Graz University of Technology

Institute for Computer Graphics and Vision

PhD Thesis

Combining Descriptive and Discriminative Information for Person Re-Identification

Martin Hirzer

Graz, Austria, May 2014

Thesis supervisors

Prof. Dr. Horst Bischof

Dr. François Brémond

Abstract

A central task in many visual surveillance scenarios is person re-identification, i.e., recognizing an individual person across a network of spatially disjoint cameras. This is a very hard task for human operators and even harder for automated systems due to several challenges such as changes in viewpoint, pose, and illumination. To cope with these difficulties, most existing methods either try to find a suitable description of a person's appearance or learn a discriminative model. Since these different representational strategies capture a large extent of complementary information, in this thesis, we propose to exploit both directions.

In particular, we first introduce an application-focused approach of integrating a descriptive and a discriminative person model into a single system. Given a specific query person, we initially run a fast, descriptive stage, where appearance is captured by a set of region covariance descriptors. This allows us to quickly provide a preliminary search result to a human operator. In a second stage, the operator can then refine the thus obtained result by applying a discriminatively learned person model, which is based on boosting for feature selection. In this way, we can take advantage of both, the time efficiency of the descriptive as well as the improved accuracy of the discriminative model.

The second part of this thesis is devoted to metric learning, a relatively new direction in the field of person re-identification. Although it provides a very elegant and mathematically principled fusion of descriptive and discriminative techniques, most existing metric learning approaches are not adapted to the task at hand and additionally suffer from high computational costs. Hence, in our work, we address these shortcomings and develop methods that are not only much more efficient, but also less prone to over-fitting, thus, enhancing their practical applicability in realistic, large-scale camera networks.

In order to demonstrate the benefits of our combined strategy, we present results on several publicly available benchmark datasets of different complexity. We show that having two complementary information cues capturing diverse aspects of a person's appearance is advantageous for the given problem, and that metric learning can achieve state-of-the-art or even better performance, however, requiring much less computational power compared to many other person re-identification approaches.

Keywords: inter-camera person re-identification, descriptive methods, discriminative methods, efficient metric learning.

Kurzfassung

Die Wiedererkennung von Personen in Kameranetzwerken zählt zu den Kernaufgaben vieler visueller Überwachungssysteme. Ausgehend von der Sichtung einer gesuchten Person in einem Kamerabild, sollen möglichst rasch sämtliche Erscheinungen derselben Person in weiteren Kameras des Netzwerkes gefunden werden. Sowohl für Menschen als auch für automatische Systeme stellt dies eine äußerst schwierige Aufgabe dar, da sich die Abbildung einer Person zwischen zwei verschiedenen Kameras sehr stark unterscheiden kann, beispielsweise aufgrund von Veränderungen im Blickwinkel, der Körperhaltung und der Beleuchtung. Viele der existierenden Systeme zur automatischen Personensuche setzen daher entweder auf eine deskriptive Strategie, versuchen also eine robuste, ganzheitliche Personenbeschreibung zu generieren, oder verfolgen einen diskriminativen Ansatz, um spezifische Details einer bestimmten Person zu extrahieren. Da diese beiden komplementären Richtungen ganz unterschiedliche Aspekte eines Personenbildes erfassen können, schlagen wir in dieser Dissertation vor, beide für die Personensuche zu verwenden.

Um dies zu erreichen, stellen wir zuerst einen anwendungsorientierten Ansatz vor, der beide Strategien in einem System vereint. Wird eine Person zur Suche ausgewählt, so beginnen wir mit einem schnellen, deskriptiven Suchverfahren, bei dem verschiedene visuelle Merkmale mit Hilfe einer Kovarianzbeschreibung erfasst werden. Dadurch ist es unserem System möglich, dem Benutzer sehr rasch ein erstes Suchergebnis zu präsentieren. Falls nötig, kann dieses Ergebnis in einem zweiten Schritt dann noch weiter verfeinert werden. Dazu verwenden wir ein auf Boosting basierendes, diskriminatives Suchverfahren. Bezogen auf das Gesamtsystem bedeutet diese zweistufige Vorgehensweise, dass wir sowohl die geringere Rechenzeit des deskriptiven, als auch die höhere Genauigkeit des diskriminativen Modells ausnutzen können.

Im zweiten Teil dieser Dissertation beschäftigen wir uns mit verschiedenen Metrik-Lernverfahren, einem relativ neuen Forschungsgebiet im Bereich der visuellen Personenwiedererkennung. Obwohl Metrik-Lernverfahren eine sehr elegante und mathematisch fundierte Verbindung von deskriptiven und diskriminativen Techniken erlauben, so sind die meisten vorhandenen Ansätze nicht an die speziellen Herausforderungen, die bei der Personensuche auftreten, angepasst und benötigen darüber hinaus noch eine hohe Rechenleistung. Um diese Einschränkungen zu beseitigen und damit die praktische Anwendbarkeit der Lernverfahren zu erhöhen, untersuchen wir in dieser Arbeit Methoden zum Lernen von Metriken die nicht nur sehr viel effizienter, sondern auch robuster sind als existierende Ansätze.

Im letzten Teil demonstrieren wir schließlich die Vorteile unserer kombinierten Strategie auf mehreren öffentlich zugänglichen Personendatenbanken unterschiedlicher Komplexität. Die Ergebnisse zeigen, dass sich die komplementären Aspekte, die von deskriptiven und diskriminativen Modellen beschrieben werden, äußerst nutzbringend miteinander verbinden lassen. Dies trifft im Besonderen auf Metrik-Lernverfahren zu, welche nicht nur hervorragende Resultate erzielen, sondern im Vergleich zu anderen Ansätzen auf dem Gebiet der visuellen Personenwiedererkennung auch um ein Vielfaches effizienter sind.

Schlüsselwörter: kameraübergreifende Personenwiedererkennung, deskriptive Methoden, diskriminative Methoden, effiziente Metrik-Lernverfahren.

Statutory Declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Place

Date

Signature

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Ort

Datum

Unterschrift

Acknowledgments

At this point, I would like to thank all those people who contributed to this thesis in various different ways. First of all, I would like to express my gratitude to my supervisor, Horst Bischof, for awakening my interest in the exciting field of computer vision and giving me the opportunity to do research in this area. He supported and guided me throughout the years with his knowledge while giving me the freedom to develop and follow my own ideas. Next, I want to thank my second supervisor, François Brémond, for his comments on this thesis.

Special thanks go to our Learning, Recognition, and Surveillance (LRS) group leader Peter M. Roth, the Boss, who had great influence on all of my research activities. I am very grateful for your continuous assistance during my work, your guidance and advices, the proof-reading of this thesis, and all the funny conversations we had inside and outside of computer vision. Thank you! Concerning research, I also want to thank all my co-authors for countless fruitful discussions, for providing me with new insights, and for supporting me in finishing our papers in time.

Moreover, I want to express my gratitude to all members of the Institute for Computer Graphics and Vision for making this a very pleasant and friendly working place, where social life is an essential part. In this context, I would like to particularly thank all members of the LRS group, Michael Donoser for organizing the reading group, the tech talks, and the awesome social events, the StripeMaster, the soccer team, and last but not least, the brave Mensa Men, who could not be "gnocched" down by whatsoever the "cooks" served.

Finally, I want to thank my family for giving me the opportunity to choose an educational career according to my liking, and for always supporting me during my studies. I would also like to thank my fellow students for the successful collaboration in many lectures, and my friends who always supported me.

This work has been supported by Siemens AG Österreich, Corporate Technology (CT T CEE), KiwiSecurity, the Austrian Research Promotion Agency (FFG) projects ECV (815105) under the COMET program, and MobiTrick (8258408) and Outlier (820923) under the FIT-IT program, and the project CultAR (601139) under the European Union's Seventh Framework Programme.

Contents

List of Figures

List of Tables

1 Introduction

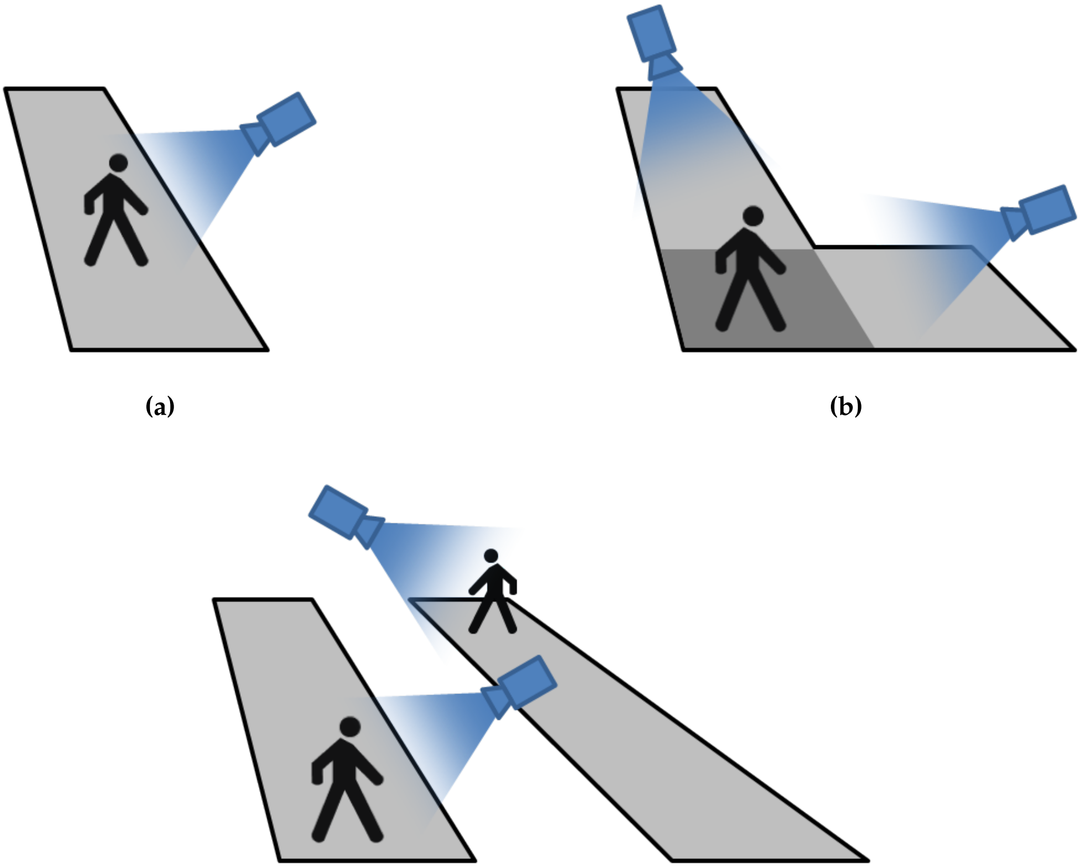
Contents

1.1 Problem Statement

Due to ceaseless advances in the research in semi-conductors, communication technologies, and image sensors, installations of camera networks nowadays are widespread. They occur in various domains ranging from rather small home surveillance applications in private areas, to medium-sized and large installations for monitoring public areas, e.g., shopping centers, airports, train stations, public transportation, sports centers, and so on. Since more and more public areas become subject to video surveillance, it becomes infeasible to analyze the ever growing amount of recorded video data. Hence, efficient, automatic systems are required in order to reduce the load on human operators.

This especially applies for person re-identification, a central task in many surveillance scenarios. It can be described as recognizing an individual in different locations across a network of non-overlapping cameras, thereby enabling person tracking within the entire observed scene. In contrast, most tracking approaches encountered in the field of computer vision are targeted at single-view applications and work on a frame-to-frame basis, which means that they are able to maintain the identity of a person only within the video feed of one camera. While a single camera might be sufficient for monitoring very small areas, additional cameras quickly become necessary as larger scenes have to be observed. If the dimensions and layout of the site allow deploying a camera network with enough overlapping fields of view, inter-camera tracking can be achieved by spatio-temporal reasoning. However, in most realistic settings, the required number of cameras and associated costs would be too high, so that the coverage is rather sparse, leaving "blind gaps". Since there is a significant uncertainty in the behavior of a person in a blind gap, taking advantage of space-time proximity is often impossible. Indeed, a person can leave one camera view, enter an uncovered area, and then re-appear in a completely different view after an unknown period of time. Establishing correspondences between person appearances in such conditions is the main target of person re-identification. In other words, re-identification extends tracking beyond blind gaps. An overview of the different setups is depicted in Figure 1.1.

In this thesis, we focus on person re-identification across blind gaps, i.e., the unconstrained setting. In particular, we address the task of selecting a person in one camera view and then recognizing the same person in any other view without easing the problem by spatio-temporal reasoning based on the scene layout. Furthermore, we assume that persons have already been detected in all camera views, i.e., we do not tackle the detection problem. Figure 1.2 shows a typical use case, where a human operator selects a specific person in one camera view, and the system should automatically find the same person in another view. Usually, the images of the query persons are referred to as *probe images*, and the images searched trough are termed *gallery images* [30]. Depending on the setup, we also have to differentiate between systems that only provide a single image per person and camera view (*single-shot*) and systems where multiple images per person are available (*multi-shot*). It is obvious that the latter type provides richer information, since each person is captured in several different poses. But on the other hand, such systems require more sophisticated algorithms to be able to exploit the additional information, so that the runtime is considerably higher than in the single-shot case.



(a)

(b)

(c)

Figure 1.1: From single-view tracking to person re-identification: (a) Single-view tracking algorithms typically maintain the identity of an individual from one video frame to the next. (b) If possible, spatio-temporal constraints can facilitate the task of establishing correspondences between camera views. (c) In the general case of person re-identification, no assumptions about the behavior of a person in unobserved areas can be made.

1.2 Applications

As already mentioned, person re-identification plays a crucial role in many surveillance scenarios. Being able to automatically browse hundreds of camera views for a specific person can be very helpful in many different situations and leads to a number of useful applications. Examples include specific re-identification scenarios such as tracking suspects and criminals over multiple cameras or finding children who lost their parents, but also anonymous tasks such as analyzing crowd movements in public places or gath-

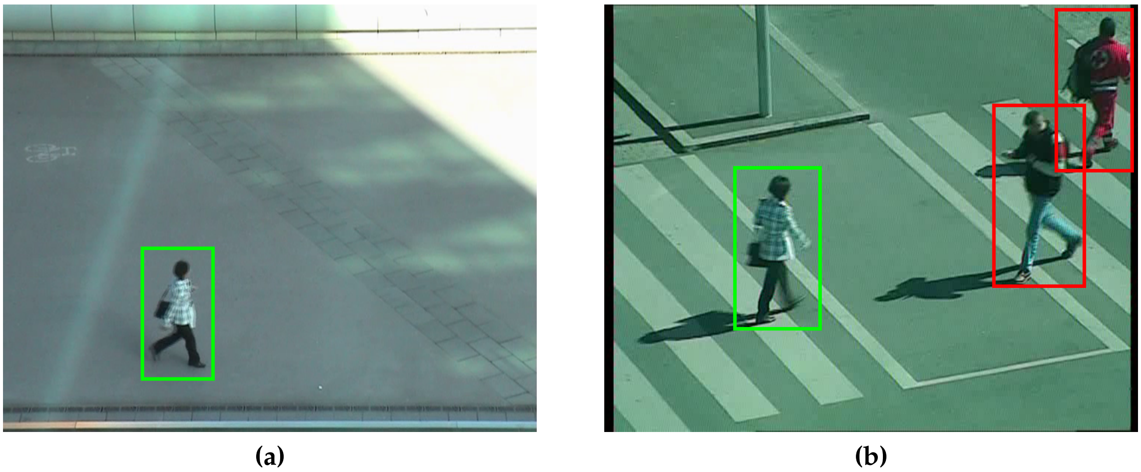


Figure 1.2: An exemplary person re-identification task: (a) Starting from bounding boxes returned by some person detector, a certain individual is selected in one camera view, yielding the probe image. (b) The same person should then be automatically identified in another view containing the gallery images. Matching instances are indicated by green bounding boxes, non-matching ones by red bounding boxes.

ering information about the flow of customers in shops and shopping centers, which is achieved by identifying individual persons. These latter examples show that person re-identification can also be used to gather information about the behavior of people, thus, facilitating scene understanding, the ultimate goal of any surveillance system.

While basically being applicable to all setups that involve more than one camera, person re-identification is especially targeted at large-scale installations consisting of dozens of cameras. Monitoring a high number of different cameras can quickly over-challenge a human operator and lead to lack in concentration, tiredness, and faults. To prevent this, the workload has to be shifted to (semi-) automatic re-identification systems. Even if such systems are sometimes not able to recognize the correct person with absolute confidence, providing a list with the most likely matches is often entirely sufficient to significantly reduce the effort for a user. This is the main purpose of most person re-identification systems: assisting human operators in recognizing individual persons throughout an entire camera network, so that they can efficiently and reliably perform critical security tasks.

1.3 Challenges

Although person re-identification appears to be a simple task at first glance, in fact it is very challenging for several reasons, even for humans. As already mentioned, in most practical scenarios there are no or only weak spatial and temporal constraints that can ease the task of finding a certain person in a network of cameras, and the gaps between camera views are quite large. Furthermore, since images of typical surveillance cameras have rather low resolution and are taken from long distances, biometric information such as a person's face or gait cannot be used either. This leaves the appearance of a person, more precisely the appearance of the clothing, as the only reasonable cue that can be exploited. In this context, person re-identification is often also termed appearance-based person re-identification in order to better distinguish it from strategies that, for instance, exploit space-time proximity or biometric cues. However, the appearance of an individual can vary extremely across a network of cameras, mainly due to the following difficulties:

- Changes in viewpoint
- Changes in pose
- Different illumination
- Different camera characteristics
- Shadows
- (Self-) Occlusions
- Loose or wrinkled clothes

Some of these challenges are pre-determined by the layout of the camera network, e.g., variations in viewpoint and camera characteristics, others can be addressed on the algorithmic level. For example, the choice of an appropriate feature space or normalization of feature responses can help in alleviating effects that arise from illumination changes and shadows. Furthermore, if available, a multi-shot setup can be used to gain some robustness to pose variations, occlusions, and appearance changes caused by loose or wrinkled clothes. Another important problem typically encountered is the potentially high number of similar persons that can occur in busy public places (e.g., many people wear rather dark clothes in winter). All these challenges make the task of person re-identification very difficult. Actually, it might even happen that different persons look very similar, while the same person can have very diverse appearances across a network of cameras (e.g., due to camera noise, scene geometry, or different viewpoints: frontal versus back view), rendering the task somewhat ill-posed.

1.4 Strategies

Motivated by the high number of practical applications and still unresolved problems, there has been a considerable scientific interest in person re-identification within the last years. Due to the characteristics of the problem mentioned before, most of the existing works concentrate on appearance-based methods. Under the assumption that people do not change their clothes between different sightings in a network, these approaches extract information from person images in order to generate human signatures, which are then matched between individual camera views.

In general, appearance-based techniques can be roughly divided into two groups, namely descriptive and discriminative methods. Descriptive methods first extract visual features that should be both, distinctive and stable under changing viewing conditions between different cameras, and then use a standard distance measure for matching. On the other hand, discriminative methods take advantage of class labels to exploit the discriminative information given by the data. Further details about the descriptive and the discriminative strategy are presented in Sections 1.4.1 and 1.4.2, respectively. Finally, a comparison between both is drawn in Section 1.4.3.

1.4.1 Descriptive Methods

Descriptive approaches tackle the person re-identification problem by seeking a very distinctive and at the same time stable feature representation for describing a person's appearance under changing viewing conditions between cameras. Usually, they rely on hand-crafted features extracted from pre-defined image regions, e.g., upper and lower body regions, horizontal stripes, or some other spatial layout related to the human body (see, e.g., [28, 7, 11]). Since descriptive methods do not know a priori which features are important and which features should be discarded, they produce a rather holistic feature representation capturing the whole person. After feature extraction, a standard distance measure such as L^1-norm [79], L^2-norm [38, 39], or Bhattacharyya distance [18] is applied in order to compare person representations. However, due to the large intra-class appearance variations typically encountered in person re-identification, the computation of features that are both, distinctive and stable under realistic conditions, is very difficult if not impossible at all. Another drawback is the application of a standard distance measure in the matching step, as it treats all features equally, regardless of the importance and reliability of individual features in certain situations.

1.4.2 Discriminative Methods

Since descriptive approaches often are not distinctive enough and rely on hand-crafted features, other methods aim at learning discriminative models. In general, their goal is to find those features out of a set of possible features that are best suited to discriminate a specific person from the remaining people. Usually, this is achieved using some kind of machine learning algorithm, e.g., boosting [31, 6], support vector machines [64], or subspace learning [70]. Hence, in contrast to descriptive methods, discriminative approaches require a training stage and labeled training data before they can be applied to person matching. Another differentiation is the kind of features that is typically generated. Unlike the holistic feature representation produced by descriptive approaches, discriminative methods select only those features that are really discriminative, i.e., they produce a representation consisting of rather local features.

1.4.3 Descriptive versus Discriminative Methdods

From the description of the two strategies it is quite obvious that both have their advantages and disadvantages. Descriptive methods, for example, can be considered rather robust due to the used holistic feature representation. However, they fail in capturing fine details of a person's appearance, e.g., a red bag, a blue scarf, and so on. Such details are usually underrepresented, despite the fact that they often contain very valuable information to discriminate a specific person. As a result, descriptive methods are not distinctive enough in situations when we have to deal with a lot of similar people. On the other hand, discriminative methods are well suited to capture local appearance features with high saliency. This seems very reasonable at first glance, but involves the danger of focusing too much on these details and ignoring the remaining appearance. For example, a red bag visible in one camera view would be very discriminative, however, if it is occluded in the other view it becomes impossible to find the specific person again. Models that just focus on the red bag would fail in this case. In general, special

Descriptive Methods	Discriminative Methods
hand-crafted features	learned features
holistic feature representation	local feature representation
captures coarse appearance	captures fine details
matching using standard distance measure	matching using learned model
robust	error-prone, over-fitting
low computational effort	high computational effort

Table 1.1: Typical characteristics of descriptive and discriminative methods.

care has to be taken in order to prevent the learned model from over-fitting the training data. Another important aspect is the usually higher computational effort of discriminative methods as a consequence of the additional model learning phase. The typical characteristics of both strategies are outlined in Table 1.1.

1.5 Contributions

As described in the previous sections, descriptive and discriminative methods tackle the task of person re-identification from two different directions, each having its advantages and disadvantages. While descriptive approaches usually run quite fast and robust, they can easily be outperformed by discriminative models when it comes to capturing fine details, however, at the cost of increased computational complexity. Thus, to benefit from these complementary representational strategies and to get the best of both, we propose to combine them for person re-identification, which is the main topic of this thesis. We show how to exploit the advantages of both strategies, but avoid their respective drawbacks, leading to an improved re-identification performance. In particular, this can be achieved by either applying two separate models and coupling their results in an appropriate way, or using an approach that inherently merges descriptive and discriminative techniques, such as metric learning.

1.5.1 Combination of Descriptive and Discriminative Models

The first contribution we present in this thesis is an application-focused combination of descriptive and discriminative techniques for person re-identification. Specifically, our goal is to obtain a system that, on the one hand, runs very efficiently and has a low response time, but on the other hand, also benefits from the generally higher performance of discriminative methods. Hence, to be able to rapidly browse through a set of gallery images for a specific person, we first apply a fast, descriptive model, allowing our system to return an initial search result with low latency. If necessary, the thus obtained result can then be refined using a more sophisticated, discriminative model in a second step. This way, not only can distinctive persons be found very quickly, but harder cases are also manageable, with additional effort, though.

1.5.2 Metric Learning

Another possibility to combine the two complementary strategies is metric learning, which can be seen as a midway between descriptive and discriminative approaches. The idea is similar to descriptive methods, i.e., the data is modeled by a set of descriptive features specifically designed for person re-identification. However, instead of matching them directly in the feature space using some standard distance measure, a discriminative metric is learned from labeled samples, typically originating from different camera views. Under the estimated metric, the distance between image descriptors of the same person should become very small, while it should become large between image descriptors that stem from different persons. Hence, the obtained metric emphasizes those feature directions that are best suited to discriminate matching persons from non-matching ones. Since the learned metric inherently describes the transition in the feature space between two camera views, such approaches are very well suited for real world scenarios.

Moreover, a key advantage of metric learning approaches is their computational efficiency during evaluation once the metric has been learned. However, as a drawback, calculating the metric typically involves solving complex optimization problems (see, e.g., [32, 15, 80, 81, 16]), which severely limits the practical applicability of existing methods. Considering the fact that usually one metric is learned for each camera transition, i.e., each camera pair, and that real world setups consist of tens or hundreds of cameras, also the training time is highly relevant. Thus, investigating more efficient algorithms to learn suitable metrics is the second contribution of this thesis.

1.6 Outline

First, in Chapter 2, we give an overview of existing works in the field of person re-identification, organized according to the strategy they use to tackle the task. Most methods follow one of the two prevalent directions, i.e., they either try to find a very distinctive and at the same time stable feature representation (Section 2.2) or aim at learning a discriminative model (Section 2.3). However, in recent years, metric learning has emerged as a midway between these two complementary strategies. Since this is a relatively new direction in the field of person re-identification, only few such methods exist so far, which are described in Section 2.4. Finally, in order to provide a broader view, we also discuss some approaches that do not directly address the person re-identification task, but can ease the problem (Section 2.5).

Next, in Chapter 3, we introduce our combined descriptive and discriminative person re-identification system, where we take advantage of both representational strategies, but avoid their respective drawbacks to some extent. Specifically, we first run a fast, descriptive method, which is described in Section 3.3. If necessary, the thus obtained result can then be refined using a discriminatively learned model, as presented in Section 3.4.

Chapter 4 is devoted to metric learning, which is a recent trend in the field of person re-identification. We start by introducing the general idea of Mahalanobis metric learning in Section 4.2, where we also summarize related metric learning approaches that are later used for comparison in our study. As most of these approaches suffer from high computational costs, we then present our more efficient algorithms in Section 4.3, with the objective to enhance the practical usability of such methods in large-scale camera networks. Additionally, in Section 4.4, we describe our modular, three-stage metric learning framework, which allows us to easily run all presented algorithms in a plug-and-play manner.

In Chapter 5, we perform thorough evaluations of the methods presented in this thesis. First, we introduce the performance measure typically used to analyze the matching capability of person re-identification methods (Section 5.2). Next, in Section 5.3, the employed benchmark datasets and their individual characteristics are described in detail, and then, in Section 5.4, the results obtained using our person re-identification system as well as metric learning are presented and discussed.

Finally, in Chapter 6, we summarize the thesis and make some concluding remarks. Furthermore, we provide an outlook to possible future work.

2 Related Work

Contents

2.1 Overview

In recent years, there has been a considerable scientific interest in person re-identification. On the one hand, there are numerous practical applications in visual surveillance, but on the other hand, existing systems still suffer from unresolved problems. These circumstances led to an increased number of scientific works proposing different strategies to tackle the task of person re-identification. Approaches range from simple template matching techniques using hand-crafted feature representations to more sophisticated methods that fit structural models on person images or extract discriminative information using labeled training data.

In literature, one can find many different taxonomies that try to partition the field of person re-identification in a meaningful way. Some of them classify methods according to the number of person images they use, i.e., single-shot or multi-shot, others differentiate between holistic and part-based feature representations. And still others characterize methods as either following a direct, descriptive strategy or using a learning-based, discriminative strategy. In this thesis, we follow the latter criterion, since it provides the most fundamental and complementary classification.

Related work that falls into the descriptive category is presented in Section 2.2, while approaches following the discriminative direction are outlined in Section 2.3. Section 2.4 is devoted to methods that apply metric learning, a research direction that came up quite recently in person re-identification, and Section 2.5 describes additional works that cannot be entirely classified using the aforementioned categories, e.g., methods that exploit spatial and temporal constraints or approaches that include contextual information. Finally, a discussion of the presented methods is given in Section 2.6.

2.2 Descriptive Methods

In this section, person re-identification methods that follow the descriptive strategy, as introduced in Section 1.4.1, are presented. Their goal is to compute a very distinctive and at the same time stable feature representation capable of describing a person's appearance under changing viewing conditions between different cameras. In most cases, hand-crafted features are extracted from pre-defined image regions covering the whole human body. Since people are seen from different viewpoints and have different poses when moving through a network of cameras, establishing correct feature correspondences between different sightings is a core issue addressed by the majority of these works.

2.2.1 Graph-Based Matching

Gheissari et al. [28] treat person re-identification as a two-stage problem. First, they try to establish correspondences between body parts in foreground segmented images to cope with pose variations. They propose two ways of finding such correspondences, an interest point detector and a decomposable triangulated graph [2, 19] modeling body articulations. Compared to a baseline method that does not use any correspondence information, the improvement achieved with the interest point operator is negligible. However, fitting a decomposable triangulated graph on each person image leads to a significant gain in performance. The second stage is the generation of signatures that are invariant to appearance variations between cameras, especially those that stem from loose or wrinkled clothing. Gheissari et al. combine color and structural information using modified hue [72] and saturation as well as salient edgel histograms. The latter are generated using a spatio-temporal segmentation algorithm that rejects temporally unstable edges, i.e., spurious edges typically caused by clothing. Finally, signatures from corresponding body regions are compared using histogram intersection [72] and an overall matching distance is computed. Unfortunately, the approach of Gheissari et al. is limited to people seen from similar viewpoints, an assumption that is violated in most realistic setups.

2.2.2 Shape and Appearance Context

Wang et al. [79] introduce the concept of shape and appearance context by modeling the spatial distribution of appearance relative to body regions. In particular, they densely compute Histograms of Oriented Gradients (HOGs) [14] in the log-RGB color space [27], which are then aggregated in order to build an appearance model. Wang et al. propose a model that relies on a pre-defined appearance dictionary. This is similar to a bag-of-features approach, however, instead of computing histograms of dictionary labels, they seek a more distinctive description that also incorporates the distribution of the labels in space. Hence, following the idea of shape context [5, 59], they additionally build a shape dictionary besides the one for appearances. This allows them to generate two labeled person images, one segmenting the image into regions according to their appearance, and one providing regions that are loosely associated with specific body parts or background areas. The final image description is obtained by computing the co-occurrence matrix over these two labeled images. Descriptor matching is done using the L^1-norm distance. Wang et al. show that the additional information coming from the distribution of colors leads to an improved performance compared to a simple bag-of-features model. However, like the method presented in [28], their approach works well only if people are seen from very similar viewpoints, a restriction that makes it infeasible in most practical scenarios.

2.2.3 Spatial Covariance Regions

Bak et al. [7] seek for efficient features that can reliably build human signatures under different camera viewpoints. They propose an approach using spatial covariance regions of human body parts. First, person images are automatically extracted from video data by applying a HOG-based detector adapted from the face detection technique in [13]. The same detector is then trained on individual body parts in order to localize the top, torso, legs, and the left and right arm of a person. To build a human signature, Bak et al. use the idea of spatial pyramid matching. They apply a three-level pyramid representing the full body region, the body parts, and subregions inside the body parts. A covariance matrix capturing pixel location, color, and gradient information is computed for each of these regions. As similarity measure between corresponding regions the distance defined in [24] is used. Finally, matching of different signatures is done by combining the individual distances in a weighted sum.

2.2.4 Perceptual Principles of Symmetry and Asymmetry

In order to gain some robustness to pose and viewpoint variations, Farenzena et al. [18] present an appearance-based approach that exploits perceptual principles of symmetry and asymmetry when extracting features. In particular, their method consists of three phases. First, the silhouette of a person is obtained by either using Structure Element (STEL) component analysis [45] for the single-shot case, or applying background subtraction in case of video sequences. After removing the background, two horizontal axes of asymmetry are found by comparing the local foreground appearance and area above and below the axes. Typically, these axes are placed such that the person is approximately divided into a head, an upper body, and a lower body region. The head region is discarded since it usually contains only few pixels. For each of the remaining two regions, a vertical axis of symmetry is estimated, again, by comparing the appearance and area of the local neighborhood around the axis. Thus, four body parts are roughly captured: left and right upper body, and left and right lower body. Since the extraction of these parts is based on the visual and positional information of the clothes, they are robust to pose and viewpoint variations to some extent. In the second phase of the approach, three kinds of features capturing complementary aspects of the human appearance are extracted separately from each body part. The chromatic content of each part is captured by a weighted HSV color histogram. The weighting is applied in order to emphasize pixels near the vertical symmetry axes. Furthermore, Maximally Stable Color Regions (MSCRs) [23] are extracted, and their statistical properties are added to the feature description. Hence, not only the chromatic content, but also the spatial distribution of colors is included. The third feature consists of Recurrent Highly Structured Patches (RHSPs), i.e., local patches with salient texture information that are highly recurrent. As for the HSV color histograms, when extracting MSCRs and RHSPs the focus lies on regions near the vertical symmetry axes. The final phase is the feature matching between two person instances, where the three specific feature distances are weighted and unified to a joint matching distance.

