Person Re-identification by Probabilistic Relative Distance Comparison*

Wei-Shi Zheng[1,2], Shaogang Gong[2], and Tao Xiang[2]
[1]School of Information Science and Technology, Sun Yat-sen University, China
[2]School of Electronic Engineering and Computer Science, Queen Mary University of London, UK
wszheng@ieee.org, sgg@eecs.qmul.ac.uk, txiang@eecs.qmul.ac.uk

Abstract

Matching people across non-overlapping camera views, known as person re-identification, is challenging due to the lack of spatial and temporal constraints and large visual appearance changes caused by variations in view angle, lighting, background clutter and occlusion. To address these challenges, most previous approaches aim to extract visual features that are both distinctive and stable under appearance changes. However, most visual features and their combinations under realistic conditions are neither stable nor distinctive thus should not be used indiscriminately. In this paper, we propose to formulate person re-identification as a distance learning problem, which aims to learn the optimal distance that can maximises matching accuracy regardless the choice of representation. To that end, we introduce a novel Probabilistic Relative Distance Comparison (PRDC) model, which differs from most existing distance learning methods in that, rather than minimising intra-class variation whilst maximising intra-class variation, it aims to maximise the probability of a pair of true match having a smaller distance than that of a wrong match pair. This makes our model more tolerant to appearance changes and less susceptible to model over-fitting. Extensive experiments are carried out to demonstrate that 1) by formulating the person re-identification problem as a distance learning problem, notable improvement on matching accuracy can be obtained against conventional person re-identification techniques, which is particularly significant when the training sample size is small; and 2) our PRDC outperforms not only existing distance learning methods but also alternative learning methods based on boosting and learning to rank.

1. Introduction

There has been an increasing interest in matching people across disjoint camera views in a multi-camera system, known as the person re-identification problem [10, 7, 14, 8, 3]. For understanding behaviour of people in a large area of public space covered by multiple no-overlapping cameras, it is critical that when a target disappears from one view, he/she can be identified in another view among a crowd of people. Despite the best efforts from computer vision researchers in the past 5 years, the person re-identification problem remains largely unsolved. Specifically, in a busy uncontrolled environment monitored by cameras from a distance, person verification relying upon biometrics such as face and gait is infeasible or unreliable. Without accurate temporal and spatial constraints given the typically large gaps between camera views, visual appearance features alone, extracted mainly from clothing, are intrinsically weak for matching people (e.g. most people in winter wear dark clothes). In addition, a person's appearance often undergoes large variations across different camera views due to significant changes in view angle, lighting, background clutter and occlusion (see Fig. 1), resulting in different people appearing more alike than that of the same person across different camera views (see Figs. 4 and 5).

Figure 1. Typical examples of appearance changes caused by cross-view variations in view angle, lighting, background clutter and occlusion. Each column shows two images of the same person from two different camera views.

Most existing studies have tried to address the above problems by seeking a more distinctive and stable feature representation of people's appearance, ranging widely from color histogram [10, 7], graph model [4], spatial co-occurrence representation model [14], principal axis histogram [8], rectangle region histogram [2], to combinations of multiple features [7, 3]. After feature extraction, existing methods simply choose a standard distance mea-

*Most of this work was done when the first author was at QMUL.

sure such as l_1-norm [14], l_2-norm based distance [8], or Bahattacharyya distance [7]. However under severe viewing condition changes that can cause significant intra-object appearance variation (e.g. view angle, lighting, occlusion), computing a set of features that are both distinctive and stable under all condition changes is extremely hard if not impossible under realistic conditions. Moreover, given that certain features could be more reliable than others under a certain condition, applying a standard distance measure is undesirable as it essentially treats all features equally without discarding bad features selectively in each individual matching circumstance.

In this paper, we propose to formulate person re-identification as a distance learning problem which aims to learn the optimal distance metric that can maximise matching accuracy regardless the choice of representation. To that end, a novel Probabilistic Relative Distance Comparison (PRDC) model is proposed. The objective function used by PRDC aims to maximise the probability of a pair of true match (i.e. two true images of person A) having a smaller distance than that of a pair of related wrong match (i.e. two images of person A and B respectively). This is in contrast with that of a conventional distance learning approach, which aims to minimise intra-class variation in an absolute sense (i.e. making all images of person A more similar) whilst maximising inter-class variation (i.e. making all images of person A and B more dissimilar). Our approach is motivated by the nature of our problem. Specifically, the person re-identification problem has three characteristics: 1) the intra-class variation can be large and importantly can be significantly varied for different classes as it is caused by different condition changes (see Fig. 1); 2) the inter-class variation also varies drastically across different pairs of classes; and 3) annotating matched people across camera views is tedious and typically only limited number of classes (people) are available for training with each class containing only a handful of images of a person from different camera views (i.e. under-sampling for building a representative class distribution). By exploring a relative distance comparison model probabilistically, our model is more tolerant to the large intra/inter-class variation and severe overlapping of different classes in a multi-dimensional feature space. Furthermore, due to the third characteristics of under-sampling, a model could be easily over-fitted if it is learned by minimising intra-class distance and maximising inter-class distance simultaneously by brutal force. In contrast, our approach is able to learn a distance with much reduced complexity thus alleviating the over-fitting problem, as validated by our extensive experiments.

Related work–Although it has not been exploited for person re-identification, distance learning is a well-studied problem with a large number of methods reported in the literature [16, 5, 17, 7, 17, 12, 15, 9, 1]. However, most of them suffer from the over-fitting problem as explained above. Recently, a few approaches attempt to alleviate the problem by incorporating the idea of relative distance comparison as our PRDC model [12, 15, 9]. However, in these works, the relative distance comparison is not quantified probabilistically, and importantly is used as an optimisation constraint rather than objective function. Therefore these approaches, either implicitly [12, 9] or explicitly [15] still aim to learn a distance by which each class becomes more compact whilst being more separable from each other in an absolute sense. We demonstrate through experiments that they remain susceptible to over-fitting for person re-identification.

