No commercial use ^_^

Reidentification by Relative Distance Comparison

Wei-Shi Zheng, *Member*, *IEEE*, Shaogang Gong, and Tao Xiang

Abstract—Matching people across nonoverlapping camera views at different locations and different times, known as person reidentification, is both a hard and important problem for associating behavior of people observed in a large distributed space over a prolonged period of time. Person reidentification is fundamentally challenging because of the large visual appearance changes caused by variations in view angle, lighting, background clutter, and occlusion. To address these challenges, most previous approaches aim to model and extract distinctive and reliable visual features. However, seeking an optimal and robust similarity measure that quantifies a wide range of features against realistic viewing conditions from a distance is still an open and unsolved problem for person reidentification. In this paper, we formulate person reidentification as a relative distance comparison (RDC) learning problem in order to learn the optimal similarity measure between a pair of person images. This approach avoids treating all features indiscriminately and does not assume the existence of some universally distinctive and reliable features. To that end, a novel relative distance comparison model is introduced. The model is formulated to maximize the likelihood of a pair of true matches having a relatively smaller distance than that of a wrong match pair in a soft discriminant manner. Moreover, in order to maintain the tractability of the model in large scale learning, we further develop an ensemble RDC model. Extensive experiments on three publicly available benchmarking datasets are carried out to demonstrate the clear superiority of the proposed RDC models over related popular person reidentification techniques. The results also show that the new RDC models are more robust against visual appearance changes and less susceptible to model overfitting compared to other related existing models.

Index Terms—Person reidentification, feature quantification, feature selection, relative distance comparison

1 Introduction

For understanding behavior of people in a large area of public space covered by multiple nonoverlapping (disjoint) cameras, it is critical that when a target disappears from one view, he/she can be reidentified in another view at a different location among a crowd of people. Solving this intercamera people association problem, known as *reidentification*, enables tracking of the same person through different camera views located at different physical sites [26], [15], [32], [17], [8].

Despite the best efforts from computer vision researchers in the past five years, the person reidentification problem remains largely unsolved. This is due to a number of reasons. First, in a busy uncontrolled environment monitored by cameras from a distance, person verification relying upon biometrics such as face and gait is infeasible and unreliable. Second, as the transition time between disjoint cameras[1] varies greatly from individual to individual with uncertainty, it is hard to impose accurate temporal and spatial constraints. Therefore, the person reidentification problem is made harder still as a model can only rely on mostly appearance features alone. Third, the visual appearance features, extracted mainly from the clothing and shapes of people, are intrinsically indistinctive for matching people (e.g., most people in winter wear dark clothes). In addition, a person's appearance often undergoes large variations across nonoverlapping 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. 6 and 7).

Given a query image of a person, in order to find the correct match among a large number of candidate images captured from different camera views, two steps need to be taken. First, a feature representation is computed from both the query and each of the gallery images. Second, the distance between each pair of potential matches is measured, which is then used to determine whether a gallery image contains the same person as the query image. Most existing studies have focused on the first step, that is, seeking a more distinctive and reliable feature representation of people's appearance, ranging widely from color histogram [26], [15], graph model [10], spatial co-occurrence representation model [32], principal axis [17], rectangle region histogram [6], part-based models [1], [4], to combinations of multiple features [15], [8]. After feature extraction, these methods simply choose a standard distance measure such as l_1-norm [32], l_2-norm-based distance [17], or Bhattacharyya distance [15]. However, under severe changes in viewing conditions that can cause significant appearance variations (e.g., view

1. The time gap between a person disappearing in one camera view and reappearing in another.

- *W.-S. Zheng is with the School of Information Science and Technology, Sun Yat-sen University, Guangzhou, Guangdong 510006, P.R. China. E-mail: wszheng@ieee.org.*
- *S. Gong and T. Xiang are with the School of Electronic Engineering and Computer Science, Queen Mary University London, Mile End Road, London E1 4NS, United Kingdom. E-mail: {sgg, txiang}@eecs.qmul.ac.uk.*

Manuscript received 7 Sept. 2011; revised 9 Feb. 2012; accepted 30 May 2012; published online 20 June 2012.
Recommended for acceptance by B. Schiele.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2011-09-0628.
Digital Object Identifier no. 10.1109/TPAMI.2012.138.