2.2.5 Pictorial Structures

Cheng et al. [11] apply pictorial structures [20] for the task of person re-identification. They follow the framework of [3], i.e., general part detectors are used to localize body parts, and prior knowledge about the human body structure is incorporated by a kinematic tree. This allows them to estimate the pose of a person, thereby gaining robustness to pose variations to some extent. Specifically, Cheng et al. fit a body configuration composed of head, chest, thighs, and legs on pedestrian images and extract per-part color information as well as the color displacement within the whole body. The color content is extracted from each part independently via HSV histograms, which are then concatenated to a single feature vector. For describing the color displacement, the MSCR operator proposed in [23] is used, extracting MSCR blobs from within the whole pictorial structure body mask. Statistical properties of these blobs then serve as description of the color distribution. Similar to [18], the color histograms and MSCR statistics are finally combined to a single joint distance measure quantifying the similarity of two individuals. If multiple images of a single person are available, Cheng et al. propose to customize the fit of the pictorial structure on that specific person, resulting in what they call custom pictorial structures. The idea is to exploit the additional visual information through spatio-temporal reasoning, leading to an improved fitting and consequently to higher re-identification performance.

2.3 Discriminative Methods

As described in Section 1.4.2, discriminative approaches take advantage of class labels in order to exploit the discriminative information given by the data. The goal is to generate a feature representation that is best suited for distinguishing a specific person from the remaining people. To achieve this, the methods presented in this section use different strategies, e.g., directly selecting discriminative features out of a set of possible features, using class-aware dimensionality reduction, applying pairwise dissimilarity profiles in nearest neighbor classification, or relative ranking using support vector machines.

2.3.1 Boosting for Feature Selection

Gray and Tao [31] use an ensemble of discriminative, localized features in order to describe pedestrians for person re-identification. Instead of designing specific features by hand to solve the matching problem, they just define a suitable feature space using intuition, and then let a machine learning algorithm select the most relevant out of all possible features. In particular, Gray and Tao use 8 color channels (RGB, HS, YCbCr) and 19 texture channels (Gabor [22] and Schmid [69]) for feature extraction. A single feature is defined by three elements: a feature channel, an image region, and a histogram bin, i.e., a value range. To reduce the overall search space during feature selection, regions are restricted to horizontal stripes. For every instance of a pedestrian, the proposed feature definition provides a probability of a pixel from the specified channel and region being in the specified value range. To be able to compare different instances, there is a likelihood ratio test associated with each of the features. This test is performed on the absolute difference between corresponding feature values. Specifically, given labeled training data, the distribution of feature differences is estimated for matching and non-matching image pairs, e.g., considering Exponential, Gamma, or Gaussian parametrization, and a likelihood ratio test is computed. To find good features and model parameters, Gray and Tao apply AdaBoost [26]. In every iteration, the algorithm selects the most suited feature and model parameters based on the current distribution of positive, i.e., matching, and negative, i.e., non-matching, person examples. The final output of the proposed method is a similarity function consisting of an ensemble of likelihood ratio tests, each corresponding to a specific, local feature.

Another work that also relies on AdaBoost is presented by Bak et al. [6]. First, people are detected and tracked in video data using a modified version of the HOG-based technique presented in [13]. This results in a set of accumulated images for each individual. In the next step, a color-based foreground-background separation method [8] is applied in order to remove the influence of the background region. The segmented images are then used in a boosting step. Similar to [31], Bak et al. employ AdaBoost in order to find the most discriminative features that separate one person from the rest of the detected people. Hence, for each individual, a classifier has to be trained. Images of the corresponding person serve as positive training samples, images from other people build the set of negative samples. Finally, the resulting feature representation provides a human signature that can be matched utilizing a distance similarity function. Concerning the type of features, Bak et al. propose using Haar-like features and MPEG-7 dominant color descriptors [83]. To enhance discriminability, the latter are extracted separately from the upper and lower body region.

2.3.2 Partial Least Squares

To reduce ambiguities often encountered in person re-identification, Schwartz and Davis [70] propose using a rich set of features extracted from overlapping image blocks. Besides commonly used color histograms, they also incorporate co-occurrence matrices capturing texture information [34], and HOGs representing local shape. Since the extraction of such rich descriptors from overlapping image blocks produces extremely high-dimensional feature vectors, some sort of dimensionality reduction is needed. For this task, Schwartz and Davis propose using Partial Least Squares (PLS) regression [82, 66], which, in contrast to the commonly used Principal Component Analysis (PCA) [62, 37], considers class information during reduction. In this way, a low-dimensional, discriminative subspace is created for every person in the training set using a one-against-all scheme. This can be seen as a feature selection procedure. Each subspace emphasizes those image features that are best suited to distinguish the corresponding person from the rest of the training set. After estimating a PLS model for all training appearances, each training sample is projected onto its own subspace. The resulting low-dimensional feature representations are then used in the matching stage. During testing, a query sample is projected onto each of the previously estimated subspaces and classified as belonging to the same person as the training sample having the smallest Euclidean distance. However, a drawback of this method is that it is based on the given data, i.e., it has to be re-computed if new samples are added.

2.3.3 Pairwise Dissimilarity Profiles

Lin and Davis [52] propose to learn pairwise dissimilarity profiles that can be applied to nearest neighbor classification. They formulate person re-identification as a multi-class classification problem. Each person is considered to be a separate class containing his or her appearances as class samples. The idea now is to combine the direct distance measure typically used in nearest neighbor classification with a dissimilarity distance measure which also considers relations between different classes. This newly proposed distance is based on the Kullback-Leibler divergence between appearances that are modeled as kernel probability density functions in a joint 4D color-height space, i.e., 3D color vector plus vertical image coordinate. The log-likelihood ratio function used in the distance calculation quantitatively reflects discriminating features between two appearances. Lin and Davis assume that such pairwise profiles are very similar between all appearance variations of two specific individuals. Hence, the profiles model those properties that are very discriminative between two persons, but ignore intra-class variations caused by pose, viewpoint, and illumination changes to some extent. In the matching stage, a modified nearest neighbor classification scheme is used to assign person identities to query samples. While conventional nearest neighbor classification calculates only the direct distance between query sample and stored prototypes, Lin and Davis additionally incorporate an indirect, discriminative distance computed by comparing pairwise profiles. This means that they do not only consider information between query and prototypes, but also inter-relations between different training samples. However, similar to [70], the method is based on the persons to re-identify and has to be re-computed if new samples are added.

2.3.4 SVM-Based Relative Ranking

Inspired by document retrieval concepts, Prosser et al. [64] treat person re-identification as a relative ranking problem rather than an absolute scoring problem of correct versus incorrect matches. Their goal is to learn a subspace where the potential true match is given the top rank. In general, person re-identification typically suffers from largely overlapping feature distributions of different persons in a multi-dimensional feature space. To overcome this problem, Prosser et al. use a Support Vector Machine (SVM) [12] based ranking method, where features become more separable once mapped into a higher-dimensional feature space via the kernel trick. However, existing methods such as Ranking SVM (RankSVM) [44] do not scale very well on large datasets due to high computational costs and memory requirements. While computational costs can be alleviated by using Primal RankSVM (PRSVM) [10], a sped up version of the original RankSVM algorithm, the high memory consumption calls for a more advanced organization of the data. To address this issue, Prosser et al. apply ensemble learning, i.e., they train weak PRSVMs on small subsets of the data and then combine them to form a strong ranker using a boosting principle. In this way, memory costs can be significantly reduced without loss in performance. Another advantage compared to the original formulation is the automatic selection of optimal model parameters by the ensemble learning scheme, rendering time consuming cross validation in the training stage unnecessary. Similar to [31], Prosser et al. use 8 color channels (RGB, HS and YCbCr) and 21 texture filters (Gabor [22] and Schmid [69]) applied to the luminance channel as feature representation. To take the spatial layout into account, feature histograms are extracted from six equally-sized horizontal stripe regions, roughly capturing the head, upper and lower torso, and upper and lower legs of a person.

2.4 Metric Learning

Metric learning represents a midway between descriptive and discriminative approaches. As mentioned in Section 1.5.2, the appearance of a person is modeled by a set of descriptive features specifically designed for the task of person re-identification. However, instead of matching them directly in the feature space using some standard distance measure, an appropriate metric is learned, emphasizing discriminative feature directions, i.e., those directions that are best suited to distinguish matching persons from non-matching ones. Since metric learning for person re-identification came up quite recently, only few works follow this direction so far. Hence, there are still open issues that need to be addressed, especially concerning the complex optimization schemes usually involved. At this point, three related works are presented, one estimating a metric by optimizing over local neighborhoods in the feature space, one using a probabilistic distance learning formulation, and one computing a projection into a low-dimensional space.

2.4.1 Local Neighborhood-Based Metric

In order to tackle the person re-identification problem, Dikmen et al. [16] learn a Mahalanobis metric by optimizing over local neighborhoods in the feature space using the Large Margin Nearest Neighbor (LMNN) method presented by Weinberger et al. [80, 81]. Specifically, the goal of LMNN is to minimize the distance between each training point and its k nearest neighbors that are equally labeled (target neighbors). At the same time, points having a different label, but which are closer than the aforementioned neighbors plus a constant margin (impostors), should be pushed away. This leads to a metric that optimizes distances in local neighborhoods, i.e., is optimal for k nearest neighbors classification. Based on the iterative framework of Weinberger et al., Dikmen et al. additionally introduce a rejection option, which enables the LMNN classifier to return no matches if all nearest neighbors are beyond a certain distance, thereby indicating that the searched person does not appear in the selected scene. In order to be able to define a universal rejection threshold, Dikmen et al. replace the original distance criterion that defines impostors locally, i.e., depending on the neighboring training points, by a global average. Further details about the learned metric can be found in Section 4.2.6. As features, they employ simple RGB and HSV color histograms, which are extracted from overlapping, rectangular regions and concatenated to a single feature vector per image. To be applicable to the metric learning framework, the dimensionality of these feature vectors is reduced via PCA.

2.4.2 Probabilistic Relative Distance Comparison

Zheng et al. [87] also use metric learning, but formulate it in a probabilistic manner. While conventional distance learning usually tries to minimize intra-class variability and maximize inter-class variability in an absolute sense, their probabilistic approach is motivated by the nature of the person re-identification task. Typically, one has to deal with large intra-class and inter-class variability as well as only very limited number of training samples. These characteristics can easily lead to over-fitting if a model is learned by minimizing intra-class distances and maximizing inter-class distances with brute force. In contrast, Zheng et al. propose a relative distance comparison model similar to the ranking approach presented in [64]. However, unlike [64], they formulate it in a probabilistic manner, making it more tolerant to the challenges encountered in person re-identification. In particular, Zheng et al. seek a Mahalanobis distance that maximizes the probability of a matching pair having a smaller distance than a non-matching pair. This is achieved using an iterative optimization algorithm. More details about the learned distance are given in Section 4.2.7. For describing the appearance of a person, Zheng et al. follow the feature setup used in [31]. A mixture of color (RGB, HSV and YCbCr) and texture (Gabor [22] and Schmid [69]) histograms is extracted from six horizontal stripe regions.

2.4.3 Pairwise Constrained Component Analysis

Mignon and Jurie [58] present Pairwise Constrained Component Analysis (PCCA), which, in contrast to most other metric learning approaches, can cope with high-dimensional input data and small training sets. In particular, the proposed algorithm learns a projection into a low-dimensional space from sparse, pairwise distance constraints imposed on the training samples. In the thus obtained space, the Euclidean distance between matching person pairs should become smaller than a certain threshold, while it should become larger than that threshold for non-matching pairs. To compute such a mapping, an objective function penalizing training pairs which do not meet these requirements is minimized, as described in more detail in Section 4.2.8. Similar to [31], as image representation, Mignon and Jurie use a mixture of color (RGB, HSV and YCbCr) and texture (Local Binary Pattern (LBP) [60]) histograms extracted from six horizontal stripe regions.

2.5 Other Methods

This section describes approaches that either do not directly tackle the person re-identification problem, but instead aim at solving a related task, or do not follow our strategy, i.e., appearance-based person re-identification, but instead exploit other information cues, such as spatial and temporal constraints, or visual context information coming from surrounding people. However, as already mentioned, some of these alternative strategies are based on assumptions hardly met in realistic, large-scale environments, so that they are only of limited practical use.

2.5.1 Spatio-Temporal Modeling

Rahimi et al. [65] describe a method for simultaneous calibration and tracking with a network of non-overlapping cameras. Although not specifically targeted at the person re-identification task, since object correspondences between cameras are assumed to be known, their approach can be used to automatically determine the spatial layout of a camera network. In particular, using a dynamics model and sporadic observations from non-overlapping cameras, they show how to concurrently calibrate the network and track a person. In contrast to previous works based on simple velocity extrapolation, i.e., linear motion, Rahimi et al. apply a Gaussian Markov chain in order to model a person's trajectory in areas that are not covered by any camera, allowing to also handle non-linear paths and speed changes. The final solution is obtained via a joint maximum a posteriori estimation over the person's trajectories and the pose parameters of the cameras.

A quite similar work is presented by Makris et al. [57], who aim at automatically constructing a probabilistic, tempo-topographical model of a camera network from a large set of observed persons moving around the scene. Entry and exit zones of individual camera views are represented as nodes, and the links between them describe the corresponding transition probabilities. In order to extend their method beyond approaches that are able to combine person tracks only in overlapping fields of view, Makris et al. also incorporate virtual nodes modeling the behavior of people in uncovered areas between cameras. For example, a person can disappear from one camera view, enter an unobserved region, and then re-appear in a second view after some time. However, a person can also leave the scene within the unobserved area and never appear in the second camera view, or a new person that has not passed the first view might directly enter the second view from the unwatched region. An advantage of this method over [65] is that it works fully unsupervised, i.e., it does not require labeling of track correspondences between cameras. Hence, the system can be applied in a plug-and-play manner to automatically learn a probabilistic model of the network layout, which can then be utilized to facilitate other security tasks such as person re-identification.

While the works of Rahimi et al. [65] and Makris et al. [57] do not use any appearance cues, Javed et al. [42] demonstrate that this kind of information can supplement spatio-temporal approaches. Since the appearance of a person can vary significantly when viewed from different cameras, Javed et al. apply brightness transfer functions (BTFs) in order to model these visual changes. Thus, they follow the same idea as metric learning, i.e., estimating the transition between cameras in the feature space. This appearance mapping from one view to the other is not unique and depends on various parameters, e.g., current illumination, possibly dynamic scene geometry, exposure time, and so on. Nevertheless, the authors show that all BTFs for a given pair of cameras lie in one low-dimensional subspace. During a training phase, the subspace for each pair of cameras is learned from known person correspondences using probabilistic PCA [74]. Once learned, this allows estimating the probability that the mapping between appearances observed in two different camera views lies in the corresponding subspace, i.e., the observations belong to the same person. Finally, the proposed appearance-based matching scheme is combined with space-time features such as location, time, and velocity. For this purpose, a maximum a posteriori framework is applied, where the posterior probability is maximized using a graph-theoretic approach [41]. As appearance features, Javed et al. extract normalized histograms of brightness values, which provide some robustness to changes in pose [72].

2.5.2 Visual Group Context

Zheng et al. [86] exploit visual context information coming from surrounding people in order to enhance person re-identification in crowded public spaces. Their approach is based on the assumption that people often walk in groups, either with people they know or strangers. At first glance, associating groups of people over a network of cameras may seem to be easy compared to common person re-identification, since typically more and richer visual content is given. In addition, a group as a whole is usually less affected by occlusions than individual persons. However, a severe challenge is the highly non-rigid structure of a group, as people are likely to constantly change their positions relative to each other. Moreover, in contrast to a single upright person, also the aspect ratio of a group's shape can vary significantly. To address these problems, Zheng et al. utilize a novel people group representation. After background subtraction, they compute local color (mean RGB color vector) and gradient information (scale-invariant feature transform [54] vectors extracted from each RGB channel separately) at every foreground pixel of the group image, and concatenate them to one feature vector per pixel. All extracted feature vectors are then quantized into clusters by k-means clustering to build a code book of visual words. Hence, a group of people can be represented by the distribution of visual words contained in the image, e.g., modeled by a histogram. However, such a representation would not be very distinctive due to the lack of spatial information. Taking into account the special characteristics of group images, i.e., global spatial relations between people can be highly unstable, while local spatial relations between small patches may be stable, Zheng et al. propose two descriptors. The first one is extracted from rectangular ring regions, thus, gaining some invariance to rotations. In particular, it aims at describing the ratio information of visual words within and between different ring regions. The second descriptor also captures spatial relationships between visual words, but on a smaller scale, i.e., based on local image blocks. As a consequence, the described spatial relations are much simpler, but also more robust to illumination changes, occlusions, and so on. Finally, the overall group representation is obtained by combination of both descriptors, and matching is performed based on L^1-norm.

2.6 Discussion

Most of the works described in this chapter fall either into the descriptive or discriminative category. Considering descriptive methods, common problems are their limited distinctiveness or dependence on assumptions that are hard to fulfill in realistic scenarios. For example, Gheissari et al. [28] and Wang et al. [79] try to increase the descriptive power of their feature representations by establishing correspondences between individual body regions. However, their approaches are restricted to people seen from similar viewpoints. Another example is the work of Farenzena et al. [18], who aim for roughly capturing the human body configuration by exploiting perceptual principles of symmetry and asymmetry. A drawback of this method is the required foreground-background segmentation, which is a very challenging pre-processing step considering typical illumination changes, shadows, and camera noise.

On the other hand, the discriminative approaches presented in this chapter try to generate a feature representation that is either specifically targeted at describing a single person (e.g., [6, 70]) or is designed to distinguish matching from non-matching person instances (e.g., [31, 64]). Characteristic problems of these works are generally higher runtimes, a consequence of the learning stage needed, and the tendency to over-fit the training data. The methods of Schwartz and Davis [70] as well as Lin and Davis [52] even have to be re-computed if new samples are added to the system, a fact that further increases their runtime.

From this point of view, it seems quite reasonable to combine descriptive and discriminative strategies to tackle the task of person re-identification. The idea is to exploit the advantages of both strategies, but to avoid their respective drawbacks. For instance, a descriptive approach can be used to quickly browse through a set of gallery images for a specific person, thus, providing a very efficient mechanism to obtain an initial search result. If the person is not found in this way, a discriminative model can be run subsequently in order to achieve a better result, however, at the cost of additional computational complexity. For further details, the reader is referred to Chapter 3, where we present a combined descriptive and discriminative system involving a human operator. Besides such an application-focused approach, another possibility to jointly apply descriptive and discriminative techniques is metric learning. Only little work has been devoted to this latter direction yet, and existing approaches mainly suffer from complex optimization schemes during the learning phase. In Chapter 4, we address this shortcoming and show how to obtain metrics suitable for person re-identification much more efficiently.

3 Combined Descriptive and Discriminative Person Re-Identification

Contents

3.1 Overview

As discussed in the previous chapters, most existing person re-identification methods try to either find a suitable description of a person's appearance or learn a discriminative model. Since these different representational strategies capture a large extent of complementary information, we propose to combine them both in an interactive system, so that we can benefit from the advantages of each strategy, but avoid their respective drawbacks. This is achieved by first running a fast, descriptive method and then, if necessary, refining the result by applying a discriminatively learned model.

A general view of our combined person re-identification system is given in Section 3.2. Next, the descriptive and the discriminative person model are presented in more detail in Sections 3.3 and 3.4, respectively. Finally, Section 3.5 summarizes the approach, discusses the benefits of using both strategies in a single system, and takes a closer look at the role of the human operator involved.

3.2 Combined Person Re-Identification System

With our person re-identification system, we follow the common task of assisting human operators in recognizing individual persons across a network of cameras. In order to reduce the required workload, a specific person selected in one camera view should be (semi-) automatically found in any other view. The main purpose of such a system is a significant reduction in search time the user has to spend. Instead of manually browsing through all camera views within a certain time frame, the user runs a re-identification system that generates a ranked list of potential matches, and only these candidates are further examined. Hence, the goal of our system is to produce such a ranked list of people that are most similar to the selected query person, with the correct match showing up among the first few ranks.

In particular, given a specific query image, we first run a fast, descriptive person model, where appearance is captured by a set of region covariance descriptors. This allows us to quickly get an initial ranking of all gallery images. The top ranked ones are then shown to a human operator, who decides whether the searched person has been found or not. If the true match gets a low rank, ideally rank one, we consider the task to be accomplished. However, if the descriptive person model fails so that the true match receives a high rank, i.e., it does not appear in the list shown to the user, a second stage has to be applied in order to get a refined result. During this second phase, we learn and evaluate a discriminative person model using boosting for feature selection. Since this model captures different aspects of an individual, focusing on details best separating the selected person from the rest, there is a good chance that it can improve the ranking. Furthermore, if requested, our setup also supports user interaction during search. Specifically, one can apply the discriminative system multiple times in an iterative manner. Each time, the human operator selects positive, i.e., similar, and negative, i.e., dissimilar, person images from the current ranking, thus, trying to adapt the learned person model to the searched individual. An overview of the proposed system is provided in Figure 3.1.

The descriptive model is based on a feature representation designed by hand, hence, it can be estimated for any given single image. The discriminative model, however, is learned for each instance requiring positive and negative training data. Since we focus on person re-identification in real world surveillance scenarios, where usually whole person trajectories containing multiple images are available (multi-shot scenario), we can use these images as training samples. If just one probe image is available (single-

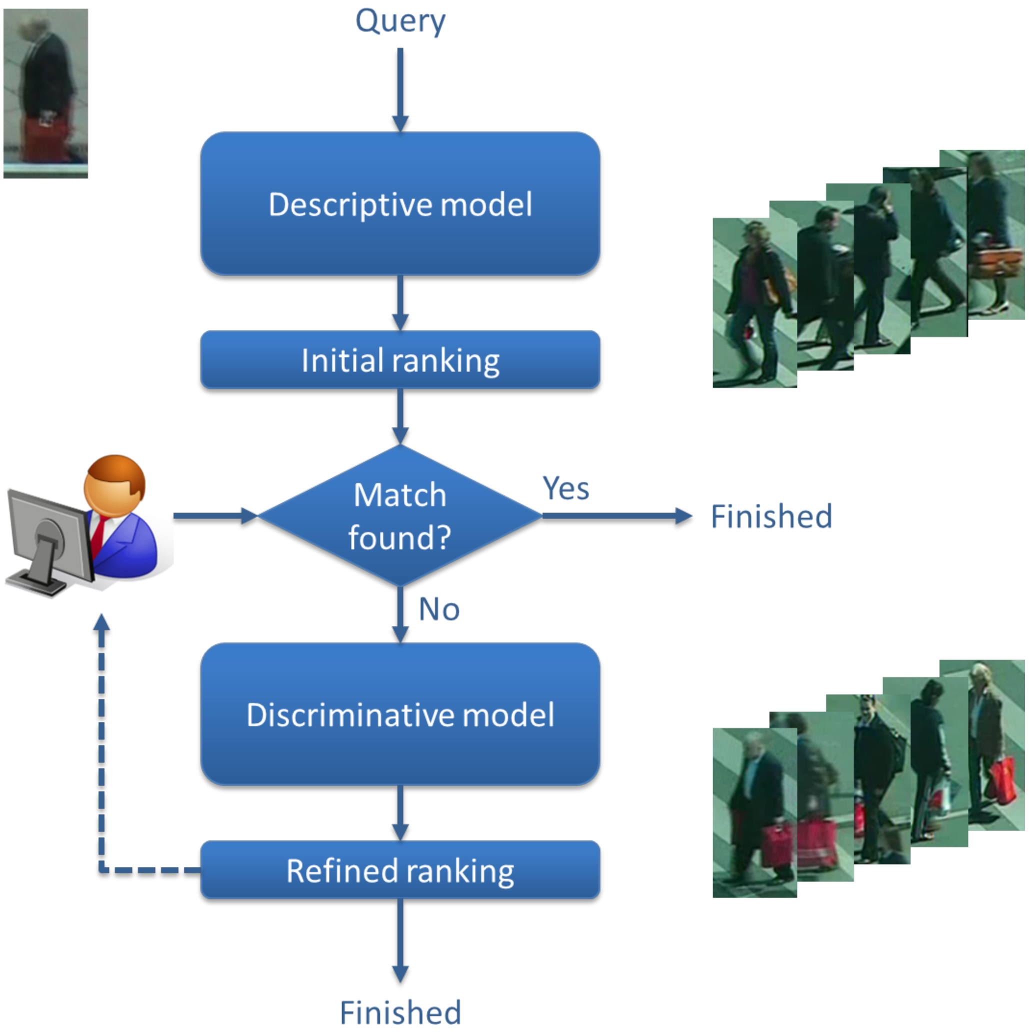


Figure 3.1: Overview of the combined descriptive and discriminative system: After applying a descriptive model to obtain an initial ranking, a discriminative model can be used to refine the result. As an option, the discriminative model can be iteratively applied with the human in the loop (dashed line).

shot scenario), we can still apply the discriminative model by generating virtual samples using geometric transformations and displacements. Hence, obtaining positive training samples is not much of a problem. In contrast, for generating negative training samples, a more sophisticated sampling mechanism is required. For this purpose, we use our descriptive model as starting point. As described before, applying this model already generates an initial ranked list of person images. Thus, we sample the negative images from the end of the list. Assuming that the descriptive person model provides a "good" ranking, those images should be most dissimilar to the searched person, making them a safe choice for drawing negative training samples. The overall principle is illustrated in Figure 3.2.

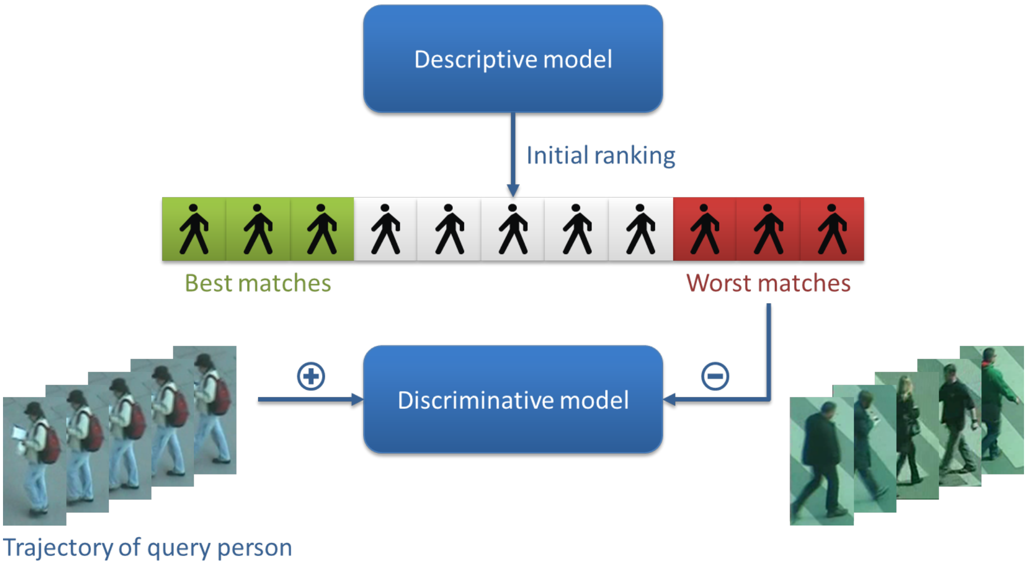


Figure 3.2: Sampling of training images for the discriminative model: Positive samples are obtained from the trajectory (multi-shot) or virtual samples (single-shot) of the query person, negative samples are drawn from the worst matches of the initial ranking provided by the descriptive model.

3.3 Descriptive Person Model

In the first stage of our person re-identification system, we generate a descriptive statistical model which encodes visual appearance information. Considering the given task, the employed representation must meet requirements of specificity, invariance, and computational efficiency. This implies that on the one hand, the visual description must encompass distinctive visual information, and on the other hand, it must remain mostly unaffected in presence of photometric, view, and pose changes. Moreover, for practical applicability, the representation should be computed and matched rapidly at small memory requirements.

Since the descriptive model targets at comparing hand-crafted feature representations to each other, it is perfectly suited for scenarios that just offer one image per person in each camera view, i.e., single-shot setups. In contrast, when working with multi-shot systems, further mechanisms are required in order to be able to exploit the additional data. For instance, if we assume that we are given whole trajectories for the query person as well as the gallery persons, a straightforward approach would be to use all possible image pairs when comparing two trajectories. However, this would lead to a runtime behavior that is quadratic in the number of person images per trajectory. Hence, we restrict our system to just comparing the one query person image that has been selected by the human operator to five equidistantly sampled images from each gallery trajectory, resulting in a drastically reduced computation time. Thus, we obtain five distance measurements for each gallery person, and the final rank is determined by the minimum distance achieved.

3.3.1 Feature Representation

Describing person appearances by extraction of color and texture information is a natural choice and widely used in the field of person re-identification. Especially popular are histogram representations of color mappings, gradient magnitudes, and texture filter responses, as for example proposed in [28, 79, 31, 64, 87]. Given a certain image region, a histogram models the distribution of feature values inside that region, which can then be used as region descriptor. However, histograms are not very robust in presence of illumination changes, and the joint representation of several different feature channels via histograms is exponential in the number of channels. Hence, we seek a more robust and efficient representation for our purpose.

In order to meet these requirements, we employ the region covariance descriptor introduced by Tuzel et al. [75], which is capable of combining multiple complementary cues, efficient to compute via integral images, and which generates a compact signature. Furthermore, it is robust to illumination changes, and noise that corrupts individual points is largely filtered out. Since the descriptor discards any information regarding the number and ordering of points inside a region, it is also scale and rotation invariant unless such information is presented by the feature values themselves. However, in the case of person re-identification, it is usually desired to also capture the spatial arrangement of appearance features, as this provides valuable information when matching person images. Hence, spatial information is normally added to the set of feature channels, e.g., in form of pixel coordinates, making the descriptor rotation variant. This is not a problem in person re-identification scenarios, though, as people are almost always captured in upright position. Nevertheless, since the descriptor aggregates several visual features, structural information of the human appearance, such as the brightness relationship between upper and lower body halves, often still remains underrepresented. In order to enhance the structural specificity of the representation, we use a set of covariance descriptors computed from multiple horizontal stripes covering the entire area of the person image patch. This strategy is similar to the multiple region scheme used by Tuzel et al., the principal axis histogram signature employed in [38, 39], and the region setup proposed in [31].

In particular, for a given bounding box B with dimensions $W \times H$, a set of region covariance descriptors is computed in the following manner: The image within the bounding box $I_B(x,y)$ is used to calculate diverse features capturing location, intensity, color, and texture information. Specifically, in our case, the applied feature set consists of the vertical pixel coordinate y, the L, a, b color channels, and the horizontal and vertical derivative of the luminance channel. Hence, we obtain a feature vector

$$\mathbf{f} = \left[y, L, a, b, \left| \frac{\partial L}{\partial x} \right|, \left| \frac{\partial L}{\partial y} \right| \right]^\top \tag{3.1}$$

with $d = 6$ dimensions. Note that throughout this thesis, we consider vectors to be column vectors. The x-component of the pixel coordinates is excluded, since the horizontal location of appearance features can vary significantly due to changes in viewpoint and pose. Next, in order to increase the structural specificity, the bounding box B is divided into $N = 7$ equally large horizontal stripes $\{S_l\}_{l=1,\ldots,N}$, and within each stripe, the covariance descriptor is computed as

$$\boldsymbol{\Sigma}^l = \frac{1}{n-1} \sum_{k=1}^{n} (\mathbf{f}_k^l - \boldsymbol{\mu}^l)(\mathbf{f}_k^l - \boldsymbol{\mu}^l)^\top , \tag{3.2}$$

where $\boldsymbol{\Sigma}^l$ denotes the covariance matrix of size $d \times d$ computed over n feature vectors within the l-th stripe, and $\boldsymbol{\mu}^l$ is the corresponding mean vector. As can be seen from Equation (3.2), the covariance matrix provides an elegant way of fusing multiple features into a compact descriptor that can be computed very efficiently. Each diagonal entry represents the variance of the corresponding feature, and the off-diagonal entries represent the correlations between features. The previously mentioned robustness to illumination changes and noise stem from the subtraction of the data mean and the averaging over all data points, respectively. Furthermore, since the matrix is symmetric, it contains only $(d^2 + d)/2$ different feature values. Thus, it produces a rather low-dimensional description compared to joint histograms, which require b^d dimensions, where b is the number of histogram bins used for each feature.

Finally, the obtained set of covariance matrices $\{\mathbf{\Sigma}^l\}_{l=1,\ldots,N}$ defines a compact description which encodes the interdependence between individual features computed inside the bounding box B. A coarse structural information is captured by calculating covariances from multiple horizontal stripe regions, and a weak spatial dependence is given by the only slightly specific variation of the y-coordinate feature within each stripe.

