Relaxed Pairwise Learned Metric for Person Re-identification

Martin Hirzer, Peter M. Roth, Martin Köstinger, and Horst Bischof

Institute for Computer Graphics and Vision
Graz University of Technology

Abstract. Matching persons across non-overlapping cameras is a rather challenging task. Thus, successful methods often build on complex feature representations or sophisticated learners. A recent trend to tackle this problem is to use metric learning to find a suitable space for matching samples from different cameras. However, most of these approaches ignore the transition from one camera to the other. In this paper, we propose to learn a metric from pairs of samples from different cameras. In this way, even less sophisticated features describing color and texture information are sufficient for finally getting state-of-the-art classification results. Moreover, once the metric has been learned, only linear projections are necessary at search time, where a simple nearest neighbor classification is performed. The approach is demonstrated on three publicly available datasets of different complexity, where it can be seen that state-of-the-art results can be obtained at much lower computational costs.

1 Introduction

Person re-identification, *i.e.*, recognizing an individual across spatially disjoint cameras, is becoming one of the major challenges in visual surveillance. Typical applications include but are not limited to tracking criminals, analyzing crowd movements in public places, and finding children who lost their parents. Since the number of public areas that become subject to video surveillance is ever growing, efficient, automatic systems are required to reduce the load on human operators.

In general, person re-identification is very challenging for several reasons. First, the appearance of an individual can vary extremely across a network of cameras due to changing view points, illumination, different poses, etc. Second, there is a potentially high number of "similar" persons (*e.g.*, people wear rather dark clothes in winter). Third, in contrast to similar large scale search problems typically no accurate temporal and spatial constraints can be exploited to ease the task. Thus, motivated by the high number of practical applications and still unresolved difficulties there has been an increased scientific interest (*e.g.*, [1–10]) in recent years, and also various benchmark datasets [6, 8, 11] have been published.

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part VI, LNCS 7577, pp. 780–793, 2012.

Due to the complexity of the task different strategies have been proposed to solve it. The methods can be roughly divided into the following groups: (a) descriptive methods (*e.g.*, [1, 2, 12, 13]), (b) discriminative methods (*e.g.*, [3–8, 14]), and (c) metric learning methods (*e.g.*, [10, 15]). The main idea of descriptive methods is to extract visual features that are both, distinctive and stable under changing conditions between different cameras. After feature extraction, a standard distance measure is applied to compare different person representations. However, large intra-class appearance variations often prevent the computation of distinctive and stable features under realistic conditions. To overcome these limitations, discriminative methods additionally take advantage of class information to exploit the discriminative information given by the data. But, as a drawback, such methods tend to overfit to the training data. Moreover, they are often based on local image descriptors, which might be a severe disadvantage. For instance, a red bag visible in on view would be very discriminative, however, if it is not visible in the other view it becomes impossible to find the specific person again.

A midway between both approaches is to use metric learning [10, 15]. Such methods are similar to descriptive methods – the data is modeled by some kind of descriptive feature. However, they differ as the descriptors are not directly compared in the feature space, but instead a non-Euclidean distance is used. Since this has to be estimated beforehand, in contrast to descriptive models a training stage is necessary. In fact, such metrics describe the transition in feature space between two camera views, which makes these approaches much more suitable for real world scenarios. Moreover, during evaluation metric learning approaches are very efficient since additionally to the feature extraction and the matching only a linear projection has to be computed. However, there are two main drawbacks considering the training stage. First, existing metric learners such as Large Margin Nearest Neighbor (LMNN) [16], Information Theoretic Metric Learning (ITML) [17], and Logistic Discriminant Metric Learning (LDML) [18] build on complex optimization schemes resulting in high computational costs. Second, these methods typically assume a multi-class classification problem, which is not the case for person re-identification. In fact, we are only given image pairs, so existing methods have to be adapted. There are only a few methods such as [19, 20] which directly indent learning a metric from data pairs. But again, these methods build on complex numerical methods, making them infeasible in practice.

The goal of this work is to benefit from the advantages of metric learning, however, reducing the computational effort. In fact, we introduce a pairwise metric learning approach taking advantage of the structure of the data, however, aim to reduce the computational effort. This is realized by relaxing the original hard constraints, getting a simpler problem and thus avoiding iterative procedures. In this way, we finally obtain state-of-the-art or even significantly better results on three different datasets of varying size and complexity. This is in particular of interest, since, compared to existing methods, only rather simple image descriptors extracting color (HSV and Lab) and texture information (LBP) are

used. Moreover, we give a detailed comparison to other methods, an analysis of runtime, and show the influence of the number of training samples.