0162-8828/13/$31.00 © 2013 IEEE Published by the IEEE Computer Society

No commercial use ^_^

Fig. 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.

angle and lighting condition changes, occlusion), computing a set of features that are both distinctive and reliable is extremely hard if not implausible. 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 focus on the second step of person reidentification. That is, given a set of features extracted from each person image, we seek to quantify and differentiate these features by learning the optimal distance measure that is most likely to give correct matches. This is significantly different from most existing approaches in that it requires model learning from a set of training data. In essence, images of each person in a training set form a class. This learning problem can be framed as a distance learning problem which always searches for a distance that minimizes intraclass distances while maximizing interclass distances. However, the person reidentification problem has four characteristics.

1. The intraclass variation can be large and, more importantly, can vary significantly for different classes as it is caused by large and unpredictable viewing condition changes (see Fig. 1).
2. The interclass variation also varies drastically across different pairs of classes and there are often severe overlaps between classes in a feature space due to similar appearance (e.g., clothing) of different people.
3. The training set for learning the model consists of images of matched people across different camera views. In order to capture the large intra and intervariations, the number of classes is necessarily large, typically on the order of hundreds. This represents a large scale learning problem that challenges existing machine learning algorithms.
4. Annotating a large number of matched people across camera views is not only tedious, but also inherently limited in its usefulness.

Typically, each annotated class contains only a handful of images of a person from different camera views, i.e., the data are inherently undersampled for building a representative class distribution. Due to these intrinsic characteristics of the reidentification problem, especially the problem of a large number of undersampled classes, a learning model could easily be overfitted and/or be intractable if it is learned by minimizing intraclass distance and maximizing interclass distance simultaneously by brute-force, as is typically done by existing popular distance learning techniques.

To alleviate this inherently ill-posed distance learning problem in person reidentification, we formulate the problem as a relative distance comparison (RDC) problem. That is, we perform feature quantification by learning a relative distance comparison model. More specifically, a novel relative distance comparison model is formulated in order to differentiate the similarity score of a pair of true match (i.e., two images of person A) from that of a pair of related wrong match (i.e., two images of different people A and B, respectively) so that the latter one can always be smaller. In other words, the model aims to learn an optimal distance in the sense that for a given query image, the true match is desired to be ranked higher than the wrong matches among the gallery image set. The model cares less about how large the absolute distance between the pair of images for the true match. This differs conceptually from a conventional distance learning approach which aims to minimize intraclass variation in an absolute sense (i.e., making all images of person A more similar or closer in a features space) while maximizing interclass variation (i.e., making two images of person A and B more dissimilar). A conventional approach thus attempts to maximize the margin between two classes or, in the context of person reidentification, enforces a harder discriminant constraint that the true match is not only ranked higher but also has as small a distance to the query image as possible compared to that of wrong matches. One of the key advantages of our relative distance comparison-based method is that our model is not easily biased by large variations across many undersampled classes as it aims to seek an optimized individual comparison between any two data points rather than comparison among data distribution boundaries or among clusters of data. This alleviates the overfitting problem in person identification given undersampled training data.

Computationally, learning the proposed relative distance comparison model can be a nonconvex optimization problem. It is also a large scale learning problem even given a moderate training data size. This is because the distance between each pair of images in a training set needs be compared exhaustively during model learning and the feature space for person reidentification is typically of high dimension. To address this problem, a novel iterative optimization algorithm is developed in this work for learning the RDC model. The algorithm is theoretically validated and its convergence is guaranteed.

Furthermore, in order to alleviate the large space complexity (memory usage cost) and the local optimum learning problem due to the proposed iterative algorithm for solving high-order nonlinear optimization criterion, we develop an ensemble RDC in this work. The aim is to learn a set of weak RDC models, each computed on a small subset of data, and then combine them into a stronger RDC using ensemble learning.

Extensive experiments are conducted on three publicly available large person reidentification datasets, including the ETHZ [7], i-LIDS [37], and VIPeR [14] datasets. The results demonstrate that 1) by formulating the person reidentification problem as a relative distance comparison

