haar-based signature needs to keep training examples for computing the similarity between signatures, which may result in unreachable storage requirement (see section 5.1.4). The latter contribution is derived from the fact that our best descriptors are based on covariance matrix, which distance operator algorithm consists of parts that can be easily parallelized. Before going into details of the proposed GPU-based implementation, we present a short description of a general-purpose parallel computing architecture.

8.1.1 GPU: a general-purpose parallel computing architecture

'Driven by the insatiable market demand for real-time, high-definition 3D graphics, the programmable Graphic Processor Unit or GPU has evolved into a highly parallel, multi-threaded, many-core processor with tremendous computational horsepower and very high memory bandwidth' [NVIDIA 2011], as illustrated in figure 8.2.

GPU is specialized for intensive and highly parallel computation, driven by requirements of graphics rendering. It was designed as a high performance computing unit in which the emphasis is put on the computing units instead of data caching and control flow units in contrast with CPU (see figure 8.3). The memory is not cached, so to obtain the maximum performance it is crucial that the programmer ensures the coalesced memory accesses.

GPU is especially well-suited to address problems that can be expressed as data-parallel computations. The same program is executed on many data elements in parallel with high arithmetic intensity (the ratio of arithmetic operations to memory operations). As the same program is executed for each data element, there is a lower requirement for sophisticated flow control. In the result, it is executed on many data elements and has high arithmetic intensity, hiding the memory access latency with calculations instead of big data caches.



(a) Floating-Point Operations (b) Memory Bandwidth

Figure 8.2: Floating-Point Operations per Second and Memory Bandwidth for CPU and GPU [NVIDIA 2011].

Data-parallel processing maps data elements to parallel processing threads. Many applications that process large data sets can use a data-parallel programming model to speed up the computations. In $3D$ rendering, large sets of pixels and vertices are mapped to parallel threads. Similarly, image and media processing applications such as post-processing of rendered images, video encoding and decoding, image scaling, stereo vision, and pattern recognition can map image blocks and pixels to parallel processing threads. In fact, many algorithms outside the field of image rendering and processing are accelerated by data-parallel processing, from general signal processing or physics simulation to computational finance or computational biology. Actually, as GPU fits very well to *biologically-inspired* approaches [Pinto 2011], it is a fundamental component in *brain-inspired* computing.

8.1.1.1 A scalable programming model

The CUDA parallel programming model is designed to transparently scale the application software using the increasing number of processor cores. The scaling is provided at the abstraction level, requiring minimum knowledge about parallel programming from programmers familiar with standard programming languages such as C/C++. Three key abstractions: (1) a hierarchy of thread groups, (2) shared memories and (3) barrier synchronization, are simply exposed to the programmer as a minimal set of language extensions.

These abstractions guide the programmer to divide the problem into coarse sub-problems that can be solved independently in parallel by blocks of threads. Further, each sub-problem is divided into finer pieces that can be solved cooperatively in parallel by all threads within the block.

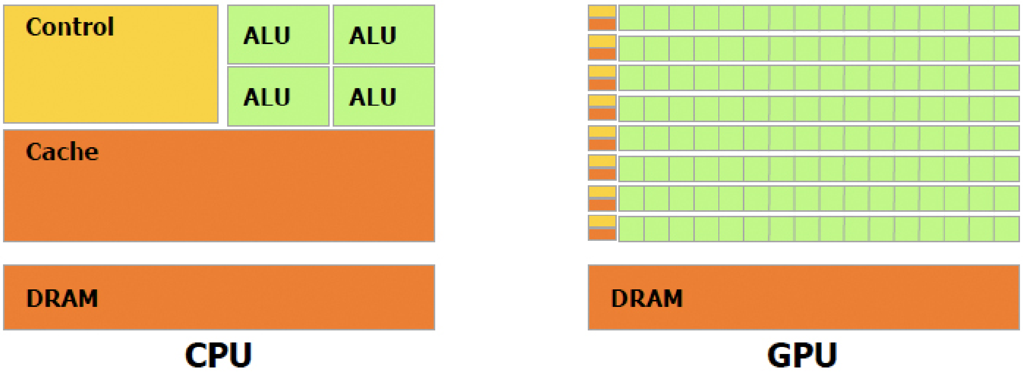


Figure 8.3: GPU devotes more transistors to data processing rather than to data caching and to flow control [NVIDIA 2011].

This decomposition enables automatic scalability and preserves language expressivity by allowing threads to cooperate while solving each sub-problem. Indeed, each block of threads can be scheduled on any of the available processor cores, in any order, concurrently or sequentially, so that a compiled CUDA program can execute on any number of processor cores as illustrated by figure 8.4, and only the runtime system needs to know the number of physical processors.

8.1.1.2 Thread and memory hierarchy

Threads can form a one-dimensional, two-dimensional, or three-dimensional *thread block*. There is a limit to the number of threads per block, since all threads of a block are expected to reside on the same processor core and must share the limited memory resources of that core. On current GPUs, a thread block may contain up to 1024 threads. Blocks are organized into a one-dimensional, two-dimensional, or three-dimensional *grid* of thread blocks. The number of thread blocks in a grid is usually dictated by the size of the data being processed or the number of processors in the system, which it can greatly exceed.

Threads work in a SIMD model (Single Instruction Multiple Data). All threads of a block reside on the same processor core. Threads within one block are split into basic scheduling units called warps, consisting of 32 threads. A warp executes one common instruction at a time, so full efficiency is realized when all 32 threads of a warp agree on their execution path.

CUDA threads may access data from multiple memory spaces during their execution as illustrated by figure 8.5. Each thread has private *local* memory. Each thread block has *shared* memory visible to all threads of the block and with the same lifetime as the block. Shared memory is organized in banks, and if the data is accessed

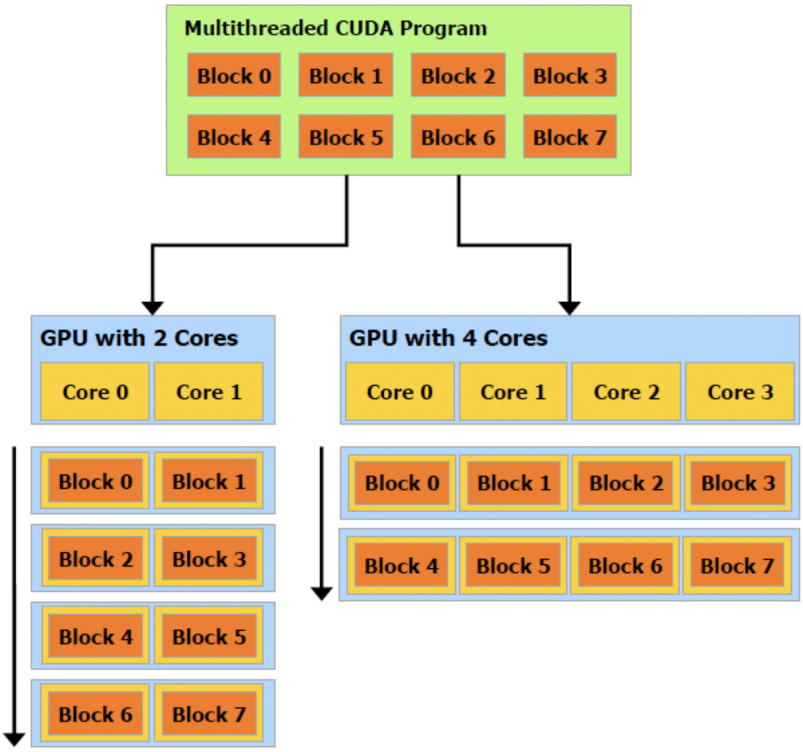


Figure 8.4: Automatic scalability: a multithreaded program is partitioned into blocks of threads that execute independently from each other, so that a GPU with more cores will automatically execute the program in less time than a GPU with fewer cores [NVIDIA 2011].

so that each thread of a block accesses a different bank (the addresses must meet alignment criteria), shared memory is as fast as registers. All threads have access to the same *global* memory. Global memory is a slow off-chip memory, which can be accessed by both CPU and GPU. It is also used to synchronize data between threads in different blocks. In order to have a low latency, it should be accessed in a coalesced fashion (consecutive threads in a warp must read memory in their order). Coalescing memory requests boost performance significantly over separate requests. The large thread count, together with support for many load requests, helps to cover load-to-use latency for *local* and *global* memory.

8.1.2 GPU database of signatures

We suggest storing signatures directly in GPU's global memory. This solution offloads CPU and minimize processing time during searching the most similar signatures (when the distance between signatures is calculated, the data is already stored on a proper unit).