There have been a couple of feature selection based methods proposed specifically for person re-identification [7, 11]. Gray et al. [7] proposed to use boosting to select a subset of optimal features for matching people. However, in a boosting framework, good features are only selected sequentially and independently in the original feature space where different classes can be heavily overlapped. Such selection may not be globally optimal. Rather than selecting features individually and independently (local selection), we aim to learn an optimal distance measure for all features jointly via distance learning (global selection). An alternative global selection approach was developed based on RankSVM [11]. By formulating the person re-identification as a ranking problem, the RankSVM approach shares the spirit of relative comparison in our model. Nevertheless, our approach is more principled and tractable than the RankSVM in that 1) PRDC is a second-order feature selection approach whereas RankSVM is a first-order one which is not able to exploit correlations of different features; 2) although RankSVM alleviates the over-fitting problem by fusing a ranking error function with a large margin function in its objective function, the probabilistic formulation of our objective function makes PRDC more tolerant to large intra- and inter-class variations and data sparsity; 3) tuning the critical free parameter of RankSVM that determines the weight between the margin function and the ranking error function is computationally costly and can be sub-optimal given limited data. In contrast, our PRDC model does not such a problem. We demonstrate the advantage of our approach over both the Boosting [7] and RankSVM [11] based methods through experiments.

The main contributions of this work are two-fold. 1) We formulate the person re-identification problem as a distance learning problem, which leads to noteworthy improvement on re-identification accuracy. To the best of our knowledge, it has not been investigated before. 2) We propose a probabilistic relative distance comparison based method that overcomes the limitations of existing distance learning methods when applied to person re-identification.

2. Probabilistic relative distance comparison for person re-identification

Let us formally cast the person re-identification problem into the following distance learning problem. For an image $\mathbf{z}$ of person A, we wish to learn a re-identification model to successfully identify another image $\mathbf{z}'$ of the same person captured elsewhere in space and time. This is achieved by learning a distance function $f(\cdot,\cdot)$ so that $f(\mathbf{z},\mathbf{z}') < f(\mathbf{z},\mathbf{z}'')$, where $\mathbf{z}''$ is an image of any other person except A. To that end, given a training set $\{(\mathbf{z}_i, y_i)\}$, where $\mathbf{z}_i \in \mathcal{Z}$ is a multi-dimensional feature vector representing the appearance of a person in one view and y_i is its class label (person ID), we define a pairwise set $\mathbb{O} = \{\mathbb{O}_i = (\mathbf{x}_i^p, \mathbf{x}_i^n)\}$, where each element of a pair-wise data $\mathbb{O}_i$ itself is computed using a pair of sample feature vectors. More specifically, $\mathbf{x}_i^p$ is a difference vector computed between a pair of relevant samples (of the same class/person) and $\mathbf{x}_i^n$ is a difference vector from a pair of related irrelevant samples, i.e. only one sample for computing $\mathbf{x}_i^n$ is one of the two relevant samples for computing $\mathbf{x}_i^p$ and the other is a mis-match from another class. The difference vector $\mathbf{x}$ between any two samples $\mathbf{z}$ and $\mathbf{z}'$ is computed by

$$\mathbf{x} = d(\mathbf{z},\mathbf{z}'),\ \mathbf{z},\mathbf{z}' \in \mathcal{Z} \tag{1}$$

where d is an entry-wise difference function that outputs a difference vector between $\mathbf{z}$ and $\mathbf{z}'$. The specific form of function d will be described in Sec. 2.3.

Given the pairwise set $\mathbb{O}$, a distance function f can be learned based on relative distance comparison so that a distance between a relevant sample pair ($f(\mathbf{x}_i^p)$) is smaller than that between a related irrelevant pair ($f(\mathbf{x}_i^n)$). That is $f(\mathbf{x}_i^p) < f(\mathbf{x}_i^n)$ for each pair-wise data $\mathbb{O}_i$. To this end, we measure the probability of the distance between a relevant pair being smaller than that of a related irrelevant pair as:

$$P(f(\mathbf{x}_i^p) < f(\mathbf{x}_i^n)) = \left(1+\exp\left\{f(\mathbf{x}_i^p)-f(\mathbf{x}_i^n)\right\}\right)^{-1}. \tag{2}$$

We assume the events of distance comparison between a relevant pair and an irrelevant pair, i.e. $f(\mathbf{x}_i^p) < f(\mathbf{x}_i^n)$, are independent[1]. Then, based on the maximum likelihood principle, the optimal function f can be learned as follows:

$$\begin{aligned} f &= \arg\min_f r(f,\mathbb{O}), \\ r(f,\mathbb{O}) &= -\log\Big(\prod\nolimits_{\mathbb{O}_i} P(f(\mathbf{x}_i^p) < f(\mathbf{x}_i^n))\Big). \end{aligned} \tag{3}$$

The distance function f is parameterised as a Mahanalobis (quadratic) based distance function:

$$f(\mathbf{x}) = \mathbf{x}^T\mathbf{M}\mathbf{x},\ \mathbf{M} \succeq 0 \tag{4}$$

where $\mathbf{M}$ is a semidefinite matrix. The distance learning problem thus becomes learning $\mathbf{M}$ using Eqn. (3). Directly learning $\mathbf{M}$ using semidefinite program techniques is computationally expensive for high dimensional data [15]. In particular, we found out in our experiments that given a dimensionality of thousands, typical for visual object representation, a distance learning method based on learning $\mathbf{M}$ becomes intractable. To overcome this problem, we perform eigenvalue decomposition on $\mathbf{M}$:

$$\mathbf{M} = \mathbf{A}\mathbf{\Lambda}\mathbf{A}^T = \mathbf{W}\mathbf{W}^T,\ \mathbf{W} = \mathbf{A}\mathbf{\Lambda}^{\frac{1}{2}}, \tag{5}$$

where the columns of $\mathbf{A}$ are orthonormal eigenvectors of $\mathbf{M}$ and the diagonals of $\mathbf{\Lambda}$ are the corresponding eigenvalues. Note that $\mathbf{W}$ is orthogonal. Therefore, learning a function f is equivalent to learning an orthogonal matrix $\mathbf{W} = (\mathbf{w}_1, \cdots, \mathbf{w}_l, \cdots, \mathbf{w}_L)$ such that

$$\begin{aligned} &\mathbf{W} = \arg\min_{\mathbf{W}} r(\mathbf{W},\mathbb{O}),\ s.t.\ \mathbf{w}_i^T\mathbf{w}_j = 0,\ \forall i \neq j \\ &r(\mathbf{W},\mathbb{O}) = \sum\nolimits_{\mathbb{O}_i} \log(1+\exp\left\{||\mathbf{W}^T\mathbf{x}_i^p||^2 - ||\mathbf{W}^T\mathbf{x}_i^n||^2\right\}). \end{aligned} \tag{6}$$

2.1. An iterative optimisation algorithm

It is important to point out that our optimisation criterion (6) may not be a convex optimisation problem against the orthogonal constraint due to the relative comparison modelling. It means that deriving an global solution by directly optimising $\mathbf{W}$ is not straightforward. In this work we formulate an iterative optimisation algorithm to learn an optimal $\mathbf{W}$, which also aims to seek a low rank (non-trivial) solution automatically. This is critical for reducing the model complexity thus overcoming the overfitting problem given sparse data.