2 Related Work

Many of the proposed approaches try to 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. For instance, Gheissari *et al.* [1] try to fit a triangulated graph to each person to cope with pose variations. However, their approach works only if people are seen from similar viewpoints, which is not the case in most practical setups. The same restriction applies for the work of Wang *et al.* [2]. The image of a person is divided into regions and their color spatial structure is captured by a co-occurrence matrix. In [12], Farenzena *et al.* combine multiple features to describe the appearance of a person by exploiting perceptual principles. After obtaining a person's silhouette through a segmentation step, symmetry and asymmetry axes are found and used for accumulating color and texture feature responses. Cheng *et al.* [21] apply Pictorial Structures to person re-identification. They fit a body configuration composed of chest, head, thighs, and legs on pedestrian images and extract per-part color information as well as color displacement within the whole body. The extracted descriptors are then used in a matching step. Additionally, the authors introduce a method to customize the fit of Pictorial Structures on a specific person in cases when more images are available.

In contrast to such descriptive approaches relying on hand crafted features, other methods aim at learning a more discriminative feature model. For instance, Bak *et al.* [4] first apply a person detector, and then use AdaBoost to generate a visual signature consisting of Haar-like features. Gray and Tao [3] use AdaBoost to select the most relevant out of a set of color and texture features. To compare corresponding features they additionally estimate a likelihood ratio test providing a similarity function. Lin and Davis [5] propose to learn pairwise dissimilarities that can be applied for nearest neighbor classification. Schwartz and Davis [6] use Partial Least Squares reduction to project high dimensional signatures onto a low dimensional discriminant space. Another method is presented by Prosser *et al.* [7]. Here, the person re-identification problem is formulated as a ranking problem. The authors introduce Ensemble RankSVM, a method that learns a subspace where the potential true match gets the highest rank. Hirzer *et al.* [8] combine both of the aforementioned strategies, *i.e.*, they apply a descriptive and a discriminative model in parallel, showing that using the complementary information captured by both models leads to improved performance.

Exploiting geometry was proposed by Baltieri *et al.* [22]. They generate a highly sophisticated 3D human body model from foreground segmented person images, which can then be matched using a histogram based distance. The authors put strong assumptions on the input data and require a sufficiently accurate tracker, capable of extracting foreground segmented images as well as position

and orientation data of persons. To improve classification results some methods also exploit additional cues besides visual information. Makris *et al.* [23] and Rahimi *et al.* [13] simplify the problem by applying temporal constraints based on the spatial layout of the observed scene. Javed *et al.* [24] try to learn transitions between cameras to cope with illumination changes, and Zheng *et al.* [9] use contextual, visual information that comes from surrounding people.

A new direction that has recently been pursued is metric learning. Dikmen *et al.* [10] learn a Mahalanobis distance that is optimal for k-nearest neighbor classification using a maximum margin formulation. In fact, they extend the work of Weinberger and Saul [16] by introducing a rejection option into the LMNN framework. This option enables the LMNN classifier to return no matches if all nearest neighbors are beyond a certain distance, thereby telling the user that a searched person does not occur in a certain scene. Similarly, Zheng *et al.* [15] also use metric learning, but formulate it in a probabilistic manner. They seek a distance that maximizes the probability of a matching pair having a smaller distance than a non-matching pair.

3 Relaxed Pairwise Metric Learning

Metric learning for person re-identification has been previously addressed [10, 15]. However, as a main drawback, these methods require computationally complex optimization schemes, which hamper a practical application in large scale scenarios. Thus, the goal of this paper is to introduce a more efficient, still effective metric learning approach, which will be derived in the following.

One prominent approach for metric learning is Mahalanobis distance learning. Given n data points $x_i \in \mathbb{R}^m$, the goal is to estimate a matrix $\mathbf{M}$ such that

$$d_{\mathbf{M}}(x_i, x_j) = (x_i - x_j)^\top \mathbf{M}(x_i - x_j) \tag{1}$$

describes a pseudo-metric. In fact, this is assured if $\mathbf{M}$ is positive semi-definite, *i.e.*, $\mathbf{M} \succeq 0$. If $\mathbf{M} = \mathbf{\Sigma}^{-1}$, *i.e.*, the inverse of the covariance matrix $\mathbf{\Sigma}$, the distance defined by Eq. (1) is referred to as the Mahalanobis distance.