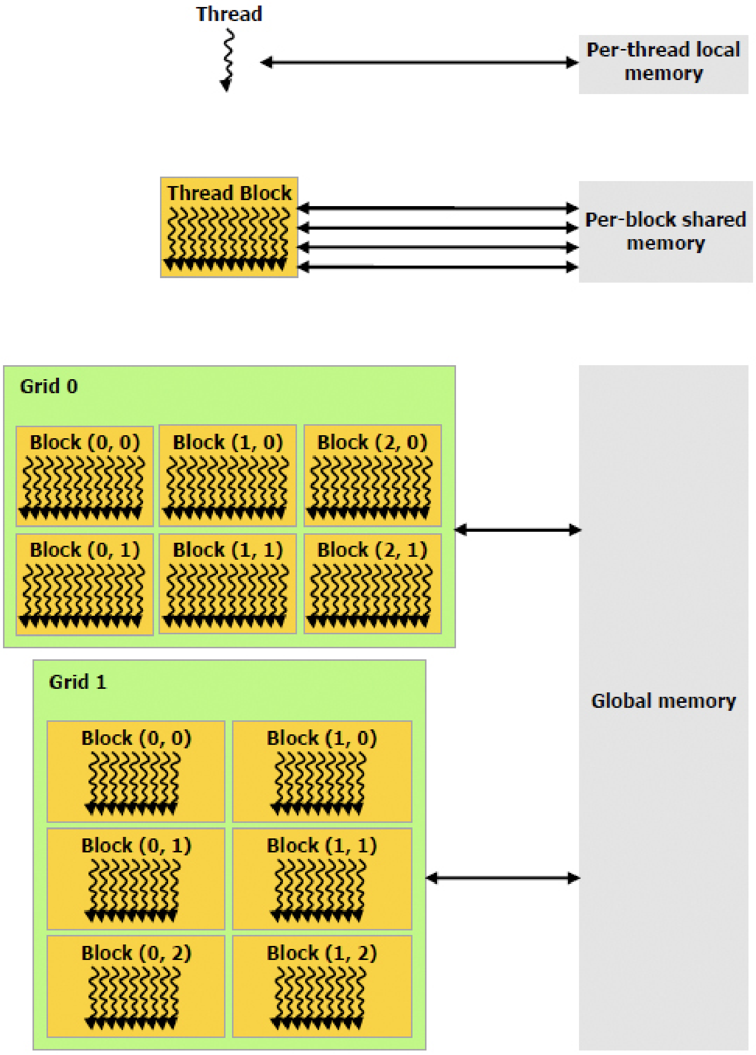


Figure 8.5: Memory hierarchy [NVIDIA 2011].

In our experimental setup, we use a Tesla S1070 with four GPU units. Each GPU unit has available 4GB of global memory. Taking into account the free space needed for calculations of a query to a database, we are able to store more than 2.000.000 covariance matrices. In the result, depending on the size of signature, 20.000 - 80.000 signatures can be stored in the database on a single GPU unit, which is sufficient for the purpose of the re-identification system (we do not expect more than such amount of signatures into the database as the re-identification approaches are constraint to the one day period: the strong assumption about the same clothes).

1. Calculate the *Cholesky decomposition*

$$B = LL^T. \tag{8.4}$$

2. Calculate *forward* and *backward substitution* to get

$$C = L^{-1}AL^{-T}. \tag{8.5}$$

3. *Tridiagonalize* the symmetric matrix C to prepare for solving the eigenvalue problem.
4. Use the *bisection algorithm* to find the eigenvalues of a symmetric tridiagonal matrix C.

Figure 8.6: The algorithm for solving a generalized eigenvalues problem.

8.2 Generalized eigenvalue problem on GPU

To compute the distance between two covariance matrices, it is necessary to solve a *generalized eigenvalue problem.* This problem is computationally heavy, making the signature distance computation time consuming. The generalized eigenvalue problem corresponds to finding vector $\mathbf{x}$ that obeys

$$A\mathbf{x} = \lambda B\mathbf{x}, \tag{8.1}$$

where A and B are matrices. In our case, A and B are positive definite symmetric matrices. λ denotes a vector of eigenvalues and $\mathbf{x}$ is the eigenvector. This equation can be decomposed to equation

$$(L^{-1}AL^{-T})(L^T\mathbf{x}) = \lambda(L^T\mathbf{x}) \tag{8.2}$$

where L is a lower triangular matrix calculated as $B = LL^T$. We can notice that the decomposed equation (8.2) already corresponds to the original *eigenvalue problem*:

$$C\mathbf{v} = \lambda\mathbf{v}. \tag{8.3}$$

To find eigenvalue vector λ, we solve a generalized eigenvalue problem. Our algorithm is presented in figure 8.6.

Let us note that there are many methods to calculate the eigenvalues of a symmetric matrix. On CPU we use a Jacobi algorithm, which does not need to perform the tridiagonalization first, however this algorithm is not easy to parallelize. For the GPU implementation we first *tridiagonalize* the matrix and then we use the *bisection algorithm*, which can be parallelized much more efficiently.

8.2.1 Porting the algorithm to GPU architecture

To compute the similarity of two signatures, we need to solve a generalized eigenvalue problem for a set of pairs of covariance matrices (distance computation). This can be naturally processed in parallel, providing the first source of parallelism in our algorithm. The second source involves the parallelism extracted from each step of the algorithm (figure 8.6) performed on a given pair of covariance matrices. We will now briefly describe how we use the parallelism of each procedure to efficiently port it to the GPU architecture.

8.2.1.1 Cholesky decomposition

In *Cholesky decomposition*, the formula to calculate each element of lower triangular matrix L is given as

$$L_{j,j} = \sqrt{B_{j,j} - \sum_{k=1}^{j-1} L_{j,k}^2}, \tag{8.6}$$

$$L_{i,j} = \frac{1}{L_{j,j}}(A_{i,j} - \sum_{k=1}^{j-1} L_{i,k}L_{j,k}) \quad for \;\; i > j. \tag{8.7}$$

Matrix B is loaded from the *global* memory to the *shared* memory in a coalesced way. After the calculations, matrix L is loaded back to the *global* memory, also assuring coalescence.

For the calculations we use the *Cholesky-Crout* algorithm, which starts from the upper left corner of matrix L and proceeds to calculate the matrix column by column, as processing data in column order ensures the memory coalescence. To calculate an element of a column, only the elements from the columns to the left are needed. So, all the elements in one column can be processed in parallel.

For matrices of size $d \times d$, we can therefore use d threads to calculate each column in parallel, using d iterations to calculate the whole matrix. The natural decomposition of a problem would be to assign one matrix to a block. However, d threads can be much less than the size of a warp (32), which is the smallest scheduling unit. In the result, assigning $\lfloor \frac{32}{d} \rfloor$ matrices to a block, we enable to process $\lfloor \frac{32}{d} \rfloor$ matrices without the increase of processing time, as $\lfloor \frac{32}{d} \rfloor \times d$ threads still form only one warp scheduled in one operation.

8.2.1.2 Forward and backward substitution

Using *forward* and *backward substitution* we can compute matrix C in the following way:

$$backward: \quad YL^T = A \quad \Rightarrow \quad Y = AL^{-T}, \tag{8.8}$$

and

$$forward: \quad LC = Y \quad \Rightarrow \quad C = L^{-1}Y = L^{-1}AL^{-T}, \tag{8.9}$$

without inverting matrices. Both substitutions are processed together to avoid copying intermediate data to the global memory.

In *forward substitution*, for each column k in matrix C, we can assign a system of linear equations:

$$\begin{array}{ccccccccc} l_{1,1}c_{1,k} & & & & & & & = & y_{1,k} \\ l_{2,1}c_{1,k} & + & l_{2,2}c_{2,k} & & & & & = & y_{2,k} \\ \vdots & & \vdots & & \ddots & & & & \vdots \\ l_{m,1}c_{1,k} & + & l_{m,2}c_{2,k} & + \cdots + & & & l_{m,m}c_{m,k} & = & y_{m,k}, \end{array}$$

where l, c, y represent corresponding matrix elements.