Starting from an empty matrix, after iteration ℓ, a new estimated column $\mathbf{w}_\ell$ is added to $\mathbf{W}$. The algorithm terminates after L iterations when a stopping criterion is met. Each iteration consists of two steps as follows:

Step 1. Assume that after ℓ iterations, a total of ℓ orthogonal vectors $\mathbf{w}_1, \cdots, \mathbf{w}_\ell$ have been learned. To learn the next orthogonal vector $\mathbf{w}_{\ell+1}$, let

$$a_i^{\ell+1} = \exp\{\sum\nolimits_{j=0}^{\ell} ||\mathbf{w}_j^T\mathbf{x}_i^{p,j}||^2 - ||\mathbf{w}_j^T\mathbf{x}_i^{n,j}||^2\}, \tag{7}$$

where we define $\mathbf{w}_0 = \mathbf{0}$, and $\mathbf{x}_i^{p,\ell}$ and $\mathbf{x}_i^{n,\ell}$ are the difference vectors at the ℓ-th iteration defined as follows:

$$\begin{aligned} &\mathbf{x}_i^{s,\ell} = \mathbf{x}_i^{s,\ell-1} - \tilde{\mathbf{w}}_{\ell-1}\tilde{\mathbf{w}}_{\ell-1}^T\mathbf{x}_i^{s,\ell-1}, \\ &\quad s \in \{p,n\},\ i = 1,\cdots,\left|\mathbb{O}\right|,\ \ell \geq 1, \\ &\text{where}\ \ \tilde{\mathbf{w}}_{\ell-1} = \mathbf{w}_{\ell-1}/||\mathbf{w}_{\ell-1}||. \end{aligned} \tag{8}$$

Note that we define $\mathbf{x}_i^{s,0} = \mathbf{x}_i^s$, $s \in \{p,n\}$, and $\tilde{\mathbf{w}}_0 = \mathbf{0}$.

Step 2. Obtain $\mathbf{x}_i^{p,\ell+1}$, $\mathbf{x}_i^{n,\ell+1}$ by Eqn. (8). Let $\mathbb{O}^{\ell+1} = \{\mathbb{O}_i^{\ell+1} = (\mathbf{x}_i^{p,\ell+1}, \mathbf{x}_i^{n,\ell+1})\}$. Then, learn a new optimal projection $\mathbf{w}_{\ell+1}$ on $\mathbb{O}^{\ell+1}$ as follows:

$$\mathbf{w}_{\ell+1} = \arg\min_{\mathbf{w}} r_{\ell+1}(\mathbf{w},\mathbb{O}^{\ell+1}),\ \text{where} \tag{9}$$

$$\begin{aligned} &r_{\ell+1}(\mathbf{w},\mathbb{O}^{\ell+1}) \\ &= \sum\nolimits_{\mathbb{O}_i^{\ell+1}} \log(1+a_i^{\ell+1}\exp\left\{||\mathbf{w}^T\mathbf{x}_i^{p,\ell+1}||^2 - ||\mathbf{w}^T\mathbf{x}_i^{n,\ell+1}||^2\right\}). \end{aligned}$$

[1] Note that we do not assume the data are independent.

We seek an optimal solution by a gradient descent method:

$$\mathbf{w}_{\ell+1} \leftarrow \mathbf{w}_{\ell+1} - \lambda \cdot \frac{\partial r_{\ell+1}}{\partial \mathbf{w}_{\ell+1}}, \ \ \lambda \geq 0, \tag{10}$$

$$\frac{\partial r_{\ell+1}}{\partial \mathbf{w}_{\ell+1}} = \sum_{\mathbb{O}_i^{\ell+1}} \frac{2 \cdot a_i^{\ell+1} \cdot \exp\{||\mathbf{w}_{\ell+1}^T \mathbf{x}_i^{p,\ell+1}||^2 - ||\mathbf{w}_{\ell+1}^T \mathbf{x}_i^{n,\ell+1}||^2\}}{1 + a_i^{\ell+1} \cdot \exp\{||\mathbf{w}_{\ell+1}^T \mathbf{x}_i^{p,\ell+1}||^2 - ||\mathbf{w}_{\ell+1}^T \mathbf{x}_i^{n,\ell+1}||^2\}}$$
$$\times (\mathbf{x}_i^{p,\ell+1} \mathbf{x}_i^{p,\ell+1^T} - \mathbf{x}_i^{n,\ell+1} \mathbf{x}_i^{n,\ell+1^T}) \mathbf{w}_{\ell+1}.$$

where λ is a step length automatically determined at each gradient update step. According to the descent direction in Eqn. (10) the initial value of $\mathbf{w}_{\ell+1}$ for the gradient descent method is set to

$$\mathbf{w}_{\ell+1} = |\mathbb{O}^{\ell+1}|^{-1} \sum\nolimits_{\mathbb{O}_i^{\ell+1}} (\mathbf{x}_i^{n,\ell+1} - \mathbf{x}_i^{p,\ell+1}). \tag{11}$$

Note that the update in Eqn. (8) deducts information from each sample $\mathbf{x}_i^{s,\ell-1}$ affected by $\mathbf{w}_{\ell-1}$ as $\mathbf{w}_{\ell-1}^T \mathbf{x}_i^{s,\ell} = 0$, so that the next learned vector $\mathbf{w}_\ell$ will only quantify the part of the data left from the last step, i.e. $\mathbf{x}_i^{s,\ell}$. In addition, $a_i^{\ell+1}$ indicates the trends in the change of distance measures for $\mathbf{x}_i^p$ and $\mathbf{x}_i^n$ over previous iterations and serve as *a priori* weight for learning $\mathbf{w}_\ell$.