An alternative formulation for Eq. (1), which is more intuitive, is given via the squared distance

$$d_{\mathbf{L}}(x_i, x_j) = ||\mathbf{L}(x_i - x_j)||^2 \,, \tag{2}$$

which is easily obtained from

$$(x_i - x_j)^\top \mathbf{M}(x_i - x_j) = (x_i - x_j)^\top \underbrace{\mathbf{L}^\top \mathbf{L}}_{\mathbf{M}}(x_i - x_j) = ||\mathbf{L}(x_i - x_j)||^2 \,. \tag{3}$$

If additionally class labels are given, not only the generative structure of the data but also the discriminative information can be exploited. However, the person re-identification problem is lacking class labels, but we can exploit the information that the data is given in form of pairs. Thus, we break down the original multi-class problem into a two-class problem in two steps. First, we

transform the samples from the data space to the label agnostic difference space, which is inherently given by the metric definition in Eqs. (1) and (2). Second, the original class labels are discarded and the samples are arranged using pairwise equality and inequality constraints, where we obtain the class *same* $\mathcal{S}$ if two data points x_i and x_j share the same label and the class *different* $\mathcal{D}$ otherwise. In our case sharing a label means that the samples x_i and x_j describe views of the same person in different cameras.

In the following, we introduce an efficient but still effective solution for this task. We build on the simple observation that distances $d_{\mathbf{L}}(x_i, x_j)$ for $(x_i, x_j) \in \mathcal{S}$ should be small whereas they are expected to be large for $(x_i, x_j) \in \mathcal{D}$. This can be realized by using the following objective function:

$$\mathcal{L}(\mathbf{L}) = \frac{1}{|\mathcal{S}|} \sum_{(i,j)\in\mathcal{S}} ||\mathbf{L}(x_i - x_j)||^2 - \frac{1}{|\mathcal{D}|} \sum_{(i,j)\in\mathcal{D}} ||\mathbf{L}(x_i - x_j)||^2 \,. \tag{4}$$

Let $||\cdot||$ be the Frobenius norm, then we can re-write the squared distances:

$$\begin{aligned} ||\mathbf{L}(x_i - x_j)||^2 &= \langle \mathbf{L}(x_i - x_j), \mathbf{L}(x_i - x_j) \rangle = tr\left((x_i - x_j)^\top \mathbf{L}^\top \mathbf{L}(x_i - x_j)\right) \\ &= tr\left(\mathbf{M}(x_i - x_j)(x_i - x_j)^\top\right) \,, \end{aligned} \tag{5}$$

where $\langle \cdot, \cdot \rangle$ indicates the inner product and $tr(\cdot)$ the matrix' trace. In particular, we are exploiting the fact that the inner product can be transferred to a trace formulation and that the trace is invariant under cyclic permutations. Thus, we can re-write the objective function Eq. (4) to

$$\begin{aligned} \mathcal{L}(\mathbf{M}) = &\frac{1}{|\mathcal{S}|} \sum_{(i,j)\in\mathcal{S}} tr\left(\mathbf{M}(x_i - x_j)(x_i - x_j)^\top\right) \\ &- \frac{1}{|\mathcal{D}|} \sum_{(i,j)\in\mathcal{D}} tr\left(\mathbf{M}(x_i - x_j)(x_i - x_j)^\top\right) \,. \end{aligned} \tag{6}$$

Finally, let

$$\mathbf{\Sigma}_S = \frac{1}{|\mathcal{S}|} \sum_{(i,j)\in\mathcal{S}} (x_i - x_j)(x_i - x_j)^\top \tag{7}$$

$$\mathbf{\Sigma}_D = \frac{1}{|\mathcal{D}|} \sum_{(i,j)\in\mathcal{D}} (x_i - x_j)(x_i - x_j)^\top \tag{8}$$

be the sample covariance matrices of $\mathcal{S}$ and $\mathcal{D}$, respectively, and taking into account the linear properties of the trace, we get the objective function

$$\mathcal{L}(\mathbf{M}) = tr(\mathbf{M}\mathbf{\Sigma}_S) - tr(\mathbf{M}\mathbf{\Sigma}_D) = tr(\mathbf{M}(\mathbf{\Sigma}_S - \mathbf{\Sigma}_D)) \,. \tag{9}$$