The resulting formulas are:

$$\begin{array}{rcl} c_{1,k} & = & \frac{y_{1,k}}{l_{1,1}}, \\ c_{2,k} & = & \frac{y_{2,k} - l_{2,1}c_{1,k}}{l_{2,2}}, \\ & \vdots & \\ c_{m,k} & = & \frac{y_{m,k} - \sum_{i=1}^{m-1} l_{m,i}c_{i,k}}{l_{m,m}} \end{array}$$

All columns of C can be calculated independently and in parallel. However, elements in each column $(c_{1,k}, c_{2,k}, \ldots, c_{m,k})$ are calculated sequentially.

Analogously, for *backward substitution*, we follow the same scheme, applied for computing rows in parallel. In this case, as we have upper triangular matrix L^T, we first compute $y_{k,m}$ row element, and then substitute that back into the previous equation for solving $y_{k,m-1}$, and repeating till $y_{k,1}$.

Similarly to *Cholesky decomposition*, we use d threads to calculate matrices, and we assign $\lfloor \frac{32}{d} \rfloor$ matrices to a block.

8.2.1.3 Tridiagonalization

Tridiagonalization is the part of the generalized eigenvalue algorithm which has the lowest level of parallelism. We use the *Householder transformation* [Householder 1958] to tridiagonalize the symmetric matrix. In this algorithm, $d-2$ iterations need to be performed sequentially; in each iteration the appropriate elements in the kth row and column are zeroed. In one iteration some of the computations, such as matrix multiplication and vector-vector multiplications can be parallelized. To do this, we use $d \times d$ threads for each matrix. We have also tested the version in which there are only d threads, each calculating one column of the multiplied matrices, but it was less efficient. As $d \times d$ threads is more than a size of a warp, we can assign one matrix per block without loosing the efficiency.

8.2.1.4 Bisection algorithm

Bisection algorithm is used to calculate the eigenvalues of a symmetric tridiagonal matrix C with a given approximation (for details see *e.g.* [Demmel 1997]).

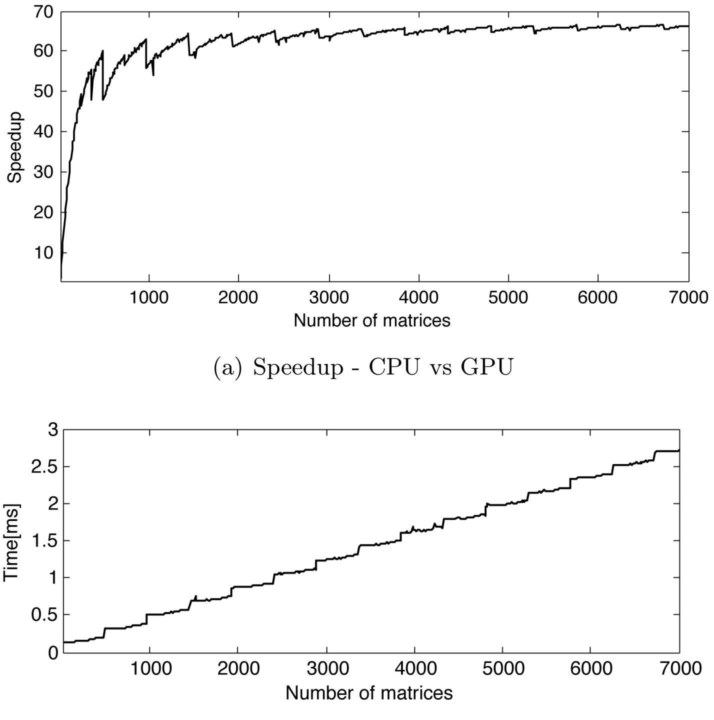
The core function of this algorithm is the `Count()` procedure returning the number of eigenvalues present in a given interval. The algorithm starts with the initial interval constructed using *Gerschgorin's* theorem. Then, it is divided in two and `Count()` procedure returns a number of eigenvalues in each subset. If it returns zero, the node is abandoned, otherwise it is further subdivided into two subsets, until the size of a subset is not bigger than the assumed approximation. For our purposes it is enough to use the approximation equal to 10e-6.

The main source of parallelism comes from the fact that the `Count()` function in each node can be calculated independently. As only d eigenvalues can be found, there are only up to d nodes containing at least one eigenvalue on each level of the binary tree. We therefore use d threads to calculate one matrix. Again, we assign $\lfloor \frac{32}{d} \rfloor \times d$ threads to a block, calculating $\lfloor \frac{32}{d} \rfloor$ matrices. *Gerschgorin's* procedure cannot be parallelized, so we process it sequentially, but a parallelism is obtained by calculating all the procedures for different matrices in parallel.

8.3 Experimental results

Experimental setup

In our experimental setup we use the NVIDIA Tesla S1070 with 4 graphic cards, each consisting of 30 SMs (Streaming Multiprocessors) with 8 scalar processors ($4 \times 30 \times 8$). The price of NVIDIA Tesla S1070 is around $7,000$\$ but NVIDIA Tesla C1060, which contains only a single graphic card with 240 cores costs around $2,000$\$.



(a) Speedup - CPU vs GPU

(b) Time of Bisection algorithm

Figure 8.7: Speedup for finding generalized eigenvalues and time of Bisection algorithm.

We perform comparison with CPU using only one graphic card, thus the total number of available processors is therefore equal to 240 cores.

We calculated the performance for matrices number ranging from 10 to 3000. Each matrix was of size 11 × 11, reflecting covariance C_1 (see section 7.3.2). As we have described in section 8.1.2, we assume that the database of signatures is stored directly on GPU. In our time estimation, we do not take into account the time of the data transfer, because the reference signatures already reside in the device memory, and the time of transferring the query signature is negligible.

Results

Below we present the speedup obtained with comparison to the optimal version on CPU, *i.e.* a version with *Jacobi* algorithm (implementation from LTI library), as on CPU this version is faster than the calculation of tridiagonalization and then the bisection algorithm.

The results are presented in figure 8.7 (a). We can easily notice that the speedup grows with the number of matrices, and reaches its maximum (66) from about 1500 matrices. Distributing the database of signatures between 4 GPU cards of Tesla,

N	Cholesky	Forward.S	Tridiag.	Bisect.	Total.GPU	Total.CPU
200	0.037	0.035	0.140	0.147	0.359	16
400	0.049	0.071	0.272	0.187	0.579	32
600	0.082	0.078	0.389	0.325	0.874	48
800	0.093	0.112	0.522	0.352	1.08	64
1000	0.126	0.120	0.657	0.505	1.41	80

Table 8.1: Time[ms] of component procedures.

and performing the calculations for a query signature on all cards in parallel, would result in further speedup improvement.

Table 8.1 presents the time of the component procedures for different number of matrices (N). The *Cholesky* and *forward/backward substitution* times are the lowest, while the biggest impact on the total time comes from the tridiagonalization procedure, as it has the lowest degree of parallelism.

In figure 8.7 (a) we can notice the *'stairs'* (drop of speedup) which occur in regular intervals every 480 matrices. This is a result of a similar effect observed in component functions (see *e.g. bisection algorithm* in figure 8.7 (b)) and is correlated to the number of processors. The architecture is the most efficiently used when all processors (240 for Tesla S1070) have blocks assigned, *i.e.* when $k \times 480$ matrices are processed (each block calculates two matrices). The worst case is for $k \times 480 + 1$ matrices - then, the time is almost equal to processing $(k + 1) \times 480$ matrices, which results in sudden drop of speedup in these points.

8.4 Conclusion

Recently, GPU has become popular tool not only for graphics application but also in physics simulation, computational finance, biology and image processing. Driven by demand of a large data processing for real-time, GPU has evolved into a high performance computing.

As human re-identification requires large data processing and fast response from a database of signatures, we have employed GPU in developing re-identification system. We have focussed on porting the *generalized eigenvalue* problem to the GPU architecture as this problem is computationally intensive and must be repeated constantly in our re-identification system during the browsing of signatures of interest. We have also proposed to store the database of signatures in the GPU memory, offloading CPU and minimizing data transfer while browsing signatures. In the result we have reached 66 speed-up with comparison to CPU implementation using a single GPU card in Tesla S1070.

Chapter 9

Conclusions and perspectives

"*A conclusion is simply the place where someone got tired of thinking.*"
(Arthur Block)

This thesis presented and evaluated several novel methods for appearance-based human re-identification. We have demonstrated that the proposed methods gain state of the art performance. We conclude our work pointing out the key contributions (section 9.1) and their limitations (section 9.2). Finally, we discuss future perspectives (section 9.3), indicating interesting directions for future research in this field.