3.3.2 Distance Calculation

The space of covariance matrices does not form a Euclidean space, e.g., it is not closed under multiplication with negative scalars and does not fulfill the symmetry requirement. Instead, covariance matrices lie on the Riemannian manifold of symmetric, positive definite matrices, which means that most of the common machine learning methods are not directly applicable. Hence, many approaches exploit Riemannian geometry, where covariance matrices are first mapped to a linear tangent space of the manifold, and then, distance calculations are performed in this tangent space. We follow this direction here and estimate the similarity between two human appearances by applying an approximated distance function through manifold mappings according to Förstner and Moonen [24]. In particular, the distance between two covariance matrices is defined as the sum of squared logarithms of the generalized eigenvalues:

$$\rho(\mathbf{\Sigma}_i^l, \mathbf{\Sigma}_j^l) = \sqrt{\sum_{k=1}^{d} \ln^2 \lambda_k(\mathbf{\Sigma}_i^l, \mathbf{\Sigma}_j^l)} \, , \tag{3.3}$$

where $\mathbf{\Sigma}_i^l$ and $\mathbf{\Sigma}_j^l$ are computed for two different person images i and j, but using the same stripe element l, and $\{\lambda_k\}_{k=1,\ldots,d}$ denotes the set of generalized eigenvalues of $\mathbf{\Sigma}_i^l$ and $\mathbf{\Sigma}_j^l$. The eigenvalues are obtained from solving the generalized eigenvalue problem

$$\lambda_k \mathbf{\Sigma}_i^l \mathbf{x}_k - \mathbf{\Sigma}_j^l \mathbf{x}_k = \mathbf{0} \, , \quad k = 1, \ldots, d \, , \tag{3.4}$$

where $\mathbf{x}_k \neq \mathbf{0}$ are the generalized eigenvectors. Finally, the overall covariance-based distance between two human appearances is given by

$$\bar{\rho}_{ij} = \frac{1}{N} \sum_{l=1}^{N} \rho(\mathbf{\Sigma}_i^l, \mathbf{\Sigma}_j^l) \, , \tag{3.5}$$

i.e., the mean distance measure obtained from N stripe versus stripe comparisons, as shown in Figure 3.3.

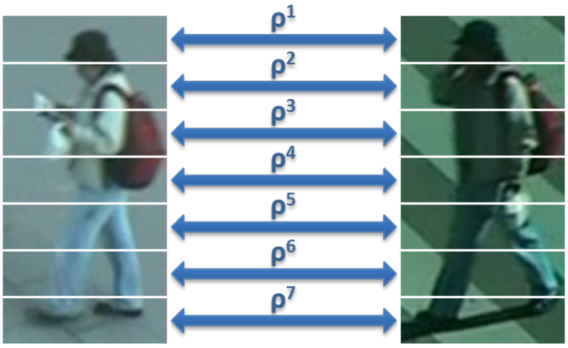


Figure 3.3: Descriptor matching: Descriptors are extracted from horizontal stripe regions and compared using a metric for covariance matrices. Finally, the overall distance for the descriptive model is obtained from averaging over the individual stripe distances.

When matching person appearances, a specific probe image is compared to all gallery images using Equation (3.5), resulting in a set of distances. These distances are then sorted in order to generate a ranking of the gallery with respect to the probe, which serves as starting point for the discriminative model presented in Section 3.4.

3.4 Discriminative Person Model

In the second stage of our system, we apply a discriminative model, which is estimated by boosting for feature selection [73]. Thus, similar to [31, 6], the goal is to select the most discriminative features for a specific individual from an over-complete feature set. However, unlike these methods, our approach does not involve any labeling of training data by hand, since positive and negative samples are automatically provided by the system. As already described in Section 3.2, positive samples are obtained from the query person, i.e., either images are extracted from the corresponding trajectory in case of a multi-shot setup, or virtual samples are generated by applying geometric transformations and displacements in the single-shot case. On the other hand, negative samples are obtained from the initial ranking produced by the descriptive model. Moreover, in contrast to [31, 6], the goal is not to learn a similarity function between image pairs, but to finally generate a ranking of all gallery images. This is similar to [64], where an SVM-based ranking method is used to sort the gallery.

In particular, when applying the discriminative stage, a specific feature model is computed for the chosen query person using AdaBoost [26], a boosting algorithm that has shown great performance for the task of feature selection over the years, as briefly outlined in the next section. After the training phase, the obtained model is evaluated on all gallery images, which are then sorted according to their confidence values: a higher confidence results in a lower rank. Like in the descriptive stage, we evaluate the learned model only on a subset of the gallery images in case of multi-shot setups, thus, reducing the computational load. Again, five images are equidistantly extracted from each person trajectory, and the final rank is determined by the maximum confidence value achieved.

3.4.1 Boosting for Feature Selection

In machine learning, a learner tries to identify an unknown concept based on a set of samples of that concept. How well rules that have been learned from given training data can be applied to unseen, new data is characterized by the generalization ability, a crucial property of any learning algorithm. Over the years, a lot of research has been devoted to new ideas that can improve the generalization ability of learning algorithms. One of the most successful paradigms developed so far is ensemble learning. The idea is to combine a set of weak learners to a strong learner, leading to improved performance compared to each of the individual learners.

This is related to an interesting question raised by Kearns and Valiant [47], whether "weakly" learnable and "strongly" learnable problems, i.e., classes of different complexity, are in fact equal. The answer is "yes", as proven by Schapire [68]. In particular, a concept class is weakly learnable if, given access to a set of samples of an unknown concept, a learner can produce a hypothesis that performs only slightly better than random guessing. On the other hand, a concept class is strongly learnable if a learner is very likely to output a hypothesis that is correct on all but an arbitrarily small portion of the samples. Schapire shows that these two notions of learnability are equivalent by constructing the first boosting algorithm: If they are independent, i.e., make errors on different samples of the input data, weak learners that just perform little better than random guessing can be boosted into a strong learner of arbitrary accuracy. The main idea of this initial boosting algorithm is to increase the performance of a single weak learner by two additional weak learners trained on different versions of the input data. Their outputs are combined using majority voting, and the whole procedure is carried out recursively. A simpler and more efficient boosting method is presented by Freund [25], who connects many weak learners via a single majority gate, leading to improved performance compared to [68]. However, as a drawback, both approaches require prior knowledge of the accuracies achieved by the weak learners, information that is usually not available in practice. Hence, in order to overcome this limitation, further research has been devoted to the investigation of new boosting algorithms, finally leading to the development of AdaBoost by Freund and Schapire [26]. AdaBoost is an adaptive boosting algorithm that does not need those unknown error bounds, an advantage that implies a very broad practical applicability. A detailed description of the algorithm is given in Section 3.4.2.

A particularly useful application of the AdaBoost algorithm is boosting for feature selection, first introduced by Tieu and Viola [73] for the task of image retrieval. Their approach is based on the assumption that each image consists of a sparse set of visual causes, and that causes are shared between visually similar images. The causes can be seen as a simplifying structure in the distribution of images, where the goal of a learner is to discover this structure. Hence, Tieu and Viola propose to generate an over-complete set of highly selective features, i.e., features that respond only to a small number of given database images. They compute over 46,000 such features by three levels of filtering with 25 simple linear features like oriented edges, center surround, and bar filters. Given a few query images showing the same content, the idea now is that only a small number of highly selective features, e.g., 20 to 50, is necessary to find database images similar to the exemplars. For the task of selecting such an appropriate subset of features, Tieu and Viola apply AdaBoost, where each weak learner corresponds to one feature out of the over-complete set. Thus, as the algorithm consecutively adds weak learners to finally obtain a strong learner, it implicitly also selects the most suitable features.

Inspired by this work, Viola and Jones present a real-time object detector that utilizes boosting for feature selection. Using an exhaustive set of Haar-like features, they achieve excellent performance on the task of face [77] and person detection [78], however, much faster than previous approaches. The obtained real-time capability of the detector results from two contributions. First, Viola and Jones introduce the "integral image", a new image representation that allows a very efficient calculation of pixel sums needed for evaluating Haar-like features. Second, the proposed detector combines several classifiers with increasing complexities in a cascade structure, which is especially useful if a lot of image patches have to be processed, such as in case of sliding window techniques. The first few stages contain classifiers of quite limited complexity, so that a majority of image patches can be rejected very rapidly. In this way, only promising patches are forwarded to computationally more expensive classifiers, resulting in a drastically reduced runtime.

Due to its success, there has been a considerable scientific interest in boosting for feature selection. Following the works of Viola and Jones, various different approaches have been proposed over the years, which mainly differ in the type of features they employ, e.g., rotated Haar-like features [51], Gabor filters [71], edge orientation histograms [50], PCA features [40], or boundary fragments [61].

As it is applicable to practically all kinds of features that can be computed reasonably efficiently, boosting for feature selection is perfectly suited for our purpose, person re-identification, which is very similar to its original task of image retrieval. Given a set of highly selective features, we want to select a sufficiently discriminative subset in order to distinguish a specific person from the rest of the people.

3.4.2 AdaBoost

AdaBoost, introduced by Freund and Schapire [26], is one of the most popular boosting algorithms, and (referring to a NIPS workshop) was called the "best off-the-shelf classifier in the world" by Breiman in 1996. In contrast to previous boosting methods, AdaBoost re-weights the training samples instead of re-sampling them, so that the weak learners can be trained with respect to this weight distribution. Furthermore, it adapts to the weak learners' accuracies and outputs a weighted majority hypothesis, where the weight of each individual weak hypothesis is related to its training error. Freund and Schapire also prove that if the weak learners have errors just below 50%, i.e., they perform just better than random guessing in binary classification tasks, then the training error of the final hypothesis drops to zero exponentially fast.

In this work, we focus on discrete AdaBoost, where each weak learner returns a binary decision. In particular, the algorithm consists of several steps as follows. Assume we are given a set of M labeled training samples $\mathcal{X} = \{(\mathbf{x}_m, y_m)\}_{m=1,\ldots,M}$, where $\mathbf{x}_m \in \mathbb{R}^d$ is an arbitrary feature vector and $y_m \in \mathcal{Y} = \{-1, +1\}$ is its corresponding class label, further a weak learning algorithm $\mathcal{L}$ that outputs a hypothesis $h : \mathbb{R}^d \to \mathcal{Y}$, i.e., a weak classifier predicting a label, and finally a weight distribution D over the training set $\mathcal{X}$ initialized uniformly to $D(m) = 1/M$.

Now in each iteration $t = 1, \ldots, T$ of the AdaBoost algorithm, the training samples $\mathcal{X}$ are fed to the weak learner $\mathcal{L}$, which generates a hypothesis h_t that hopefully has a small training error ϵ_t, at least below 50%, with respect to the current distribution D_t:

$$\epsilon_t = \sum_{m:h_t(\mathbf{x}_m) \neq y_m} D_t(m) \; . \tag{3.6}$$

Related to the training error, the weight α_t of the generated hypothesis is calculated according to the following formula:

$$\alpha_t = \frac{1}{2} \ln \left(\frac{1 + \epsilon_t}{\epsilon_t} \right) \; . \tag{3.7}$$

Before the next iteration starts, the weight distribution D_t is updated such that weights are increased for misclassified and decreased for correctly classified samples:

$$D_{t+1}(m) = \begin{cases} \dfrac{D_t(m)}{Z_t} \exp(-\alpha_t) & h_t(\mathbf{x}_m) = y_m \\ \dfrac{D_t(m)}{Z_t} \exp(\alpha_t) & h_t(\mathbf{x}_m) \neq y_m \end{cases} , \tag{3.8}$$

where Z_t is a normalization factor chosen such that D_{t+1} is a distribution. Since we assume binary class labels from the set $\{-1, +1\}$, Equation (3.8) can be re-written to

$$D_{t+1}(m) = \frac{D_t(m)}{Z_t} \exp(-\alpha_t h_t(\mathbf{x}_m) y_m) \; . \tag{3.9}$$

Updating the weight distribution in this way forces subsequent hypotheses to focus on the difficult samples, i.e., those samples that have been misclassified so far, in order to achieve small training errors.

The whole process is repeated until either a maximum number of T iterations is reached, or a certain stopping criterion, e.g., a training error of 0%, is met. Finally, a strong classifier H is computed as a weighted linear combination of all generated hypotheses, i.e., weak classifiers, h_t:

$$H(\mathbf{x}) = \sum_{t=1}^{T} \alpha_t h_t(\mathbf{x}) \; , \tag{3.10}$$

which returns a real confidence value. If a discrete label is required, a discrete strong classifier $\tilde{H}$ can be obtained by simply applying the sign function:

$$\tilde{H}(\mathbf{x}) = \operatorname{sgn}(H(\mathbf{x})) \; . \tag{3.11}$$

Since we are interested in person re-identification, and our goal is to generate a similarity-based ranking of person images, we stick to Equation (3.10). This allows us to calculate a confidence value for each of the images in the gallery, which is then sorted accordingly. The individual steps of the discrete AdaBoost algorithm are summarized in Algorithm 3.1.

3.4.3 Features

As mentioned previously, due to its popularity, many different feature types, e.g., rotated Haar-like features [51], Gabor filters [71], edge orientation histograms [50], PCA features [40], or boundary fragments [61], have been proposed for the application with boosting for feature selection. These features have shown excellent performance for various vision tasks such as object recognition, detection, and tracking. In the particular case of person re-identification, we found that the most important information queues are intensity changes between the upper and lower body of a person and essential color attributes. Thus, for our task, we use a combination of horizontally divided Haar-like features, which model vertical intensity changes, and covariance features, which are capable of fusing multiple color channels into a compact descriptor. Moreover, to avoid that too much background information is modeled by the rather local features, we prohibit features that are placed close to the image borders. Since Haar-like features are well known in the context of boosting (e.g., [77]), in the following, we focus on the discussion of the covariance features, and how to efficiently compute them in order to be applicable in a boosting framework.

As described in Section 3.3.1, covariance matrices, in general, provide an elegant way of integrating various different feature channels, in our case RGB color channels, into one compact representation. They capture the variances of these channels and the correlations between them. However, since covariance matrices do not lie on Euclidean space, they cannot directly be used in a boosting framework. While we could simply draw on the distance measure defined in Equation (3.3), i.e., calculate the similarity between covariance matrices on Riemannian manifolds, this would lead to a significant increase in runtime, as a lot of features have to be evaluated during training. Hence, in order to overcome this limitation, we follow the approach described in [48, 36], allowing to represent individual covariance matrices directly in a Euclidean vector space. In this way, computationally costly processing steps on manifolds can be avoided.

Specifically, given a covariance descriptor computed according to Equation (3.2), we want to map the first and second order moments of the underlying distribution of data points, i.e., mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, to Euclidean vector space. This can

Algorithm 3.1 Discrete AdaBoost

```
Input: set of M labeled training samples $\mathcal{X} = \{(\mathbf{x}_m, y_m)\}_{m=1,\dots,M}$, weak learning algorithm $\mathcal{L}$, maximum number of iterations $T$
Output: strong classifier $H(\mathbf{x})$
Initialize weight distribution:
    $D_1(m) = \frac{1}{M}$
for $t = 1, \dots, T$ do
    Call weak learner to obtain a hypothesis that has a small training error with respect to the current weight distribution:
        $h_t = \mathcal{L}(\mathcal{X}, D_t)$
    Calculate training error:
        $\epsilon_t = \sum_{m: h_t(\mathbf{x}_m) \neq y_m} D_t(m)$
    if $\epsilon_t = 0$ then
        Set $T = t$ and abort loop
    end if
    if $\epsilon_t \geq \frac{1}{2}$ then
        Set $T = t - 1$ and abort loop
    end if
    Calculate hypothesis weight:
        $\alpha_t = \frac{1}{2} \ln \left( \frac{1 + \epsilon_t}{\epsilon_t} \right)$
    Update weight distribution:
        $D_{t+1}(m) = \begin{cases} \frac{D_t(m)}{Z_t} \exp(-\alpha_t) & h_t(\mathbf{x}_m) = y_m \\ \frac{D_t(m)}{Z_t} \exp(\alpha_t) & h_t(\mathbf{x}_m) \neq y_m \end{cases}$
    ($Z_t$ is a normalization factor chosen such that $D_{t+1}$ is a distribution.)
end for
return $H(\mathbf{x}) = \sum_{t=1}^{T} \alpha_t h_t(\mathbf{x})$
```

efficiently be done utilizing the Unscented Transform (UT) proposed by Julier et al. [46], where the key insight is that it is generally easier to approximate a single distribution than to approximate an arbitrary non-linear transformation. Hence, a given distribution is estimated by selecting a representative set of test samples, which is constructed such that it captures the distribution's mean and covariance. These test points are then mapped via a given non-linear transformation, resulting in a new set of points, referred to as *sigma points*. This is beneficial, since the desired non-linear transformation can *exactly* be applied to each of the test points, as opposed to approaches that try to approximate the transformation itself, e.g., using linearization, such as the extended Kalman filter [43]. Finally, the mean and covariance of the transformed points are computed, representing an estimation of the non-linear transformation of the original distribution.

In particular, for the d-dimensional case, a set of $2d+1$ sigma points $\mathbf{s}_i \in \mathbb{R}^d$, $i = 0, \ldots, 2d$, is constructed as follows:

$$\mathbf{s}_0 = \boldsymbol{\mu} \,, \tag{3.12}$$

$$\mathbf{s}_j = \boldsymbol{\mu} + \alpha \left(\sqrt{\boldsymbol{\Sigma}}\right)_j \,, \tag{3.13}$$

$$\mathbf{s}_{j+d} = \boldsymbol{\mu} - \alpha \left(\sqrt{\boldsymbol{\Sigma}}\right)_j \,, \tag{3.14}$$

with $j = 1, \ldots, d$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ being the data's mean vector and covariance matrix, respectively, and $\left(\sqrt{\boldsymbol{\Sigma}}\right)_j$ being the j-th column of the covariance matrix square root, which can efficiently be calculated using the Cholesky decomposition. The scalar α is a constant weighting for the elements in the covariance matrix and is set to $\alpha = \sqrt{2}$ for Gaussian data. Constructing sigma points in this way can be seen as an efficient mapping of a specified set of test vectors $\mathbf{t}_i \in \mathbb{R}^d$ that are located at the intersections of a unit hypersphere with a d-dimensional Cartesian coordinate system, with $\mathbf{t}_0 = \boldsymbol{\mu}$ representing the origin, as illustrated in Figure 3.4 for a simplified two-dimensional case. It is worth noting that the deterministic test point sampling utilized in the UT is in contrast to Monte Carlo methods, where the test points are chosen randomly. Furthermore, the statistics of the obtained sigma points accurately capture the original mean and covariance information up to third order for Gaussian and up to second order for non-Gaussian data. The final feature representation is built by concatenation of all sigma points to one vector

$$\mathbf{s} = [\mathbf{s}_0, \ldots, \mathbf{s}_{2d}] \,, \tag{3.15}$$

which describes Euclidean space, thus, allowing element-wise distance calculations between individual descriptors. As can be seen, sigma points offer a very powerful representation that is capable of integrating various different feature channels into one compact feature vector. The obtained feature representation takes advantage of all the nice properties of the covariance descriptor presented by Tuzel et al. [75], however, circumvents the disadvantage of costly computations on manifolds by providing feature vectors directly in Euclidean vector space.

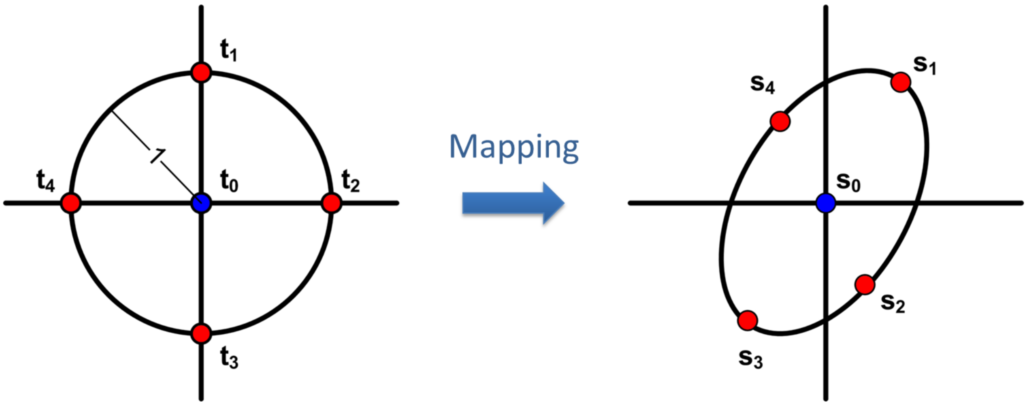


Figure 3.4: Generation of sigma points: Deterministically sampled test vectors $\mathbf{t}_i$ are mapped to sigma points $\mathbf{s}_i$ given in a second coordinate system, while the characteristics of the original covariance matrix are preserved. The image stems from [48].

With this representation, we are now able to efficiently capture local color information in our boosting framework. As for Haar-like features, we use a rectangularly shaped region for extracting color information (RGB) from an image. All pixels within the feature region are used to calculate the mean vector $\boldsymbol{\mu}$, the covariance matrix $\boldsymbol{\Sigma}$, and finally the sigma points representation. This enables us to capture very discriminative, local color features of a person (e.g., a red bag), as opposed to the descriptive statistical model described in Section 3.3, which extracts color and gradient information from regular stripe regions laid over the whole person image. Both feature types, i.e., Haar-like and covariance features, are illustrated in Figure 3.5.

3.4.4 Feature Response Statistics

As already mentioned, if boosting is applied for feature selection, then each weak learner corresponds to one feature out of an over-complete set. However, besides a specific visual feature, a weak learner also requires some sort of statistics in order to model the distribution of feature responses. This statistical model enables the weak learner

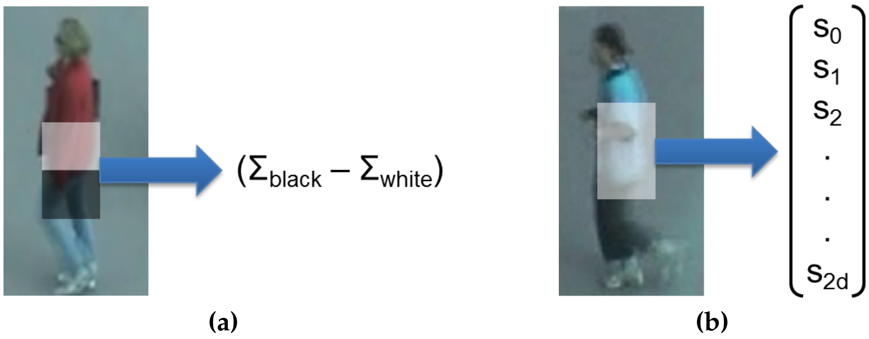


Figure 3.5: Features applied in the discriminative stage: (a) Haar-like features mainly capture intensity changes between the upper and lower body of a person, (b) covariance features extract local color information in form of vectors of sigma points.

to actually learn and make decisions based on the values obtained from the feature representation. During training, the model is updated according to the observed feature responses and corresponding label information. Afterwards, given a certain feature value, the model is able to make a decision based on the learned statistics, i.e., it can be used to classify previously unseen test samples.

In our setup, we have to distinguish between two feature types that generate different kinds of responses. The Haar-like features produce a scalar response $f_H(\mathbf{x})$, while the covariance features, i.e., sigma points, return a vector response $\mathbf{f}_s(\mathbf{x})$. In particular, to generate a hypothesis for the Haar-like features, we use a Bayesian decision criterion:

$$h_H(\mathbf{x}) = \mathrm{sgn}(P(f_H(\mathbf{x}) \mid 1) - P(f_H(\mathbf{x}) \mid -1)) \,, \tag{3.16}$$

where the probabilities for the positive and negative class are modeled via Gaussian distributions $\mathcal{N}(\mu_p, \sigma_p^2)$ and $\mathcal{N}(\mu_n, \sigma_n^2)$ estimated from the feature responses of the positive and negative training samples, respectively, as illustrated in Figure 3.6. Hence, Equation (3.16) can be re-written to

$$h_H(\mathbf{x}) = \mathrm{sgn}(g(f_H(\mathbf{x}) \mid \mu_p, \sigma_p^2) - g(f_H(\mathbf{x}) \mid \mu_n, \sigma_n^2)) \,, \tag{3.17}$$

with $g(x \mid \mu, \sigma^2)$ being the Gaussian probability density function. Assuming the distributions to be Gaussian implies that the parameters can be computed very easily.

In contrast, for the multi-dimensional sigma points representation, we apply a nearest neighbor classifier using the Euclidean distance:

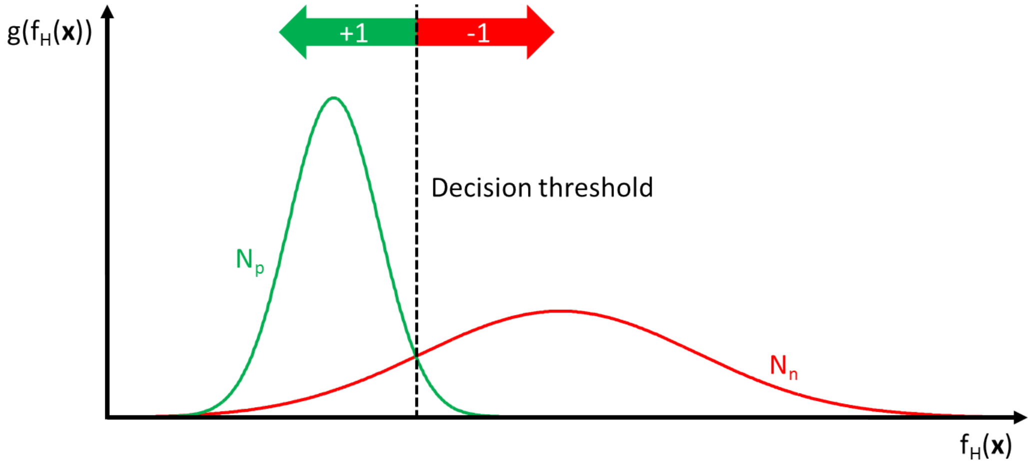


Figure 3.6: Bayesian decision criterion: Based on the positive and negative training distributions, a threshold is calculated, which is then used to classify new samples as either belonging to the positive or negative class.

$$h_s(\mathbf{x}) = \operatorname{sgn}(\|\mathbf{f}_s(\mathbf{x}) - \boldsymbol{\mu}_n\|_2 - \|\mathbf{f}_s(\mathbf{x}) - \boldsymbol{\mu}_p\|_2) \; . \tag{3.18}$$

The cluster centers for positive and negative samples, $\boldsymbol{\mu}_p$ and $\boldsymbol{\mu}_n$, are estimated by computing the mean for each feature vector entry separately.

3.5 Discussion

In our person re-identification system, we combine a descriptive and a discriminative person model in order to take advantage of the different representational strategies, which capture a large extent of complementary information. In the first step, the fast, descriptive stage is executed to obtain an initial ranking of person images very quickly. Since the feature representation in this stage is hand-crafted, i.e., fixed, the signatures of all persons in the gallery database can be pre-computed. Hence, the search time is significantly reduced, as only the signature of the selected query person has to be generated, which is then matched with the elements stored in the database. After that, a human operator examines the top ranks of the produced ranking and decides whether the searched person has been found or not. Empirically, from our experiments, we know that this initial ranking is rather robust, i.e., there are usually no severe outliers. In general, the descriptive strategy works well in cases where the query person has a distinctive appearance that is not found overly often in the gallery. But if there are a lot of similar persons, the holistic feature representation can quickly become too unspecific, so that the descriptive model is likely to fail.

In such situations, the human operator can still apply the discriminative model as a second step. Due to the different characteristics of this model, there is a good chance that it can improve the existing ranking, especially since it focuses more on finer details, which are usually underrepresented in the descriptive stage. Unfortunately, a disadvantage of the discriminative step is its higher computational complexity, which significantly increases the overall runtime, typically from a couple of seconds to a few minutes. To alleviate this drawback, the discriminative model might be started in the background right after the descriptive stage, i.e., while the user is still examining the initial ranking. As indicated in Figure 3.1, it is also possible to apply the discriminative model multiple times with the human in the loop. In each iteration, the user investigates the current ranking and labels the shown persons as being similar or dissimilar to the query person, with the goal to steer the model towards the correct match. However, during our experiments, we found that untrained users tend to distort the discriminative person model when interacting with the system. The problem seems to be the discrepancy between the human understanding of similarity and that of machine learning algorithms, especially since humans perceive a person rather holistically, while our discriminative model produces a quite local feature representation. This observation suggests that the operating personnel should be trained before running the system in an interactive mode.

Besides this qualitative analysis, in Section 5.4.1, we give a more thorough evaluation. In particular, we show that both models in our system capture different aspects of the appearance of a person, thus, providing two complementary methods that can be used to find a specific query person. However, since the decision whether or not to run the discriminative model is still made by the human operator, we seek a more principled way of combining descriptive and discriminative techniques. One possibility to achieve this is via metric learning, which is the topic of the next chapter.

4 Metric Learning for Person Re-Identification

Contents

4.1 Overview

Recently, metric learning has gained considerable scientific interest in the field of person re-identification, as it provides a very elegant fusion of the descriptive and discriminative techniques typically encountered in the community. The main idea is to build on an existing feature representation, which is usually designed to generate a descriptive signature of the whole person appearance, and then to learn a suitable metric that reflects the visual camera-to-camera transition. This is similar to the idea of inter-camera color calibration (see, e.g., [63]), where feature space transitions between various camera views are modeled using labeled samples. Hence, in contrast to methods that match features directly in the feature space using some standard distance measure, metric learning has the advantage that even less distinctive features, which need not capture the visual invariance between different camera views, are sufficient for achieving high matching performance. Moreover, since the learned metric inherently emphasizes or attenuates directions in the feature space based on their importance for the given task, it can also be seen as a discriminative feature selector. Just like in the case of discriminative methods, to estimate such a metric, a training stage is necessary. However, once learned, metric learning approaches are very efficient during evaluation, since additionally to the feature extraction and the matching, only linear projections have to be computed.

In the special case of person re-identification, we have to cope with three main difficulties concerning existing metric learning approaches. First, in order to capture all relevant information, often high-dimensional feature representations are needed. Thus, widely used metric learners such as Large Margin Nearest Neighbor (LMNN) [80, 81], Information-Theoretic Metric Learning (ITML) [15], and Logistic Discriminant Metric Learning (LDML) [32], i.e., methods that build on complex optimization schemes, run into high computational costs and memory requirements. As a consequence, they are infeasible in practical scenarios consisting of dozens of cameras. Second, most of these methods assume a multi-class classification problem, which is not the case for person re-identification. In fact, we are typically given image pairs showing the same person and image pairs showing different persons captured by two distinct cameras. However, since usually no information about the inter-similarity between individual image pairs, i.e., persons, is given, we do not want to treat them as separate classes, as this could distort the metric. Hence, existing methods have to be adapted for our purpose. There are only a few methods such as those presented in [29, 1] which intend to learn a metric directly from data pairs. Third, as already mentioned, we have to deal with a partially ill-posed problem. In fact, two images showing the same person might not look similar at all (e.g., due to camera noise, scene geometry, or different viewpoints: frontal versus back view). On the other hand, images not showing the same person can be very similar (e.g., many people wear rather dark clothes in winter). Thus, for standard methods, there is a high tendency to over-fit the training data, yielding insufficient results during testing.

The goal of this chapter now is twofold. First, we analyze the applicability of metric learning to the task of person re-identification from a more general point of view. Hence, we review the main idea of Mahalanobis metric learning and give an overview of selected approaches which are targeted at the problem of discriminative metric learning via different strategies. In particular, we concentrate on established methods applied to diverse visual classification tasks (e.g., LMNN, ITML, and LDML), as well as approaches that have been developed specifically for person re-identification (e.g., Large Margin Nearest Neighbor with Rejection (LMNN-R) [16], Probabilistic Relative Distance Comparison (PRDC) [87], and Pairwise Constrained Component Analysis (PCCA) [58]). Since these methods typically rely on computationally expensive optimization techniques, in the second part of this chapter, we present new solutions on how to obtain metrics that are especially suited for our purpose, but can be computed much more efficiently. First, we show how to balance the influence of matching and non-matching person pairs during optimization via an iterative, but very fast approach in order to achieve state-of-the-art re-identification performance. To further increase efficiency, we then propose learning a metric similar to LMNN, i.e., optimized over local neighborhoods in the feature space, however, via a closed-form solution. Thus, the runtime is significantly reduced compared to the iterative optimization scheme of the original LMNN algorithm.

The rest of the chapter is organized as follows. First, in Section 4.2, Mahalanobis metric learning in general is introduced, and the methods used for comparison in our study are summarized. Then, in Section 4.3, we present our more efficient metric learning approaches that are particularly designed for the task at hand. In Section 4.4, the framework for applying metric learning to person re-identification is described, and finally, in Section 4.5, the gained insights are discussed.

4.2 Mahalanobis Metric Learning

In this section, we first provide a definition of metrics, introduce the general idea of Mahalanobis metric learning, and then present an overview of the approaches used for comparison in our study. These include generic methods that have shown good performance for diverse visual classification tasks, as well as specific methods that have been developed for person re-identification. Moreover, to give a broader analysis, the selected methods tackle the same problem from different points of view, e.g., using generative data analysis, statistical inference, information-theoretic aspects, or discriminative learning. Additionally, we consider the standard Mahalanobis metric and Linear Discriminant Analysis (LDA) [21], which can be regarded as simple baselines.

4.2.1 What is a Metric?

In mathematics, a *metric* is a function that defines a distance between elements of a set, which is called a metric space in this case. Hence, it is also referred to as *distance function*, or simply *distance*. A more formal definition is given by the following constraints, where d_m denotes a metric on set $\mathcal{X}$, i.e., $d_m : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$, and x_i, x_j, x_k are elements of $\mathcal{X}$:

$$d_m(x_i, x_j) = 0 \quad \text{if and only if} \quad x_i = x_j \; , \tag{4.1}$$

$$d_m(x_i, x_j) = d_m(x_j, x_i) \; , \tag{4.2}$$

$$d_m(x_i, x_k) \leq d_m(x_i, x_j) + d_m(x_j, x_k) \; , \tag{4.3}$$

$$d_m(x_i, x_j) \geq 0 \; . \tag{4.4}$$

The first constraint, Equation (4.1), is the coincidence axiom, which ensures that the distance between two points can only be zero if the points are identical. Further, Equation (4.2) requires the distance to be symmetric, and Equation (4.3) describes the triangle inequality. Finally, the non-negativity constraint, Equation (4.4), is implied by the other three.