No commercial use ^_^

TABLE 1
Main Development of Person Reidentification

Authors	Year	Image Features	Using Temporal Information	Representation
Javed el al. [19]	2005	colour	Yes	colour appearance with colour brightness transform
Gilbert el al. [11]	2006	colour	Yes	consensus-colour conversion of munsell colour space with colour transformation matrix
Gheissari, et al. [10]	2006	colour and shape	Yes	graph partition based representation
Hu et al. [17]	2007	geometry	Yes	principal axis with segmentation
Wang et al. [32]	2007	colour, gradient, and shape	No	co-occurrence spatial context
Chen et al. [3] & Prosser et al. [27]	2008	colour	Yes	colour appearance with temporal colour brightness transform and spatial information
Javed et al. [18]	2008	colour	Yes	colour appearance with spatial temporal colour brightness transform and spatial information
Gray and Tao [15]	2008	colour, gradient, filters	No	selected histogram features by Adaboost
Zheng et al. [37]	2009	colour and gradient	No	grouping as dynamic spatial context
Bak et al. [1] & Cheng et al. [4]	2010/2011	colour	No	covariance matrix between parts or pictorial structures modelling
Prosser et al. [28]	2010	colour, gradient, filters	No	quantified histogram feature by RankSVM
Farenzena et al. [8]	2010	colour and structure	No	symmetry-based ensemble of local features with background subtraction

learning problem based on logistic function modeling, significant improvement on matching accuracy can be obtained against related popular person reidentification techniques; and 2) our RDC models outperform not only related distance learning methods but also related learning methods based on boosting and rank support vector machines (SVMs), both in terms of matching accuracy and tractability.

2 Related Works

The problem of matching people across disjoint camera views has received increasing attention in recent years. Existing works predominantly focus on the problem of feature extraction and representation with a bag-of-word representation of color and texture features being the most common choice. Table 1 summarizes the features and representations employed by existing methods reported in the literature. In addition to matching based on similarity of visual appearance, contextual cues can also be exploited. Brightness transfer function is introduced to explicitly compensate for the lighting condition changes between cameras [3], [27], [18]. However, to learn a brightness transfer function one has to not only annotate a set of matched people but also segment each person from the image, which significantly increases the already large annotation cost. The temporal relationships between camera views can be exploited for object tagging. By modeling the transition time between two camera views one can reduce the number of potential matches while also using the probability distribution of transition time as a feature [12], [25], [24], [22]. However, transition time information could be unreliable when camera views are significantly disjoint or feature a large number of moving objects. Nevertheless, when it can be obtained reliably, it has been exploited to good effect (see Table 1, column 4). Such contextual constraints can also be easily employed to the proposed RDC models either as part of the representation or a postprocessing step.

Since not all features are equally reliable and informative for person reidentification, Gray and Tao [15] propose a boosting approach based on Adaboost to select a subset of optimal features for matching people. However, in a boosting framework, good features are only selected individually 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 consider instead quantifying all features jointly (global selection). Critically, the Adaboost-based feature selection method in [15] could be biased by large variations between the appearance of people as its modeling shares similar spirit with a typical discriminant model that tries to maximize the difference between two images of different people. It is thus prone to model overfitting, as shown in our experiments (see Section 6). In contrast, the proposed RDC model can be seen as a soft discriminant approach. Our model thus is less susceptible to overfitting and more tolerant to intra and interclass variations and severe overlapping of different classes in a multidimensional feature space.

Relative distance comparison is a special case of learning to rank or machine-learned ranking. Ranking techniques such as RankSVM [16] and RankBoost [9] have been widely used in text document analysis and information retrieval. In our early work [28], the primal RankSVM [2] is applied to solve the problem of global feature quantification for person reidentification. The primal RankSVM solves the high computational cost problem for large scale constraint optimization in a standard RankSVM formulation. Compared to RankSVM and RankBoost, the proposed new model in this paper is more principled and tractable in three aspects: 1) RDC is a second-order feature quantification model, taking into account the joined effect between different features, whereas both RankSVM [2] and RankBoost [9] are a first-order model unable to exploit correlations among different features. 2) RDC utilizes a logistic function to provide a soft margin measure between the difference vectors of different types while RankSVM does not, and such a formulation of our objective function makes RDC more tolerant to large intra and interclass variations and better suited for coping with data undersampling. 3) Using a primal RankSVM, one must determine the weight between the margin function and the ranking error cost function, which is computationally costly. In contrast, our RDC model does not suffer from such a problem, leading to lower computational cost. A more detailed discussion on the

No commercial use ^_^

differences between RDC and related ranking models is given in Section 5. Extensive experiments are presented in Section 6.6 to validate the advantages of RDC over RankSVM and RankBoost.

Although it has not previously been exploited for person reidentification, distance learning in general is a well-studied problem [35], [13], [36], [34], [15], [29], [33], [20], [5]. The proposed RDC model is related to several existing distance learning methods. In particular, our model shares the same spirit with a number of recent works that exploit the idea of relative distance comparison [29], [33], [20]. However, the relative distance comparison formulations in these works are not quantified using logistic function for soft measure, and crucially they are used as an optimization constraint rather than an objective function. Therefore, as analyzed in more detail in Section 5, these approaches, either implicitly [29], [20] or explicitly [33], still aim to learn a distance by which each class becomes more compact while being more separable from each other in an absolute sense. We demonstrate through extensive experiments that, in practice, they remain susceptible to model overfitting and poor tractability for person reidentification.