9.1 Key contributions

Dense grid of mean Riemannian covariance features

We proposed a novel representation of a human appearance for re-identification. Our idea is to combine efficiently *mean Riemannian covariance* features with a spatial information carried out by a dense grid structure. We have found that *mean Riemannian covariance* is in general the best descriptor for matching appearances across disjoint camera views. The performance of the covariance features is found to be superior to other methods, as rotation and illumination changes are absorbed by the covariance matrix. Among the advantages of our solutions, the following ones are noteworthy:

- Covariance descriptor together with applying *histogram equalization* produce color invariant representation to camera changes.
- Averaging covariances, we blend appearance information from multiple images, providing partially pose invariant representation.
- Designing signature as a dense grid structure of features we are able to handle partial occlusions.
- Driven by the idea that different regions of the object appearance ought to be matched using different strategies, we formulate the appearance matching problem as the task of learning a model that selects the most descriptive features for a specific class of objects (*e.g.* humans).

Discriminative approaches

We investigated discriminative approaches by boosting human appearances in *one-against-all* learning scheme and by introducing discriminant analysis. In these approaches we enhance distinctive characteristics of a specific appearance by using information from the appearance of other individuals.

Body part driven re-identification

We developed two approaches which extract human appearance from a single image using body-part detection methods. Detecting body parts we overcome issues related to pose changes. Localization of body parts is based either on an asymmetry of a human body or on learned characteristics using *histogram of oriented gradients* (HOG).

Efficient implementation

Covariance matrix is a computationally intensive descriptor. Thus, we proposed an efficient and distinctive representation of the object appearance by using a combination of small covariance matrices (4×4) between a few relevant features. Moreover, we design and implement covariance distance operator on GPU architecture. Our implementation significantly speeds up computation with comparison to the optimal CPU implementation of Jacobi algorithm (we reached 66 speed-up).

Datasets

We extracted two new image sets of individuals from i-LIDS data to study more carefully advantages of using multiple images in generating human signature. These two publicly available datasets fully satisfy requirements of *multiple-shot* person re-identification.

9.2 Limitations of our approaches

The proposed approaches have still a number of limitations. A few of them could be investigated as possible extensions in the near future, while others still are open issues. This section presents the limitations and the next section provides a discussion of the future work, suggesting short-term and long-term perspectives.

Pose alignment

Although the proposed approaches give good results for many datasets, they are still basically a $2D$ template matching approaches that encode object appearance using image position. While matching appearance, we require a geometrically well aligned poses (in multiple-shot case, we believe that a sufficiently large number of training images can overcome this issue but it is still dependent on training data). As upright people are well structured objects, a fixed matching model can perform relatively well. However, significant pose changes (*e.g.* matching upright person with bending person) are still difficult to match (*the correspondence problem*), especially when body part detectors return inaccurate results.

Human detection and body part localization

As shown in section 7.6.3, our approaches are directly dependent on human detection accuracy. In the case of inaccurate detections, recognition performance significantly goes down.

Low resolution images

In section 7.6.5, we have also found that covariance descriptor does not handle low resolution images. In the result, performance of covariance matrix is relatively low comparing to simple color-based descriptors.

Time complexity

Despite covariance matrix can be computed very fast using *integral images*, it is still the computationally heavy part of our framework. Moreover, a distance operator between covariance matrices requires solving the *generalized eigenvalues problem*, which is also an intensive task. Although we propose two solutions for this problem (GPU-based implementation or/and the usage of small covariance matrices), the run-time is still significantly higher than a similar framework using simple color histograms.

9.3 Future Work

9.3.1 Short-Term Perspectives

Learning on a manifold

In section 6.2.2.1 we discuss problem related to learning on a manifold, where model is usually created using tangent planes over the Karcher means of the positive training data which can lead to the overfitted classifier. In the result, we have decided to use a selection method to find the most descriptive features for a specific class of objects. However, we have to stress that selection methods are usually used as preprocessing steps for machine learning algorithms. Although learning on a manifold is still an open issue, there are already approaches which explore a manifold space either by clustering methods [Goh 2008a] or by *control points* [Sen 2008]. Further studies should look for machine learning methods which could prosper in finding subspaces on a manifold determining the metric for appearance matching.

Human detection and body part localization

As we have already mentioned, inaccurate detections can significantly deteriorate recognition performance. Thus, future study should consider either improving accuracy of detection results or filtering methods which could ensure correctness of the data. Unfortunately, we did not take full advantage of silhouette of a person due to inaccurate results of the background subtraction algorithm. However, currently the market provides new kinds of sensors which are able to extract depth information (*e.g.* Kinect sensor) and could be used to extract only the features which surely belong to foreground regions.

Motion and shape information

As our background subtraction algorithm does not provide accurate extraction of a silhouette, shape descriptors could not be applied. However, thanks to such sensors as Kinect, currently we are able to extract depth information enabling extraction of shape and motion of an object. Shape characteristics might be used for pruning set of candidates by gender classification. Moreover, integrating the notion of motion in the recognition framework, we would allow to distinguish individuals using their behavioral characteristics (*e.g.* gait features).

9.3.2 Long-Term Perspectives

Context information

It would be interesting to explore top-down approaches, where the challenging issue is the exploitation of the general context. This general context could be related to different objects surrounding a person of interest. In section 2.2.1.3, we have already presented few approaches which employ visual information coming from surrounding people to reduce ambiguity in person re-identification [Zheng 2009, Cai 2010]. One alternative in further studies might be to use specialized object detectors (*e.g.* luggage detectors in airports) to determine presence of particular objects in close neighborhood of an individual and then to take advantage of this information to generate distinctive description of a person of interest.

Content-dependent signature

We expected that covariance matrix would be the best descriptor for appearance matching across disjoint camera views for every image/video. However, the experimental results on CAVIAR dataset indicate that it is not always correct assumption. It appears that simple features can perform better than covariance descriptors in the case of low resolution images. One direction in further studies might be to use content-dependent signature generation. Appearance representation could be extracted dependent on cues available in an image/video (*e.g.* different kinds of features could be used depending on image resolution). For that we would need a *smart* module which could decide which features should be extract w.r.t. to an image/video, to ensure the highest recognition accuracy. Different combinations of different kinds of features should also be considered in future research.

Camera invariant features

Although we have proposed techniques which cope with variations due to camera changes, further research should consider different descriptor representations. Currently, we can observe tendency of looking for metrics which handle camera variations. However, these approaches focus on learning a function which transfers features space from the first camera to the second one, introducing requirement of training $\left(\begin{array}{c} c \\ 2 \end{array}\right) = \frac{c!}{2!(c-2)!}$ models for c cameras. It is worth investigating if it is possible to learn a single model which could handle variations in all cameras. This also opens directions for exploring novel descriptors which would be invariant to camera changes.

Bibliography

[Andriluka 2009] Mykhaylo Andriluka, Stefan Roth and Bernt Schiele. *Pictorial Structures Revisited: People Detection and Articulated Pose Estimation.* In Proceedings of the 22nd Conference on Computer Vision and Pattern Recognition, CVPR. IEEE Computer Society, 2009. (Cited on page 29.)

[Bak 2010a] Slawomir Bak, Etienne Corvee, Francois Bremond and Monique Thonnat. *Person Re-identification Using Haar-based and DCD-based Signature.* In Proceedings of the 2nd Workshop on Activity Monitoring by Multi-Camera Surveillance Systems in conjunction with 7th IEEE International Conference on Advanced Video and Signal-Based Surveillance, AMMCSS. IEEE Computer Society, 2010. (Cited on page 7.)

[Bak 2010b] Slawomir Bak, Etienne Corvee, Francois Bremond and Monique Thonnat. *Person Re-identification Using Spatial Covariance Regions of Human Body Parts.* In Proceedings of the 7th IEEE International Conference on Advanced Video and Signal-Based Surveillance, AVSS. IEEE Computer Society, 2010. (Cited on page 7.)

[Bak 2011a] Slawomir Bak, Etienne Corvee, Francois Bremond and Monique Thonnat. *Boosted human re-identification using Riemannian manifolds.* Image and Vision Computing, 2011. (Cited on page 7.)

[Bak 2011b] Slawomir Bak, Etienne Corvee, Francois Bremond and Monique Thonnat. *Multiple-shot Human Re-Identification by Mean Riemannian Covariance Grid.* In Proceedings of the 8th IEEE International Conference on Advanced Video and Signal-Based Surveillance, AVSS. IEEE Computer Society, 2011. (Cited on page 7.)

[Bak 2011c] Slawomir Bak, Krzysztof Kurowski and Krystyna Napierala. *Human Re-identification System On Highly Parallel GPU and CPU Architectures.* In Proceedings of the 4th International Conference on Multimedia Communications, Services and Security, MCSS. Communications in Computer and Information Science, Springer, 2011. (Cited on page 7.)