4.2.2 Mahalanobis Metric

Since practically all machine learning algorithms rely on some sort of distance calculation between individual data points, a prominent and widely used approach for improving classification results is Mahalanobis metric learning, where the idea is to learn a suitable distance function by exploiting the structure of the data. Such metrics form a whole class of distance functions that is based on the standard Mahalanobis distance, introduced by Mahalanobis in 1936 [56]. The Mahalanobis distance is one of the most fundamental distance measures for classification. In contrast to the Euclidean distance, it is scale-invariant and takes the correlations within the data set into account. Specifically, given n feature vectors $\mathbf{x}_i \in \mathbb{R}^d$ and assuming a Gaussian structure of the data, the Mahalanobis distance is obtained by first calculating the covariance matrix

$$\mathbf{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top \; , \tag{4.5}$$

and then computing the squared distance

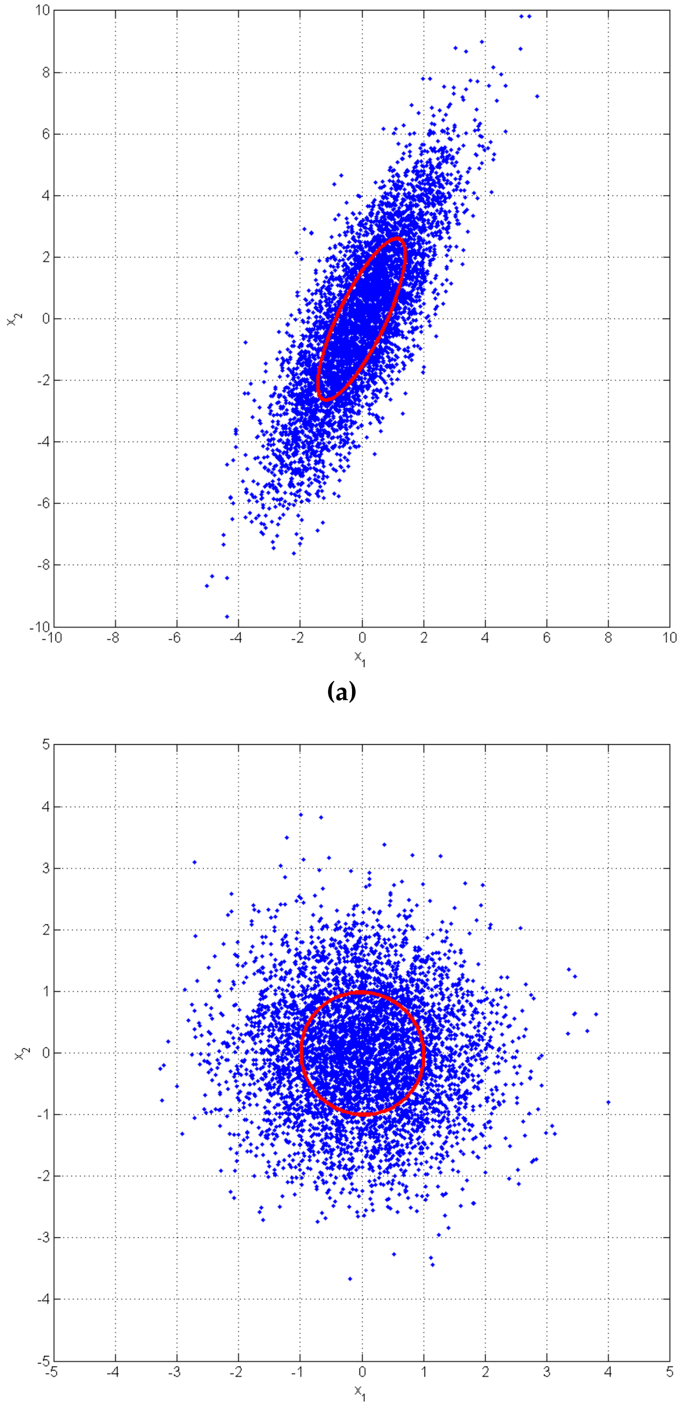
$$d^2_{\mathbf{\Sigma}^{-1}}(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{\Sigma}^{-1} (\mathbf{x}_i - \mathbf{x}_j) \tag{4.6}$$

between individual feature vectors, which is parametrized by the inverse of the covariance matrix. This is equivalent to first applying a whitening transformation to the data, yielding a set of points that are uncorrelated and have a variance of one, and then measuring Euclidean distances in the transformed space, as illustrated in Figure 4.1 for a two-dimensional example. While in the original space, points having the same Mahalanobis distance from the data's mean lie on ellipses that follow the Gaussian model, in the new space, these points lie on circles around the mean. This relation can also easily be seen from the covariance matrix of the transformed points, which equals the identity matrix. Hence, the Mahalanobis distance and the Euclidean distance are identical in the transformed space.

While using the Mahalanobis distance already improves classification performance if the data distribution is shaped properly, for some applications, even better results can be obtained by computing a more specific metric that is adapted to the problem at hand. Instead of just taking the distribution of data points into account, i.e., the generative structure, such metrics also incorporate discriminative information, e.g., in form of class labels or equivalence constraints. In particular, this is achieved by Mahalanobis metric learning, where the inverse covariance matrix in Equation (4.6) is replaced by a positive (semi-) definite matrix $\mathbf{M} \in \mathbb{R}^{d \times d}$ learned specifically for the given task:

$$d^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{M} (\mathbf{x}_i - \mathbf{x}_j)\ . \tag{4.7}$$

In our scenario, we want the distance between instances of the same person to become very small, while the distance between instances of different persons should become rather large. As can be seen from Equation (4.7), the distance function is parametrized by the matrix $\mathbf{M}$. Depending on the structure of this matrix, the influence on the distance $d^2_{\mathbf{M}}$ ranges from simple feature re-weighting in case of a diagonal matrix to a full linear transformation including scaling and rotation. By transforming the features appropriately, significant improvements for various visual classification tasks can be achieved compared to just using a standard distance. For example, if the Euclidean distance is utilized, then the largest feature dominates the computation, although that feature might not be important for classification at all. Note that in this context, i.e., feature space transformation, it is not overly important to compute a mathematically valid metric that fulfills all of the constraints defined in Equations (4.1) – (4.4), which would



(a)

(b)

Figure 4.1: Whitening transformation: (a) In the original space, the data points follow a Gaussian distribution, and points having the same Mahalanobis distance from the data's mean lie on ellipses. (b) In the transformed space, the data points are uncorrelated and have a variance of one. Hence, points that are equidistant from the center lie on circles. In both images, the unit distance from the center is highlighted in red.

require the matrix $\mathbf{M}$ to be positive definite ($\mathbf{M} \succ 0$). Indeed, in most applications, it is entirely sufficient to work with matrices that are only positive semi-definite ($\mathbf{M} \succeq 0$), hence, describe a pseudo-metric. A pseudo-metric is the most common generalization of a metric. It satisfies all metric axioms except for the first, i.e., the coincidence axiom given in Equation (4.1), which is substituted by

$$d_m(x_i, x_i) = 0 \,, \tag{4.8}$$

thus, requiring the distance between identical points to be zero, but on the other hand, allowing a distance of zero also between two different points. Put another way, points in a pseudo-metric space need not be distinguishable.

To reveal the feature transformation characteristic of metrics, an alternative, but more intuitive formulation for Equation (4.7) can be derived using matrix decomposition. Since the matrix $\mathbf{M}$ is assumed to be positive (semi-) definite, a factorization of the following form exists:

$$\mathbf{M} = \mathbf{L}\mathbf{L}^\top \,, \tag{4.9}$$

where $\mathbf{L} \in \mathbb{R}^{d \times d}$ can be computed via Cholesky decomposition, for example. If we now substitute $\mathbf{M}$ by $\mathbf{L}\mathbf{L}^\top$ in Equation (4.7), we obtain:

$$\begin{aligned} d^2_{\mathbf{L}}(\mathbf{x}_i, \mathbf{x}_j) &= (\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{L}\mathbf{L}^\top (\mathbf{x}_i - \mathbf{x}_j) \\ &= (\mathbf{L}^\top \mathbf{x}_i - \mathbf{L}^\top \mathbf{x}_j)^\top (\mathbf{L}^\top \mathbf{x}_i - \mathbf{L}^\top \mathbf{x}_j) \\ &= \|\mathbf{L}^\top \mathbf{x}_i - \mathbf{L}^\top \mathbf{x}_j\|_2^2 \\ &= \|\mathbf{L}^\top (\mathbf{x}_i - \mathbf{x}_j)\|_2^2 \,. \end{aligned} \tag{4.10}$$

As indicated by Equation (4.10), measuring distances under a certain metric $\mathbf{M}$ is equivalent to first projecting the data points into a transformed feature space using $\mathbf{L}$, i.e., $\mathbf{x}'_i = \mathbf{L}^\top \mathbf{x}_i$, and then measuring Euclidean distances there. Thus, either directly a positive (semi-) definite matrix $\mathbf{M}$ or a linear transformation $\mathbf{L}$ can be estimated from the data. While $\mathbf{L}$ uniquely defines $\mathbf{M}$, $\mathbf{M}$ defines $\mathbf{L}$ only up to rotation, which, however, has no influence on the calculation of distances. Note that if the transformation matrix $\mathbf{L}$ has full rank, it induces a positive definite matrix $\mathbf{M}$, i.e., a valid metric that fulfills all axioms defined in Equations (4.1) – (4.4), and a pseudo-metric otherwise.

In order to be able to actually learn a suitable metric, some sort of meta information describing the relations between individual data samples has to be provided, allowing to exploit not only the generative structure of the data, but also discriminative cues. Hence, in most visual classification tasks, additionally to each training sample $\mathbf{x}_i$ the corresponding class label y_i is given. However, for some problems, explicit class labels are not available or desired, so that other forms of meta information are needed. This also applies to person re-identification, where we typically learn from a training set consisting of image pairs $(\mathbf{x}_i, \mathbf{x}_j)$, each showing a specific person in two different camera views A and B, i.e., $y_i = y_j$. Although explicit person labels are provided in most cases, so that we could simply treat each image pair, i.e., person, as a separate class and formulate a multi-class problem, a more natural choice is to transform the original problem into a two-class problem, especially since usually no information about the inter-similarity between individual person pairs is given. Thus, we generate two sets of image pairs $\mathcal{S}$ and $\mathcal{D}$ such that each pair in $\mathcal{S}$ has two images describing the *same* person and each pair in $\mathcal{D}$ has two images describing *different* persons:

$$\mathcal{S} = \{(\mathbf{x}_i, \mathbf{x}_j) \mid i \in \mathcal{I}_A, j \in \mathcal{I}_B, y_i = y_j\} \;, \tag{4.11}$$

$$\mathcal{D} = \{(\mathbf{x}_i, \mathbf{x}_j) \mid i \in \mathcal{I}_A, j \in \mathcal{I}_B, y_i \neq y_j\} \;, \tag{4.12}$$

where $\mathcal{I}_A$ and $\mathcal{I}_B$ are the sets of sample indices of camera view A and B, respectively. Note that samples of $\mathcal{D}$ can easily be obtained by proper re-combination of the pairs given in $\mathcal{S}$. Clearly, for the task of person re-identification, the goal is to learn a metric that brings each of the image pairs in $\mathcal{S}$ (usually referred to as matching, similar, or positive pairs) closer together and increases the distance between each of the image pairs in $\mathcal{D}$ (usually referred to as non-matching, dissimilar, or negative pairs) at the same time. Since in our case, a person pair typically consists of two images captured by different cameras A and B, i.e., distances are measured between two specific camera views, in the following, we skip the camera notation for simplicity.

Having arranged image pairs using such pairwise equality and inequality constraints, we now transform the samples from the data space into the label-agnostic difference space

$$\mathcal{X} = \{\mathbf{x}_{ij} = \mathbf{x}_i - \mathbf{x}_j\} \;, \tag{4.13}$$

which is inherently given by the metric definitions in Equations (4.7) and (4.10). Also note that $\mathcal{X}$ is invariant to the actual sample location in the feature space. Furthermore, in order to increase readability in the following sections, we introduce the outer product matrix of pairwise differences

$$\mathbf{C}_{ij} = (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top \tag{4.14}$$

and the similarity variable

$$y_{ij} = \begin{cases} 1 & y_i = y_j \\ 0 & y_i \neq y_j \end{cases}, \tag{4.15}$$

where the latter indicates whether a given image pair $(\mathbf{x}_i, \mathbf{x}_j)$ is an element of $\mathcal{S}$ or $\mathcal{D}$.

4.2.3 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) [21], sometimes also referred to as Fisher LDA, is a linear dimensionality reduction technique which projects binarily labeled data samples from a d-dimensional space onto a one-dimensional subspace while preserving as much of the class discriminatory information as possible. Put another way, it computes that projection vector in the feature space which is most suited to discriminate data samples from two different classes, assuming that both are normally distributed and have equal covariance. In contrast to the unsupervised Principal Component Analysis (PCA) [62, 37], which finds the most accurate data representation in a lower-dimensional subspace by projecting the samples onto those directions having the largest variances, LDA exploits class label information, i.e., it works supervised. The difference between PCA and LDA is illustrated in Figure 4.2 for a two-dimensional example with two classes. As can be seen, the projection obtained by PCA captures most of the data variance, however, is quite inappropriate for classification. LDA, on the other hand, tries to find a representation where individual classes are compact and well separated from each other. Thus, it is much better suited for classification tasks.

In particular, assume we have a set of samples $\mathbf{x}_i \in \mathbb{R}^d$ and corresponding class labels $y_i \in \{1, 2\}$, i.e., a binary classification problem. Then the goal of LDA is to compute a projection $\mathbf{L}$ onto a line, i.e., $x'_i = \mathbf{L}^\top \mathbf{x}_i$, which is optimal for discriminating samples from the given classes. Hence, we first compute the within-class scatter matrix

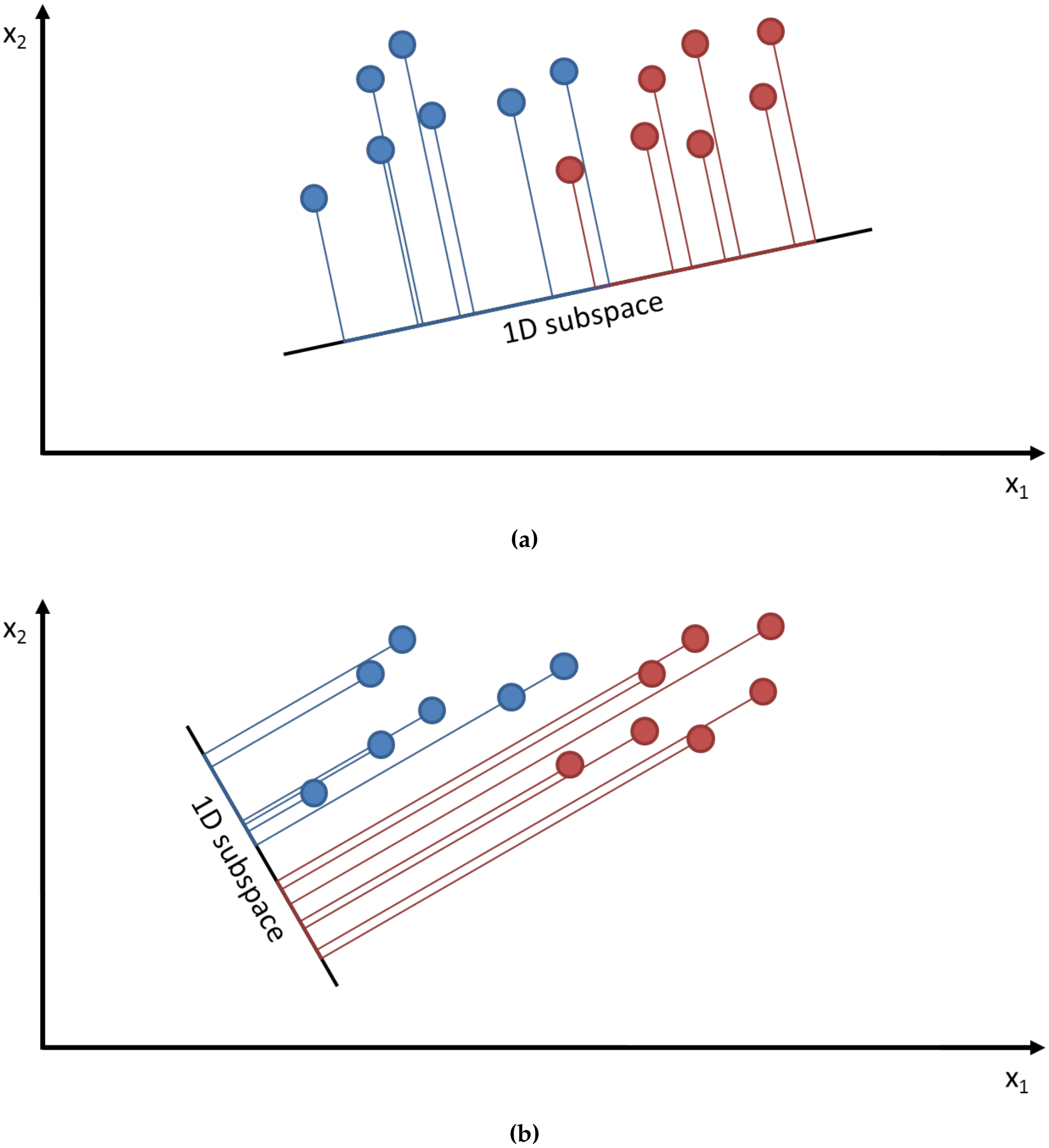


Figure 4.2: Difference between PCA and LDA: Given two-dimensional data samples from two classes 'blue' and 'red', the goal is to find a one-dimensional subspace, i.e., a line, to project onto. (a) Since PCA does not take labels into account, the obtained projection yields classes that are fairly spread out and even overlap. (b) In contrast, LDA computes a one-dimensional subspace where individual classes are as compact and separated from each other as possible.

$$\mathbf{S}_W = \sum_{c=1}^{2} \sum_{i:y_i=c} (\mathbf{x}_i - \boldsymbol{\mu}_c)(\mathbf{x}_i - \boldsymbol{\mu}_c)^\top \tag{4.16}$$

and the between-class scatter matrix

$$\mathbf{S}_B = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \tag{4.17}$$

of the data. Equation (4.16) expresses the scatter of the data samples around their respective class means $\boldsymbol{\mu}_c$, a measure that should be minimized by the optimal projection. On the other hand, the separation of the two class means defined in Equation (4.17) should be maximized. Thus, the optimal projection can be found by maximizing the ratio between $\mathbf{S}_B$ and $\mathbf{S}_W$ in the projected space, also referred to as Fisher-criterion:

$$z(\mathbf{L}) = \frac{|\mathbf{L}^\top \mathbf{S}_B \mathbf{L}|}{|\mathbf{L}^\top \mathbf{S}_W \mathbf{L}|} \; . \tag{4.18}$$

Typically, a solution that maximizes the Fisher-criterion is either obtained via solving the generalized eigenvalue problem

$$\mathbf{S}_B \mathbf{L} = \lambda \mathbf{S}_W \mathbf{L} \; , \tag{4.19}$$

or, if $\mathbf{S}_W$ has full rank, i.e., its inverse exists, by directly computing the eigenvectors of $\mathbf{S}_W^{-1}\mathbf{S}_B$.

Although Fisher's original formulation considers only two classes, there is a generalization to C classes. In this case, the goal is to compute at most $C-1$ projection vectors, i.e., a projection from the original space onto a discriminative subspace with up to $C-1$ dimensions. Naturally, Equations (4.16) and (4.17) have to be adapted accordingly:

$$\mathbf{S}_W = \sum_{c=1}^{C} \sum_{i:y_i=c} (\mathbf{x}_i - \boldsymbol{\mu}_c)(\mathbf{x}_i - \boldsymbol{\mu}_c)^\top \; , \tag{4.20}$$

$$\mathbf{S}_B = \sum_{c=1}^{C} n_c (\boldsymbol{\mu}_c - \boldsymbol{\mu})(\boldsymbol{\mu}_c - \boldsymbol{\mu})^\top \; , \tag{4.21}$$

where n_c is the number of samples of class c and $\boldsymbol{\mu}$ is the total mean of all data samples. However, it is known that the Fisher-criterion given in Equation (4.18) is only optimal in Bayes' sense for two classes (see, e.g., [53]).

Since we frame the task of person re-identification as a binary classification problem between matching and non-matching person image pairs anyway, we stick to the original LDA formulation, but replace the within-class and the between-class scatter matrix by two mean outer product matrices based on the sets defined in Equations (4.11) and (4.12):

$$\mathbf{C}_{\mathcal{S}} = \frac{1}{|\mathcal{S}|} \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{S}} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top , \tag{4.22}$$

$$\mathbf{C}_{\mathcal{D}} = \frac{1}{|\mathcal{D}|} \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{D}} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top . \tag{4.23}$$

Thus, we finally obtain the following modified version of the Fisher-criterion:

$$z(\mathbf{L}) = \frac{|\mathbf{L}^\top \mathbf{C}_{\mathcal{D}} \mathbf{L}|}{|\mathbf{L}^\top \mathbf{C}_{\mathcal{S}} \mathbf{L}|} , \tag{4.24}$$

which we want to maximize. In other words, we try to decrease the distance between similar and increase the distance between dissimilar pairs by projecting them onto the optimal line in the feature space. This can be regarded as a simple baseline method for our comparisons.

4.2.4 Logistic Discriminant Metric Learning

A similar idea is followed by Logistic Discriminant Metric Learning (LDML) introduced by Guillaumin et al. [32] for the task of face identification. They also seek a metric under which positive pairs, i.e., face images showing the same person, have a smaller distance than negative pairs. However, Guillaumin et al. tackle the problem from a probabilistic point of view. Thus, to estimate a Mahalanobis distance, the a posteriori probability p_{ij} that a pair $(\mathbf{x}_i, \mathbf{x}_j)$ belongs to the positive class is modeled as

$$p_{ij} = p(y_{ij} = 1 \mid \mathbf{x}_i, \mathbf{x}_j; \mathbf{M}, b) = \sigma(b - d^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j)) , \tag{4.25}$$

where y_{ij} is the similarity variable defined in Equation (4.15), $\sigma(x) = (1 + \exp(-x))^{-1}$ is a sigmoid function, and b is a bias term that directly serves as classification threshold and is learned together with the metric. Since Equation (4.25) describes a standard linear logistic model, $\mathbf{M}$ can be optimized by maximizing the log-likelihood

$$z(\mathbf{M}) = \sum_{ij} y_{ij} \ln(p_{ij}) + (1 - y_{ij}) \ln(1 - p_{ij}) \ . \tag{4.26}$$

The optimal solution is then obtained by iteratively updating the metric matrix $\mathbf{M}$ via gradient ascent in the following direction:

$$\frac{\partial z(\mathbf{M})}{\partial \mathbf{M}} = \sum_{ij} (y_{ij} - p_{ij}) \mathbf{C}_{ij} \ , \tag{4.27}$$

where $\mathbf{C}_{ij}$ is the outer product matrix of pairwise differences, as defined in Equation (4.14). Hence, the influence of each pair on the gradient direction depends on its current probability p_{ij} of being positive and the given label y_{ij}. In short, if a positive pair has a low probability, then an update is performed in direction of $\mathbf{C}_{ij}$, and if a negative pair has a high probability, then an update is performed in the negative direction of $\mathbf{C}_{ij}$.

Finally, note that in their experiments, Guillaumin et al. do not impose any further constraints on the problem. In particular, $\mathbf{M}$ is not required to be positive (semi-) definite. This makes their approach fairly prone to over-fitting the training data, especially in our scenario, which is somewhat ill-posed.

4.2.5 Information-Theoretic Metric Learning

In contrast to LDML, Information-Theoretic Metric Learning (ITML) presented by Davis et al. [15] provides a regularized solution, making it more robust to the difficulties typically encountered in person re-identification. Specifically, Davis et al. search for a Mahalanobis metric that trades off satisfying distance constraints against being close to a pre-defined metric. The key idea is that often certain knowledge about the task at hand is available, and that this knowledge can be incorporated into the solution via an appropriately chosen metric prior.

To achieve this, Davis et al. exploit the existence of a bijection between the set of Mahalanobis distances and the set of equal-mean, multivariate Gaussian distributions. More precisely, let $d^2_{\mathbf{M}}$ be a Mahalanobis distance, then its corresponding multivariate Gaussian is defined by

$$g(\mathbf{x}, \mathbf{M}) = \frac{1}{Z} \exp\left(-\frac{1}{2} d^2_{\mathbf{M}}(\mathbf{x}, \boldsymbol{\mu})\right) \ , \tag{4.28}$$

where Z is a normalization factor, $\boldsymbol{\mu}$ is the data's mean, and the covariance matrix is given by $\mathbf{M}^{-1}$. Thus, if they are represented as Gaussian distributions, the difference between individual Mahalanobis distance functions can be measured using the Kullback-Leibler divergence between their distributions. This leads to the following constrained optimization problem, where the goal is to minimize the relative entropy between the searched metric $\mathbf{M}$ and the chosen prior $\mathbf{M}_0$:

$$\min_{\mathbf{M}} \quad D_{KL}(g(\mathbf{x}, \mathbf{M}_0) \mid\mid g(\mathbf{x}, \mathbf{M})) \tag{4.29}$$

$$\text{s. t.} \quad d^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) \leq u \quad (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S} \ , \tag{4.30}$$

$$d^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) \geq l \quad (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{D} \ . \tag{4.31}$$

In the above formulation, D_{KL} denotes the Kullback-Leibler divergence, and the constraints in Equations (4.30) and (4.31) enforce that on the one hand, the distance between similar pairs is smaller than a given upper bound u, while on the other hand, the distance between dissimilar pairs is larger than a certain lower bound l.

As the information-theoretic objective defined in Equations (4.29) – (4.31) can be expressed via Bregman divergence [9], starting from $\mathbf{M}_0$, the Mahalanobis distance matrix $\mathbf{M}$ can be obtained by applying the following iterative update scheme:

$$\mathbf{M}_{t+1} = \mathbf{M}_t + \beta \mathbf{M}_t \mathbf{C}_{ij} \mathbf{M}_t \ , \tag{4.32}$$

where $\mathbf{C}_{ij}$ is defined according to Equation (4.14), and β is a projection parameter computed by the optimization algorithm encoding both, the pair label and the step size. For similar pairs, β is positive, causing an update in direction of $\mathbf{C}_{ij}$, and vice versa for dissimilar pairs.

4.2.6 Large Margin Nearest Neighbor

Weinberger et al. [80, 81] present Large Margin Nearest Neighbor (LMNN) classification, a distance learning approach that computes a Mahalanobis metric by optimizing over local neighborhoods in the feature space. In particular, for each instance, a local perimeter surrounding its k nearest neighbors sharing the same label (*target neighbors*) is established. Samples having a different label that invade this perimeter (*impostors*) are penalized. More technically, for a target pair $(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S}$, i.e., $y_{ij} = 1$, any sample $\mathbf{x}_l$ with $y_{il} = 0$ (which implies $y_{jl} = 0$) is an impostor if

$$d^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_l) < d^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) + 1 \,, \tag{4.33}$$

i.e., if it lies closer than the local neighborhood perimeter around $\mathbf{x}_i$ plus a constant margin. As illustrated in Figure 4.3, the goal of LMNN now is to learn a metric that makes the distance between samples and their target neighbors small while pushing impostors out of the local perimeters at the same time. This is realized via minimizing the following objective function:

$$z(\mathbf{M}) = \sum_{i,j \rightsquigarrow i} \left[d^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) + \mu \sum_{l} (1 - y_{il}) \xi_{ijl}(\mathbf{M}) \right] \,, \tag{4.34}$$

where $j \rightsquigarrow i$ indicates that $\mathbf{x}_j$ is a target neighbor of $\mathbf{x}_i$, μ is a weighting factor that trades off decreasing the distance between target neighbors against pushing impostors away, and the slack variable

$$\xi_{ijl}(\mathbf{M}) = [1 + d^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) - d^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_l)]_+ \tag{4.35}$$

denotes the amount by which an impostor $\mathbf{x}_l$ invades the perimeter of $\mathbf{x}_i$ and $\mathbf{x}_j$. In Equation (4.35), positivity is ensured using the hinge function $[x]_+ = \max(x, 0)$, however, at the cost of non-differentiability.

Nevertheless, since the objective function is convex, its subgradient can be computed, allowing to use standard descent algorithms in order to find its minimum and estimate a metric. Weinberger et al. apply an iterative solver, which reduces the objective function by taking one step at a time along the subgradient. After each step, the matrix $\mathbf{M}$ is projected back onto the cone of positive semi-definite matrices. Thus, at the t-th iteration, the gradient can be expressed as:

$$\left. \frac{\partial z(\mathbf{M})}{\partial \mathbf{M}} \right|_{\mathbf{M}^t} = \sum_{i,j \rightsquigarrow i} \mathbf{C}_{ij} + \mu \sum_{(i,j,l) \in \mathcal{N}^t} (\mathbf{C}_{ij} - \mathbf{C}_{il}) \,, \tag{4.36}$$

where $\mathbf{C}_{ij}$ is defined according to Equation (4.14), and $\mathcal{N}^t$ describes the set of triplet indices that correspond to a positive slack at iteration t, i.e., $\xi_{ijl}(\mathbf{M}^t) \; > \; 0$.

From the above description, it is apparent that the obtained metric is optimal for k nearest neighbors classification. However, as already mentioned in Section 2.4.1, Dikmen et al. [16] apply LMNN for the task of person re-identification based on image pairs, i.e., one nearest neighbor classification. They use a slightly modified version

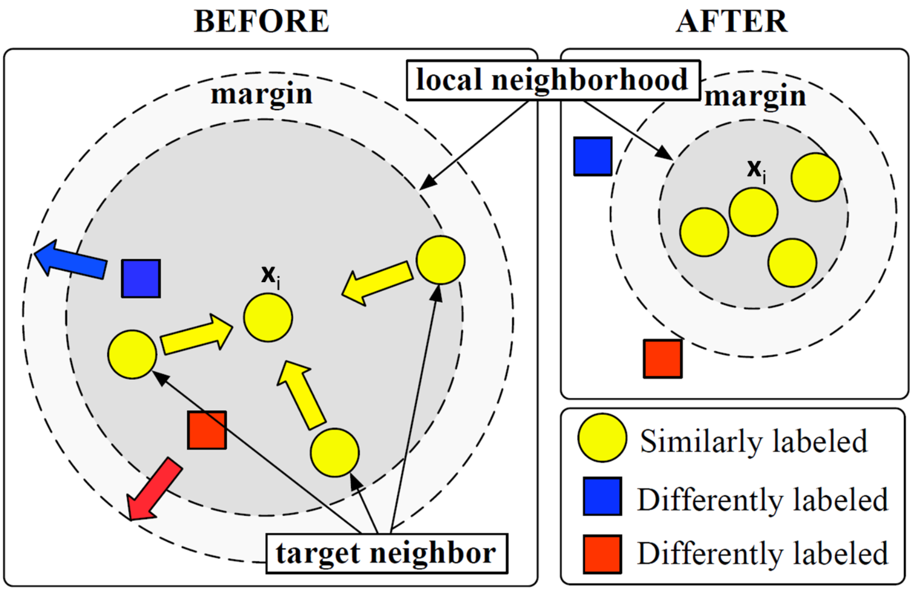


Figure 4.3: LMNN learning dynamics: The left image shows a sample $\mathbf{x}_i$ and its $k = 3$ target neighbors before training. After training, the target neighbors lie closer to the sample, while impostors have been pushed out of the local neighborhood, with a margin of at least one unit distance. The image stems from [80].

of LMNN that additionally contains a rejection option not returning a match if all nearest neighbors are beyond a certain threshold: LMNN-R. In particular, in order to be able to define a universal rejection threshold, they replace the original distance criterion that defines impostors locally, i.e., depending on the neighboring training samples, by the average distance of all k nearest neighbor pairs:

$$R = \frac{1}{nk} \sum_{i,j \rightsquigarrow i} d^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) \;, \tag{4.37}$$

where n is the number of training images in one camera view, and, since only image pairs are given, $k = 1$ in their experiments. Thus, LMNN-R computes the optimal metric in the same way as LMNN, however, using a slightly modified version of Equation (4.35):

$$\xi^R_{ijl}(\mathbf{M}) = [1 + R - d^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_l)]_+ \;. \tag{4.38}$$

4.2.7 Probabilistic Relative Distance Comparison

As mentioned in Section 2.4.2, Zheng et al. [87] propose a probabilistic metric learning formulation in order to tackle the person re-identification task. Motivated by the nature of the problem, i.e., large intra-class and inter-class variability as well as only very limited number of training samples, they use a relative distance comparison model, denoted as Probabilistic Relative Distance Comparison (PRDC). In particular, Zheng et al. seek a Mahalanobis distance that maximizes the probability of a matching person pair having a smaller distance than a non-matching one.