The iteration of the algorithm (for $\ell > 1$) is terminated when the following criterion is met:

$$r_\ell(\mathbf{w}_\ell, \mathbb{O}^\ell) - r_{\ell+1}(\mathbf{w}_{\ell+1}, \mathbb{O}^{\ell+1}) < \varepsilon. \tag{12}$$

where ε is a small tolerance value set to 10^{-6} in this work. The algorithm is summarised in Algorithm 1.

Algorithm 1: Learning the PRDC model

Data: $\mathbb{O} = \{\mathbb{O}_i = (\mathbf{x}_i^p, \mathbf{x}_i^n)\}$, $\varepsilon > 0$
begin
  $\mathbf{w}_0 \longleftarrow \mathbf{0}$, $\tilde{\mathbf{w}}_0 \longleftarrow \mathbf{0}$;
  $\mathbf{x}_i^{s,0} \longleftarrow \mathbf{x}_i^s, s \in \{p, n\}$, $\mathbb{O}^0 \longleftarrow \mathbb{O}$;
  $\ell \longleftarrow 0$;
  while *1* **do**
    Compute $a_i^{\ell+1}$ by Eqn. (7);
    Compute $\mathbf{x}_i^{s,\ell+1}$, $s \in \{p, n\}$ by Eqn. (8);
    $\mathbb{O}^{\ell+1} \longleftarrow \{\mathbb{O}_i^{\ell+1} = (\mathbf{x}_i^{p,\ell+1}, \mathbf{x}_i^{n,\ell+1})\}$;
    Estimate $\mathbf{w}_{\ell+1}$ using Eqn. (9);
    $\tilde{\mathbf{w}}_{\ell+1} = \frac{\mathbf{w}_{\ell+1}}{||\mathbf{w}_{\ell+1}||}$;
    if $(\ell > 1) \& (r_\ell(\mathbf{w}_\ell, \mathbb{O}^\ell) - r_{\ell+1}(\mathbf{w}_{\ell+1}, \mathbb{O}^{\ell+1}) < \varepsilon)$ **then**
      break;
    end
    $\ell \longleftarrow \ell + 1$;
  end
end
Output: $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_\ell]$

2.2. Theoretical validation

The following two theorems validate that the proposed iterative optimisation algorithm learns a set of orthogonal projections $\{\mathbf{w}_\ell\}$ that iteratively decrease the objective function in Criterion (6).

Theorem 1. *The learned vectors* $\mathbf{w}_\ell$, $\ell = 1, \cdots, L$, *are orthogonal to each other.*

Proof. Assume that $\ell - 1$ orthogonal vectors $\{\mathbf{w}_j\}_{j=1}^{\ell-1}$ have been learned. Let $\mathbf{w}_\ell$ be the optimal solution of Criterion (9) at the ℓ iteration. First, we know that $\mathbf{w}_\ell$ is in the range space[2] of $\{\mathbf{x}_i^{p,\ell}\} \cup \{\mathbf{x}_i^{n,\ell}\}$ according to Eqns. (10) and (11), i.e. $\mathbf{w}_\ell \in span\{x_i^{s,\ell}, i = 1, \cdots, |\mathbb{O}|, s \in \{p, n\}\}$. Second, according to Eqn. (8), we have

$$\begin{aligned} &\mathbf{w}_j^T \mathbf{x}_i^{s,j+1} = 0, \ s \in \{p, n\}, \ j = 1, \cdots, \ell - 1 \\ &span\{x_i^{s,\ell}, i = 1, \cdots, |\mathbb{O}|, s \in \{p, n\}\} \\ &\quad \subseteq span\{x_i^{s,\ell-1}, i = 1, \cdots, |\mathbb{O}|, s \in \{p, n\}\} \\ &\quad \subseteq \cdots \subseteq span\{x_i^{s,0}, i = 1, \cdots, |\mathbb{O}|, s \in \{p, n\}\} \end{aligned} \tag{13}$$

Hence, $\mathbf{w}_\ell$ is orthogonal to $\mathbf{w}_j, j = 1, \cdots, \ell - 1$. □

Theorem 2. $r(\mathbf{W}^{\ell+1}, \mathbb{O}^{\ell+1}) \leq r(\mathbf{W}^\ell, \mathbb{O}^\ell)$, *where* $\mathbf{W}^\ell = (\mathbf{w}_1, \cdots, \mathbf{w}_\ell)$, $\ell \geq 1$. *That is, the algorithm iteratively decreases the objective function value.*

Proof. Let $\mathbf{w}_{\ell+1}$ be the optimal solution of Eqn. (9). By Theorem 1, it is easy to prove that for any $j \geq 1$, $\mathbf{w}_j^T \mathbf{x}_i^{s,j} = \mathbf{w}_j^T \mathbf{x}_i^{s,0} = \mathbf{w}_j^T \mathbf{x}_i^s$, $s \in \{p, n\}$. Hence we have

$$\begin{aligned} &r_{\ell+1}(\mathbf{w}_{\ell+1}, \mathbb{O}^{\ell+1}) \\ &= \sum_{\mathbb{O}_i^{\ell+1}} \log(1 + a_i^{\ell+1} \exp\{||\mathbf{w}_{\ell+1}^T \mathbf{x}_i^{p,\ell+1}||^2 - ||\mathbf{w}_{\ell+1}^T \mathbf{x}_i^{n,\ell+1}||^2\}) \\ &= r(\mathbf{W}^{\ell+1}, \mathbb{O}) \end{aligned}$$

Also $r_{\ell+1}(\mathbf{0}, \mathbb{O}^{\ell+1}) = r(\mathbf{W}^\ell, \mathbb{O})$. Since $\mathbf{w}_{\ell+1}$ is the minimal solution, we have $r_{\ell+1}(\mathbf{w}_{\ell+1}, \mathbb{O}^{\ell+1}) \leq r_{\ell+1}(\mathbf{0}, \mathbb{O}^{\ell+1})$, and therefore $r(\mathbf{W}^{\ell+1}, \mathbb{O}) \leq r(\mathbf{W}^\ell, \mathbb{O})$. □

Since Criterion (9) may not be convex, a local optimum could be obtained in each iteration of our algorithm. However, even if the computation was trapped in a local minimum of Eqn. (9) at the $\ell + 1$ iteration, Theorem 2 is still valid if $r_{\ell+1}(\mathbf{w}_{\ell+1}, \mathbb{O}^{\ell+1}) \leq r_\ell(\mathbf{w}_\ell, \mathbb{O}^\ell)$, otherwise the algorithm will be terminated by the stopping criterion (12). To alleviate the local optimum problem at each iteration, multiple initialisations could also be deployed in practice.