[Bak 2012a] Slawomir Bak, Guillaume Charpiat, Etienne Corvee, Francois Bremond and Monique Thonnat. *Learning to Match Appearances by Correlations in a Covariance Metric Space.* In submitted to Proceedings of the 12th European Conference on Computer Vision,, ECCV. IEEE Computer Society, 2012. (Cited on page 7.)

[Bak 2012b] Slawomir Bak, Duc-Phu Chau, Julien Badie, Etienne Corvee, Francois Bremond and Monique Thonnat. *Multi-target tracking by discriminative analysis on Riemannian manifold.* In Proceedings of the 19th International

Conference on Image Processing, ICIP. IEEE Computer Society, 2012. (Cited on page 7.)

[Bauml 2010] Martin Bauml, Keni Bernardin, Mika Fischer, Hazim Kemal Ekenel and Rainer Stiefelhagen. *Multi-pose Face Recognition for Person Retrieval in Camera Networks.* In Proceedings of the 7th IEEE International Conference on Advanced Video and Signal-Based Surveillance. AVSS, 2010. (Cited on page 14.)

[Bay 2008] Herbert Bay, Andreas Ess, Tinne Tuytelaars and Luc Van Gool. *Speeded-Up Robust Features (SURF).* Computer Vision and Image Understanding, vol. 110, pages 346–359, June 2008. (Cited on page 27.)

[Bazzani 2010] Loris Bazzani, Marco Cristani, Alessandro Perina, Michela Farenzena and Vittorio Murino. *Multiple-Shot Person Re-identification by HPE Signature.* In Proceedings of the 20th International Conference on Pattern Recognition, ICPR, pages 1413–1416. IEEE Computer Society, 2010. (Cited on pages 18, 28, 33, 104, 125, 126 and 128.)

[Belhumeur 1997] P. N. Belhumeur, J. P. Hespanha and D. J. Kriegman. *Eigenfaces vs. Fisherfaces: recognition using class specific linear projection.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pages 711–720, July 1997. (Cited on pages 14 and 19.)

[Belongie 2002] S. Belongie, J. Malik and J. Puzicha. *Shape Matching and Object Recognition Using Shape Contexts.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 4, pages 509–522, April 2002. (Cited on page 21.)

[Bird 2005] Nathaniel D. Bird, Osama Masoud, Nikolaos P. Papanikolopoulos and Aaron Isaacs. *Detection of loitering individuals in public transportation areas.* IEEE Transactions on Intelligent Transportation Systems, pages 167–177, 2005. (Cited on pages 28, 32 and 33.)

[Bjork 1996] E.L. Bjork and R.A. Bjork. Memory. Handbook of perception and cognition. Academic Press, 1996. (Cited on page 6.)

[Boykov 2001] Y. Y. Boykov and M. P. Jolly. *Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images.* In Proceedings of the 8th IEEE International Conference on Computer Vision, ICCV, pages 105–112. IEEE Computer Society, 2001. (Cited on page 46.)

[Buchsbaum 1980] G. Buchsbaum. *A spatial processor model for object colour perception.* Journal of the Franklin Institute, vol. 310, no. 1, pages 1 – 26, 1980. (Cited on pages 49 and 50.)

[Cai 2008] Yinghao Cai, Kaiqi Huang and Tieniu Tan. *Human appearance matching across multiple non-overlapping cameras.* In Proceedings of the 19th International Conference on Pattern Recognition, ICPR. IEEE Computer Society, 2008. (Cited on pages 19, 20 and 33.)

[Cai 2010] Yinghao Cai, Valtteri Takala and Matti Pietikainen. *Matching Groups of People by Covariance Descriptor.* In Proceedings of the 20th International Conference on Pattern Recognition, ICPR, pages 2744–2747. IEEE Computer Society, 2010. (Cited on pages 24 and 150.)

[Canny 1986] J. Canny. *A computational approach to edge detection.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 8, no. 6, pages 679–698, November 1986. (Cited on page 19.)

[Catanzaro 2008] Bryan Catanzaro, Narayanan Sundaram and Kurt Keutzer. *Fast support vector machine training and classification on graphics processors.* In Proceedings of the 25th International Conference on Machine learning, ICML, pages 104–111. ACM, 2008. (Cited on page 83.)

[Chellappa 2007] Rama Chellappa, Amit K. Roy-Chowdhury and Amit Kale. *Human Identification using Gait and Face.* In Proceedings of the 20th Conference on Computer Vision and Pattern Recognition, CVPR, pages 1–2. IEEE Computer Society, 2007. (Cited on page 16.)

[Cheng 2011] Dong Seon Cheng, Marco Cristani, Michele Stoppa, Loris Bazzani and Vittorio Murino. *Custom Pictorial Structures for Re-identification.* In Proceedings of the British Machine Vision Conference, BMVC, pages 68.1–68.11. BMVA Press, 2011. (Cited on pages 29, 30, 33, 101, 102, 104, 108, 125, 126, 128, 129, 130 and 131.)

[Cortes 1995] Corinna Cortes and Vladimir Vapnik. *Support-Vector Networks.* Machine Learning, vol. 20, pages 273–297, September 1995. (Cited on pages 26 and 37.)

[Corvee 2009] E. Corvee and F. Bremond. *Combining face detection and people tracking in video sequences.* In Proceedings of the 3rd International Conference on Imaging for Crime Detection and Prevention, ICDP. IEEE Computer Society, 2009. (Cited on page 40.)

[Corvee 2010] Etienne Corvee and Francois Bremond. *Body parts detection for people tracking using trees of Histogram of Oriented Gradient descriptors.* In Proceedings of the 7th IEEE International Conference on Advanced Video and Signal-Based Surveillance, AVSS. IEEE Computer Society, 2010. (Cited on pages 40, 41 and 54.)

[Corvee 2012] Etienne Corvee, Slawomir Bak and François Bremond. *People detection and re-identification for multi surveillance cameras.* In Proceedings of the 7th International Conference on Computer Vision Theory and Applications, VISAPP. INSTICC Press, 2012. (Cited on pages 7, 106, 107 and 118.)

[Cressie 1999] Noel Cressie and Hsin cheng Huang. *Classes of Nonseparable, Spatio-temporal Stationary Covariance Functions.* Journal of the American Statistical Association, pages 1330–1340, 1999. (Cited on page 76.)

[Dalal 2005] Navneet Dalal and Bill Triggs. *Histograms of Oriented Gradients for Human Detection.* In Proceedings of the 18th Conference on Computer Vision and Pattern Recognition, CVPR, pages 886–893. IEEE Computer Society, 2005. (Cited on pages 38, 39 and 86.)

[Daugman 2002] J. Daugman. *How iris recognition works.* In Proceedings of the 2002 International Conference on Image Processing, ICIP, pages 33 – 36, 2002. (Cited on pages 12 and 13.)

[Demmel 1997] J.W. Demmel and M.T. Heath. *Applied Numerical Linear Algebra.* In Society for Industrial and Applied Mathematics. SIAM, 1997. (Cited on page 143.)

[Deng 2001] Yining Deng, B.S. Manjunath, C. Kenney, M.S. Moore and H. Shin. *An efficient color representation for image retrieval.* IEEE Transactions on Image Processing, vol. 10, no. 1, pages 140 –147, jan 2001. (Cited on page 50.)

[Dikmen 2010] Mert Dikmen, Emre Akbas, Thomas S. Huang and Narendra Ahuja. *Pedestrian recognition with a learned metric.* In Proceedings of the 10th Asian Conference on Computer Vision, ACCV, pages 501–512. IEEE Computer Society, 2010. (Cited on pages 18, 23, 33, 108, 129 and 131.)

[Doretto 2011] 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, vol. 2, pages 127–151, 2011. (Cited on page 17.)

[Duan 2009] Genquan Duan, Chang Huang, Haizhou Ai and Shihong Lao. *Boosting Associated Pairing Comparison Features for Pedestrian Detection.* In Proceedings of the 12th International Conference on Computer Vision Workshops, ICCV Workshop. IEEE Computer Society, 2009. (Cited on page 39.)

[Ess 2007] A. Ess, B Leibe and L Van Gool. *Depth and Appearance for Mobile Scene Analysis.* In Proceedings of the 11th International Conference on Computer Vision, ICCV, pages 1–8. IEEE Computer Society, Oct. 2007. (Cited on page 102.)