First, they define a pairwise set of difference vectors

$$\mathbb{O} = \{\mathbb{O}_i = (\mathbf{x}_i^p, \mathbf{x}_i^n)\} \; , \tag{4.39}$$

where $\mathbf{x}_i^p$ is a difference vector computed between a pair of relevant samples, i.e., samples of the same person, and $\mathbf{x}_i^n$ is a difference vector from a pair of related irrelevant samples. This means that one sample for computing $\mathbf{x}_i^n$ is equal to a sample for computing $\mathbf{x}_i^p$, but the other is a mismatch from another person. Given this pairwise set of difference vectors, the goal now is to learn a distance function d_m based on relative comparison, so that for each entry $\mathbb{O}_i$, the distance between the relevant sample pair is smaller than that between the related irrelevant pair, i.e., $d_m(\mathbf{x}_i^p) < d_m(\mathbf{x}_i^n)$. The probability for such an outcome is calculated as

$$P(d_m(\mathbf{x}_i^p) < d_m(\mathbf{x}_i^n)) = \frac{1}{1 + \exp(d_m(\mathbf{x}_i^p) - d_m(\mathbf{x}_i^n))} \; , \tag{4.40}$$

where the events of distance comparison between a relevant and an irrelevant pair are assumed to be independent. The optimal solution is then computed based on the maximum likelihood principle, i.e., by minimizing

$$z(\mathbf{L}) = -\log(\prod_{\mathbb{O}_i} P(d_m(\mathbf{x}_i^p) < d_m(\mathbf{x}_i^n))) \; . \tag{4.41}$$

Zheng et al. parametrize the function d_m using the Mahalanobis distance formulation as defined in Equation (4.10), however, with the additional constraint that the vectors of the transformation matrix $\mathbf{L} = [\mathbf{l}_1 \cdots \mathbf{l}_L]$ have to be pairwise orthogonal. Thus, after some conversions, they finally obtain the following conditioned minimization problem:

$$\min_{\mathbf{L}} \quad \sum_{\mathbb{O}_i} \log(1 + \exp(\|\mathbf{L}^\top \mathbf{x}_i^p\|_2^2 - \|\mathbf{L}^\top \mathbf{x}_i^n\|_2^2)) \tag{4.42}$$

$$\text{s. t.} \quad \mathbf{l}_i^\top \mathbf{l}_j = 0 \quad \forall \, i \neq j \; . \tag{4.43}$$

As the optimization criterion in Equation (4.42) may not be convex against the orthogonality constraint due to the involved relative distance comparison model, Zheng et al. propose an iterative optimization algorithm, which automatically reduces the complexity of the final solution by seeking a low-rank matrix.

4.2.8 Pairwise Constrained Component Analysis

As already outlined in Section 2.4.3, Mignon and Jurie [58] introduce Pairwise Constrained Component Analysis (PCCA), a method that tries to solve the person re-identification problem by mapping high-dimensional feature vectors to a low-dimensional space that is adapted to the given task. Their goal is to learn a projection such that in the new space, the Euclidean distance between matching person pairs becomes smaller than a certain threshold t, while it becomes larger than that threshold for non-matching pairs. As t only fixes the scale of the obtained distances, it can be simply set to $t = 1$.

Specifically, following the Mahalanobis distance formulation given in Equation (4.10), Mignon and Jurie seek an optimal transformation matrix $\mathbf{L}$ by minimizing

$$z(\mathbf{L}) = \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{S}} \ell_\beta(d^2_{\mathbf{L}}(\mathbf{x}_i,\mathbf{x}_j) - 1) + \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{D}} \ell_\beta(1 - d^2_{\mathbf{L}}(\mathbf{x}_i,\mathbf{x}_j)) \ , \tag{4.44}$$

where $\ell_\beta(x) = \dfrac{1}{\beta}\log(1+\exp(\beta x))$ is the generalized logistic loss function, which is a smooth approximation of the hinge loss.

To solve this optimization problem, a line search approach based on gradient descent is applied, with the gradient being given as

$$\frac{\partial z(\mathbf{L})}{\partial \mathbf{L}} = 2\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{S}} \sigma_\beta(1 - d^2_{\mathbf{L}}(\mathbf{x}_i,\mathbf{x}_j))\mathbf{L}\mathbf{C}_{ij} - 2\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{D}} \sigma_\beta(d^2_{\mathbf{L}}(\mathbf{x}_i,\mathbf{x}_j) - 1)\mathbf{L}\mathbf{C}_{ij} \ . \tag{4.45}$$

In the above formulation, $\sigma_\beta(x) = \dfrac{1}{1+\exp(-\beta x)}$ and $\mathbf{C}_{ij}$ is the outer product matrix of pairwise differences, as defined in Equation (4.14). The sharpness parameter β is fixed during optimization, and, together with the dimensionality of the output space, it is estimated using cross validation.

4.3 Efficient Metric Learning

In this section, we present new methods for computing metrics based on equivalence constraints that are particularly suited for person re-identification, but require much less computational power than the approaches presented in the previous sections (except for the very simple baselines given by the standard Mahalanobis distance and LDA). Specifically, we want to avoid complex optimization schemes that make algorithms such as LDML, ITML, LMNN (and LMNN-R), PRDC, and PCCA impracticable in realistic scenarios consisting of dozens of cameras. Furthermore, since we focus on efficient techniques, i.e., methods that are not over-sophisticated, the risk of over-fitting the training data is also reduced, resulting in better generalization ability. This is especially important considering the partially ill-posed nature of the problem at hand.

We start by formulating an objective function based on the definitions of similar and dissimilar person pairs given in Equations (4.11) and (4.12), respectively. Naturally, we want to learn a metric that minimizes the distance between similar person images, i.e., those stored in set $\mathcal{S}$, and pushes dissimilar ones, i.e., those stored in set $\mathcal{D}$, apart. This can be realized via minimizing

$$z(\mathbf{M}) = \frac{1}{|\mathcal{S}|} \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{S}} d^2_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_j) - \frac{1}{|\mathcal{D}|} \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{D}} d^2_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_j)\ , \tag{4.46}$$

where we normalize each term by the corresponding set cardinality to keep the influence of similar and dissimilar pairs balanced. As we are interested in efficient algorithms in order to increase the practical applicability of person re-identification systems, we seek a closed-form solution for computing $\mathbf{M}$. Thus, after expanding the objective function given in Equation (4.46) to

$$\begin{aligned} z(\mathbf{M}) = \frac{1}{|\mathcal{S}|} &\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{S}} (\mathbf{x}_i-\mathbf{x}_j)^\top \mathbf{M}(\mathbf{x}_i-\mathbf{x}_j) - \\ &\frac{1}{|\mathcal{D}|} \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{D}} (\mathbf{x}_i-\mathbf{x}_j)^\top \mathbf{M}(\mathbf{x}_i-\mathbf{x}_j)\ , \end{aligned} \tag{4.47}$$

we exploit Equation (4.9) and the fact that an inner product can be expressed in terms of a matrix trace, i.e.,

$$\begin{aligned}(\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{M} (\mathbf{x}_i - \mathbf{x}_j) &= (\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{L}\mathbf{L}^\top (\mathbf{x}_i - \mathbf{x}_j) \\ &= \langle \mathbf{L}^\top (\mathbf{x}_i - \mathbf{x}_j), \mathbf{L}^\top (\mathbf{x}_i - \mathbf{x}_j) \rangle \\ &= \mathrm{tr}((\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{L}\mathbf{L}^\top (\mathbf{x}_i - \mathbf{x}_j)) \\ &= \mathrm{tr}((\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{M} (\mathbf{x}_i - \mathbf{x}_j)) \ . \end{aligned} \tag{4.48}$$

Hence, we can re-write Equation (4.47) to:

$$\begin{aligned} z(\mathbf{M}) = &\frac{1}{|\mathcal{S}|} \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S}} \mathrm{tr}((\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{M} (\mathbf{x}_i - \mathbf{x}_j)) - \\ &\frac{1}{|\mathcal{D}|} \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{D}} \mathrm{tr}((\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{M} (\mathbf{x}_i - \mathbf{x}_j)) \ . \end{aligned} \tag{4.49}$$

Considering the invariance of the trace under cyclic permutations and its linearity property, i.e.,

$$\mathrm{tr}(\mathbf{ABC}) = \mathrm{tr}(\mathbf{CAB}) = \mathrm{tr}(\mathbf{BCA}) \ , \tag{4.50}$$

$$\mathrm{tr}(\mathbf{A} \pm \mathbf{B}) = \mathrm{tr}(\mathbf{A}) \pm \mathrm{tr}(\mathbf{B}) \ , \tag{4.51}$$

$$\mathrm{tr}(c\mathbf{A}) = c\, \mathrm{tr}(\mathbf{A}) \ , \tag{4.52}$$

where $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{R}^{d \times d}$ and $c \in \mathbb{R}$, Equation (4.49) can be further re-shaped to

$$
\begin{aligned}
z(\mathbf{M}) = \frac{1}{|\mathcal{S}|} &\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{S}} \mathrm{tr}(\mathbf{M}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top) - \\
&\frac{1}{|\mathcal{D}|} \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{D}} \mathrm{tr}(\mathbf{M}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top) \\
= {}& \mathrm{tr}\left(\frac{1}{|\mathcal{S}|} \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{S}} \mathbf{M}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top\right) - \\
&\mathrm{tr}\left(\frac{1}{|\mathcal{D}|} \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{D}} \mathbf{M}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top\right) \\
= {}& \mathrm{tr}\left(\mathbf{M}\,\frac{1}{|\mathcal{S}|} \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{S}} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top\right) - \\
&\mathrm{tr}\left(\mathbf{M}\,\frac{1}{|\mathcal{D}|} \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{D}} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top\right) . \qquad (4.53)
\end{aligned}
$$

Finally, using the definitions of $\mathbf{C}_\mathcal{S}$ and $\mathbf{C}_\mathcal{D}$ given in Equations (4.22) and (4.23), respectively, we can simplify Equation (4.53) to

$$
\begin{aligned}
z(\mathbf{M}) &= \mathrm{tr}(\mathbf{M}\mathbf{C}_\mathcal{S}) - \mathrm{tr}(\mathbf{M}\mathbf{C}_\mathcal{D}) \\
&= \mathrm{tr}(\mathbf{M}(\mathbf{C}_\mathcal{S} - \mathbf{C}_\mathcal{D})) . \qquad (4.54)
\end{aligned}
$$

However, in order to obtain a suitable metric when minimizing the objective function defined in Equation (4.54), constraints on $\mathbf{M}$ have to be imposed. Typically, the first such constraint requires $\mathbf{M}$ to be positive semi-definite, i.e., $\mathbf{M} \succeq 0$, ensuring a valid metric. Moreover, to regularize the metric and prevent it from collapsing to the trivial solution where all distances become zero, further constraints are necessary. Naturally, we are especially interested in those constraints that give rise to mathematically efficient solutions, as discussed in more detail in the next sections.

In particular, first, in Section 4.3.1, we examine the influence of the two contrary terms in Equation (4.54), i.e., $\mathbf{C}_\mathcal{S}$ and $\mathbf{C}_\mathcal{D}$, in order to find a solution that optimally balances the effects of positive and negative pairs. Second, in Section 4.3.2, we build on the thus gained insights and introduce an approach that follows the idea of LMNN, i.e., it optimizes a metric over local neighborhoods in the feature space. In contrast to the iterative optimization scheme used in LMNN, though, we derive a closed-form solution based on Equation (4.54), making our method very efficient not only during evaluation, but also during training.

4.3.1 Balancing the Influence of Positive and Negative Pairs

An important aspect when minimizing the objective function defined in Equation (4.54) is the influence of positive and negative pairs on the final metric. Special care has to be taken for the negative ones, as unbounded pair distances can easily degenerate the metric. Consequently, we add a balancing factor λ to Equation (4.54), which trades off the effect of positive against that of negative pairs:

$$z(\mathbf{M}) = \operatorname{tr}(\mathbf{M}(\mathbf{C}_{\mathcal{S}} - \lambda \mathbf{C}_{\mathcal{D}})) \ . \tag{4.55}$$

Based on the thus obtained objective function, we now formulate the following constrained optimization problem:

$$\begin{aligned} \min_{\mathbf{M}} \quad & \operatorname{tr}(\mathbf{M}(\mathbf{C}_{\mathcal{S}} - \lambda \mathbf{C}_{\mathcal{D}})) && (4.56)\\ \text{s. t.} \quad & \mathbf{M} \succeq 0 \ , && (4.57)\\ & \operatorname{tr}(\mathbf{M}) = 1 \ , && (4.58) \end{aligned}$$

where the first constraint ensures a positive semi-definite matrix $\mathbf{M}$, i.e., a valid metric, and the second one involving the matrix trace prevents the solution from collapsing to the trivial case where all distances equal zero. By converting the objective function such that it explicitly solves for a transformation matrix $\mathbf{L}$ according to Equation (4.10), i.e.,

$$z(\mathbf{L}) = \operatorname{tr}(\mathbf{L}^{\top}(\mathbf{C}_{\mathcal{S}} - \lambda \mathbf{C}_{\mathcal{D}})\mathbf{L}) \ , \tag{4.59}$$

positive semi-definiteness of $\mathbf{M} = \mathbf{L}\mathbf{L}^{\top}$ is inherently guaranteed, with the result that we can skip the first constraint. Furthermore, as the second condition defined in Equation (4.58) would lead to a rank-one solution, we replace it by a more sophisticated constraint that also prohibits the trivial case, but avoids rank-one solutions and additionally ensures uncorrelated directions in the new feature space. Thus, the optimization problem becomes

$$\min_{\mathbf{L}} \quad \operatorname{tr}(\mathbf{L}^\top(\mathbf{C}_\mathcal{S} - \lambda \mathbf{C}_\mathcal{D})\mathbf{L}) \tag{4.60}$$

$$\text{s. t.} \quad \mathbf{L}^\top \mathbf{L} = \mathbf{I}_n \;, \tag{4.61}$$

where $\mathbf{I}_n$ denotes the identity matrix of size $n \times n$. As can be seen, the parameter n allows controlling the rank of the obtained solution $\mathbf{L} \in \mathbb{R}^{d \times n}$ and thus the dimension of the resulting feature space. Usually, n is chosen to be much smaller than the original feature dimension d, i.e., $n \ll d$, in order to avoid over-fitting the training data.

The goal now is to compute the optimal balancing factor λ and transformation matrix $\mathbf{L}$, which can be achieved by finding the zero point of the trace difference function

$$f(\lambda) = \min_{\mathbf{L}^\top \mathbf{L} = \mathbf{I}_n} \operatorname{tr}(\mathbf{L}^\top(\mathbf{C}_\mathcal{S} - \lambda \mathbf{C}_\mathcal{D})\mathbf{L}) \;, \tag{4.62}$$

as shown by Guo et al. [33]. In particular, they propose an iterative algorithm that first adapts λ using the bisection method to find the zero point of $f(\lambda)$, and then calculates the best matrix $\mathbf{L}$ via an eigenvalue problem. The second step is discussed in more detail in Section 4.3.2. This approach works very well for person re-identification and converges within only a few iterations, making it much faster than most of the metric learning approaches described in Section 4.2.

Nevertheless, since we want to avoid any iterative techniques and directly provide a closed-form solution, we have to eliminate the balancing factor λ from Equation (4.55). In order to still be able to balance the influence of positive and negative pairs, though, we propose using a more advanced data sampling technique instead, which is presented in the next section.

4.3.2 Efficient Impostor-Based Metric Learning

In this section, we refine the metric learning method developed in the previous section to further increase its efficiency during training and thus its applicability to person re-identification in large-scale camera networks. Specifically, we want to get rid of the iterative procedure involved in balancing the influence of positive and negative pairs and directly come up with a closed-form solution. This can be achieved by an appropriate data sampling strategy, where only a meaningful subset of the negative person pairs is used in the optimization.

In particular, we do not want to compute a metric by minimizing the distance between matching and maximizing the distance between non-matching person pairs with brute force and regardless of a pair's influence on the final classification result. Instead, inspired by the ideas of LMNN classification, our goal is to estimate a metric using more advanced, local sampling techniques. A key insight here is that already well separable samples have only little influence on the optimization result and thus can be skipped. Hence, we restrict our method to take only meaningful samples in the feature space into account, i.e., those samples which actually provide information that can increase the classification performance. As a consequence, an additional benefit of the proposed approach is its improved generalization ability, since most of the unnecessary distance constraints are filtered out. Note that we still obtain a global metric that is constant across the entire feature space. However, the optimization itself is carried out locally, i.e., based on the local neighborhood around matching person image pairs. Due to its nature, we call our technique Efficient Impostor-Based Metric Learning (EIML).

To derive our metric, we stick to the concept of impostors introduced by Weinberger et al. in their LMNN formulation. Similar to the sets $\mathcal{S}$ and $\mathcal{D}$ containing image pairs describing the same person and image pairs describing different persons, respectively (Equations (4.11) and (4.12)), we define a new set $\mathcal{I}(i,j)$ based on the impostors that invade the perimeter of a target pair $(\mathbf{x}_i, \mathbf{x}_j)$ as follows:

$$\begin{aligned}\mathcal{I}(i,j) = \{(\mathbf{x}_i, \mathbf{x}_l) \mid \|\mathbf{x}_i - \mathbf{x}_l\|_2^2 < \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 + 1\} \cup \\ \{(\mathbf{x}_j, \mathbf{x}_l) \mid \|\mathbf{x}_j - \mathbf{x}_l\|_2^2 < \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 + 1\}\,,\end{aligned} \tag{4.63}$$

i.e., the perimeter is established around both samples of a target pair with an additional margin of one unit distance. As with the definitions of $\mathcal{S}$ and $\mathcal{D}$, image pairs are only constructed across two specific camera views A and B. Finally, the union of all such obtained local sets yields the overall impostor set

$$\mathcal{I} = \bigcup_{ij} \mathcal{I}(i,j)\,. \tag{4.64}$$

Note that in contrast to the iterative approach proposed by Weinberger et al., in our formulation, impostors are defined only once at the beginning, i.e., in the original feature space.

The goal now is to minimize the distance between similar person images, i.e., those stored in set $\mathcal{S}$, and push impostors out of the local perimeters at the same time. This can again be realized via minimizing the objective function defined in Equation (4.54), however, with a slight modification. In order to incorporate impostors, we replace the involved matrix $\mathbf{C}_{\mathcal{D}}$ by

$$\mathbf{C}_{\mathcal{I}} = \frac{1}{|\mathcal{I}|} \sum_{(\mathbf{x}_i, \mathbf{x}_l) \in \mathcal{I}} w_{il} (\mathbf{x}_i - \mathbf{x}_l)(\mathbf{x}_i - \mathbf{x}_l)^\top , \tag{4.65}$$

resulting in the following new objective function:

$$z(\mathbf{M}) = \mathrm{tr}(\mathbf{M}(\mathbf{C}_{\mathcal{S}} - \mathbf{C}_{\mathcal{I}})) . \tag{4.66}$$

Compared to Equation (4.55) defined in the previous section, no balancing parameter is needed in Equation (4.66) due to the more advanced data sampling via $\mathbf{C}_{\mathcal{I}}$, which allows deriving a closed-form solution. Furthermore, when computing the influence of impostors in Equation (4.65), we additionally apply a weighting factor

$$w_{il} = \exp\left(- \frac{\|\mathbf{x}_i - \mathbf{x}_l\|_2}{\|\mathbf{x}_i - \mathbf{x}_j\|_2} \right) , \tag{4.67}$$

which takes into account how much an impostor $\mathbf{x}_l$ invades the perimeter of a target pair $(\mathbf{x}_i, \mathbf{x}_j)$. Thus, impostors that already lie close to the border of the perimeter have less influence than impostors that lie close to the target neighbors themselves, as illustrated in Figure 4.4. In this way, we can mimic the behavior of the original LMNN algorithm, which consecutively updates the impostor set based on the current metric estimation, however, without using cumbersome, iterative optimization techniques. Instead, we come up with a closed-form solution.

For that purpose, we impose the same conditions on the solution as in Section 4.3.1, i.e., ensuring positive semi-definiteness of $\mathbf{M} = \mathbf{L}\mathbf{L}^\top$ by explicitly solving for a transformation matrix $\mathbf{L}$ according to Equation (4.10), and avoiding the trivial solution by requiring $\mathbf{L}^\top\mathbf{L} = \mathbf{I}_n$ to be fulfilled. Again, $\mathbf{I}_n$ denotes the identity matrix of size $n \times n$, so that the dimension of the resulting feature space can be controlled via the parameter n. This yields the following constrained optimization problem:

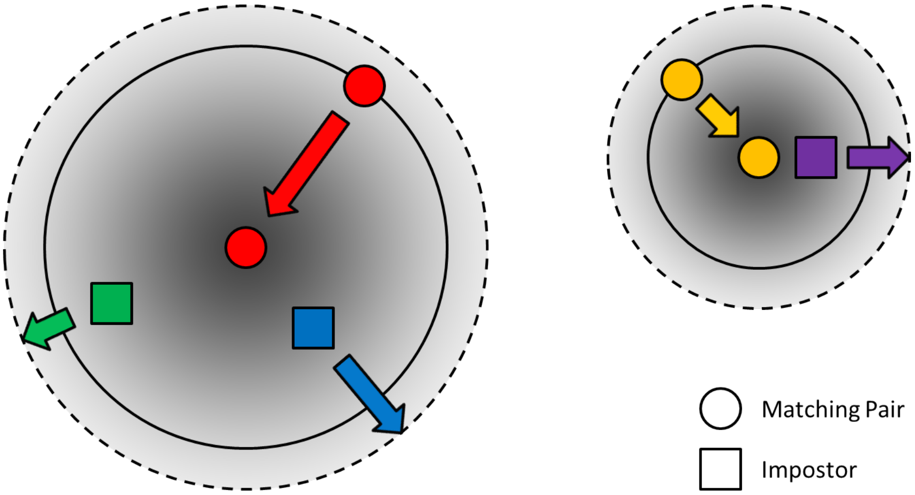


Figure 4.4: EIML learning dynamics: The distance between matching pairs should be minimized, while impostors should be pushed out of the local neighborhoods at the same time. In order to simulate the iterative behavior of the LMNN algorithm, impostors are weighted accordingly: The closer they lie to the perimeter's center, the higher is their weight, as indicated by the gray shading.

$$\min_{\mathbf{L}} \quad \mathrm{tr}(\mathbf{L}^\top(\mathbf{C}_\mathcal{S} - \mathbf{C}_\mathcal{I})\mathbf{L}) \tag{4.68}$$

$$\text{s. t.} \quad \mathbf{L}^\top\mathbf{L} = \mathbf{I}_n \ . \tag{4.69}$$

Since the matrix $(\mathbf{C}_\mathcal{S} - \mathbf{C}_\mathcal{I})$ is real and symmetric, having eigenvalues $\lambda_1 \leq \cdots \leq \lambda_d$ and associated orthonormal eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_d$, the optimal solution is given by

$$\mathbf{L} = [\mathbf{v}_1 \cdots \mathbf{v}_n] \ , \tag{4.70}$$

i.e., the eigenvectors corresponding to the n smallest eigenvalues of $(\mathbf{C}_\mathcal{S} - \mathbf{C}_\mathcal{I})$ (see, e.g., [55]). The related minimum cost equals $\lambda_1 + \cdots + \lambda_n$. This means that the obtained solution captures the n most informative directions according to the objective function defined in Equation (4.66). However, note that a transformation matrix $\mathbf{L}$ consisting of all d eigenvectors would only represent another orthonormal basis of the Euclidean space, which would rotate the feature space, but preserve the original distances. Thus, the result can be seen as a dimensionality reduction technique, where we project all data points onto a lower-dimensional subspace that is adapted to the problem at hand, i.e., a subspace which minimizes the distance between similar pairs stored in $\mathcal{S}$ and maximizes the distance between impostor pairs stored in $\mathcal{I}$. Furthermore, in contrast to other, more complex metric learning algorithms, the derived solution is quite robust to over-fitting due to the orthogonality constraint.

In summary, our approach exploits local information given by labeled training data similar to LMNN, i.e., based on matching and impostor pairs, in order to compute a metric suitable for person re-identification. In contrast to LMNN, though, the proposed method avoids complex optimization schemes by directly providing a closed-form solution, making it not only very efficient during evaluation, but also during training. As a consequence, it can effectively be used in even large-scale networks consisting of dozens of cameras, which is a big advantage over most other metric learning techniques that typically suffer from long training times.

4.4 Metric Learning Framework

In this section, we introduce our proposed metric learning framework for person re-identification consisting of three stages: feature extraction, metric learning, and classification. The overall system is illustrated in Figure 4.5. After computing the feature representation for all training samples, a metric between two cameras as defined in Equation (4.7) or (4.10) is estimated in order to facilitate inter-camera person queries. Once learned, the metric can be applied very efficiently when calculating the distance between a probe image and the gallery images stored in a database, which are then sorted accordingly. The three steps are discussed in more detail in the next sections.

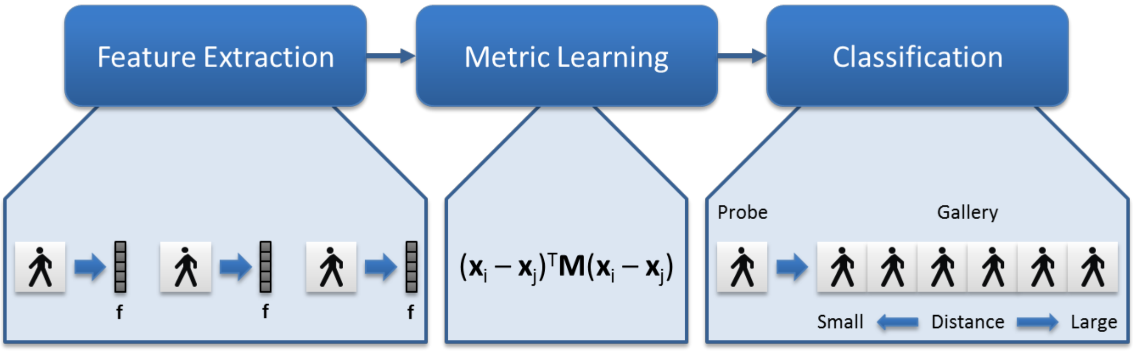


Figure 4.5: Overview of our metric learning framework consisting of three stages: First, color and texture features are densely sampled from the images. Second, a metric is learned in order to decrease the distance between similar and increase the distance between dissimilar person image pairs. Finally, in the classification step, the gallery images are ranked according to their distance to the probe sample under the learned metric.

4.4.1 Feature Representation

Color and texture features have proven to be successful for the task of person re-identification (see, e.g., [28, 79, 18, 31, 70, 64]). We use HSV and Lab color channels as well as Local Binary Patterns (LBPs) [60] to capture the appearance of a person. The features are extracted from 16 × 8 rectangular patches sampled from the image with a grid of 8 × 4 pixels, i.e., 50% overlap in both directions, as illustrated in Figure 4.6.

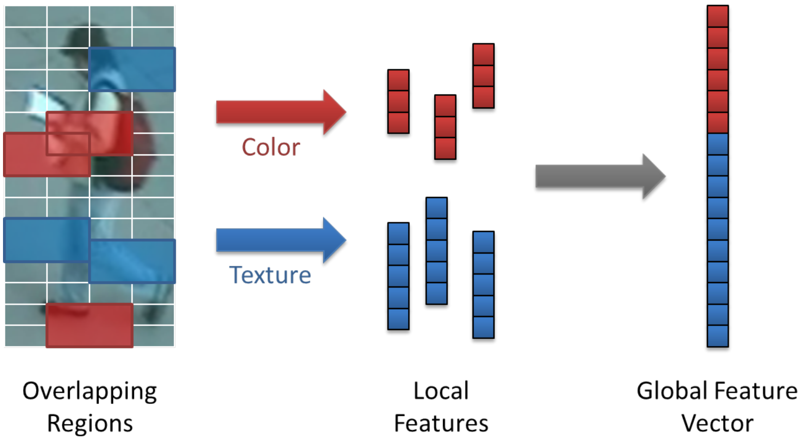


Figure 4.6: Dense sampling of features: Different local features (HSV, Lab, and LBP) are extracted from overlapping regions and then concatenated to a single feature vector. The sampling grid is laid over the image in white color.

In particular, within each rectangular patch, we calculate the mean values per color channel, where the value range is discretized to 0 to 40 bins in order to gain some illumination invariance. Additionally, a histogram of LBP codes is generated from a gray value representation of the patch. The obtained values are then put together to form a local feature vector. Finally, the vectors from all regions are concatenated to generate a single representation for the whole image. Depending on the image size, the overall feature vector can be quite high-dimensional, so that we run a PCA step to reduce the dimensionality and remove some noise. In general, this step is not overly critical. However, in order to avoid over-fitting, the number of principal components should be roughly adjusted according to the number of given training samples.

As can be seen, our feature representation covers the whole person image, i.e., it is similar to the holistic feature configurations typically encountered in descriptive person re-identification approaches. Unlike these methods, though, we employ a rather simple representation that captures only essential color and texture information and is not hand-crafted for the given task. Instead of requiring the features themselves to cope with all the difficulties arising from inter-camera appearance matching, we shift this task to the much more powerful concept of metric learning and show that even such a basic feature setup is sufficient for achieving state-of-the-art or better performance.

4.4.2 Metric Learning

In this step, a Mahalanobis metric is computed from labeled training data in order to obtain distances that are specifically suited for matching person appearances between the two involved camera views. Since all metric learning methods presented in this thesis are based on this principle, they can be simply plugged in at this stage with only little effort. Depending on the particular kind of optimization algorithm, either a transformation matrix **L** is calculated, which allows projecting all data points into a new feature space and measuring Euclidean distances there (Equation (4.10)), or a metric matrix **M** is obtained, which parametrizes the distance calculation during testing (Equation (4.7)). Thus, in both cases, applying the metric in the evaluation stage is very efficient, requiring only a matrix multiplication besides the actual cost for generating the gallery ranking.

4.4.3 Classification

In the classification step, the goal is to recognize a certain person across two different, non-overlapping camera views. In particular, a query person, i.e., a probe image, selected in the first camera view should be found in the second view containing gallery images. In other words, we want to classify each gallery image as either showing the same person as the probe image or not. However, as person re-identification is an extremely challenging task, such an evaluation procedure would be too restrictive in most practical scenarios. Thus, the hard classification scheme is usually replaced by a soft one, where the gallery images are ranked based on their feature distance to the probe sample under the learned metric. Images having a small distance are considered to be possible matches and appear in the front of the ranking, that part which is then presented to a human operator for further examination.

4.5 Discussion

In general, metric learning has shown great potential in many different fields of computer vision, e.g., stereo matching [85], face identification [32, 49], and clustering [84, 15].

In the specific case of person re-identification, it provides a very elegant tool for merging the descriptive and discriminative strategy typically encountered in this field of research. Based on a descriptive feature setup designed to capture the appearance of a person, the goal is to learn a metric that reflects the visual camera-to-camera transition. As the learned metric inherently emphasizes or attenuates directions in the feature space depending on their capability to discriminate between matching and non-matching person pairs, it can also be seen as a discriminative feature selector.

Another advantage of metric learning approaches is their efficiency during evaluation, since additionally to the feature extraction and the matching, only linear projections have to be computed. However, since usually one metric has to be learned between each pair of cameras in a network, the training time is very important too. Unfortunately, most of the existing metric learners are not particularly targeted at the task of person re-identification and build on computationally complex optimization schemes, which severely limits their practical applicability. In contrast, our more efficient algorithms introduced in this chapter require only a fraction of the training time of established methods such as LDML, ITML, LMNN (and LMNN-R), PRDC, and PCCA. Moreover, by avoiding over-sophisticated optimization techniques, we also reduce the risk of overfitting the training data, resulting in better generalization ability. This is especially important considering the somewhat ill-posed nature of the task at hand.

Finally, in order to verify our strategy, we show results on various benchmark datasets, where we achieve state-of-the-art or better performance at much lower computational costs compared to other person re-identification approaches. Detailed results are presented in the next chapter, where a thorough evaluation of all methods is given, including those of Chapter 3.

5 Experimental Results

Contents

5.1 Overview

In this chapter, we evaluate and compare the person re-identification approaches presented in this thesis. In particular, we show results on five publicly available and widely used benchmark datasets, each focusing on different aspects of the problem at hand, i.e., recognizing person appearances across disjoint camera views. For analysis, we follow the common procedure of searching different probe samples in a set of gallery images, each time recording the ranking performance of a method. This evaluation procedure is perfectly suited to measure the practical applicability of an approach: The lower the rank of the correct match is on average, the less potential candidates have to be presented to a human operator in a real world system, thus, decreasing search effort and time. Moreover, besides the pure recognition scores, we also compare the runtimes of the individual methods, as this is an important issue as well. Especially in large-scale camera networks, computation time plays a crucial role and can be the deciding factor whether or not to use a certain system.

The remainder of this chapter is structured as follows. First, in Section 5.2, we introduce the performance measure typically used to analyze the matching capability of person re-identification methods. After that, in Section 5.3, we describe the employed datasets and their individual characteristics. Finally, in Section 5.4, the obtained results are presented and discussed. In particular, we examine our combined descriptive and discriminative person re-identification system introduced in Chapter 3 as well as the metric learning approaches from Chapter 4.