In summary, the main contributions of this work are three-fold.

1. For the first time, the person reidentification problem is formulated as a relative distance comparison learning problem, with strong rationale both conceptually and computationally.
2. We propose a novel logistic function-based relative distance comparison model for feature quantification which overcomes the limitations of existing distance learning techniques given undersampled data with large intra and interclass variations.
3. A novel iterative optimization algorithm and an ensemble RDC model are proposed to improve the tractability of the RDC model and make it more suitable for large scale learning.

An early version of this work appeared in [38]. In addition to giving a more detailed description of the RDC model, the main changes include 1) an ensemble RDC model proposed to improve the scalability and tractability of the original RDC model, 2) more in-depth discussion and analysis on its relationship to alternative learning methods, and 3) more extensive experimental evaluations including the introduction of a new dataset.

3 Quantifying Features for Person Reidentification

3.1 Proposed Relative Distance Comparison Learning

We formally cast the person reidentification problem into the following distance comparison problem, where we assume each instance of a person is represented by a feature set (e.g., the representation described in Section 6.2). For an instance $\mathbf{z}$ of person A, we wish to learn a reidentification model to successfully identify another instance $\mathbf{z}'$ 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 instance of any other person except A. To this end, given a training set $\mathcal{Z} = \{(\mathbf{z}_i, y_i)\}_{i=1}^{N}$, where $\mathbf{z}_i \in \mathcal{R}^q$ is a multidimensional 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 mismatch from another class (e.g., $\mathbf{x}_i^p$ and $\mathbf{x}_i^n$ share the same $\mathbf{z}$ in the following (1), while they have different $\mathbf{z}'$). The difference vector $\mathbf{x}$ between any two samples $\mathbf{z}$ and $\mathbf{z}'$ is computed by

$$\mathbf{x} = d(\mathbf{z}, \mathbf{z}'), \quad \mathbf{z}, \mathbf{z}' \in \mathcal{R}^q, \tag{1}$$

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 Section 3.4.

Given the pairwise set $\mathbb{O}$, a distance function f will take the difference vector as input and can be learned based on relative distance comparison so that a distance between a relevant sample pair ($f(\mathbf{x}_i^p)$) is wished to be smaller than that between a related irrelevant pair ($f(\mathbf{x}_i^n)$). In order to differentiate these two types of difference vectors, we propose a logistic function based modeling to describe how a distance between a relevant pair differs from the one between a related but irrelevant pair as follows:

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

We assume the events of distance comparison between a relevant pair and a related irrelevant pair are independent[2]. Then, we wish to minimize the risk of learning f via all the above relative distance comparisons as follows:

$$\min_f r(f, \mathbb{O}),\ r(f, \mathbb{O}) = -\log\left(\prod_{\mathbb{O}_i} C_f(\mathbf{x}_i^p, \mathbf{x}_i^n)\right). \tag{3}$$

The distance function f is parameterized as a Mahalanobis (quadratic) distance function:

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

where $\mathbf{M}$ is a semidefinite matrix. The distance learning problem thus becomes learning $\mathbf{M}$ using (3). Directly learning $\mathbf{M}$ using semidefinite program techniques is computationally expensive for high-dimensional data [33]. 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 $\mathbf{M}$:

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

where the columns of $\mathbf{A}$ are orthonormal eigenvectors of $\mathbf{M}$ and the leading diagonal of $\mathbf{\Lambda}$ contains the corresponding nonzero eigenvalues. Note that the columns of $\mathbf{W}$ form a set

2. Note that we do not assume the data are independent.

No commercial use ^_^

of orthogonal vectors. Therefore, learning a function f is equivalent to learning such a matrix $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_l, \ldots, \mathbf{w}_L)$ such that

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

We call this relative distance comparison learning for person reidentification. RDC is based on a logistic function ranging from 0 to 1 in value. This is designed to avoid dramatic changes in the response to different relative distance comparisons.