[Farenzena 2010] M. Farenzena, L. Bazzani, A. Perina, V. Murino and M. Cristani. *Person re-identification by symmetry-driven accumulation of local features.* In Proceedings of the 23th Conference on Computer Vision and Pattern Recognition, CVPR, pages 2360–2367. IEEE Computer Society, 2010. (Cited on pages 18, 27, 29, 33, 102, 104, 108, 125, 126, 128, 130 and 131.)

[Fayyad 1993] Usama M. Fayyad and Keki B. Irani. *Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning.* In Proceedings of the International Joint Conference on Uncertainty in AI, IJCAI, pages 1022–1027, 1993. (Cited on page 93.)

[Fisher 1936] Ronald A. Fisher. *The use of multiple measurements in taxonomic problems.* Annals of Eugenics, vol. 7, pages 179–188, 1936. (Cited on pages 14 and 26.)

[Fogel 1989] I. Fogel and D. Sagi. *Gabor filters as texture discriminator.* Biological Cybernetics, vol. 61, no. 2, pages 103–113, June 1989. (Cited on page 22.)

[Forssén 2007] Per-Erik Forssén. *Maximally Stable Colour Regions for Recognition and Matching.* In Proceedings of the 20th Conference on Computer Vision and Pattern Recognition, CVPR. IEEE Computer Society, 2007. (Cited on pages 28, 29 and 129.)

[Förstner 1999] Wolfgang Förstner and Boudewijn Moonen. *A Metric for Covariance Matrices.* In Quo vadis geodesia ...?, Festschrift for Erik W. Grafarend on the occasion of his 60th birthday, TR Dept. of Geodesy and Geoinformatics, Stuttgart University, 1999. (Cited on pages 55, 92 and 96.)

[Freund 1995] Yoav Freund and Robert E. Schapire. *A decision-theoretic generalization of on-line learning and an application to boosting.* In Proceedings of the 2nd European Conference on Computational Learning Theory, EuroCOLT, pages 23–37. Springer-Verlag, 1995. (Cited on page 65.)

[Freund 1997] Yoav Freund and Robert E. Schapire. *A decision-theoretic generalization of on-line learning and an application to boosting.* Journal of Computer and System Sciences, vol. 55, pages 119–139, August 1997. (Cited on page 37.)

[Freund 2003] Yoav Freund, Raj Iyer, Robert E. Schapire and Yoram Singer. *An efficient boosting algorithm for combining preferences.* Journal of Machine Learning Research, vol. 4, pages 933–969, December 2003. (Cited on page 23.)

[Fuentes 2006] Montserrat Fuentes. *Testing for separability of spatial-temporal covariance functions.* Journal of Statistical Planning and Inference, no. 2, pages 447 – 466, 2006. (Cited on page 76.)

[Funt 2002] B. V. Funt and G. D. Finlayson. *Color constant color indexing.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 17, no. 5, pages 522–529, August 2002. (Cited on page 21.)

[Gallagher 2008] Andrew C. Gallagher and Tsuhan Chen. *Clothing cosegmentation for recognizing people.* In Proceedings of the 21st Conference on Computer Vision and Pattern Recognition, CVPR, pages 1–8. IEEE Computer Society, 2008. (Cited on pages 19 and 33.)

[Gevers 2004] Theo Gevers and Harro Stokman. *Robust Histogram Construction from Color Invariants for Object Recognition.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 26, pages 113–117, 2004. (Cited on pages 49 and 50.)

[Gheissari 2006] Niloofar Gheissari, Thomas B. Sebastian and Richard Hartley. *Person Reidentification Using Spatiotemporal Appearance.* In Proceedings of the 19th Conference on Computer Vision and Pattern Recognition, CVPR, pages 1528–1535. IEEE Computer Society, 2006. (Cited on pages 26, 27 and 33.)

[Goh 2008a] A. Goh and R. Vidal. *Clustering and dimensionality reduction on Riemannian manifolds.* In Proceedings of the 21st Conference on Computer Vision and Pattern Recognition, CVPR, pages 1–7. IEEE Computer Society, june 2008. (Cited on pages 97 and 149.)

[Goh 2008b] Alvina Goh and René Vidal. *Unsupervised Riemannian Clustering of Probability Density Functions.* In Proceedings of the 2008 European Conference on Machine Learning and Knowledge Discovery in Databases - Part I, ECML PKDD, pages 377–392, Berlin, Heidelberg, 2008. Springer-Verlag. (Cited on page 74.)

[Gonzalez 2001] Rafael C. Gonzalez and Richard E. Woods. Digital image processing. Addison-Wesley Longman Publishing Co., Boston, MA, USA, 2001. (Cited on page 50.)

[Grauman 2005] Kristen Grauman and Trevor Darrell. *The pyramid match kernel: Discriminative classification with sets of image features.* In Proceedings of the 10th IEEE International Conference on Computer Vision, ICCV, pages 1458–1465. IEEE Computer Society, 2005. (Cited on page 57.)

[Gray 2007] D. Gray, S. Brennan and H. Tao. *Evaluating Appearance Models for Recognition, Reacquisition, and Tracking.* In Proceedings of the IEEE International Workshop on Performance Evaluation for Tracking and Surveillance, PETS. IEEE Computer Society, 2007. (Cited on pages 99 and 108.)

[Gray 2008] Douglas Gray and Hai Tao. *Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features.* In Proceedings of the 10th European Conference on Computer Vision, ECCV, pages 262–275. Springer-Verlag, 2008. (Cited on pages 22, 23, 33, 108, 120, 126 and 131.)

[Hall 1999] Mark A. Hall. *Correlation-based Feature Subset Selection for Machine Learning.* PhD thesis, Department of Computer Science, University of Waikato, 1999. (Cited on page 93.)

[Hamdoun 2008] O. Hamdoun, F. Moutarde, B. Stanciulescu and B. Steux. *Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences.* In Proceedings of the 2nd ACM/IEEE International Conference on Distributed Smart Cameras, ICDSC, pages 1–6. IEEE Computer Society, Sept. 2008. (Cited on pages 27 and 33.)

[Hirzer 2011] Martin Hirzer, Csaba Beleznai, Peter M. Roth and Horst Bischof. *Person Re-identification by Descriptive and Discriminative Classification.* In Proceedings of the 17th Scandinavian Conference on Image Analysis, SCIA, pages 91–102, Berlin, Heidelberg, 2011. Springer-Verlag. (Cited on pages 30, 32, 33, 90, 93 and 108.)

[Hordley 2005] S. D. Hordley, G. D. Finlayson, G. Schaefer and G. Y. Tian. *Illuminant and device invariant colour using histogram equalisation.* Pattern Recognition, 2005. (Cited on pages 31, 49, 50 and 89.)

[Householder 1958] Alston S. Householder. *Unitary Triangularization of a Nonsymmetric Matrix.* Journal of the ACM, pages 339–342, October 1958. (Cited on page 143.)

[Hu 2008] Lei Hu, Yizhou Wang, Shuqiang Jiang, Qingming Huang and Wen Gao. *Human reappearance detection based on on-line learning.* In Proceedings of the 19th International Conference on Pattern Recognition, ICPR, pages 1 –4. IEEE Computer Society, 2008. (Cited on pages 30 and 33.)

[Huang 1997] Jing Huang, S. Ravi Kumar, Mandar Mitra, Wei-Jing Zhu and Ramin Zabih. *Image Indexing Using Color Correlograms.* In Proceedings of the 10th Conference on Computer Vision and Pattern Recognition, CVPR, pages 762–769. IEEE Computer Society, 1997. (Cited on page 30.)

[Huang 2009] Chung-Hsien Huang, Yi-Ta Wu and Ming-Yu Shih. *Unsupervised Pedestrian Re-identification for Loitering Detection.* In Advances in Image and Video Technology, Lecture Notes in Computer Science, pages 771–783. Springer Berlin / Heidelberg, 2009. (Cited on pages 26, 28 and 33.)

[Huang 2010] Chang Huang and R. Nevatia. *High performance object detection by collaborative learning of Joint Ranking of Granules features.* In Proceedings of the 23th Conference on Computer Vision and Pattern Recognition, pages 41 –48, june 2010. (Cited on page 39.)

[Jacobi 1846] C.J.G. Jacobi. *Über ein leichtes Verfahren, die in der Theorie der Säkularstörungen vorkommenden Gleichungen numerisch aufzulösen.* Journal für reine und angewandte Mathematik, vol. 30, pages 51–95, 1846. (Cited on page 7.)