5.2 Evaluation Measure

In order to evaluate the recognition performance of person re-identification systems, Cumulative Matching Characteristic (CMC) curves are commonly used (see, e.g., [79, 30, 31, 52, 70]). They provide a summary of the ranking behavior obtained from several person queries carried out during the test phase. Specifically, a CMC curve represents the expectation of the true gallery match being found within the first r ranks, i.e., the matching rate versus the number of ranks shown to a user. An exemplary CMC curve is shown in Figure 5.1.

To generate such a curve, we start with a flat line where all values are zero. Then, after each test query, the rank of the matching gallery image r_{match} is reported, and all curve values corresponding to a rank greater than or equal to r_{match} are increased by one. Finally, the thus obtained curve is normalized by the number of performed test queries, yielding a matching rate in percentage terms. From this description, it is clear that a CMC curve is always monotonically increasing, and that the maximum matching rate of 100% is reached at least for the last rank, if the number of considered ranks equals the size of the gallery set.

As a CMC curve indicates the average matching rate for any desired number of ranks, it is perfectly suited to evaluate and compare person re-identification systems. The higher the matching rate is for lower rank numbers, the better is the overall performance. In the ideal case, a system would achieve a matching rate of 100% already for the first rank, meaning that no ranking of potential matches would have to be generated, since the system would actually be capable of re-detecting a person within a camera network. However, such a high level of performance can hardly be achieved under realistic conditions given the challenges that come along with the re-identification task. Thus, the ranking-based evaluation strategy is much more practical. For example, if a certain mean matching rate is required, the information about how many ranks have to

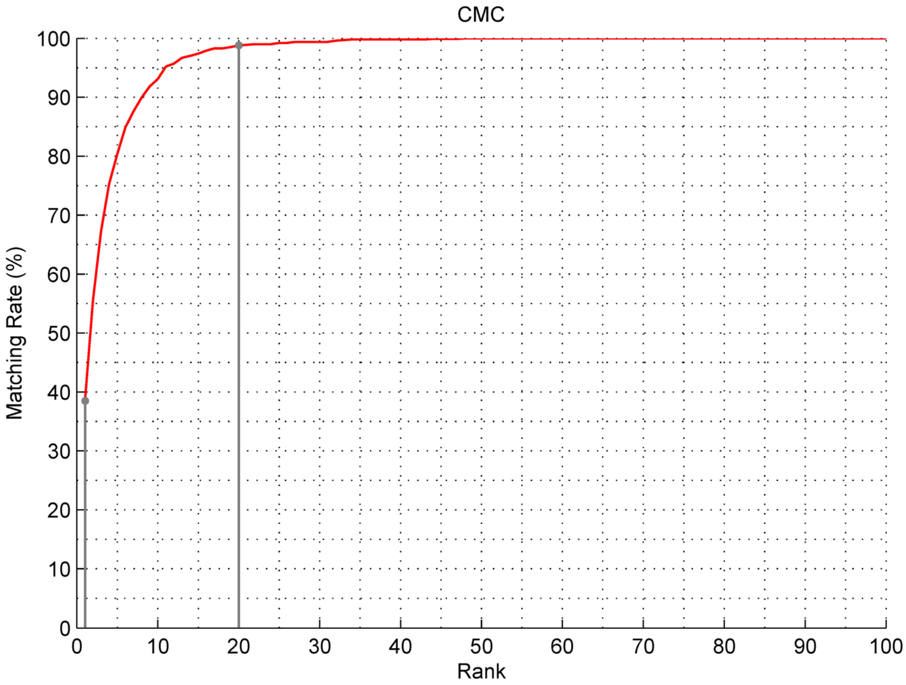


Figure 5.1: Exemplary CMC curve: In the above example, the considered system achieves a rank one matching rate of about 38%. Thus, if just the best match is shown to the human operator, the correct person is only found in 38% of the cases. However, if the system is extended to present the top 20 ranks to the user, the correct match is among them in 99% of the cases.

be examined by a human operator, i.e., how much effort and time have to be spent, can easily be read from a CMC curve.

5.3 Datasets

In the following, we give an overview of the datasets used in our evaluation and explain the corresponding setups. In particular, we run experiments on *VIPeR* [30], the de facto standard benchmark dataset for single-shot methods, *PRID 2011* [35], which provides a single-shot and a multi-shot setup, *PRID 450S* [67], which is single-shot only but additionally provides part-level segmentations, *ETHZ* [70], a commonly used but not really task-specific dataset, and finally *CAVIAR4REID* [11], a multi-shot dataset. The selected datasets are publicly available and widely used for benchmarking person re-identification methods, making them the perfect choice for this study.

Note that also other datasets have been proposed for evaluating person re-identification approaches. However, most of them are either not very realistic or kept under restricted access, thus, limiting their benefit. For example, some authors test their methods on person patches cropped from the *i-LIDS* dataset [76], which is not publicly available. Moreover, there exist different versions of the thus obtained subsets, so that it is difficult to give fair comparisons.

5.3.1 VIPeR

The VIPeR dataset [30] contains 632 person image pairs taken from two different camera views, i.e., it provides a single-shot setup. Each image has a size of 48 × 128 pixels. Changes in viewpoint, illumination, and pose are the most prominent sources of appearance variation between the two instances of a person, as can be seen from the examples presented in Figure 5.2. These challenges make person re-identification very difficult on this dataset.

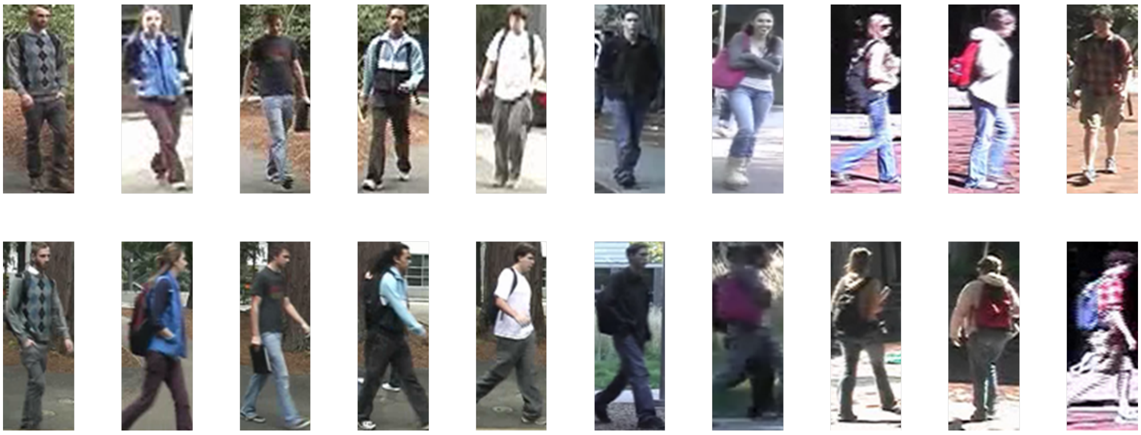


Figure 5.2: Examples from the VIPeR dataset: Most of the image pairs contain a viewpoint change of about 90 degrees as well as significant changes in illumination and pose. Upper and lower row correspond to appearances of the same person in different camera views.

For evaluation, we follow the procedure described in [31]. The set of 632 image pairs is randomly split into two sets of 316 image pairs each, one for training and one for testing. In the test case, the two images of an image pair are randomly assigned to the probe and the gallery set. A single image from the probe set is then selected and compared to all images from the gallery set. This process is repeated for every image in the probe set in order to generate a CMC curve. Finally, the whole evaluation procedure is carried out multiple times, and the average result is reported.

5.3.2 PRID 2011

The PRID 2011 dataset [35] consists of person images recorded from two different static surveillance cameras. Since these images have been extracted from trajectories, additionally to the single-shot case, also a multi-shot scenario providing roughly 50 to 100 images per person and camera view is available. The raw person images are typically 100 to 200 pixels tall, depending on their location within the overall scene image that has a resolution of 720 × 576 pixels. However, for maximum usability, they have been re-scaled to a uniform size of 64 × 128 pixels. Characteristic challenges on this dataset are viewpoint and pose changes as well as significant differences in illumination, background, and camera characteristics. Some exemplary images are depicted in Figure 5.3.

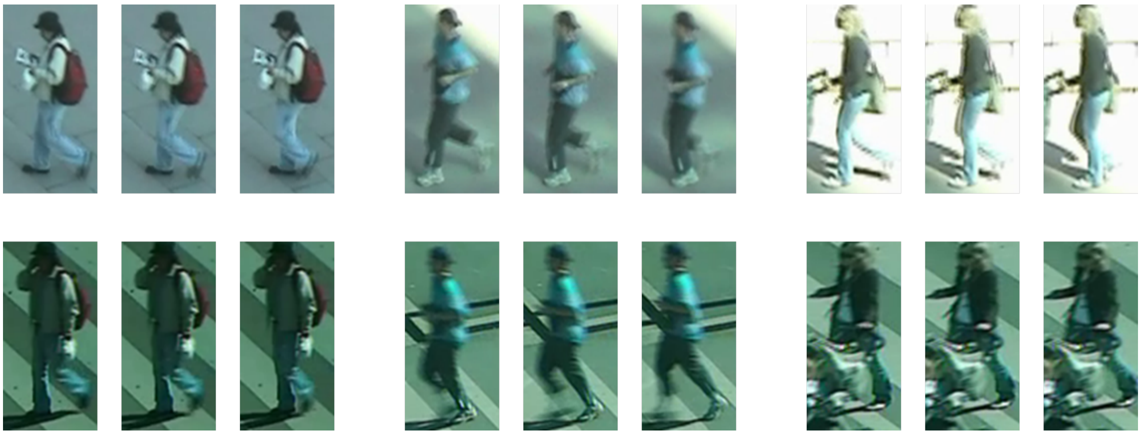


Figure 5.3: Examples from the PRID 2011 dataset: The images depict the multi-shot case, i.e., they show a subset of the trajectory images of three persons. Upper and lower row correspond to appearances of the same person in different camera views.

Camera view A contains 385 persons, camera view B contains 749 persons, with 200 of them appearing in both views. Hence, there are 200 person image pairs in the dataset. These image pairs are randomly split into a training and a test set of equal size. For evaluation on the test set, we follow the procedure described in [35], i.e., camera A is used for the probe set and camera B is used for the gallery set. Thus, each of the 100 persons in the probe set is searched in a gallery set of 649 persons (all images of camera view B except the 100 training samples). Again, the whole procedure is repeated several times and the result is reported in form of an average CMC curve.

5.3.3 PRID 450S

The PRID 450S dataset [67] builds on PRID 2011 and, as a consequence, shows similar characteristics. However, it is arranged according to the VIPeR dataset by image pairs and contains more linked samples than PRID 2011. In particular, the dataset contains 450 single-shot image pairs depicting walking humans captured by two spatially disjoint cameras. In addition, for each image, a part-level segmentation is provided containing the following regions: head, torso, legs, carried object at torso level (if any), and carried object below torso (if any). The union of all these parts yields a foreground segmentation of the whole person. Exemplary images and corresponding segmentations are illustrated in Figure 5.4.

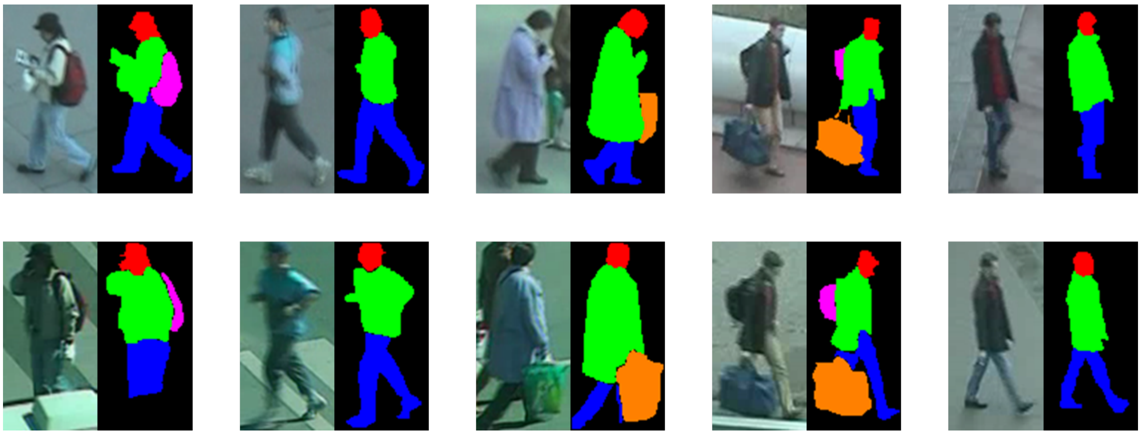


Figure 5.4: Examples from the PRID 450S dataset: Person images and part-level segmentations are shown next to each other. Upper and lower row correspond to appearances of the same person in different camera views.

The evaluation protocol for this dataset follows that used for VIPeR, i.e., the set of image pairs is randomly split into a training and a test set of equal size. Hence, to compute a CMC curve, 225 probe images are searched in a gallery containing also 225 samples. Since the camera setup is very similar to PRID 2011, we apply the same strategy when forming the probe and the gallery set, i.e., the probe images are taken from camera view A and the gallery images are taken from view B. Finally, as for the other datasets, the result is averaged over different training and test set splits.

5.3.4 ETHZ

The ETHZ dataset, originally proposed for pedestrian detection [17] and later modified for benchmarking person re-identification methods [70], consists of person images extracted from three video sequences of urban scenes captured by a moving camera. Thus, it provides multiple images per person, i.e., a multi-shot setup, which is structured as follows: SEQ. #1 contains 83 persons (4,857 images), SEQ. #2 contains 35 persons (1,961 images), and SEQ. #3 contains 28 persons (1,762 images). All images have been re-sized to 32 × 64 pixels. The most challenging aspects of the ETHZ dataset are illumination changes, occlusions, and a very low image resolution. However, since the person images stem from a single moving camera, the dataset does not provide a realistic scenario for person re-identification with multiple disjoint cameras, different viewpoints, different camera characteristics, etc. Figure 5.5 shows some exemplary images.

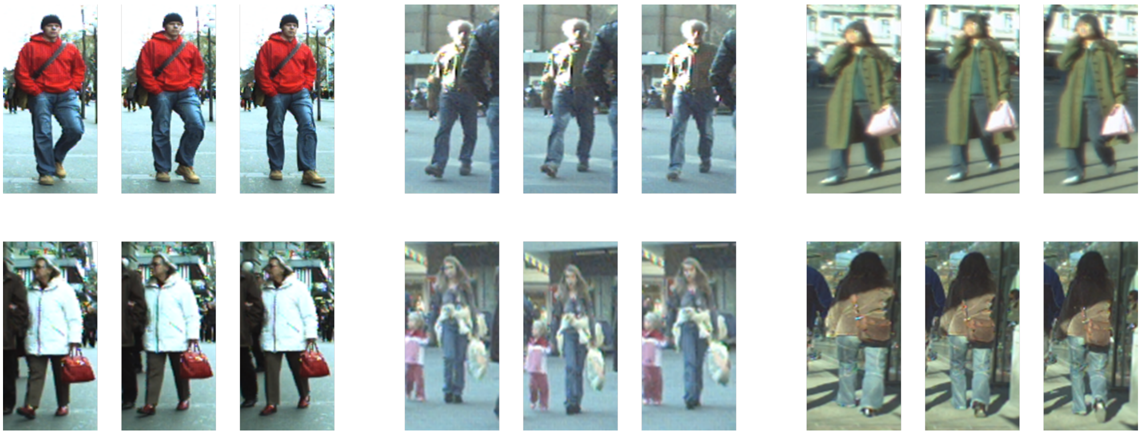


Figure 5.5: Examples from the ETHZ dataset: The left column shows images of two persons from SEQ. #1, the middle column corresponds to SEQ. #2, and the right column corresponds to SEQ. #3.

Despite this limitation, it is commonly used for evaluating person re-identification approaches, so we also perform tests on this dataset. Similar to [70] and [18], we use a single-shot evaluation strategy, i.e., we randomly sample two images per person to build a training pair, and another two images to build a test pair. The images of the test pairs are then assigned to the probe and the gallery set. In order to compute an average CMC curve, the experiment is repeated multiple times.

5.3.5 CAVIAR4REID

The CAVIAR4REID dataset [11] contains images of 72 individuals captured by two different cameras in a shopping center, where the original images have been re-sized to 64 × 128 pixels. 50 of them appear in both camera views, the remaining 22 only in one view. Since we are interested in person re-identification in different cameras, we only use individuals appearing in both views in our experiments. Each person is represented by 10 instances per camera view, yielding a multi-shot scenario. Typical challenges on this dataset are viewpoint and pose changes, different illumination conditions, different camera characteristics, occlusions, and a low image resolution. The resulting appearance variations can be seen very well from the examples depicted in Figure 5.6.

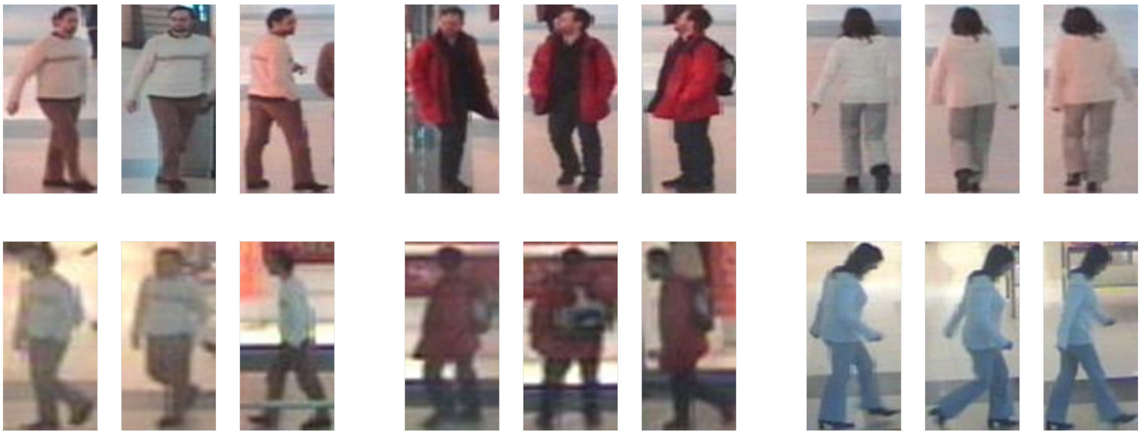


Figure 5.6: Examples from the CAVIAR4REID dataset: Especially the differences in lighting, camera characteristics, and the low resolution make person re-identification very challenging on this dataset. Upper and lower row correspond to appearances of the same person in different camera views.

To compare different methods, we use a multi-shot evaluation strategy similar to [4]. The set of 50 persons is randomly split into a training set of 42 persons and a test set of 8 persons. Since every individual is represented by 10 images per camera view, we can generate 100 different image pairs between the views of two individuals. During training, we use all possible combinations of positive pairs showing the same person and negative pairs showing different persons. When comparing two individuals in the evaluation stage, we again use all possible combinations in order to calculate the mean distance between the two persons. As for the other datasets, the final result is computed by averaging over different training and test set splits.

5.4 Results

In this section, we present the results and discuss the insights obtained after performing experiments on the datasets described in Section 5.3. Specifically, first, in Section 5.4.1, our combined descriptive and discriminative person re-identification system introduced in Chapter 3 is evaluated on two datasets, one representing the single-shot and the other representing the multi-shot case. The second part, Section 5.4.2, is then devoted to the analysis of the metric learning methods presented in Chapter 4. Using our metric learning framework, we are able to run all approaches under equal conditions, i.e., applying the same training and test set splits, the same features, etc.

5.4.1 Combined Descriptive and Discriminative System

Even though our combined system is targeted at multi-shot scenarios, we first run a single-shot experiment on the VIPeR dataset in order to give an estimate of the performance under such conditions. For the second case, i.e., the multi-shot experiment, we use PRID 2011, since this dataset provides a very realistic setup containing whole person trajectories.

Note that in this section, we just present results achieved with our system. For a comparison to other methods that also tackle the person re-identification problem, we refer the reader to Section 5.4.2.

5.4.1.1 Single-Shot Evaluation

As described in Section 5.3.1, the evaluation protocol for VIPeR involves splitting the data into a training and test set of equal size. Thus, despite the fact that our system does not require a dedicated training set, so that we could use all samples for testing, we evaluate it only on a subset of 316 randomly selected image pairs. This allows a fair comparison to other methods that also follow the common protocol.

Moreover, since the involved discriminative model is based on automatically gathered positive and negative training data, where positive instances are usually obtained from multi-shot person trajectories, an alternative sampling strategy is needed in the single-shot case. Hence, we generate virtual samples by applying geometric transformations and displacements to the single positive sample that is available. However, it is clear that this strategy can only mimic a true multi-shot configuration, but not provide the same rich information. Further details about the sampling of training instances for the discriminative model can be found in Section 3.2.

Figure 5.7 and Table 5.1 show the performance of the proposed system averaged over five runs on randomly selected subsets of the VIPeR dataset. In particular, three results are presented: one for the descriptive model, one for the discriminative model, and one simulating the combined system by picking that model for each query that returns the better result, i.e., the lower matching rank. As can be seen, both models show similar performance, with a negligible advantage for the discriminative one. This quite small gap is not really surprising, though, as the power of the discriminative model comes from its ability to extract distinct features given diverse samples of the same individual, i.e., a multi-shot setup. However, more important than the performance of the individual models is the outcome of their combination. While the absolute numbers are not directly comparable to other methods, since we only simulate the system behavior by always choosing the better model, the key insight here is that both models capture different aspects of the appearance of a person, and that their combination can lead to superior performance compared to using just one of them. Thus, integrating both models into a single system is beneficial for the task of person re-identification, especially if they are arranged in an application-focused way as in our setup. As described in Section 3.1, we first run the fast, descriptive person model in order to quickly compute an initial ranking of the gallery images, and only if necessary, the computationally more complex, discriminative model is initiated.

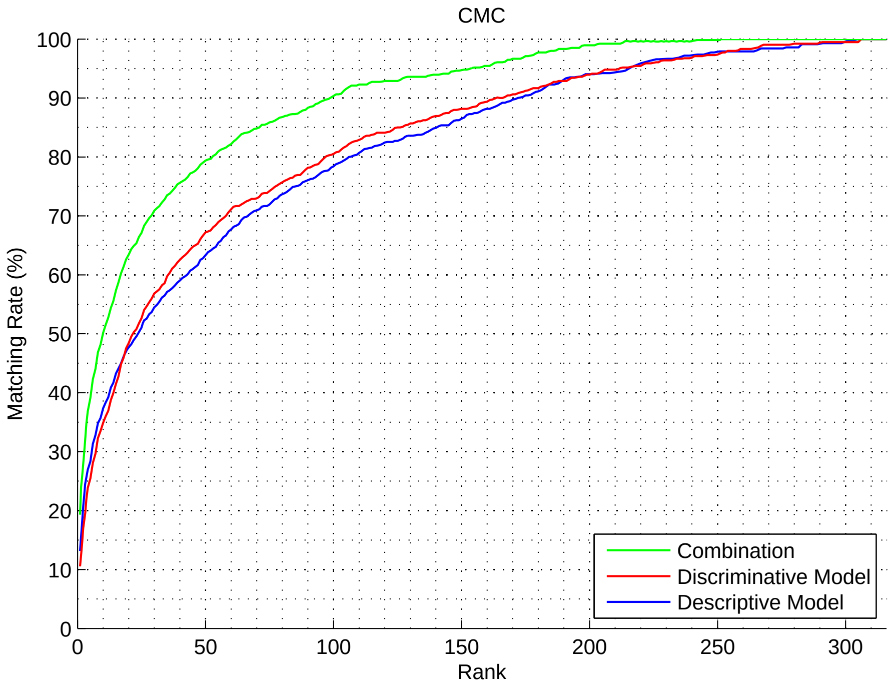


Figure 5.7: Average CMC curves of our system on VIPeR: Although both models show similar matching rates, their combination achieves superior performance compared to using just one of them.

Finally, we also give a performance comparison of different feature types that can be applied in the discriminative stage. While the focus of the descriptive model lies on

Method	$r = 1$	10	20	50	100
Combination	19	50	63	79	90
Discriminative Model	11	35	48	67	81
Descriptive Model	13	37	48	63	78

Table 5.1: Average matching rates (%) for different ranks r on VIPeR.

features that can be computed efficiently in order to quickly generate a first ranking of the gallery, in the discriminative step, diverse feature types can be applied. As already mentioned, many different kinds of features have been proposed for the application with boosting for feature selection (see, e.g., [51, 71, 50, 40, 61]). For our specific task, we propose using Haar-like features and color-based covariance descriptors. In the following, we illustrate that exactly these features are best suited for person re-identification by evaluating different feature types on a subset of image pairs of the VIPeR dataset. In particular, we examine horizontally divided Haar-like features, Histograms of Oriented Gradients (HOGs) [14], Local Binary Patterns (LBPs) [60], covariance features using RGB channels (sigma points), as well as their combinations. The obtained results are depicted in form of CMC curves in Figure 5.8. It can clearly be seen that color, captured by covariance features, is the strongest cue, followed by Haar-like features, which particularly capture intensity changes between the upper and lower body of a person. HOGs and LBPs, on the other hand, perform rather poorly, since they concentrate on finer structures which can easily lead to over-fitting the learned model to the training data. In fact, the best performance is achieved using a combination of Haar-like and covariance features.

5.4.1.2 Multi-Shot Evaluation

In this section, we present the results obtained with our system in a multi-shot configuration, which is its intended use case. Specifically, we perform experiments on the multi-shot version of PRID 2011, which has been captured under realistic conditions. On this dataset, positive samples for learning the discriminative model can easily be extracted from the trajectory of the searched person. However, in order to get some additional variation into the positive training set, we also generate a few virtual samples, as for the single-shot case. Concerning the features applied in the discriminative step, we use the same setup as for the VIPeR dataset.

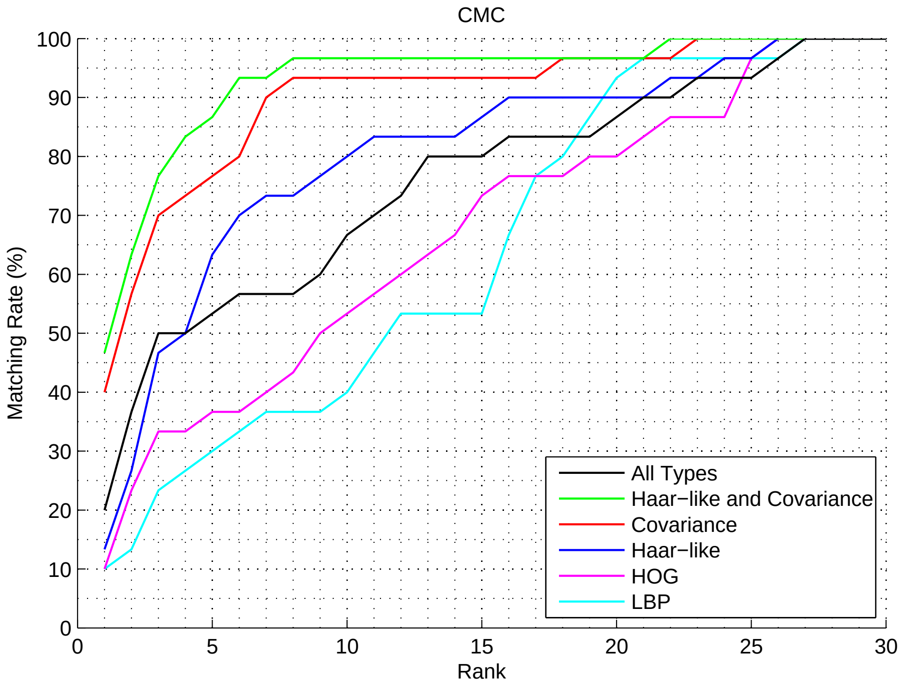


Figure 5.8: Comparison of different feature types for the discriminative model on a subset of VIPeR: As can be seen, a combination of Haar-like and covariance features yields the best result.

Method	$r = 1$	10	20	50	100
Combination	19	52	67	85	94
Discriminative Model	17	41	53	71	85
Descriptive Model	4	24	37	56	70

Table 5.2: Average matching rates (%) for different ranks r on PRID 2011.

Figure 5.9 and Table 5.2 show the average results of our approach on this dataset after three runs. As shown by the curves, in contrast to the VIPeR image pairs, the discriminatively learned model clearly outperforms the descriptive model here. This can be explained by the increased variability that is captured if positive training samples are extracted from whole trajectories. Additionally, the risk of over-fitting the training data is reduced. Finally, like on the VIPeR dataset, taking into account both models, i.e., descriptive and discriminative information, leads to superior performance.

5.4.2 Metric Learning

For the evaluation of the various metric learning approaches presented in Chapter 4, we apply the framework introduced in Section 4.4. Specifically, we examine Logis-

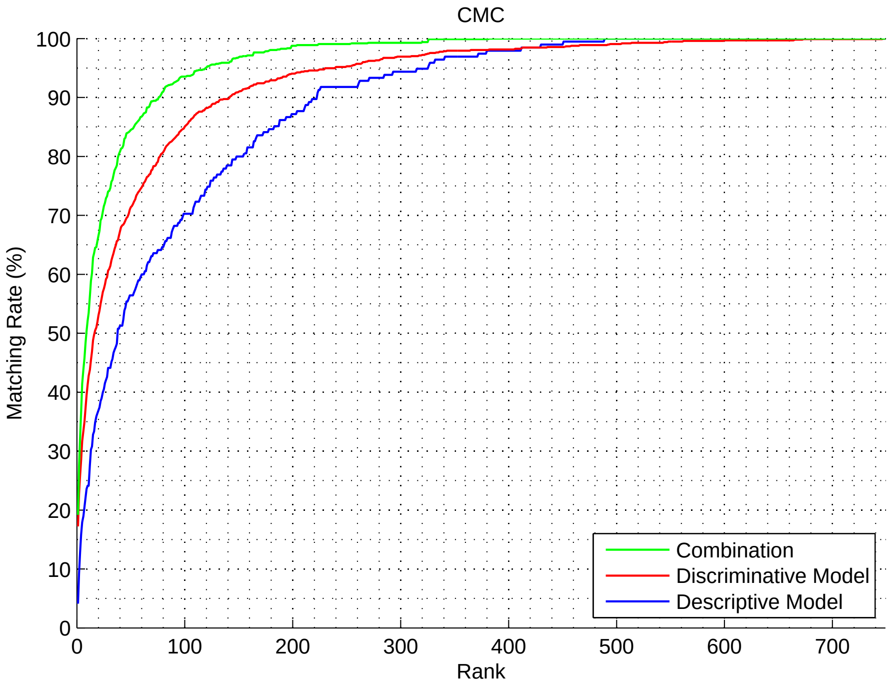


Figure 5.9: Average CMC curves of our system on PRID 2011: On this dataset, the discriminative model clearly outperforms the descriptive one. Nevertheless, combining both models increases the performance even further.

tic Discriminant Metric Learning (LDML) [32], Information-Theoretic Metric Learning (ITML) [15], Large Margin Nearest Neighbor (LMNN) [80, 81], Large Margin Nearest Neighbor with Rejection (LMNN-R) [16], Probabilistic Relative Distance Comparison (PRDC) [87], Pairwise Constrained Component Analysis (PCCA) [58], Efficient Impostor-Based Metric Learning (EIML), the balanced trace difference as described in Section 4.3.1, as well as the standard Mahalanobis distance and the simple baseline given by Linear Discriminant Analysis (LDA) [21]. Note that for the standard Mahalanobis distance, only the generative structure of the matching person pairs is considered, i.e., the metric matrix $\mathbf{M}$ is computed by inverting $\mathbf{C}_{\mathcal{S}}$ defined in Equation (4.22).

In order to give a fair comparison, we run all approaches under equal conditions, i.e., applying the same training and test set splits, the same features, etc. Moreover, since we are not interested in manually tuning the feature representation for each of the datasets, we use mostly the same setup, i.e., that presented in Section 4.4.1. The only main difference between the individual feature setups is the number of principal components used in the dimensionality reduction step, which has to be adapted to the number of training samples in order to avoid over-fitting. However, there are two exceptions, PRID 2011 and PRID 450S, where we obtain better results by skipping the texture description, i.e., LBP codes, and thus, only use the color features.

5.4.2.1 VIPeR

We start with the results obtained on VIPeR, which can be considered the de facto standard benchmark dataset for single-shot person re-identification scenarios. As it is the most widely used benchmark in this field of research, we provide a more thorough evaluation on this dataset by including another common feature setup [87] besides our own. Furthermore, we show that our metrics can also be learned specifically for each query sample due to their efficiency, resulting in better matching performance.

First, the CMC curves of the different metric learning approaches utilizing our feature setup as presented in Section 4.4.1, computed with the PCA dimensionality set to 80 and averaged over 10 runs, are shown in Figure 5.10. It can be seen that except for LDA, which does not have enough discriminative power, and LDML and PRDC, which are not designed to operate on such a simple feature representation, all approaches significantly outperform the Euclidean distance.