3.2 An Iterative Optimization Algorithm

It is important to point out that our optimization criterion (6) may not be a convex optimization problem against the orthogonal constraint due to the logistic function-based relative comparison modeling. It means that deriving an global solution by directly optimizing $\mathbf{W}$ is not straightforward. In this work, we formulate an iterative optimization algorithm to learn an optimal $\mathbf{W}$, which also aims to seek a low-rank and nontrivial solution automatically. This is critical for reducing the model complexity, thus alleviating the overfitting problem given a large number of undersampled classes.

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 $\mathbf{w}_1, \ldots, \mathbf{w}_\ell$ have been learned. To learn the next orthogonal vector $\mathbf{w}_{\ell+1}$, let

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

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 ℓth iteration defined as follows:

$$\mathbf{x}_i^{s,\ell} = \mathbf{x}_i^{s,\ell-1} - \tilde{\mathbf{w}}_{\ell-1} \tilde{\mathbf{w}}_{\ell-1}^T \mathbf{x}_i^{s,\ell-1},\quad s \in \{p, n\},\ i = 1, \ldots, \left|\mathbb{O}\right|, \tag{8}$$

where $\ell \geq 1$ and $\tilde{\mathbf{w}}_{\ell-1} = \mathbf{w}_{\ell-1}/\|\mathbf{w}_{\ell-1}\|$. 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 (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:

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

where

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

We seek a solution by a gradient descent method

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

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

where λ is a step length automatically determined at each gradient update step using similar strategy in [23]. According to the descent direction in (10), the initial value of $\mathbf{w}_{\ell+1}$ for the gradient descent method is set to

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

Note that the update in (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 summarized in Algorithm 1.

Algorithm 1: Learning the RDC 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 Eq. (7);
    Compute $\mathbf{x}_i^{s,\ell+1},\ s \in \{p, n\}$ by Eq. (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 Eq. (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: $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_\ell]$

3.3 Theoretical Validation

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

Theorem 1. *The learned vectors $\mathbf{w}_\ell$, $\ell = 1, \ldots, L$, 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[3] of $\{\mathbf{x}_i^{p,\ell}\} \cup \{\mathbf{x}_i^{n,\ell}\}$ according to (10) and (11), i.e., $\mathbf{w}_\ell \in span\{x_i^{s,\ell}, i = 1, \ldots, |\mathbb{O}|, s \in \{p, n\}\}$. Second, according to (8), we have

3. This can also be explored by using Lagrangian equation for (9) for a nonzero $\mathbf{w}_\ell$.

No commercial use ^_^

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

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

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

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

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

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 $r_{\ell+1}(\mathbf{w}_{\ell+1},\mathbb{O}^{\ell+1})\le r_{\ell+1}(\mathbf{0},\mathbb{O}^{\ell+1})$, and therefore $r(\mathbf{W}^{\ell+1},\mathbb{O})\le r(\mathbf{W}^{\ell},\mathbb{O})$. □

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 (9) at the $\ell+1$ iteration, Theorem 2 is still valid if $r_{\ell+1}(\mathbf{w}_{\ell+1},\mathbb{O}^{\ell+1})\le r_\ell(\mathbf{w}_\ell,\mathbb{O}^{\ell})$; otherwise the algorithm will be terminated by the stopping criterion (12). To alleviate the local optimum problem at each iteration, multiple initializations could be deployed in practice. In this work, we formulate an ensemble algorithm in Section 4 to alleviate the problem of local optimum.

3.4 Learning in an Absolute Data Difference Space

To compute the data difference vector $\mathbf{x}$ defined in (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

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

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

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

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

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

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

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

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

$$\left|\left||\mathbf{x}_{ij}|-|\mathbf{x}_{ij'}|\right|\right|\le\left|\left|\mathbf{x}_{ij}-\mathbf{x}_{ij'}\right|\right|.\tag{19}$$

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 intraclass 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 than $\mathbf{z}_i$, the variation of interclass 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 (6) would yield more consistent distance comparison results, therefore benefitting our RDC model. Specially, for the same semidefinite matrix $\mathbf{M}$, by combining (19) and the Cauchy inequality, we have

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

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 RDC in an absolute data difference space in our experiments.

4 Ensemble Learning for Large Scale Computation

The proposed RDC is based on the comparison between each relevant and related irrelevant pairs and optimized by an iterative algorithm. However, there are the two following remaining issues could still hinder the tractability of the proposed model.