2.3. Learning in an absolute data difference space

To compute the data difference vector $\mathbf{x}$ defined in Eqn. (1), most existing distance learning methods use the following entry-wise difference function

$$\mathbf{x} = d(\mathbf{z}, \mathbf{z}') = \mathbf{z} - \mathbf{z}' \tag{14}$$

to learn $\mathbf{M} = \mathbf{W}\mathbf{W}^T$ in the normal data difference space denoted by $\mathcal{DZ} = \{\mathbf{x}_{ij} = \mathbf{z}_i - \mathbf{z}_j | \mathbf{z}_i, \mathbf{z}_j \in \mathcal{Z}\}$. The learned distance function is thus written as:

$$f(\mathbf{x}_{ij}) = (\mathbf{z}_i - \mathbf{z}_j)^T \mathbf{M} (\mathbf{z}_i - \mathbf{z}_j) = ||\mathbf{W}^T \mathbf{x}_{ij}||^2. \tag{15}$$

In this work, we compute the difference vector by the following entry-wise absolute difference function:

$$\mathbf{x} = d(\mathbf{z}, \mathbf{z}') = |\mathbf{z} - \mathbf{z}'|, \ \mathbf{x}(k) = |\mathbf{z}(k) - \mathbf{z}'(k)|. \tag{16}$$

[2] It can be explored by Lagrangian equation for Eqn. (9) for a non-zero $\mathbf{w}_\ell$.

where $\mathbf{z}(k)$ is the k-th element of the sample feature vector. $\mathbf{M}$ is thus learned in an absolute data difference space, denoted by $\left|\mathcal{DZ}\right| = \left\{|\mathbf{x}_{ij}| = |\mathbf{z}_i - \mathbf{z}_j| \middle| \mathbf{z}_i, \mathbf{z}_j \in \mathcal{Z}\right\}$, and our distance function becomes:

$$f(|\mathbf{x}_{ij}|) = |\mathbf{z}_i - \mathbf{z}_j|^T \mathbf{M} |\mathbf{z}_i - \mathbf{z}_j| = ||\mathbf{W}^T |\mathbf{x}_{ij}|\ ||^2. \quad (17)$$

We now explain why learning in an absolute data difference space is more suitable to our relative comparison model. First, we note that:

$$\begin{aligned} &|\mathbf{z}_i(k) - \mathbf{z}_j(k)| - |(\mathbf{z}_i(k) - \mathbf{z}_{j'}(k)| \\ &\quad \leq |(\mathbf{z}_i(k) - \mathbf{z}_j(k)) - (\mathbf{z}_i(k) - \mathbf{z}_{j'}(k))|, \end{aligned} \quad (18)$$

hence we have $|\mathbf{x}_{ij}| - |\mathbf{x}_{ij'}|. \leq |\mathbf{x}_{ij} - \mathbf{x}_{ij'}|$, where '. $\leq$' is an entry-wise '$\leq$'. As $|\mathbf{x}_{ij}|, |\mathbf{x}_{ij'}| \geq 0$, we thus can prove

$$\left|\left||\mathbf{x}_{ij}| - |\mathbf{x}_{ij'}|\right|\right| \leq \left|\left|\mathbf{x}_{ij} - \mathbf{x}_{ij'}\right|\right|. \quad (19)$$

This suggests that the variation of $|\mathbf{x}_{ij}|$ given the same sample space $\mathcal{Z}$ is always less than that of $\mathbf{x}_{ij}$. Specifically, if $\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_{j'}$ are from the same class, the intra-class variation is smaller in $|\mathcal{DZ}|$ than in $\mathcal{DZ}$. On the other hand, if $\mathbf{z}_j$ and $\mathbf{z}_{j'}$ belong to a different class as $\mathbf{z}_i$, the variation of inter-class differences is also more compact in the absolute data difference space. Since the variations of both relevant and irrelevant sample differences $\mathbf{x}^p$ and $\mathbf{x}^n$ are smaller, the learned distance function using Eqn. (6) would yield more consistent distance comparison results therefore benefitting our PRDC model. Specially, for the same semidefinite matrix $\mathbf{M}$, the Cauchy inequality suggests

$$upper(\left|\left|\mathbf{W}^T(|\mathbf{x}_{ij}| - |\mathbf{x}_{ij'}|)\right|\right|) \leq upper(\left|\left|\mathbf{W}^T(\mathbf{x}_{ij} - \mathbf{x}_{ij'})\right|\right|),$$

where $upper(\cdot)$ is the upper bound operation. This indicates that in the latent subspace induced by $\mathbf{W}$, the maximum variation of $|\mathbf{x}_{ij}|^T \mathbf{M} |\mathbf{x}_{ij}|$ is lower than that of $\mathbf{x}_{ij}^T \mathbf{M} \mathbf{x}_{ij}$. We show notable benefit of learning PRDC in an absolute data difference space in our experiments.

2.4. Feature representation

Our PRDC model can be applied regardless of the choice of appearance feature representation of people. However, in order to benefit from different and complementary information captured by different features, we start with a mixture of colour and texture histogram features similar to those used in [7] and let our model automatically discover an optimal feature distance. Specifically, we divided a person image into six horizontal stripes. For each stripe, the RGB, YCbCr, HSV color features and two types of texture features extracted by Schmid and Gabor filters were computed and represented as histograms. In total 29 feature channels were constructed for each stripe and each feature channel was represented by a 16 dimensional histogram vector. Each person image was thus represented by a feature vector in a 2784 dimensional feature space $\mathcal{Z}$. Since the features computed for this representation include low-level features widely used by existing person re-identification techniques, this representation is considered as generic and representative.