To avoid trivial solutions and an overfitting to the training data an additional regularization is required. In our case, we would like to enforce the minimization of the distances within $\mathcal{S}$ but also to avoid unbounded distances for $\mathcal{D}$, which

would degenerate the metric. Thus, we introduce a scaling on both, $\mathcal{S}$ and $\mathcal{D}$, obtaining the optimization problem

$$\begin{aligned} &\min\ \mathcal{L}(\mathbf{M}) \\ &s.t.\ \mathbf{M} \succeq 0\,,\ \ \mathbf{L}\boldsymbol{\Sigma}_S\mathbf{L}^\top = \mathbf{I}\,,\ \ \mathbf{L}\boldsymbol{\Sigma}_D\mathbf{L}^\top = \mathbf{I}\,, \end{aligned} \tag{10}$$

which is hard to solve (*i.e.*, requires a complex iteration scheme). Hence, we relax the positivity constraint $\mathbf{M} \succeq 0$. Further taking into account that

$$tr(\mathbf{M}\boldsymbol{\Sigma}_S) = tr(\mathbf{L}\boldsymbol{\Sigma}_S\mathbf{L}^\top) = tr(\mathbf{L}\boldsymbol{\Sigma}_D\mathbf{L}^\top) = tr(\mathbf{M}\boldsymbol{\Sigma}_D) = m \tag{11}$$

the optimization problem Eq. (10) can be relaxed to solve

$$tr(\mathbf{M}(\boldsymbol{\Sigma}_S - \boldsymbol{\Sigma}_D)) = 0\,. \tag{12}$$

Technically, due to the relaxation the finally obtained matrix $\mathbf{M}$ does not describe a pseudo-metric. Nevertheless, the experimental results show that the estimated solution provides a sufficient approximation for the given task and that competitive results can be obtained; however, on a much lower computational effort.

4 Person Re-ID System

In the following, we introduce our proposed person re-identification system consisting of three stages: (1) feature extraction, (2) metric learning, and (3) classification. The overall system is illustrated in Figure 1. During training the metric between two cameras is estimated, which is then used for calculating the distances between an unknown sample and the samples given in the database. The three steps are discussed in more detail in the next sections.

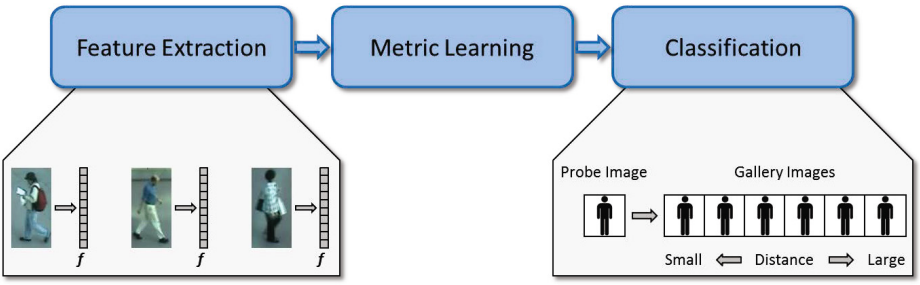


Fig. 1. Person re-identification system consisting of three stages: (1) feature extraction – dense sampling of color and texture features, (2) metric learning – exploiting the structure of similar and dissimilar pairs, (3) classification – nearest neighbor search under the learned metric.

4.1 Representation

Color and texture features have proven to be successful for the task of person re-identification. We use HSV and Lab color channels as well as Local Binary Patterns to create a person image representation. The features are extracted from 8x16 rectangular regions sampled from the image with a grid of 4x8 pixels, *i.e.*, 50% overlap in both directions, which is illustrated in Figure 2.

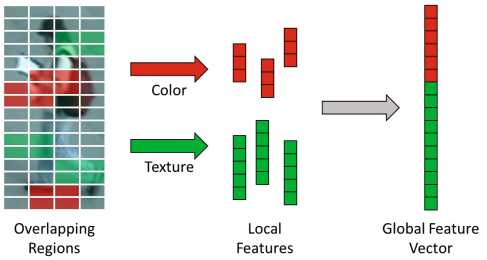


Fig. 2. Global image descriptor: different local features (HSV, Lab, LBP) are extracted from overlapping regions and are then concatenated to a single feature vector.

In each rectangular patch we calculate the mean values per color channel, which are then discretized to the range 0 to 40. Additionally, a histogram of LBP codes is generated from a gray value representation of the patch. These values are then put together to form a feature vector. Finally, the vectors from all regions are concatenated to generate a representation for the whole image.