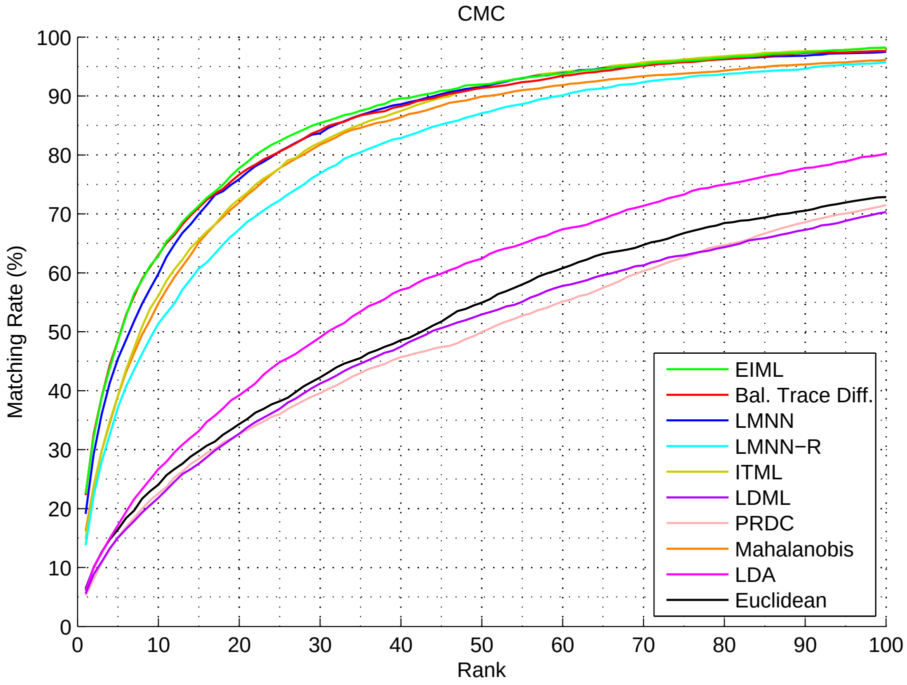


Figure 5.10: Average CMC curves of different metric learning approaches on VIPeR.

Method	$r = 1$	10	20	50	100	t_{train}
EIML	23	63	78	92	98	0.3 sec
Bal. Trace Diff.	22	63	77	91	98	0.9 sec
LMNN	19	60	76	92	98	2 min
LMNN-R	14	51	67	87	96	45 min
ITML	15	56	73	92	98	25 sec
LDML	6	22	33	53	70	0.8 sec
PRDC	5	23	33	50	71	2 min
Mahalanobis	16	55	72	90	96	0.001 sec
LDA	6	27	39	62	80	0.1 sec
Euclidean	6	24	34	55	73	–
ELF	12	43	60	81	93	5 hours
SDALF	20	50	65	85	–	–
ERSVM	13	50	67	85	94	13 min
CPS	22	57	71	87	–	–
PCCA	19	65	80	–	–	–

Table 5.3: Average matching rates (%) for different ranks r and, if available, average training times per trial on VIPeR: The results for the state-of-the-art methods in the lower part of the table stem from the respective papers.

Additionally, in Table 5.3, we compare the metric learning results to state-of-the-art methods that are particularly targeted at person re-identification, i.e., ELF [31], SDALF [18], ERSVM [64], CPS [11], and PCCA [58]. Since for many methods timings are available, we also analyze the computation time of the metric learning approaches using a Matlab implementation on a 2.83 GHz quad core CPU. The results clearly show that metric learning is able to boost the performance of the originally quite simple representation, yielding competitive results. That such a performance can also be achieved very efficiently is demonstrated by our own approaches, namely the balanced trace difference method and EIML, which outperform the other methods on this dataset, however, at dramatically reduced computational complexity.

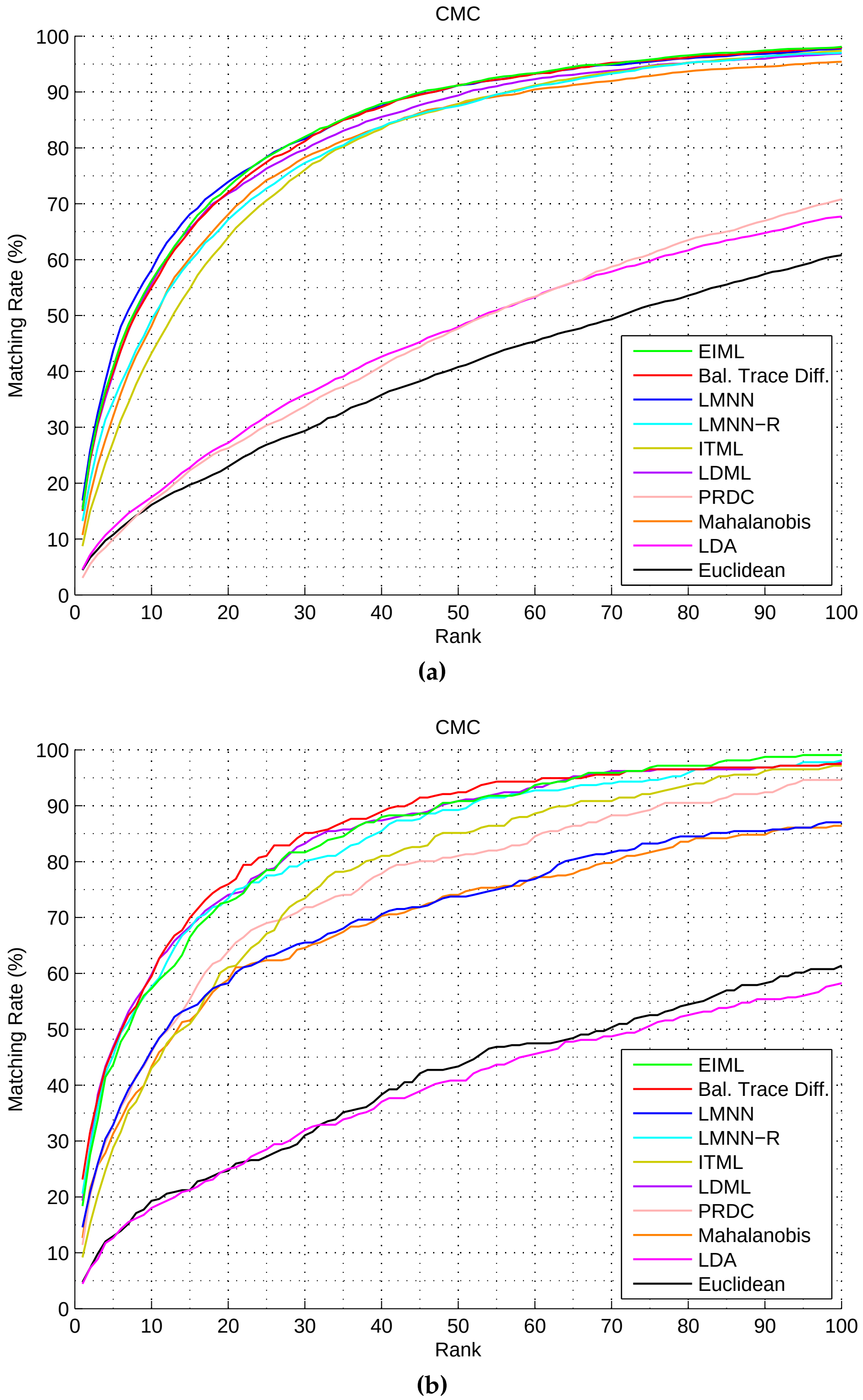
Next, we present results computed with another feature representation consisting of a mixture of color (RGB, HSV and YCbCr) and texture (Gabor [22] and Schmid [69]) histograms extracted from six horizontal stripe regions, a setup that is commonly used in person re-identification approaches (see, e.g., [31, 64, 87]). First, the results obtained after reducing the feature dimensionality to 80 via PCA are depicted in Figure 5.11a. While most metric learning approaches show a similar performance as with our feature setup, LDML is able to significantly improve its matching rates, being on par with the other methods now. Thus, LDML can clearly benefit from the more complex feature

Method	$d = 80$					$d = 2784$				
	$r = 1$	10	20	50	100	$r = 1$	10	20	50	100
EIML	15	56	73	91	98	18	57	73	91	99
Bal. Trace Diff.	15	55	72	91	98	23	59	76	92	97
LMNN	17	58	74	91	98	15	46	58	74	87
LMNN-R	13	49	67	88	97	21	58	73	89	98
ITML	9	43	64	88	97	9	43	61	85	97
LDML	16	56	72	89	97	19	60	74	91	98
PRDC	3	17	26	48	71	11	46	64	81	95
Mahalanobis	11	48	68	88	95	13	43	59	74	86
LDA	5	18	27	48	68	4	18	25	41	58
Euclidean	4	16	23	41	61	5	19	25	43	61

Table 5.4: Average matching rates (%) for different ranks r on VIPeR using the feature setup of Zheng et al. [87]: The left column shows the results obtained after reducing the feature dimensionality d to 80 via PCA, the right column shows the results obtained employing the original features (one run only).

representation, in contrast to PRDC, which still performs below average. To further investigate this behavior, we run an additional experiment on this feature setup, however, without the PCA pre-processing step, i.e., using all 2784 feature dimensions. Due to the long training times of some of the methods under these settings, just one run, i.e., one training and test set split, is carried out. As can be seen in Figure 5.11b, most of the metric learners achieve good matching rates also with such high dimensional features, at the cost of significantly increased runtime, though. Only LMNN and the standard Mahalanobis distance fall slightly back because of over-fitting. Under this setup, even PRDC performs well, which indicates that this method requires high dimensional input data in order to compute the optimal projection vectors in the feature space. The obtained results are also summarized in Table 5.4.

Finally, we present one more evaluation on VIPeR, showing that efficient metric learning algorithms can further improve the matching rates if they are applied appropriately. In particular, in this experiment, we compute a metric specifically for the chosen query sample, i.e., the searched person. In the first step, a global metric is estimated from the available training data just like in the other experiments. However, the thus obtained metric is not directly used to match the query person with the gallery. Instead, it is used to rank the training persons according to their similarity to the query, allowing us to weight them correspondingly, i.e., give higher weights to similar training samples. Then, in the second step, we re-compute the metric based on the weighted



(a)

(b)

Figure 5.11: Average CMC curves of different metric learning approaches on VIPeR using the feature setup of Zheng et al. [87]: (a) with an additional PCA step to reduce the feature dimensionality to 80 and (b) without pre-processing, i.e., using all 2784 dimensions (one run only).

training set, yielding a metric that is specifically tuned for the selected query person. The performance gain of this procedure can be seen in Figure 5.12, where we use EIML and compare the standard approach to the query-specific version. Concerning the features, we again apply our representation as described in Section 4.4.1 with the PCA dimensionality set to 80. In order to average the results, 10 runs are carried out. As a concluding remark, note that efficiency is crucial in this scenario, since the metric has to be re-calculated for every single query person. Hence, methods that rely on computationally complex optimization schemes are only of limited practical use in such a setup.

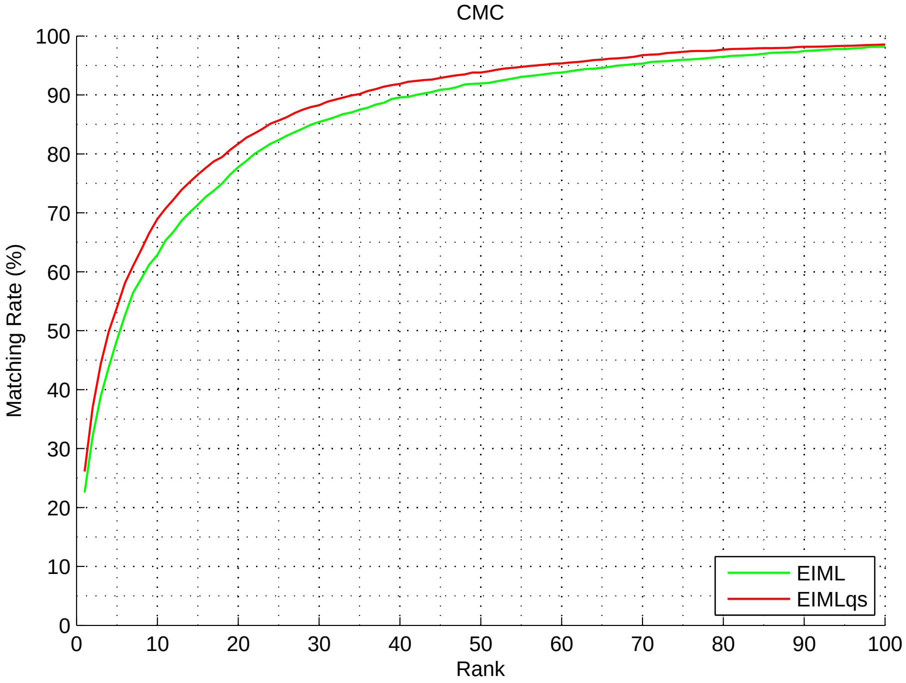


Figure 5.12: Average CMC curves of EIML and its query-specific version (EIMLqs) on VIPeR.

5.4.2.2 PRID 2011

PRID 2011 represents a very realistic, challenging scenario, where the images stem from two non-overlapping cameras. To compare the different metric learning approaches, we use the single-shot version, as most of the selected metric learners are targeted at such a setup. Compared to VIPeR, the number of training samples is quite limited and the gallery is much bigger, so that we use a reduced PCA dimensionality of 40 in order to avoid over-fitting.

Again, after performing 10 runs, it can be seen from the average CMC curves in Figure 5.13 and the average matching rates in Table 5.5 that all methods except for LDA, LDML, and PRDC show a significant gain in performance compared to the Euclidean distance. As already mentioned, LDML is not suited for our simple feature setup, and PRDC requires higher dimensional input data to work properly. Interestingly, even the standard Mahalanobis distance yields competitive results, which means that most of the valuable information is contained in the structure of the matching pairs. The only small influence of the non-matching pairs is most likely caused by the considerable amount of rather dark clothed, i.e., visually similar, persons in the dataset, making it difficult to find discriminating features.

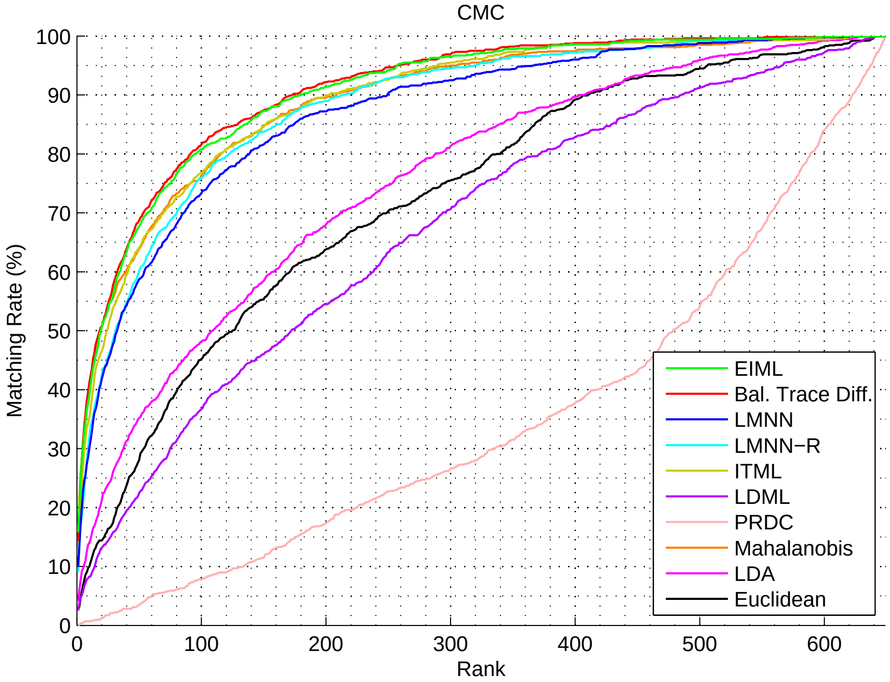


Figure 5.13: Average CMC curves of different metric learning approaches on PRID 2011.

5.4.2.3 PRID 450S

As the PRID 450S dataset builds on PRID 2011, it has similar characteristics, however, provides much more linked samples. In addition, it also contains detailed person masks, which allow us to analyze the effect of applying an exact foreground/background segmentation. Note that in our experiments, we only use the whole person mask, i.e., no part-level information. Since the training set of PRID 450S is considerably larger than

Method	$r = 1$	10	20	50	100
EIML	16	39	51	68	81
Bal. Trace Diff.	14	41	51	69	82
LMNN	10	30	42	59	73
LMNN-R	9	32	43	60	76
ITML	12	35	47	64	77
LDML	3	8	13	22	37
PRDC	0	1	1	4	8
Mahalanobis	16	41	51	64	76
LDA	4	14	21	35	48
Euclidean	3	10	14	28	45

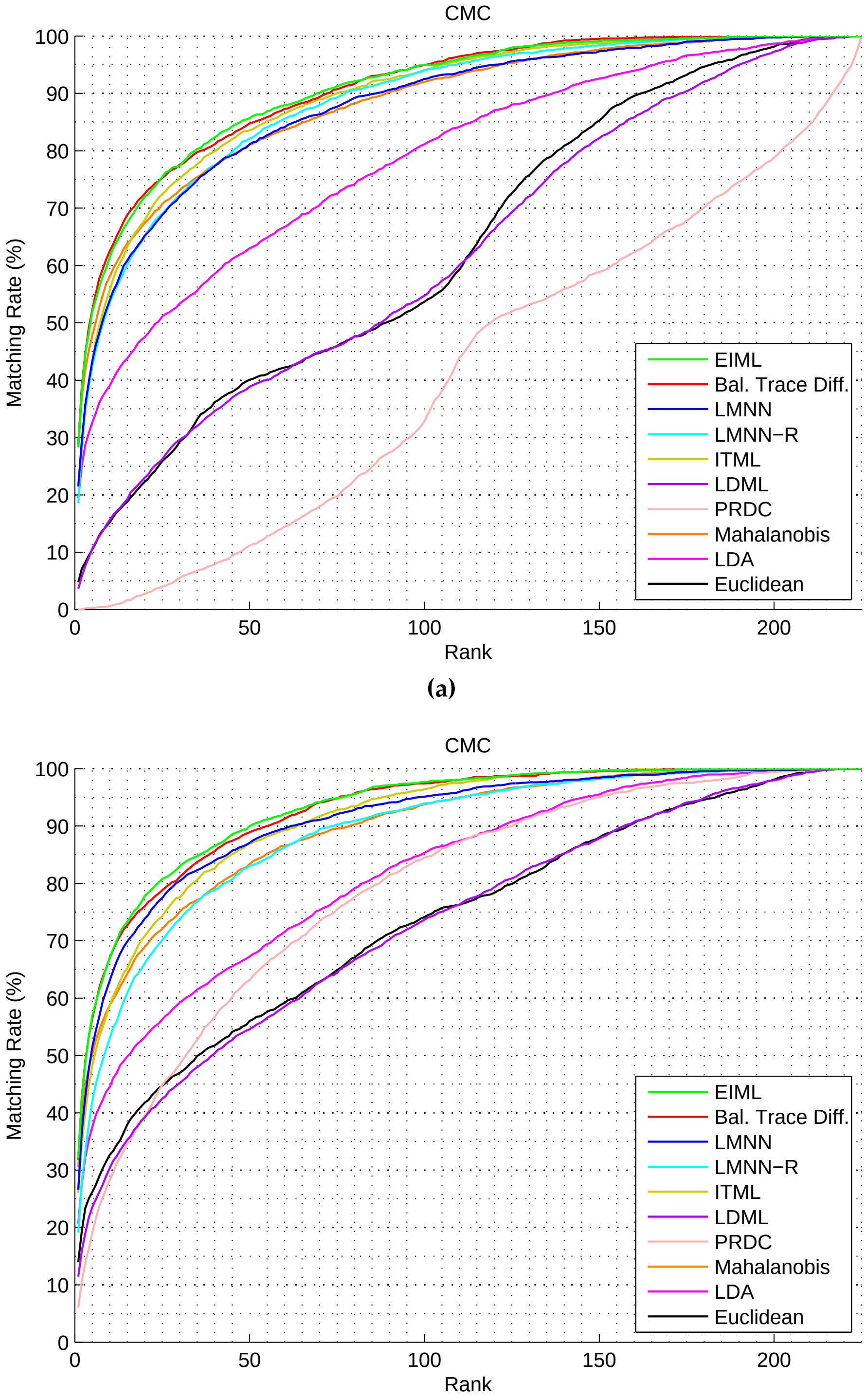
Table 5.5: Average matching rates (%) for different ranks r on PRID 2011.

Method	Without Masks					Using FG/BG Masks				
	$r = 1$	10	20	50	100	$r = 1$	10	20	50	100
EIML	28	61	72	86	95	32	67	78	90	98
Bal. Trace Diff.	29	63	72	85	95	32	67	76	89	97
LMNN	21	54	65	81	93	27	63	74	87	95
LMNN-R	19	54	65	82	94	19	53	66	83	94
ITML	21	56	68	84	94	26	59	71	87	96
LDML	4	16	23	39	55	11	30	39	55	74
PRDC	0	1	3	11	33	6	28	40	63	84
Mahalanobis	29	58	67	81	92	31	59	69	83	94
LDA	19	39	48	63	81	21	44	53	67	85
Euclidean	5	15	22	40	54	14	33	42	56	74

Table 5.6: Average matching rates (%) for different ranks r on PRID 450S.

that of PRID 2011, we also increase the PCA dimensionality, setting it to 60 for this dataset. In order to average the results, we again perform 10 runs.

The CMC curves obtained with and without the given masks are shown in Figure 5.14. Once more, it can be seen that LDML and PRDC have difficulties with our simple feature representation, and that applying LDA has only little influence on the ranking results. In contrast, all other approaches achieve significant improvements over the Euclidean distance. Furthermore, it can be recognized that using the foreground masks is beneficial for all approaches, increasing the performance by up to 5%. This can also be seen from the matching rates presented in Table 5.6.



(a)

(b)

Figure 5.14: Average CMC curves of different metric learning approaches on PRID 450S: (a) without masks and (b) using foreground/background masks.

Method	SEQ. #1							SEQ. #2							SEQ. #3						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7
EIML	73	81	84	86	88	89	90	67	76	81	84	86	88	90	86	92	95	96	97	98	98
Bal. Trace Diff.	74	82	85	87	88	90	91	67	77	81	84	86	88	90	86	93	95	97	98	98	98
LMNN	45	56	62	66	69	72	74	40	52	59	64	68	71	74	41	55	64	70	75	78	81
LMNN-R	43	55	61	66	70	72	74	45	59	65	70	75	79	81	52	67	75	81	85	88	90
ITML	71	79	82	85	87	88	89	69	78	82	85	87	89	90	88	94	96	97	98	98	98
LDML	62	70	74	77	79	81	82	58	66	72	76	79	81	83	77	86	90	92	94	95	96
PRDC	61	69	73	76	78	80	81	62	71	76	79	81	83	85	83	90	93	94	95	96	97
Mahalanobis	75	82	85	87	88	90	90	67	76	81	84	86	88	89	83	89	92	94	95	97	97
LDA	74	80	83	85	86	87	88	69	77	82	85	87	89	90	89	94	96	97	97	98	98
Euclidean	68	75	79	81	82	83	84	66	74	78	81	84	86	88	86	92	94	96	97	97	98

Table 5.7: Average matching rates (%) for the first 7 ranks on ETHZ.

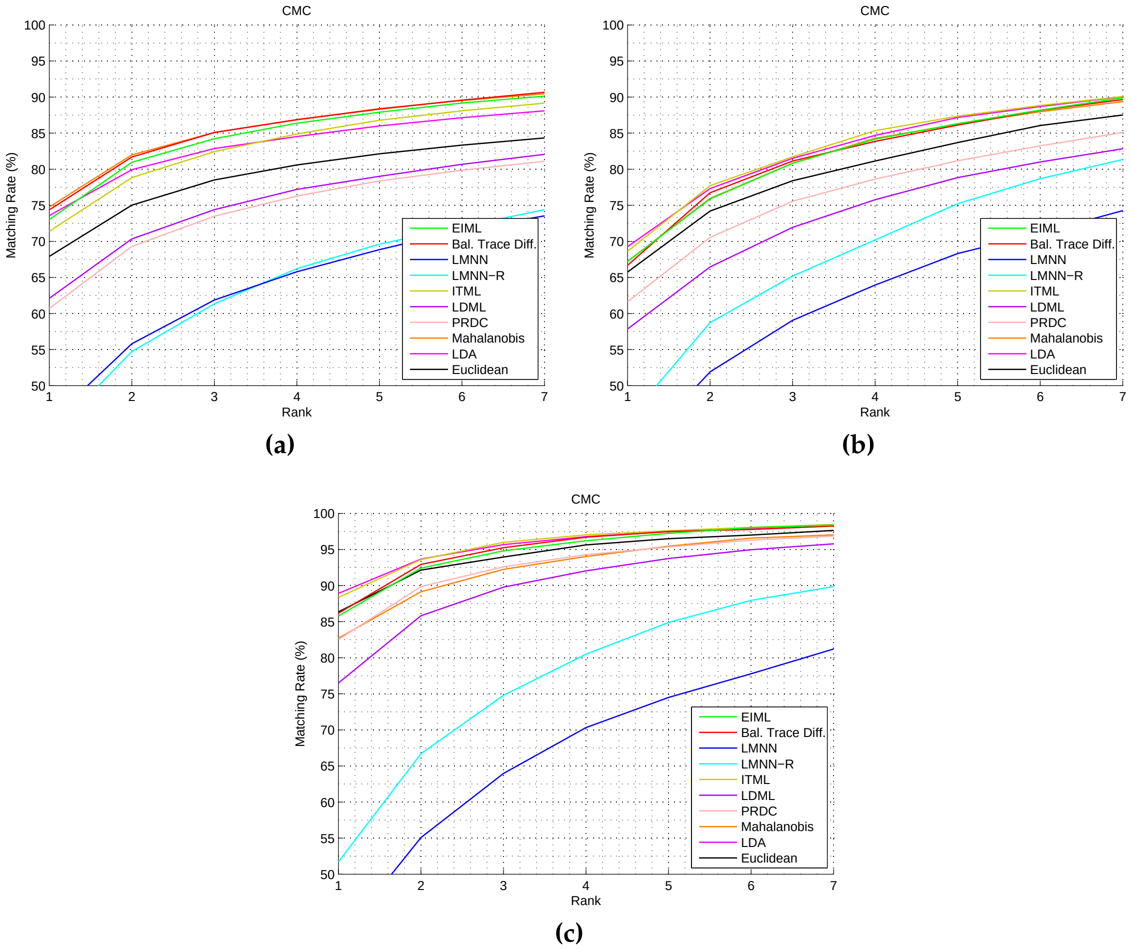
5.4.2.4 ETHZ

ETHZ is widely used for benchmarking person re-identification approaches, although it contains person trajectories captured from a single moving camera only. As a result, the individual images of a person are very similar, and metric learning has only little influence, as indicated by the CMC curves in Figure 5.15 and matching rates in Table 5.7 (averaged over 100 runs with the following PCA dimensionalities: 40 for SEQ. #1, 20 for SEQ. #2 and SEQ. #3).

Nevertheless, the matching rates for SEQ. #1, where metric learning has the largest impact, reveal that a performance gain of more than 5% can be obtained over all ranks compared to the Euclidean distance. The other two sequences, however, are not challenging enough, so that no significant improvement over the Euclidean distance can be achieved. The lower overall performance of the LMNN-based methods can be explained by the limited number of training samples, causing the computed models to over-fit the training data. Finally, LDML and PRDC are again not able to outperform the Euclidean distance, a consequence of our rather basic feature representation.

5.4.2.5 CAVIAR4REID

In this section, we show results on CAVIAR4REID [11], i.e., a multi-shot dataset, averaged over 100 runs with the PCA dimensionality set to 100. CMC curves and matching rates are shown in Figure 5.16 and Table 5.8, respectively. All approaches except for LDML show a significant improvement over the Euclidean distance. In Table 5.8, we additionally give a comparison to ICT [4], which uses the same experimental setup. Note that we exclude CPS proposed by Cheng et al. [11] and PRDC from our tests. CPS, al-



(a) **(b)** **(c)**

Figure 5.15: Average CMC curves of different metric learning approaches on ETHZ: (a) SEQ. #1, (b) SEQ. #2, and (c) SEQ. #3.

though evaluated on CAVIAR4REID, applies a completely different evaluation protocol, making a fair comparison difficult. And PRDC, on the other hand, is not able to handle the multi-shot configuration properly.

5.4.2.6 Qualitative Results

In order to also provide some qualitative results, in this section, we visualize some sample queries carried out on VIPeR. Specifically, Figure 5.17 depicts gallery rankings computed using three different methods: our own metric learning approaches, i.e., EIML and the balanced trace difference described in Section 4.3.1, as well as the Euclidean distance. The rankings clearly show what has already been indicated by the quantitative

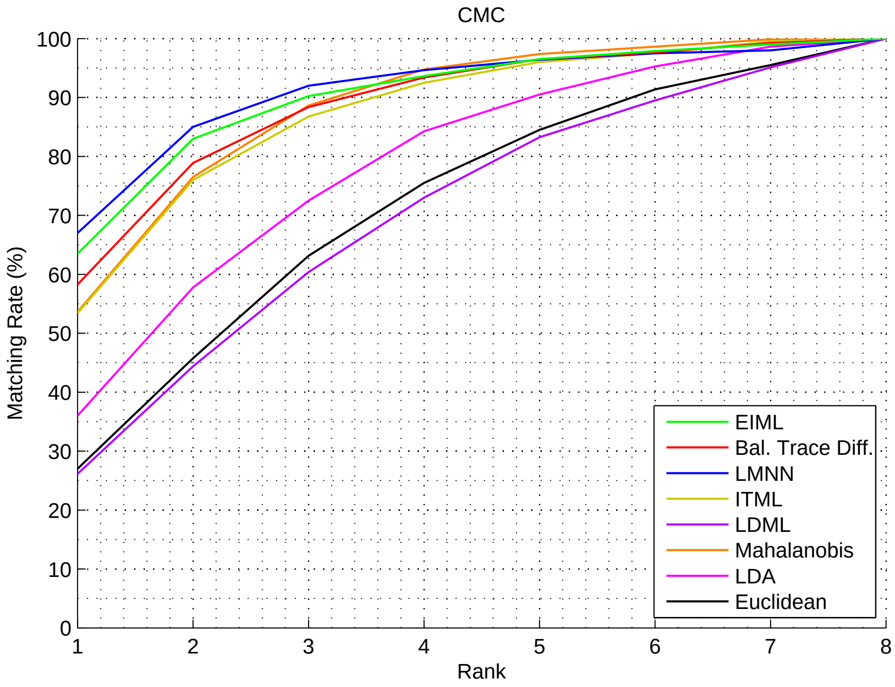
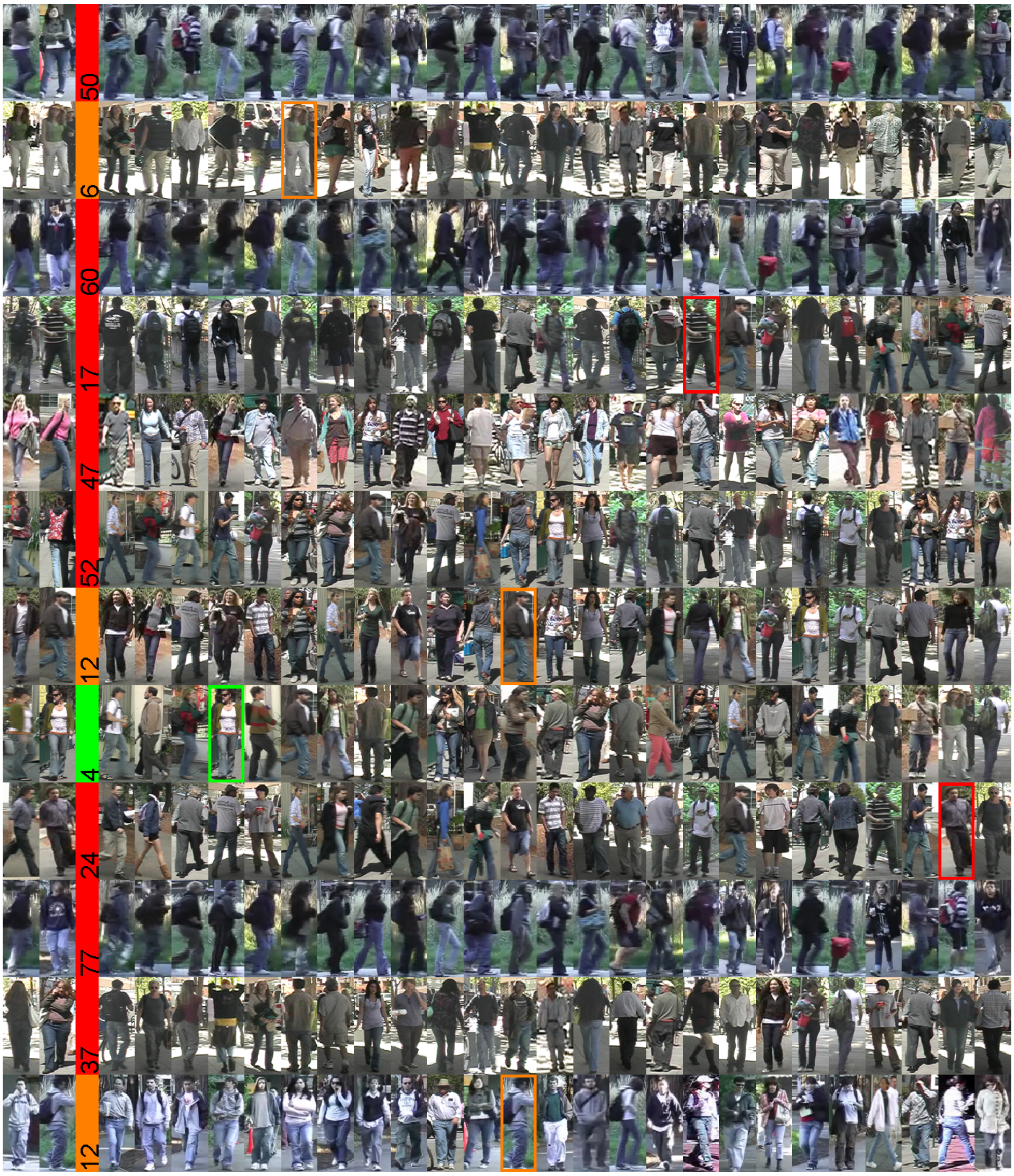


Figure 5.16: Average CMC curves of different metric learning approaches on CAVIAR4REID.