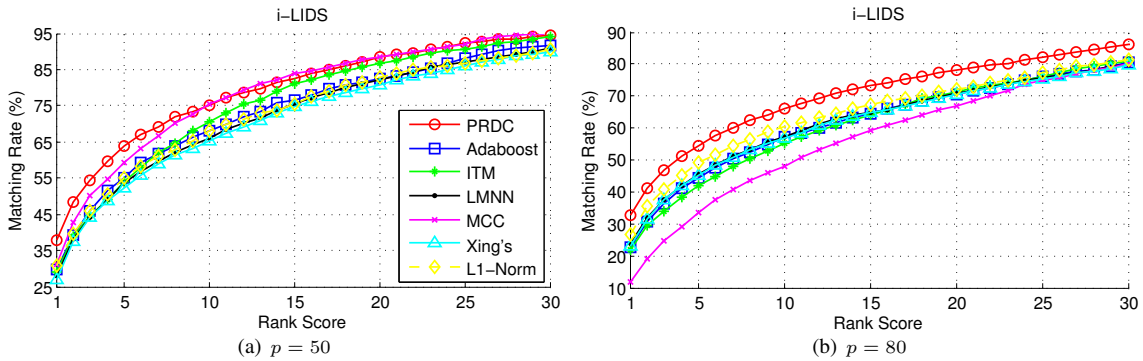
3. Experiments

Datasets and settings. Two publically available person re-identification datasets, i-LIDS Multiple-Camera Tracking Scenario (MCTS) [18, 13] and VIPeR [6], were used for evaluation. In the i-LIDS MCTS dataset, which was captured indoor at a busy airport arrival hall, there are 119 people with a total 476 person images captured by multiple non-overlapping cameras with an average of 4 images for each person. Many of these images undergo large illumination change and are subject to occlusions (see Fig. 4). The VIPeR dataset is the largest person re-identification dataset available consisting of 632 people captured outdoor with two images for each person. Viewpoint change is the most significant cause of appearance change with most of the matched image pairs containing one front/back view and one side-view (see Fig. 5).

In our experiments, for each dataset, we randomly selected all images of p people (classes) to set up the test set, and the rest were used for training. Each test set was composed of a gallery set and a probe set. The gallery set consisted of one image for each person, and the remaining images were used as the probe set. This procedure was repeated 10 times. During training, a pair of images of each person formed a relevant pair, and one image of him/her and one of another person in the training set formed a related irrelevant pair, and together they form the pairwise set $\mathbb{O}$ defined in Sec. 2.

For evaluation, we use the average cumulative match characteristic (CMC) curves [6] over 10 trials to show the ranked matching rates. A rank r matching rate indicates the percentage of the probe images with correct matches found in the top r ranks against the p gallery images. Rank 1 matching rate is thus the correct matching/recognition rate. Note that in practice, although a high rank 1 matching rate is critical, the top r ranked matching rate with a small r value is also important because the top matched images will normally be verified by a human operator [6].

PRDC vs. Non-Learning based Distances. We first compared our PRDC with non-learning based l_1-norm distance and Bhattacharyya distance which were used by most existing person re-identification work. Our results (Figs. 2 and 3, Tables 1 and 2) show clearly that with the proposed PRDC, the matching performance for both datasets is improved notably, more so when the number of people in the test pool increases (i.e. training set size decreases). The improvement is particularly dramatic on the VIPeR dataset. In particular, Table 2 shows that a 4-fold increase in correct matching rate ($r = 1$) is obtained against both l_1-norm and Bhattacharyya distances when $p = 316$. The results validate the importance of performing distance learning. Examples of matching people using PRDC for both datasets are shown in Figs. 4 and 5 respectively.



(a) $p = 50$ (b) $p = 80$

Figure 2. Performance comparison using CMC curves on i-LIDS MCTS dataset.

Methods	$p = 30$				$p = 50$				$p = 80$			
	$r = 1$	$r = 5$	$r = 10$	$r = 20$	$r = 1$	$r = 5$	$r = 10$	$r = 20$	$r = 1$	$r = 5$	$r = 10$	$r = 20$
PRDC	**44.05**	72.74	84.69	96.29	**37.83**	**63.70**	75.09	**88.35**	**32.60**	**54.55**	**65.89**	**78.30**
Adaboost	35.58	66.43	79.88	93.22	29.62	55.15	68.14	82.35	22.79	44.41	57.16	70.55
LMNN	33.68	63.88	78.17	92.64	27.97	53.75	66.14	82.33	23.70	45.42	57.32	70.92
ITM	36.37	67.99	83.11	95.55	28.96	53.99	70.50	86.67	21.67	41.80	55.12	71.31
MCC	40.24	**73.64**	**85.87**	**96.65**	31.28	59.30	**75.62**	88.34	12.00	33.66	47.96	67.00
Xing's	31.80	62.62	77.29	90.63	27.04	52.28	65.35	80.70	23.18	45.24	56.90	70.46
L1-norm	35.31	64.62	77.37	91.35	30.72	54.95	67.99	82.98	26.73	49.04	60.32	72.07
Bhat.	31.77	61.43	74.19	89.53	28.42	51.06	64.32	78.77	24.76	45.35	56.12	69.31

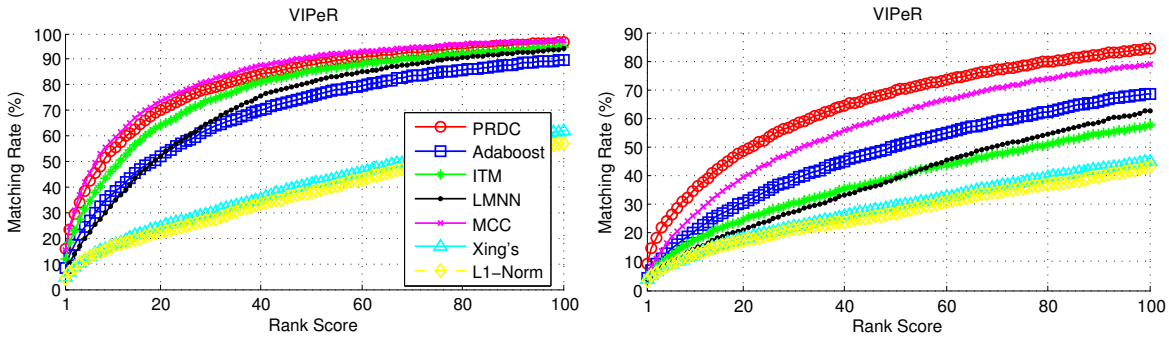
Table 1. Top ranked matching rate (%) on i-LIDS MCTS. p is size of the gallery set (larger p means smaller training set) and r is the rank.