4.2 Metric Learning

First of all, we run a PCA step, to reduce the dimensionality and for noise removal. In general, this step is not critical (the particular settings are given in Sec. 5), however, we recognized that for smaller datasets also a lower dimensional representation is sufficient. During training we learn a Mahalanobis metric $\mathbf{M}$ according to Eq. (12). Once $\mathbf{M}$ has been estimated, during evaluation the distance between two samples x_i and x_j is calculated via Eq. (1). Hence, additionally to the actual classification effort only linear projections are required.

4.3 Classification

In person re-identification we want to recognize a certain person across different, non-overlapping camera views. In the work at hand, we assume that we have already detected the persons in all camera views, *i.e.*, we do not tackle the detection problem. The goal of person re-identification now is to find a person image that has been selected in one view (*probe image*) in all the images from another view (*gallery images*). This is achieved by calculating the distances between the probe image and all gallery images using the learned metric, and returning those gallery images with the smallest distances as potential matches.

5 Experimental Results

We evaluated our approach on three publicly available datasets, the *VIPeR* dataset [11], the *PRID 2011* dataset (single shot version) [8], and the *ETHZ* dataset [6]. Examples of all three are shown in Figure 3. We chose these datasets because they provide many challenges faced in real world person re-identification applications, *e.g.*, viewpoint, pose and illumination changes, different backgrounds, image resolutions, occlusions, etc. Since the *VIPeR* dataset is widely used for evaluating person re-identification methods, we performed a more detailed analysis on this dataset. This includes an analysis of different training set sizes, comparisons to the state-of-the-art for person re-identification as well as for metric learning, and an analysis of computation times. As mentioned before, the PCA dimensions are not critical, but in particular we used 75 dimensions for *VIPeR*, 40 for *PRID 2011* and *ETHZ* SEQ. #1, 20 for *ETHZ* SEQ. #2, and 15 for *ETHZ* SEQ. #3.

(a) (b) (c)

Fig. 3. Example image pairs from the *VIPeR* (a), the *PRID 2011* (b), and the *ETHZ* dataset (c). Upper and lower row correspond to different appearances of the same person.

5.1 VIPeR Dataset

The *VIPeR* dataset contains 632 person image pairs taken from two different camera views. Changes of viewpoint, illumination and pose are the most prominent sources of appearance variation between the two images of a person. For evaluation we followed the procedure described in [3]. 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 a probe and a gallery set. A single image from the probe set is then selected and matched with all images from the gallery set. This process is repeated for all images in the probe set. The whole evaluation procedure is carried out 10 times, and the average result is reported in form of a Cumulative Matching Characteristic (CMC) curve [2], which represents the expectation of finding the true match within the first r ranks.

The thus obtained results are shown in Figure 4. In addition, we give a comparison to simple feature matching using Euclidean distance and to baseline approaches, *i.e.*, Linear Discriminant Analysis (LDA) and standard Mahalanobis

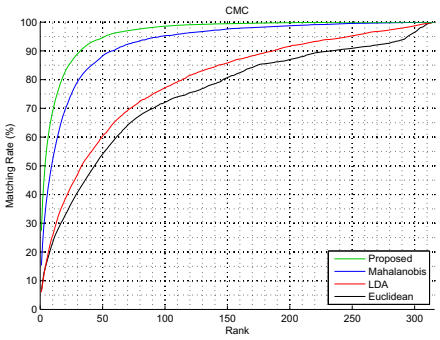


Fig. 4. Average CMC curve of our approach, Mahalanobis, LDA, and feature matching using Euclidean distance on the *VIPeR* dataset.

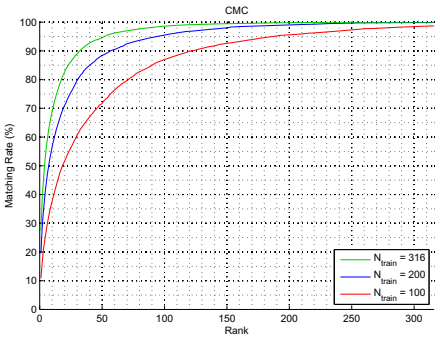


Fig. 5. Average CMC curve of our approach for different training set sizes on the *VIPeR* dataset.

metric. It is obvious that using the proposed metric leads to a huge performance gain over simple feature matching and that also the baselines can be outperformed. Moreover, in Table 1 we show a comparison of our approach to the state-of-the-art in both, person re-identification and metric learning. In the latter case, exactly the same features and training and test set splits have been used. As can be seen, our method outperforms all other methods over the whole range of ranks, however, at much reduced computational costs.