[Javed 2003] Omar Javed, Zeeshan Rasheed, Khurram Shafique and Mubarak Shah. *Tracking Across Multiple Cameras With Disjoint Views.* In Proceedings of the 9th IEEE International Conference on Computer Vision, ICCV, pages 952–957. IEEE Computer Society, 2003. (Cited on page 25.)

[Javed 2005] Omar Javed, Khurram Shafique and Mubarak Shah. *Appearance Modeling for Tracking in Multiple Non-overlapping Cameras.* In Proceedings of the 18th Conference on Computer Vision and Pattern Recognition, CVPR, pages 26–33, 2005. (Cited on pages 25 and 49.)

[Javed 2007] 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, vol. 109, pages 146–162, February 2007. (Cited on page 133.)

[Joachims 2002] Thorsten Joachims. *Optimizing search engines using clickthrough data.* In Proceedings of the 8th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD, pages 133–142, 2002. (Cited on page 23.)

[Jojic 2003] Nebojsa Jojic, Brendan J. Frey and Anitha Kannan. *Epitomic analysis of appearance and shape.* In Proceedings of the 9th IEEE International Conference on Computer Vision, ICCV, pages 34–41. IEEE Computer Society, 2003. (Cited on page 28.)

[Kang 2004] Jinman Kang, Isaac Cohen and Gerard G. Medioni. *Object Reacquisition Using Invariant Appearance Model.* In Proceedings of the 17th International Conference on Pattern Recognition, pages 759–762, 2004. (Cited on pages 20, 21 and 33.)

[Kirby 1990] M. Kirby and L. Sirovich. *Application of the Karhunen-Loeve procedure for the characterization of human faces.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, pages 103 –108, 1990. (Cited on page 14.)

[Kuhn 1955] H. W. Kuhn. *The Hungarian method for the assignment problem.* Naval Research Logistic Quarterly, vol. 2, pages 83–97, 1955. (Cited on page 42.)

[Kullback 1978] S. Kullback. *Information Theory and Statistics.* Gloucester, Mass: Peter Smith, 1978. (Cited on page 20.)

[Kuo 2010] Cheng-Hao Kuo, Chang Huang and R. Nevatia. *Multi-target tracking by on-line learned discriminative appearance models.* In Proceedings of the 23th Conference on Computer Vision and Pattern Recognition, CVPR, pages 685 –692. IEEE Computer Society, june 2010. (Cited on page 43.)

[Lazebnik 2006] Svetlana Lazebnik, Cordelia Schmid and Jean Ponce. *Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories.* In Proceedings of the 19th Conference on Computer Vision and Pattern Recognition, CVPR, pages 2169–2178. IEEE Computer Society, 2006. (Cited on page 57.)

[Lee 2003] Kuang-Chih Lee, J. Ho, Ming-Hsuan Yang and D. Kriegman. *Video-based face recognition using probabilistic appearance manifolds.* In Proceedings of the 16th Conference on Computer Vision and Pattern Recognition, CVPR, pages I–313–I–320 vol.1. IEEE Computer Society, 2003. (Cited on page 14.)

[Lienhart 2002] R. Lienhart and J. Maydt. *An extended set of Haar-like features for rapid object detection.* In Proceedings of the International Conference on Image Processing, ICIP, pages I–900–I–903 vol.1, 2002. (Cited on pages 62 and 63.)

[Lin 2008] Zhe Lin and Larry S. Davis. *Learning Pairwise Dissimilarity Profiles for Appearance Recognition in Visual Surveillance.* In Proceedings of the 4th International Symposium on Advances in Visual Computing, ISVS, pages 23–34. Springer-Verlag, 2008. (Cited on pages 18, 23, 24, 33, 82 and 88.)

[Lindholm 2008] Erik Lindholm, John Nickolls, Stuart Oberman and John Montrym. *NVIDIA Tesla: A Unified Graphics and Computing Architecture.* IEEE Micro, pages 39–55, March 2008. (Cited on page 83.)

[Lowe 2004] David G. Lowe. *Distinctive Image Features from Scale-Invariant Keypoints.* International Journal of Computer Vision, vol. 60, pages 91–110, November 2004. (Cited on pages 24, 27, 30 and 32.)

[Madden 2007a] C. Madden, M. Piccardi and S. Zuffi. *Comparison of techniques for mitigating the effects of illumination variations on the appearance of human targets.* In Proceedings of the 3rd International conference on Advances in Visual Computing, ISVC, pages 116–127, Berlin, Heidelberg, 2007. Springer-Verlag. (Cited on pages 49 and 50.)

[Madden 2007b] Christopher Madden, Eric Cheng and Massimo Piccardi. *Tracking people across disjoint camera views by an illumination-tolerant appearance representation.* Machine Vision and Applications, vol. 18, pages 233–247, 2007. (Cited on pages 26 and 33.)

[Matey 2008] James Matey, David Ackerman, James Bergen and Michael Tinker. *Iris Recognition in Less Constrained Environments.* Advances in Biometrics, vol. 1, pages 107–131, 2008. (Cited on page 13.)

[Mico 1992] L. Mico, J. Oncina and E. Vidal. *An algorithm for finding nearest neighbours in constant average time with a linear space complexity.* Pattern Recognition, vol. 2, pages 557 –560, 1992. (Cited on page 134.)

[Mitchell 1997] Tom M. Mitchell. Machine learning. McGraw-Hill, New York, 1997. (Cited on page 61.)

[Nakajima 2003] Chikahito Nakajima, Massimiliano Pontil, Bernd Heisele and Tomaso Poggio. *Full-body person recognition system.* Pattern Recognition, pages 1997–2006, 2003. (Cited on pages 31 and 33.)

[NVIDIA 2011] NVIDIA. *CUDA Programming Guide 4.0*, 2011. (Cited on pages 135, 136, 137, 138 and 139.)

[Ojala 2002] T. Ojala, M. Pietikainen and T. Maenpaa. *Multiresolution gray-scale and rotation invariant texture classification with local binary patterns.* IEEE

Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pages 971 –987, jul 2002. (Cited on page 43.)

[Oreifej 2010] Omar Oreifej, Ramin Mehran and Mubarak Shah. *Human Identity Recognition in Aerial Images.* In Proceedings of the 23th Conference on Computer Vision and Pattern Recognition, CVPR, pages 709–716. IEEE Computer Society, 2010. (Cited on pages 28, 29 and 33.)

[Papageorgiou 1998] C. P. Papageorgiou, M. Oren and T. Poggio. *A general framework for object detection.* In Proceedings of 16th International Conference on Computer Vision, ICCV, pages 555–562, 1998. (Cited on page 62.)

[Papageorgiou 2000] Constantine Papageorgiou and Tomaso Poggio. *A Trainable System for Object Detection.* International Journal of Computer Vision, vol. 38, pages 15–33, June 2000. (Cited on pages 38 and 39.)

[Park 2006] U. Park, A.K. Jain, I. Kitahara, K. Kogure and N. Hagita. *ViSE: Visual Search Engine Using Multiple Networked Cameras.* In Proceedings of the 18th International Conference on Pattern Recognition, ICPR, pages 1204–1207. IEEE Computer Society, Aug. 2006. (Cited on pages 19 and 33.)

[Pennec 2006] Xavier Pennec, Pierre Fillard and Nicholas Ayache. *A Riemannian Framework for Tensor Computing.* International Journal on Computer Vision, vol. 66, no. 1, pages 41–66, 2006. (Cited on pages 55, 73, 74, 75 and 93.)

[Pinto 2011] N. Pinto and D. D. Cox. *GPU Metaprogramming: A Case Study in Biologically-Inspired Computer Vision.* In GPU Computing Gems, Jade Edition. Morgan Kaufmann Publishers, 2011. (Cited on page 136.)

[Porikli 2003] F. Porikli. *Inter-camera color calibration by correlation model function.* In Proceedings of the 10th International Conference on Image Processing, ICIP. IEEE Computer Society, 2003. (Cited on page 49.)

[Proenca 2010] Hugo Proenca. *Iris Recognition: On the Segmentation of Degraded Images Acquired in the Visible Wavelength.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, pages 1502–1516, 2010. (Cited on page 13.)

[Prosser 2010] Bryan Prosser, Wei-Shi Zheng, Shaogang Gong and Tao Xiang. *Person Re-Identification by Support Vector Ranking.* In Proceedings of the 21st British Machine Vision Conference, BMVC, pages 21.1–21.11. BMVC Press, 2010. (Cited on pages 23, 33 and 108.)