PRDC vs. Alternative Learning Methods. We also compared PRDC with 5 alternative discriminant learning based approaches. These include 4 popular distance learning methods, namely Xing's method [16], LMNN [15], ITM [1] and MCC [5], and a method specifically designed for person re-identification based on Adaboost [7]. Among the 4 distance learning methods, only LMNN exploits relative distance comparison. But as mentioned in Sec. 1, it is used as an optimisation constraint rather than the main objective function which is also not formulated probabilistically. MCC is similar to PRDC in that a probabilistic model is used but it is not a relative distance comparison based method. Note that since MCC needs to select the best dimension for matching, we performed cross-validation by selecting its value in $\{[1 : 1 : 10], d\}$, where d is the maximum rank MCC can learn. Among the 5, the only method that learns in an absolute data different space is Adaboost.

Our results (Figs. 2 and 3, Tables 1 and 2) show clearly that our model yields the best rank 1 matching rate and overall much superior performance compared to the compared models. The advantage of PRDC is particularly apparent when a training set is small (learning becomes more difficult) and a test set is large indicated by the value of p (matching becomes harder). Table 2 shows that on VIPeR when 100 people are used for learning and 532 people for testing ($p = 532$), the correct matching rate for PRDC (and MCC) is almost more than doubled against any alternative distance learning methods. Particularly, benefiting from being a probabilistic model, MCC gives the most comparable results to PRDC when the training set is large. However, its performance degrades dramatically when the size of training data decreases (see columns under $p = 80$ in Table 1 and $p = 532$ in Table 2). This suggests that over-fitting to limited training data is the main reason for the inferior performance of the compared alternative learning approaches.

PRDC vs. RankSVM. Different from PRDC, RankSVM has a free parameter which determines the relative weights between the margin function and the ranking error function [11]. In our experiment, we cross-validated the parameter in $\{0.0001, 0.005, 0.001, 0.05, 0.1, 0.5, 1, 10, 100, 1000\}$. As shown in Tables 3 and 4, the two methods all perform very well against other compared algorithms and our PRDC yields overall better performance especially at lower rank matching rate and given less training data. The better performance of PRDC is due to the probabilistic modelling and a second-order rather than first-order feature selection. It is also noted that tuning the free parameter for RankSVM is not a trivial task and the performance can be sensitive to the tuning especially given sparse data, while PRDC does not have this problem. In addition RankSVM is computationally more expensive (see details later).

Effect of learning in an Absolute Data Difference Space. We have shown in Sec. 2.3 that in theory our relative distance comparison learning method can benefit from learning in an absolute data difference space. To validate this experimentally, we compare PRDC with PRDC$_{raw}$ which learns in the normal data difference space $\mathcal{DZ}$ (see Sec. 2.3). The result in Table 5 indicates that learning in an absolute data difference space does improve the matching performance. Note that most existing distance learning models are based on learning in the normal data difference space $\mathcal{DZ}$. It is possible to reformulate some of them in



(a) $p = 316$ (b) $p = 532$

Figure 3. Performance comparison using CMC curves on VIPeR dataset.

Methods	$p = 316$				$p = 432$				$p = 532$			
	$r = 1$	$r = 5$	$r = 10$	$r = 20$	$r = 1$	$r = 5$	$r = 10$	$r = 20$	$r = 1$	$r = 5$	$r = 10$	$r = 20$
PRDC	**15.66**	38.42	53.86	70.09	**12.64**	31.97	44.28	59.95	**9.12**	**24.19**	**34.40**	**48.55**
Adaboost	8.16	24.15	36.58	52.12	6.83	19.81	29.75	43.06	4.19	12.95	20.21	30.73
LMNN	6.23	19.65	32.63	52.25	5.14	13.13	20.30	33.91	4.04	9.68	14.19	21.18
ITM	11.61	31.39	45.76	63.86	8.38	24.54	36.81	52.29	4.19	11.11	17.22	24.59
MCC	15.19	**41.77**	**57.59**	**73.39**	11.30	**32.43**	**47.29**	**62.85**	5.00	16.32	25.92	39.64
Xing's	4.65	11.96	16.61	24.37	4.12	10.02	14.70	20.65	3.63	8.76	12.14	18.16
L1-norm	4.18	11.65	16.52	22.37	3.80	9.81	13.94	19.44	3.55	8.29	12.27	17.59
Bhat.	4.65	11.49	16.55	23.83	4.19	10.35	14.19	20.19	3.82	9.08	12.42	17.88

Table 2. Top ranked matching rate (%) on VIPeR. p is the number of classes in the testing set; r is the rank.

Rank	PRDC				RankSVM			
	$r = 1$	$r = 5$	$r = 10$	$r = 20$	$r = 1$	$r = 5$	$r = 10$	$r = 20$
$p = 30$	**44.05**	**72.74**	84.69	96.29	42.96	71.30	**85.15**	**96.99**
$p = 50$	**37.83**	**63.70**	**75.09**	**88.35**	37.41	63.02	73.50	88.30
$p = 80$	**32.60**	54.55	65.89	**78.30**	31.73	**55.69**	**67.02**	77.78

Table 3. PRDC vs. RankSVM (%) on i-LIDS.

Rank	PRDC				RankSVM			
	$r = 1$	$r = 5$	$r = 10$	$r = 20$	$r = 1$	$r = 5$	$r = 10$	$r = 20$
$p = 316$	15.66	**38.42**	**53.86**	**70.09**	**16.27**	38.23	53.73	69.87
$p = 432$	**12.64**	**31.97**	**44.28**	**59.95**	10.63	29.70	42.31	58.26
$p = 532$	**9.12**	**24.19**	**34.40**	**48.55**	8.87	22.88	32.69	45.98

Table 4. PRDC vs. RankSVM (%) on VIPeR.