In Table 1 we also analyze the computation time of our method using a Matlab implementation on a 2.83 GHz quad core CPU. The big advantage of our approach compared to others is its training time efficiency, since it does not rely on computationally complex optimization schemes. Please note, the standard Mahalanobis metric is still more efficient, however, the performance gain justifies the slightly higher computational effort. Using the learned metric during the evaluation step is efficient either, making it suitable for even large scale problems.

Moreover, we investigated the influence of the training set size on the performance. Following [15], we reduced the number of training samples from 316 to 200 and 100 respectively. A comparison of our method using the original and the reduced training sets is depicted in Figure 5. As can be seen, reducing the number of training samples to 200 has only little influence on our method. Further reducing it to 100 decreases the performance notably. With decreasing number of samples estimating a reliable metric becomes more and more difficult. Still, our method outperforms the state-of-the-art on both reduced training sets, as shown in Table 2.

5.2 PRID 2011 Dataset

The *PRID 2011* dataset consists of person images recorded from two different static surveillance cameras, where we used the single shot scenario (one image per

Table 1. Comparison of matching rates in [%] at different ranks r and, if available, average training times per trial on the *VIPeR* dataset. (* indicates that the best run was reported, which cannot be directly compared to the other results!)

Method	$r = 1$	10	20	50	100	Timings
Proposed	**27**	**69**	**83**	**95**	**99**	0.1 sec
LMNN [16]	18	59	75	91	97	75 sec
ITML [17]	14	52	71	90	98	16 sec
LDML [18]	5	21	30	51	71	0.8 sec
Mahalanobis	15	52	70	88	95	**0.003 sec**
LDA	6	26	38	60	77	0.1 sec
ELF [3]	12	43	60	81	93	5 hrs
SDALF [12]	20	50	65	85	-	-
ERSVM [7]	13	50	67	85	94	13 min
DDC [8]	19	52	65	80	91	-
PS [21]	22	57	71	87	-	-
PRDC [15]	16	54	70	87	97	15 min
LMNN-R* [10]	20	68	80	93	99	-

Table 2. Matching rates of our approach and the best methods reported in [15] in [%] at different ranks r on reduced training sets

Method	$N_{train} = 200$			$N_{train} = 100$		
	$r = 1$	10	20	$r = 1$	10	20
Proposed	**20**	**56**	**71**	**11**	**38**	**52**
Best reported in [15]	13	47	63	9	34	49

person in each camera view) for our evaluation. Typical challenges on this dataset are viewpoint and pose changes as well as significant differences in illumination, background and camera characteristics. 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 followed the procedure described in [8], *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 10 times and the result is reported in form of an average CMC curve in Figure 6. As can be seen, applying the proposed metric leads to superior performance compared to using LDA or Euclidean distance. Since the diversity in this dataset is smaller, *i.e.*, it contains a lot of similar persons, the benefits of discriminative learning cannot be fully exploited.

Table 3 compares our approach to the work of Hirzer *et al.* [8] for different ranks. Their approach uses a descriptive and a discriminative model to rank the images in the gallery set. The highest matching rates are achieved by a combination of both models. However, their system is designed for human-in-the-loop interaction, so

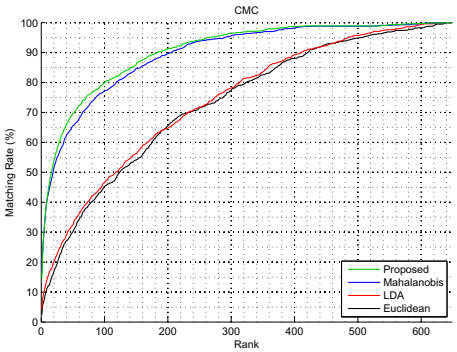


Fig. 6. Average CMC curve of our approach, Mahalanobis, LDA, and feature matching using Euclidean distance on the *PRID 2011* dataset

a human operator actually makes the decision which of both models to use. In particular, we compare our method to their descriptive model, which also uses a single shot setup, *i.e.*, person image pairs. In contrast, their discriminative model uses a multi shot setup, so that a fair comparison is not possible. As can be seen by the numbers, our method clearly outperforms their descriptive model at all ranks, especially in the middle range with nearly 20% performance gain.