[Quinlan 1986] J. R. Quinlan. *Induction of decision trees.* Machine Learning, vol. 1, no. 1, pages 81–106, March 1986. (Cited on page 67.)

[Rich 1991] Elaine Rich and Kevin Knight. Artificial intelligence. McGraw-Hill Higher Education, 1991. (Cited on page 95.)

[Rother 2004] Carsten Rother, Vladimir Kolmogorov and Andrew Blake. *"GrabCut": interactive foreground extraction using iterated graph cuts.* In ACM SIGGRAPH 2004 Papers, SIGGRAPH, pages 309–314, New York, NY, USA, 2004. ACM. (Cited on pages 46 and 127.)

[Rudin 2007] Cynthia Rudin, Robert E. Schapire and Ingrid Daubechies. *Analysis of Boosting Algorithms using the Smooth Margin Function.* The Annals of Statistics, vol. 35, no. 6, pages 2723–2768, 2007. (Cited on pages 70 and 113.)

[Schapire 1998] Robert E. Schapire, Yoav Freund, Peter Bartlett and Wee S. Lee. *Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods.* The Annals of Statistics, vol. 26, no. 5, pages 1651–1686, 1998. (Cited on pages 70 and 113.)

[Schapire 1999] Robert E. Schapire and Yoram Singer. *Improved Boosting Algorithms Using Confidence-rated Predictions.* Machine Learning, no. 3, pages 297–336, 1999. (Cited on pages 79 and 81.)

[Schmid 2001] Cordelia Schmid. *Constructing models for content-based image retrieval.* In Proceedings of the 14th Conference on Computer Vision and Pattern Recognition, CVPR, 2001. (Cited on page 22.)

[Schwartz 2009] William Robson Schwartz and Larry S. Davis. *Learning Discriminative Appearance-Based Models Using Partial Least Squares.* In Proceedings of the 22nd Brazilian Symposium on Computer Graphics and Image Processing, SIBGRAPI, pages 322–329. IEEE Computer Society, 2009. (Cited on pages 18, 24, 33, 82, 83, 88, 102, 104 and 128.)

[Sen 2008] S.K Sen. *Classification on Manifolds.* PhD thesis, The University of North Carolina at Chapel Hill. Statistics, 2008. (Cited on pages 97 and 149.)

[Sirovich 1987] L. Sirovich and M. Kirby. *Low-dimensional procedure for the characterization of human faces.* Journal of the Optical Society of America A, vol. 4, pages 519–524, 1987. (Cited on page 14.)

[Snidaro 2008] L. Snidaro, I. Visentini and G.L. Foresti. *Dynamic Models for People Detection and Tracking.* In Proceedings of the 5th IEEE International Conference on Advanced Video and Signal-Based Surveillancen, AVSS, pages 29 –35, sept. 2008. (Cited on page 42.)

[Teixeira 2009] Luis F. Teixeira and Luis Corte-Real. *Video object matching across multiple independent views using local descriptors and adaptive learning.* Pattern Recognition Letters, vol. 30, pages 157–167, January 2009. (Cited on pages 32 and 33.)

[Tieu 2004] K. Tieu and P. Viola. *Boosting Image Retrieval.* International Journal of Computer Vision, vol. 56, no. 1, pages 17–36, 2004. (Cited on page 66.)

[Tipping 1999] Michael E. Tipping and Chris M. Bishop. *Probabilistic Principal Component Analysis*. Journal of the Royal Statistical Society, Series B, vol. 61, no. 3, pages 611–622, 1999. (Cited on page 25.)

[Truong Cong 2009] Dung Nghi Truong Cong, Catherine Achard, Louahdi Khoudour and Lounis Douadi. *Video Sequences Association for People Re-identification across Multiple Non-overlapping Cameras*. In Proceedings of the 15th International Conference on Image Analysis and Processing, ICIAP, pages 179–189. Springer-Verlag, 2009. (Cited on pages 31, 32 and 33.)

[Truong Cong 2010] D. N. Truong Cong, L. Khoudour, C. Achard, C. Meurie and O. Lezoray. *People re-identification by spectral classification of silhouettes*. Signal Processing, vol. 90, pages 2362–2374, August 2010. (Cited on pages 32 and 33.)

[Tuzel 2006] Oncel Tuzel, Fatih Porikli and Peter Meer. *Region Covariance: A Fast Descriptor for Detection And Classification*. In Proceedings of the 9th European Conference on Computer Vision, ECCV, pages 589–600. Springer-Verlag, May 2006. (Cited on pages 24, 33, 55, 56 and 90.)

[Tuzel 2008] Oncel Tuzel, Fatih Porikli and Peter Meer. *Pedestrian Detection via Classification on Riemannian Manifolds*. IEEE Transactions on Pattern Analysis and Machine Intelligence, pages 1713–1727, October 2008. (Cited on pages 39 and 93.)

[Viola 2001] Paul Viola and Michael Jones. *Rapid object detection using a boosted cascade of simple features*. In Proceedings of the 14th Conference on Computer Vision and Pattern Recognition, CVPR, pages 511–518. IEEE Computer Society, 2001. (Cited on pages 38, 56, 62, 65 and 67.)

[Viola 2003] Paul Viola and Michael Jones. *Fast Multiview Face Detection*. In Proceedings of the 16th Conference on Computer Vision and Pattern Recognition, CVPR. IEEE Computer Society, 2003. (Cited on page 19.)

[Viola 2005] Paul Viola, Michael J. Jones and Daniel Snow. *Detecting Pedestrians Using Patterns of Motion and Appearance*. International Journal of Computer Vision, vol. 63, pages 153–161, July 2005. (Cited on page 38.)

[Wang 2003] Liang Wang, Tieniu Tan, Huazhong Ning and Weiming Hu. *Silhouette Analysis-Based Gait Recognition for Human Identification*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, pages 1505–1518, 2003. (Cited on pages 15 and 16.)

[Wang 2007] Xiaogang Wang, G. Doretto, T. Sebastian, J. Rittscher and P. Tu. *Shape and Appearance Context Modeling*. In Proceedings of the 11th International Conference on Computer Vision, ICCV, pages 1–8. IEEE Computer Society, Oct. 2007. (Cited on pages 21 and 33.)

[Weinberger 2009] Kilian Q. Weinberger and Lawrence K. Saul. *Distance Metric Learning for Large Margin Nearest Neighbor Classification.* The Journal of Machine Learning Research, vol. 10, pages 207–244, june 2009. (Cited on page 23.)

[Wildes 1997] R. P. Wildes. *Iris recognition: an emerging biometric technology.* Proceedings of The IEEE, vol. 85, pages 1348–1363, 1997. (Cited on page 13.)

[Xing 2009] Junliang Xing, Haizhou Ai and Shihong Lao. *Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses.* In Proceedings of the 22nd Conference on Computer Vision and Pattern Recognition, CVPR, pages 1200 –1207. IEEE Computer Society, june 2009. (Cited on page 42.)

[Yang 2008] Nai-Chung Yang, Wei-Han Chang, Chung-Ming Kuo and Tsia-Hsing Li. *A fast MPEG-7 dominant color extraction with new similarity measure for image retrieval.* Journal of Visual Communication and Image Representation, vol. 19, no. 2, pages 92–105, 2008. (Cited on pages 50, 51, 52 and 109.)

[Yianilos 1993] Peter N. Yianilos. *Data structures and algorithms for nearest neighbor search in general metric spaces.* In Proceedings of the 4th annual ACM-SIAM Symposium on Discrete algorithms, SODA, pages 311–321. Society for Industrial and Applied Mathematics, 1993. (Cited on page 134.)

[Yu 2007] Yang Yu, David Harwood, Kyongil Yoon and Larry S. Davis. *Human appearance modeling for matching across video sequences.* Machine Vision and Applications, vol. 18, pages 139–149, 2007. (Cited on pages 20 and 33.)

[Zheng 2009] Wei-Shi Zheng, Shaogang Gong and Tao Xiang. *Associating Groups of People.* In Proceedings of the 20th British Machine Vision Conference, BMVC. BMVC Press, 2009. (Cited on pages 24, 25, 104, 125, 126 and 150.)

[Zheng 2011] Wei-Shi Zheng, Shaogang Gong and Tao Xiang. *Person re-identification by probabilistic relative distance comparison.* In Proceedings of the 24th Conference on Computer Vision and Pattern Recognition, CVPR, pages 649 –656. IEEE Computer Society, 2011. (Cited on pages 18, 23, 33, 108, 125, 126, 127, 129 and 131.)