Methods	i-LIDS, ($p = 50$)				VIPeR ($p = 316$)			
	$r = 1$	$r = 5$	$r = 10$	$r = 20$	$r = 1$	$r = 5$	$r = 10$	$r = 20$
PRDC	**37.83**	**63.70**	**75.09**	**88.35**	**15.66**	**38.42**	**53.86**	70.09
PRDC$_{raw}$	19.92	50.19	68.29	86.40	12.28	37.28	53.83	**71.77**
ITM$_{abs}$	29.16	53.01	66.75	82.53	5.44	14.43	22.53	33.35
MCC$_{abs}$	5.59	23.01	43.59	70.47	1.20	3.51	5.6	9.68

Table 5. Effect of learning in an absolute data difference space.

Methods	i-LIDS MCTS			VIPeR		
	$p = 30$	$p = 50$	$p = 80$	$p = 316$	$p = 432$	$p = 532$
$rank(\mathbf{W})$	3.2	2.4	2.3	2.9	3.2	3.7

Table 6. Average Rank of $\mathbf{W}$ Learned by PRDC.

order to learn in an absolute data difference space. In Table 5 we show that when ITM and MCC are learned in the absolute data difference space $|\mathcal{DZ}|$, termed as ITM$_{abs}$ and MCC$_{abs}$ respectively, their performances become worse as compared to their results in Tables 1 and 2. This indicates that the absolute different space is more suitable for our relative comparison distance learning.

Computational cost. Though PRDC is iterative, it has relatively low cost in practice. In our experiments, for VIPeR with $p = 316$, it took around 15 minutes for an Intel dual-core 2.93GHz CPU and 48GB RAM to learn PRDC for each trial. We observed that the low cost of PRDC is partially due to its ability to seek a suitable low rank of $\mathbf{W}$ (i.e. converge within very few iterations) as shown in Table 6. For comparison, among the other compared methods, Adaboost is the most costly and took over 7 hours for each trial. For the 4 compared distance learning methods, PCA dimensionality reduction must be performed otherwise they becomes intractable given the high dimensional feature space. For the RankSVM method, each trial took around 2.5 hours due to parameter tuning.

4. Conclusion

We have proposed a new approach for person re-identification based on probabilistic relative distance comparison which aims to learn an suitable optimal distance measure given large intra and inter-class appearance variations and sparse data. Our experiments demonstrate that 1) by formulating person re-identification as a distance learning problem, clear improvement in matching performance can be obtained and the improvement is more significant when training sample size is small, and (2) our PRDC outperforms not only existing distance learning methods but also alternative learning methods based on boosting and learning to rank.

Acknowledgements

This research was partially funded by the EU FP7 project SAMURAI with grant no. 217899. Dr. Wei-Shi Zheng was also additionally supported by the 985 project in Sun Yat-sen University with grant no. 35000-3181305.

References

[1] J. Davis, B. Kulis, P. Jain, S. Sra, and I. Dhillon. Information-theoretic metric learning. In *ICML*, 2007.

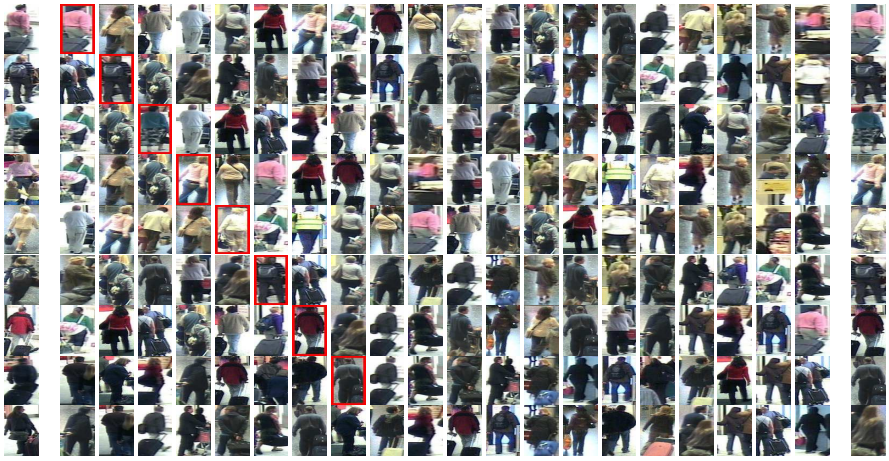


Figure 4. Examples of Person Re-identification on i-LIDS MCTS using PRDC. In each row, the left-most image is the probe, images in the middle are the top 20 matched gallery images with a highlighted red box for the correctly matched, and the right-most shows a true match

Figure 5. Examples of Person Re-identification on VIPeR using PRDC

[2] P. Dollar, Z. Tu, H. Tao, and S. Belongie. Feature mining for image classification. In *CVPR*, 2007.

[3] M. Farenzena, L. Bazzani, A. Perina, M. Cristani, and V. Murino. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010.

[4] N. Gheissari, T. Sebastian, and R. Hartley. Person reidentification using spatiotemporal appearance. In *CVPR*, 2006.

[5] A. Globerson and S. Roweis. Metric learning by collapsing classes. In *NIPS*, 2005.

[6] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *IEEE International workshop on performance evaluation of tracking and surveillance*, 2007.

[7] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008.

[8] W. Hu, M. Hu, X. Zhou, J. Lou, T. Tan, and S. Maybank. Principal axis-based correspondence between multiple cameras for people tracking. *PAMI*, 28(4):663–671, 2006.

[9] J. Lee, R. Jin, and A. Jain. Rank-based distance metric learning: An application to image retrieval. In *CVPR*, 2008.

[10] U. Park, A. Jain, I. Kitahara, K. Kogure, and N. Hagita. Vise: Visual search engine using multiple networked cameras. In *ICPR*, 2006.

[11] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by support vector ranking. In *BMVC*, 2010.

[12] M. Schultz and T. Joachims. Learning a distance metric from relative comparisons. In *NIPS*, 2004.

[13] UK. Home Office i-LIDS multiple camera tracking scenario definition. 2008.

[14] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu. Shape and appearance context modeling. In *ICCV*, 2007.

[15] K. Weinberger, J. Blitzer, and L. Saul. Distance metric learning for large margin nearest neighbor classification. In *NIPS*, 2006.

[16] E. Xing, A. Ng, M. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. In *NIPS*, 2003.

[17] L. Yang, R. Jin, R. Sukthankar, and Y. Liu. An efficient algorithm for local distance metric learning. In *AAAI*, 2006.

[18] W.-S. Zheng, S. Gong, and T. Xiang. Associating groups of people. In *BMVC*, 2009.