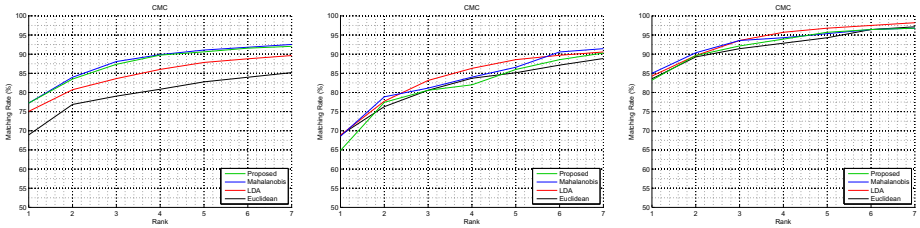
Also note that the approach of Hirzer *et al.* does not need a training phase, so the authors do not split the image pairs into a training and a test set. Instead, they use all 200 image pairs and all 749 gallery images of camera view B. This is slightly different from our setup, where we exclude the 100 training image pairs from the test phase.

Table 3. Matching rates of our approach and the descriptive model of [8] in [%] at different ranks r on the *PRID 2011* dataset

Method	$r = 1$	10	20	50	100
Proposed	**15**	**42**	**54**	**70**	**80**
Descr. Model [8]	4	24	37	56	70

5.3 ETHZ Dataset

The *ETHZ* dataset [6] contains video sequences of urban scenes captured from moving cameras. Originally proposed for pedestrian detection [25] it was later modified for benchmarking person re-identification approaches. The dataset consists of person images extracted from three video sequences 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 resized to 64x32 pixels. The most challenging aspects of the *ETHZ* dataset are illumination changes and occlusions. However, since person images are captured from a single moving camera, the dataset does not provide a realistic scenario



(a) SEQ. #1 (b) SEQ. #2 (c) SEQ. #3

Fig. 7. Average CMC curve of our approach, Mahalanobis, LDA, and feature matching using Euclidean distance on the *ETHZ* dataset. According to [6], only the first 7 ranks are shown.

for person re-identification with multiple, disjoint cameras, different viewpoints, different camera characteristics, etc.

Despite this limitation it is commonly used for person re-identification, so we also evaluated our approach on this dataset. Similar to [12] and [6] 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. After learning a metric from the training pairs, each image in the probe set is matched with all images in the gallery set. The whole procedure is repeated 10 times to generate an average CMC curve, as shown in Figure 7.

Table 4. Matching rates of our approach, SDALF (single shot), and PLS in [%] on the *ETHZ* dataset at the first 7 ranks

	SEQ. #1							SEQ. #2							SEQ. #3						
Method	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7
Proposed	77	83	**87**	**90**	**91**	**92**	**92**	65	77	**81**	82	**86**	**89**	**90**	**83**	**90**	**92**	**94**	**96**	**96**	**97**
SDALF [12]	65	73	77	79	81	82	84	64	74	79	**83**	85	87	89	76	83	86	88	90	92	93
PLS [6]	**79**	**85**	86	87	88	89	90	**74**	**79**	**81**	**83**	84	85	87	77	81	82	84	85	87	89

As already stated above, here only tracks of persons are considered but not different camera views. Hence, as expected, metric learning which mainly captures the transitions in inter-camera feature space has only little influence. Moreover, especially for SEQ. #2 and SEQ. #3 the number of training samples is very small, making metric learning quite difficult. Nevertheless, for the larger sequence, SEQ. #1, a clear performance gain can be achieved. Moreover, in Table 4 we compare our matching rates to two other methods that also use a single shot evaluation, namely SDALF [12] and PLS [6]. On SEQ. #1, we achieve state-of-the-art results comparable to [6], and outperform [12]. Interestingly, even using the Euclidean distance slightly outperforms [12]. Although on the other two sequences the power of metric learning is not fully exploited, results in the range of the state-of-the-art (SEQ. #2) or better (SEQ. #3) can be obtained.

6 Conclusion

Recently, metric learning was introduced for the task of person re-identification, which is a considerable tradeoff between descriptive and discriminative modeling. In fact, good results can be obtained, however, at high computational costs. Since for the given task even the training should be efficient, in this work we targeted a more efficient metric learning approach. In particular, we proposed to use a discriminative Mahalanobis metric learning, which can be efficiently solved after some relaxations. The benefits of the proposed method are clearly demonstrated in the experimental results, where we show state-of-the-art or even better results on three standard benchmark datasets, *i.e.*, *VIPeR*, *PRID 2011*, and *ETHZ*. This is in particular of interest, since we build on a quite simple representation. In fact, compared to the usage of Euclidean distance metric learning drastically boosts the performance. Moreover, the results reveal that the benefits of discriminative metric learning can be fully exploited for highly diverse views. Furthermore, we gave evaluations on timings and the importance of number of training samples. Future work would include the application of more sophisticated features and a more detailed study of the relaxed optimization problem.