Method	$r = 1$	2	3	4	5	6	7	8
EIML	64	83	90	94	97	98	99	100
Bal. Trace Diff.	58	79	88	93	97	98	99	100
LMNN	67	85	92	95	96	98	98	100
ITML	53	76	87	93	96	98	100	100
LDML	26	44	60	73	83	90	95	100
Mahalanobis	54	77	89	95	97	99	100	100
LDA	36	58	73	84	91	95	99	100
Euclidean	27	46	63	76	85	91	96	100
ICT	62	81	95	97	97	100	100	100

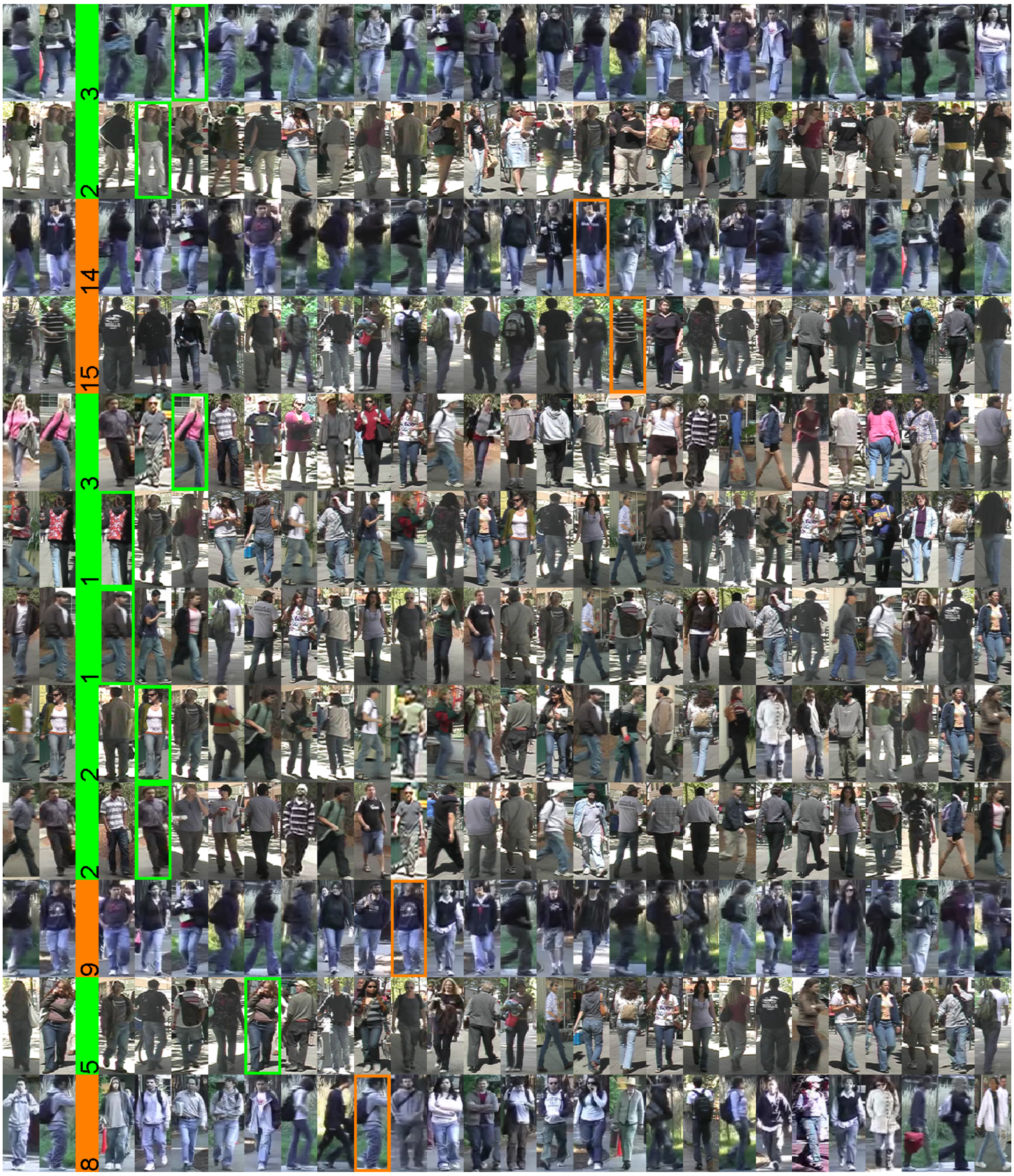
Table 5.8: Average matching rates (%) for different ranks r on CAVIAR4REID: The results for ICT stem from [4].

results presented in the previous sections: Metric learning is very beneficial for the task of person re-identification and can significantly improve search results compared to just using a standard distance measure.



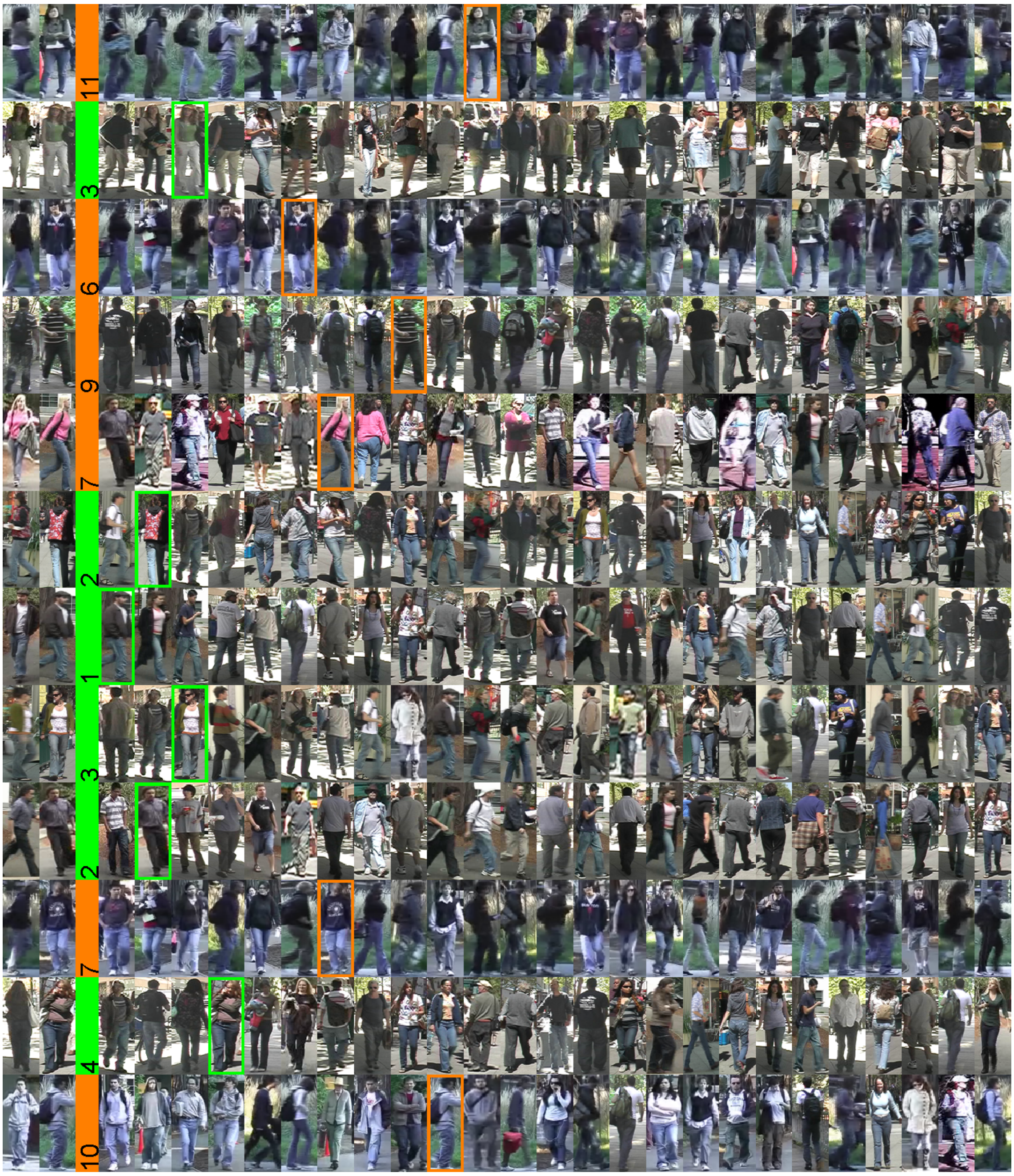
(a)

Figure 5.17: Qualitative results on VIPeR: The rankings are obtained using (a) the Euclidean distance, (b) the balanced trace difference, and (c) EIML. Each row corresponds to one person search. The first two columns show the query person pair, i.e., query sample and correct gallery match. The third column depicts the rank given to the correct match, followed by the first 25 positions of the ranking, with the matching person being highlighted if present. Additionally, the ranking performance is indicated by color (green: top 5, orange: top 15, and red: above 15).



(b)

Figure 5.17 (Cont.)



(c)

Figure 5.17 (Cont.)

6 Conclusion and Outlook

In this thesis, we presented different strategies to tackle the task of person re-identification, i.e., recognizing an individual in different locations across a network of non-overlapping cameras. Specifically, we focused on the most general setting, i.e., without easing the problem by spatio-temporal reasoning based on the scene layout. Hence, we just dealt with the appearance of a person, more precisely the appearance of the clothing, as this is the only reasonable cue that can be exploited. To match persons captured by distributed cameras, different methods have been proposed in literature, which can be roughly categorized into either implementing a descriptive or a discriminative strategy. Approaches following the first one usually try to extract (hand-crafted) features that should be both, distinctive and stable under changing viewing conditions. In contrast, those following the latter typically try to learn a discriminative model, i.e., find a representation that is well suited to distinguish a specific person from the remaining people.

As these different strategies capture a large extent of complementary information, in Chapter 3, we proposed to combine both in an interactive system, where we first run a fast, descriptive method and then, if necessary, refine the result by applying a discriminatively learned model. In this way, we are able to benefit from the lower runtime of the descriptive stage, which means that we can provide an initial search result very quickly, as well as the typically higher performance of the discriminative stage, which is required for harder cases. In Chapter 5, we demonstrated the advantage of having two models with such diverse characteristics unified in one system. As indicated by the corresponding CMC curves, they capture different aspects of a person (holistic appearance versus local details), and there is a clear tendency that one model can improve the performance in case the other one fails. Furthermore, our system also provides an iterative mode, where the discriminative model can be steered towards the correct match using person samples labeled by the operator. However, empirically, we figured out that untrained users rather tend to distort the model, which can be explained by the discrepancy between the human understanding of similarity and that of machine learning algorithms. Thus, investigating more robust algorithms for incorporating user feedback would be a promising starting point for future work.

Next, in Chapter 4, we presented another possibility to jointly apply descriptive and discriminative techniques: metric learning, which provides a very elegant and efficient tool for merging these two strategies. Specifically, starting from a descriptive feature representation, the idea is to transform the thus obtained feature space such that it emphasizes discriminative directions, i.e., those directions that are best suited to distinguish matching persons from non-matching ones. As shown in Chapter 5, even though some of the approaches are affected by over-fitting to some extent, we can conclude that metric learning can significantly boost the re-identification performance. However, since this is a relatively new direction in the field of person re-identification, only few metric learning approaches have been developed so far, and existing methods often suffer from high runtimes and memory requirements. As part of this thesis, we addressed this shortcoming and showed that also very efficient algorithms can be derived, which achieve state-of-the-art or even better performance, however, at much lower computational costs. This is particularly notable, as we only build on very basic color and texture features. Thus, depending on the available time budget in a real world application, a more sophisticated feature representation could even further improve the matching rates. Another promising direction for future work was presented in Section 5.4.2.1, where we computed a metric specifically for the chosen query sample, i.e., the searched person, resulting in a clear performance gain.

To recap, in this thesis, we studied person re-identification methods tackling the problem from different directions, i.e., applying descriptive and discriminative techniques. In order to exploit the advantages of both strategies, but avoid their respective drawbacks, we proposed to combine them. Using our interactive system as well as efficient metric learning algorithms, we saw great potential of such joint approaches, opening up new directions in this field of research. As a final note and in contrast to many established methods, in our works, we always tried to keep the algorithmic complexity on a reasonable level. We consider this property to be of high relevance, especially in the context of the challenges that come along with the task at hand. In this way, we could not only increase the robustness of our methods, but also make them applicable to large-scale camera networks.

A
List of Acronyms

CMC Cumulative Matching Characteristic

EIML Efficient Impostor-Based Metric Learning

HOG Histogram of Oriented Gradients

ITML Information-Theoretic Metric Learning

LBP Local Binary Pattern

LDA Linear Discriminant Analysis

LDML Logistic Discriminant Metric Learning

LMNN Large Margin Nearest Neighbor

LMNN-R Large Margin Nearest Neighbor with Rejection

MSCR Maximally Stable Color Region

PCA Principal Component Analysis

PCCA Pairwise Constrained Component Analysis

PLS Partial Least Squares

PRDC Probabilistic Relative Distance Comparison

PRSVM Primal RankSVM

RankSVM Ranking SVM

RHSP Recurrent Highly Structured Patch

STEL Structure Element

SVM Support Vector Machine

UT Unscented Transform

B
List of Publications

My work at the Institute for Computer Graphics and Vision led to the following peer-reviewed publications, listed in chronological order:

2009

An Automatic Hybrid Segmentation Approach for Aligned Face Portrait Images
Martin Hirzer, Martin Urschler, Horst Bischof, and Josef A. Birchbauer
In *Proc. Workshop of the Austrian Association for Pattern Recognition*
May 2009, Stainz, Austria

Saliency Driven Total Variation Segmentation
Michael Donoser, Martin Urschler, Martin Hirzer, and Horst Bischof
In *Proc. IEEE International Conference on Computer Vision*
September–October 2009, Kyoto, Japan

2011

Person Re-Identification by Descriptive and Discriminative Classification
Martin Hirzer, Csaba Beleznai, Peter M. Roth, and Horst Bischof
In *Proc. Scandinavian Conference on Image Analysis*
May 2011, Ystad, Sweden

Multi-Cue Learning and Visualization of Unusual Events
René Schuster, Samuel Schulter, Georg Poier, Martin Hirzer, Josef A. Birchbauer, Peter M. Roth, Horst Bischof, Martin Winter, and Peter Schallauer
In *Proc. IEEE International Workshop on Visual Surveillance (in conjunction with ICCV)*
November 2011, Barcelona, Spain

2012

Large Scale Metric Learning from Equivalence Constraints
Martin Köstinger, Martin Hirzer, Peter M. Roth, and Horst Bischof
In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*
June 2012, Providence, RI, USA

Person Re-Identification by Efficient Impostor-based Metric Learning
Martin Hirzer, Peter M. Roth, and Horst Bischof
In *Proc. IEEE International Conference on Advanced Video and Signal-Based Surveillance*
September 2012, Beijing, China

Dense Appearance Modeling and Efficient Learning of Camera Transitions for Person Re-Identification
Martin Hirzer, Csaba Beleznai, Martin Köstinger, Peter M. Roth, and Horst Bischof
In *Proc. IEEE International Conference on Image Processing*
September–October 2012, Orlando, FL, USA

Relaxed Pairwise Learned Metric for Person Re-Identification
Martin Hirzer, Peter M. Roth, Martin Köstinger, and Horst Bischof
In *Proc. European Conference on Computer Vision*
October 2012, Florence, Italy

2013

Pedestrian Detection, Tracking and Re-Identification for Search in Visual Surveillance Data
Csaba Beleznai, Michael Rauter, Martin Hirzer, and Peter M. Roth
In *Proc. Conference of the Hungarian Association for Image Processing and Pattern Recognition*
January–February 2013, Bakonybél, Hungary

2014

Mahalanobis Distance Learning for Person Re-Identification
Peter M. Roth, Martin Hirzer, Martin Köstinger, Csaba Beleznai, and Horst Bischof
In Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen C. Loy, editors, *Person Re-Identification*, Advances in Computer Vision and Pattern Recognition, pages 247–267
January 2014, Springer

Bibliography

[1] Babak Alipanahi, Michael Biggs, and Ali Ghodsi. Distance metric learning vs. Fisher discriminant analysis. In *Proc. AAAI Conf. on Artificial Intelligence*, 2008. (Cited on page 46.)

[2] Yali Amit and Augustine Kong. Graphical templates for model registration. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 18(3):225–236, 1996. (Cited on page 12.)

[3] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2009. (Cited on page 15.)

[4] Tamar Avraham, Ilya Gurvich, Michael Lindenbaum, and Shaul Markovitch. Learning implicit transfer for person re-identification. In *Proc. Int'l Workshop on Re-Identification (in conjunction with ECCV)*, 2012. (Cited on pages 82, 96, and 98.)

[5] Serge Belongie, Jitendra Malik, and Jan Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(4):509–522, 2002. (Cited on page 13.)

[6] Sławomir Bąk, Étienne Corvée, François Brémond, and Monique Thonnat. Person re-idendification using Haar-based and DCD-based signature. In *Proc. Workshop on Activity Monitoring by Multi-Camera Surveillance Systems (in conjunction with AVSS)*, 2010. (Cited on pages 7, 16, 24, and 32.)

[7] Sławomir Bąk, Étienne Corvée, François Brémond, and Monique Thonnat. Person re-identification using spatial covariance regions of human body parts. In *Proc. IEEE Int'l Conf. on Advanced Video and Signal-Based Surveillance*, 2010. (Cited on pages 6 and 13.)

[8] Sławomir Bąk, Sundaram Suresh, François Brémond, and Monique Thonnat. Fusion of motion segmentation with online adaptive neural classifier for robust tracking. In *Proc. Int'l Joint Conf. on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP)*, 2009. (Cited on page 16.)

[9] Lev M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967. (Cited on page 58.)

[10] Olivier Chapelle and S. Sathiya Keerthi. Efficient algorithms for ranking with SVMs. *Information Retrieval*, 13(3):201–215, 2010. (Cited on page 18.)

[11] Dong S. Cheng, Marco Cristani, Michele Stoppa, Loris Bazzani, and Vittorio Murino. Custom pictorial structures for re-identification. In *Proc. British Machine Vision Conf.*, 2011. (Cited on pages 6, 15, 77, 81, 89, and 96.)

[12] Corinna Cortes and Vladimir N. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. (Cited on page 18.)

[13] Étienne Corvée and François Brémond. Combining face detection and people tracking in video sequences. In *Proc. Int'l Conf. on Imaging for Crime Detection and Prevention*, 2009. (Cited on pages 13 and 16.)

[14] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2005. (Cited on pages 13 and 85.)

[15] Jason V. Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S. Dhillon. Information-theoretic metric learning. In *Proc. Int'l Conf. on Machine Learning*, 2007. (Cited on pages 9, 46, 57, 73, and 87.)

[16] Mert Dikmen, Emre Akbas, Thomas S. Huang, and Narendra Ahuja. Pedestrian recognition with a learned metric. In *Proc. Asian Conf. on Computer Vision*, 2010. (Cited on pages 9, 19, 47, 59, and 87.)

[17] Andreas Ess, Bastian Leibe, and Luc Van Gool. Depth and appearance for mobile scene analysis. In *Proc. IEEE Int'l Conf. on Computer Vision*, 2007. (Cited on page 80.)

[18] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2010. (Cited on pages 6, 14, 15, 24, 72, 81, and 89.)

[19] Pedro F. Felzenszwalb. Representation and detection of deformable shapes. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(2):208–220, 2005. (Cited on page 12.)

[20] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Pictorial structures for object recognition. *Int'l Journal of Computer Vision*, 61(1):55–79, 2005. (Cited on page 15.)

[21] Ronald A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936. (Cited on pages 47, 53, and 87.)

[22] Itzhak Fogel and Dov Sagi. Gabor filters as texture discriminator. *Biological Cybernetics*, 61(2):103–113, 1989. (Cited on pages 16, 19, 20, and 89.)

[23] Per-Erik Forssén. Maximally stable colour regions for recognition and matching. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2007. (Cited on pages 14 and 15.)

[24] Wolfgang Förstner and Boudewijn Moonen. A metric for covariance matrices. Technical report, Department of Geodesy and Geoinformatics, University of Stuttgart, Stuttgart, Germany, 1999. (Cited on pages 14 and 31.)

[25] Yoav Freund. Boosting a weak learning algorithm by majority. *Information and Computation*, 121(2):256–285, 1995. (Cited on page 33.)

[26] 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*, 55(1):119–139, 1997. (Cited on pages 16, 32, 34, and 35.)

[27] Brian V. Funt and Graham D. Finlayson. Color constant color indexing. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 17(5):522–529, 1995. (Cited on page 13.)

[28] Niloofar Gheissari, Thomas B. Sebastian, Peter H. Tu, Jens Rittscher, and Richard Hartley. Person reidentification using spatiotemporal appearance. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2006. (Cited on pages 6, 12, 13, 24, 29, and 72.)

[29] Ali Ghodsi, Dana F. Wilkinson, and Finnegan Southey. Improving embeddings by flexible exploitation of side information. In *Proc. Int'l Joint Conf. on Artificial Intelligence*, 2007. (Cited on page 46.)

[30] Douglas Gray, Shane Brennan, and Hai Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. IEEE Workshop on Performance Evaluation of Tracking and Surveillance*, 2007. (Cited on pages 2, 76, 77, and 78.)

[31] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Proc. European Conf. on Computer Vision*, 2008. (Cited on pages 7, 16, 18, 20, 21, 24, 29, 30, 32, 72, 76, 78, and 89.)

[32] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. Is that you? Metric learning approaches for face identification. In *Proc. IEEE Int'l Conf. on Computer Vision*, 2009. (Cited on pages 9, 46, 56, 73, and 87.)

[33] Yue-Fei Guo, Shi-Jin Li, Jing-Yu Yang, Ting-Ting Shu, and Li-De Wu. A generalized Foley-Sammon transform based on generalized Fisher discriminant criterion and its application to face recognition. *Pattern Recognition Letters*, 24(1–3):147–158, 2003. (Cited on page 67.)

[34] Robert M. Haralick, Kumarasamy S. Shanmugam, and Its'hak Dinstein. Textural features for image classification. *IEEE Trans. on Systems, Man, and Cybernetics*, 3(6):610–621, 1973. (Cited on page 17.)

[35] Martin Hirzer, Csaba Beleznai, Peter M. Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Proc. Scandinavian Conf. on Image Analysis*, 2011. (Cited on pages 77 and 79.)

[36] Xiaopeng Hong, Hong Chang, Shiguang Shan, Xilin Chen, and Wen Gao. Sigma set: A small second order statistical region descriptor. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2009. (Cited on page 37.)

[37] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6 and 7):417–441 and 498–520, 1933. (Cited on pages 17 and 53.)

[38] Min Hu, Jianguang Lou, Weiming Hu, and Tieniu Tan. Multi-camera correspondence based on principal axis of human body. In *Proc. IEEE Int'l Conf. on Image Processing*, 2004. (Cited on pages 6 and 30.)

[39] Weiming Hu, Min Hu, Xue Zhou, Tieniu Tan, Jianguang Lou, and Steve Maybank. Principal axis-based correspondence between multiple cameras for people tracking. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(4):663–671, 2006. (Cited on pages 6 and 30.)

[40] Omar Javed, Saad Ali, and Mubarak Shah. Online detection and classification of moving objects using progressively improving detectors. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2005. (Cited on pages 35, 37, and 85.)

[41] Omar Javed, Zeeshan Rasheed, Khurram Shafique, and Mubarak Shah. Tracking across multiple cameras with disjoint views. In *Proc. IEEE Int'l Conf. on Computer Vision*, 2003. (Cited on page 22.)

[42] Omar Javed, Khurram Shafique, and Mubarak Shah. Appearance modeling for tracking in multiple non-overlapping cameras. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2005. (Cited on page 22.)

[43] Andrew H. Jazwinski. *Stochastic processes and filtering theory*, volume 64 of *Mathematics in science and engineering*. Academic Press, New York, NY, USA, 1970. (Cited on page 39.)

[44] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proc. ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining*, 2002. (Cited on page 18.)

[45] Nebojsa Jojic, Alessandro Perina, Marco Cristani, Vittorio Murino, and Brendan J. Frey. Stel component analysis: Modeling spatial correlations in image class structure. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2009. (Cited on page 14.)

[46] Simon J. Julier and Jeffrey K. Uhlmann. A general method for approximating nonlinear transformations of probability distributions. Technical report, Robotics Research Group, Department of Engineering Science, University of Oxford, Oxford, United Kingdom, 1996. (Cited on page 39.)

[47] Michael J. Kearns and Leslie G. Valiant. Learning boolean formulae or finite automata is as hard as factoring. Technical Report TR-14-88, Aiken Computation Laboratory, Harvard University, Cambridge, MA, USA, 1988. (Cited on page 33.)

[48] Stefan Kluckner, Thomas Mauthner, Peter M. Roth, and Horst Bischof. Semantic classification in aerial imagery by integrating appearance and height information. In *Proc. Asian Conf. on Computer Vision*, 2009. (Cited on pages 37 and 40.)

[49] Martin Köstinger, Martin Hirzer, Paul Wohlhart, Peter M. Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2012. (Cited on page 73.)

[50] Kobi Levi and Yair Weiss. Learning object detection from a small number of examples: The importance of good features. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2004. (Cited on pages 35, 37, and 85.)

[51] Rainer Lienhart and Jochen Maydt. An extended set of Haar-like features for rapid object detection. In *Proc. IEEE Int'l Conf. on Image Processing*, 2002. (Cited on pages 35, 37, and 85.)

[52] Zhe Lin and Larry S. Davis. Learning pairwise dissimilarity profiles for appearance recognition in visual surveillance. In *Proc. Int'l Symposium on Visual Computing*, 2008. (Cited on pages 17, 24, and 76.)

[53] Marco Loog, Robert P. W. Duin, and Reinhold Haeb-Umbach. Multiclass linear dimension reduction by weighted pairwise Fisher criteria. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(7):762–766, 2001. (Cited on page 55.)

[54] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int'l Journal of Computer Vision*, 60(2):91–110, 2004. (Cited on page 23.)

[55] Helmut Lütkepohl. *Handbook of Matrices*. John Wiley & Sons, Chichester, United Kingdom, 1997. (Cited on page 70.)

[56] Prasanta C. Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2(1):49–55, 1936. (Cited on page 48.)

[57] Dimitrios Makris, Tim Ellis, and James Black. Bridging the gaps between cameras. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2004. (Cited on pages 21 and 22.)

[58] Alexis Mignon and Frédéric Jurie. PCCA: A new approach for distance learning from sparse pairwise constraints. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2012. (Cited on pages 20, 47, 62, 87, and 89.)

[59] Greg Mori and Jitendra Malik. Recovering 3d human body configurations using shape contexts. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(7):1052–1062, 2006. (Cited on page 13.)

[60] Timo Ojala, Matti Pietikäinen, and David Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51–59, 1996. (Cited on pages 21, 72, and 85.)

[61] Andreas Opelt, Axel Pinz, and Andrew Zisserman. A boundary-fragment-model for object detection. In *Proc. European Conf. on Computer Vision*, 2006. (Cited on pages 35, 37, and 85.)

[62] Karl Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6*, 2(11):559–572, 1901. (Cited on pages 17 and 53.)

[63] Fatih Porikli. Inter-camera color calibration by correlation model function. In *Proc. IEEE Int'l Conf. on Image Processing*, 2003. (Cited on page 45.)

[64] Bryan Prosser, Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Person re-identification by support vector ranking. In *Proc. British Machine Vision Conf.*, 2010. (Cited on pages 7, 18, 20, 24, 29, 32, 72, and 89.)

[65] Ali Rahimi, Brian Dunagan, and Trevor Darrell. Simultaneous calibration and tracking with a network of non-overlapping sensors. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2004. (Cited on pages 21 and 22.)

[66] Roman Rosipal and Nicole Krämer. Overview and recent advances in partial least squares. In *Proc. PASCAL Workshop on Subspace, Latent Structure and Feature Selection Techniques: Statistical and Optimization Perspectives*, 2005. (Cited on page 17.)

[67] Peter M. Roth, Martin Hirzer, Martin Köstinger, Csaba Beleznai, and Horst Bischof. Mahalanobis distance learning for person re-identification. In Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen C. Loy, editors, *Person Re-Identification*, Advances in Computer Vision and Pattern Recognition, pages 247–267. Springer, London, United Kingdom, 2014. (Cited on pages 77 and 80.)

[68] Robert E. Schapire. The strength of weak learnability. *Machine Learning*, 5(2):197–227, 1990. (Cited on page 33.)

[69] Cordelia Schmid. Constructing models for content-based image retrieval. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2001. (Cited on pages 16, 19, 20, and 89.)

[70] William R. Schwartz and Larry S. Davis. Learning discriminative appearance-based models using partial least squares. In *Proc. Brazilian Symposium on Computer Graphics and Image Processing*, 2009. (Cited on pages 7, 17, 18, 24, 72, 76, 77, 80, and 81.)

[71] Linlin Shen and Li Bai. Mutualboost learning for selecting gabor features for face recognition. *Pattern Recognition Letters*, 27(15):1758–1767, 2006. (Cited on pages 35, 37, and 85.)

[72] Michael J. Swain and Dana H. Ballard. Indexing via color histograms. In *Proc. IEEE Int'l Conf. on Computer Vision*, 1990. (Cited on pages 13 and 22.)

[73] Kinh H. Tieu and Paul A. Viola. Boosting image retrieval. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2000. (Cited on pages 32 and 34.)

[74] Michael E. Tipping and Christopher M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 61(3):611–622, 1999. (Cited on page 22.)

[75] Oncel Tuzel, Fatih Porikli, and Peter Meer. Region covariance: A fast descriptor for detection and classification. In *Proc. European Conf. on Computer Vision*, 2006. (Cited on pages 29 and 40.)

[76] UK Home Office. i-LIDS multiple camera tracking scenario definition, 2008. (Cited on page 78.)

[77] Paul A. Viola and Michael J. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2001. (Cited on pages 34 and 37.)

[78] Paul A. Viola, Michael J. Jones, and Daniel Snow. Detecting pedestrians using patterns of motion and appearance. In *Proc. IEEE Int'l Conf. on Computer Vision*, 2003. (Cited on page 34.)

[79] Xiaogang Wang, Gianfranco Doretto, Thomas B. Sebastian, Jens Rittscher, and Peter H. Tu. Shape and appearance context modeling. In *Proc. IEEE Int'l Conf. on Computer Vision*, 2007. (Cited on pages 6, 13, 24, 29, 72, and 76.)

[80] Kilian Q. Weinberger, John Blitzer, and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. In *Proc. Conf. on Neural Information Processing Systems*, 2006. (Cited on pages 9, 19, 46, 58, 60, and 87.)

[81] Kilian Q. Weinberger and Lawrence K. Saul. Fast solvers and efficient implementations for distance metric learning. In *Proc. Int'l Conf. on Machine Learning*, 2008. (Cited on pages 9, 19, 46, 58, and 87.)

[82] Herman Wold. Partial least squares. In Samuel Kotz and Norman L. Johnson, editors, *Encyclopedia of Statistical Sciences*, volume 6, pages 581–591. John Wiley & Sons, New York, NY, USA, 1985. (Cited on page 17.)

[83] 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*, 19(2):92–105, 2008. (Cited on page 17.)

[84] Jieping Ye, Zheng Zhao, and Huan Liu. Adaptive distance metric learning for clustering. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2007. (Cited on page 73.)

[85] Jie Yu, Jaume Amores, Nicu Sebe, and Qi Tian. Toward robust distance metric analysis for similarity estimation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2006. (Cited on page 73.)

[86] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Associating groups of people. In *Proc. British Machine Vision Conf.*, 2009. (Cited on page 23.)

[87] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Person re-identification by probabilistic relative distance comparison. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2011. (Cited on pages xvi, xvii, 20, 29, 47, 60, 87, 88, 89, 90, and 91.)