Acknowledgments. The work was supported by the Austrian Science Foundation (FWF) project Advanced Learning for Tracking and Detection in Medical Workflow Analysis (I535-N23) and by the Austrian Research Promotion Agency (FFG) project SHARE in the IV2Splus program.

References

1. Gheissari, N., Sebastian, T.B., Hartley, R.: Person reidentification using spatiotemporal appearance. In: Proc. CVPR (2006)
2. Wang, X., Doretto, G., Sebastian, T.B., Rittscher, J., Tu, P.H.: Shape and appearance context modeling. In: Proc. ICCV (2007)
3. Gray, D., Tao, H.: Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 262–275. Springer, Heidelberg (2008)
4. Bak, S., Corvee, E., Brémond, F., Thonnat, M.: Person re-idendification using Haar-based and DCD-based signature. In: Proc. Workshop on Activity Monitoring by Multi-Camera Surveillance Systems (2010)
5. Lin, Z., Davis, L.S.: Learning pairwise dissimilarity profiles for appearance recognition in visual surveillance. In: Advances Int'l Visual Computing Symposium (2008)
6. Schwartz, W.R., Davis, L.S.: Learning discriminative appearance-based models using partial least squares. In: Proc. Brazilian Symposium on Computer Graphics and Image Processing (2009)
7. Prosser, B., Zheng, W.S., Gong, S., Xiang, T.: Person re-identification by support vector ranking. In: Proc. BMVC (2010)
8. Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H.: Person Re-identification by Descriptive and Discriminative Classification. In: Heyden, A., Kahl, F. (eds.) SCIA 2011. LNCS, vol. 6688, pp. 91–102. Springer, Heidelberg (2011)
9. Zheng, W.S., Gong, S., Xiang, T.: Associating groups of people. In: Proc. BMVC (2009)

10. Dikmen, M., Akbas, E., Huang, T.S., Ahuja, N.: Pedestrian Recognition with a Learned Metric. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010, Part IV. LNCS, vol. 6495, pp. 501–512. Springer, Heidelberg (2011)
11. Gray, D., Brennan, S., Tao, H.: Evaluating appearance models for recognition, reacquisition, and tracking. In: Proc. PETS (2007)
12. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: Proc. CVPR (2010)
13. Rahimi, A., Dunagan, B., Darrell, T.: Simultaneous calibration and tracking with a network of non-overlapping sensors. In: Proc. CVPR (2004)
14. Chapelle, O., Keerthi, S.S.: Efficient algorithms for ranking with SVMs. Information Retrieval 13, 201–215 (2010)
15. Zheng, W.S., Gong, S., Xiang, T.: Person re-identification by probabilistic relative distance comparison. In: Proc. CVPR (2011)
16. Weinberger, K.Q., Saul, L.K.: Fast solvers and efficient implementations for distance metric learning. In: Proc. ICML (2008)
17. Davis, J.V., Kulis, B., Jain, P., Sra, S., Dhillon, I.S.: Information-theoretic metric learning. In: Proc. ICML (2007)
18. Guillaumin, M., Verbeek, J., Schmid, C.: Is that you? Metric learning approaches for face identification. In: Proc. ICCV (2009)
19. Ghodsi, A., Wilkinson, D.F., Southey, F.: Improving embeddings by flexible exploitation of side information. In: Proc. Int'l Joint Conf. on Artificial Intelligence (2007)
20. Alipanahi, B., Biggs, M., Ghodsi, A.: Distance metric learning vs. fisher discriminant analysis. In: Proc. AAAI Conf. on Artificial Intelligence (2008)
21. Cheng, D.S., Cristani, M., Stoppa, M., Bazzani, L., Murino, V.: Custom pictorial structures for re-identification. In: Proc. BMVC (2011)
22. Baltieri, D., Vezzani, R., Cucchiara, R.: SARC3D: A New 3D Body Model for People Tracking and Re-identification. In: Maino, G., Foresti, G.L. (eds.) ICIAP 2011, Part I. LNCS, vol. 6978, pp. 197–206. Springer, Heidelberg (2011)
23. Makris, D., Ellis, T., Black, J.: Bridging the gaps between cameras. In: Proc. CVPR (2004)
24. Javed, O., Shafique, K., Shah, M.: Appearance modeling for tracking in multiple non-overlapping cameras. In: Proc. CVPR (2005)
25. Ess, A., Leibe, B., Van Gool, L.: Depth and appearance for mobile scene analysis. In: Proc. ICCV (2007)