1. First, the number of comparisons can thus be very high given even a moderate training data size. Specifically, the amount of these pairwise comparison could lead to a considerably large space complexity (memory usage cost). For instance, let us assume there are N images in total in a training set belonging to L people. Assuming there are $\frac{N}{L}$ images for each person, we can learn an RDC with a space complexity of $O(q\cdot((\frac{1}{L}-\frac{1}{L^2})\cdot N^3+(\frac{1}{L}-1)\cdot N^2))$, where q is the dimension of the feature space. This high space complexity is thus caused by both the N^3 term and the typically high feature dimension q.
2. Second, although the proposed iterative optimization algorithm can effectively handle the high order nonconvex optimization problem, it could still be trapped into a local optimum.

To alleviate these two problems, rather than learning a batch mode RDC, we propose learning a set of weak RDC models, each computed using a small subset of the data, and then combining them to build a stronger RDC using ensemble learning. More specifically, by using the idea of ensemble learning, a strong RDC model $f_s(\mathbf{x})$ is constructed by a set of H weak RDC models $f_{w,i}(\mathbf{x})$ as follows:

No commercial use ^_^

$$f_s(\mathbf{x}) = \sum_{i=1}^{H} \beta_i \cdot f_{w,i}(\mathbf{x}), \tag{20}$$

where $f_{w,i}(\mathbf{x})$ are defined as in (4) and β_i is the weight of each weak RDC model.

Learning weak RDC models $f_{w,i}$—Each weak RDC model is learned using a different subset of the training samples. More specifically, to learn H weak models, the training dataset is divided into H groups. Assuming there are in total L people/classes $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_L\}$, we first equally divide them into H groups $G_1, \ldots, G_H$ without overlap, i.e., $\mathcal{C} = \bigcup_{i=1}^{H} G_i$ and $\forall i \neq j$, $G_i \bigcap G_j = \emptyset$. Subsequently, the training dataset $\mathcal{Z}$ is divided into H subsets $\mathcal{Z}_1, \ldots, \mathcal{Z}_H$ as follows:

$$\mathcal{Z}_i = \{(x_i, y_i) | y_i \in G_i\}. \tag{21}$$

Then for each subset $\mathcal{Z}_i$, another subset of samples O_i is randomly selected from the remaining samples (i.e., B percent of the data in $\mathcal{Z} - \mathcal{Z}_i$). In this paper, H and B are set to be 50 and 40, respectively. Finally, these two subsets $\mathcal{Z}_i$ and O_i are merged to form the final training set for learning the ith weak model using the batch-mode method described in Section 3.1. Note that $\mathcal{Z}_i$ and O_i are formed in a different way in that O_i is drawn randomly. By introducing a random component in the data subset we ensure that the feature space is to some extent well sampled for each weak model.

Learning β_i—Suppose H weak RDC models $\{f_{w,i}\}_{i=1}^{H}$ have been learned from the previous step. We now explore boosting to learn the weight β_i on the whole dataset $\mathcal{Z}$ iteratively (see Algorithm 2). Specifically, at the tth step, we first select the best weak distance model f_{w,k_t} that minimizes the following cost function:

$$k_t = \arg\min_i \sum_{\mathbb{O}_j} D_t^j \cdot \delta\big(f_{w,i}(\mathbf{x}_j^p) > f_{w,i}(\mathbf{x}_j^n)\big), \tag{22}$$

where D_t^j is the weight of pairwise difference vectors at the tth step, $\sum_{j=1}^{|\mathbb{O}|} D_t^j = 1$, and δ is a Boolean function. Then, D_t^j is updated as follows:

$$D_{t+1}^j = F^{-1} D_t^j \cdot \exp\left\{\alpha_t \cdot \big(f_{w,k_t}(\mathbf{x}_j^p) - f_{w,k_t}(\mathbf{x}_j^n)\big)\right\}, \tag{23}$$

where F is the normalizer such that $\sum_{j=1}^{|\mathbb{O}|} D_{t+1}^j = 1$. The weight α_t for the selected weak model f_{w,k_t} is then determined by

$$\alpha_t = 0.5 \cdot \log\frac{1+r}{1-r}, \qquad r = \sum_{j=1}^{|\mathbb{O}|} D_t^j \big(f_{w,k_t}(\mathbf{x}_j^n) - f_{w,k_t}(\mathbf{x}_j^p)\big). \tag{24}$$

According to [9], in order to ensure that the ensemble algorithm converges, each input weak RDC model $f_{w,i}$ is normalized by $\max_j \big|f_{w,i}(\mathbf{x}_j^p) - f_{w,i}(\mathbf{x}_j^n)\big|$, i.e.,

$$f_{w,i}(\cdot) \leftarrow \left(\max_j \big|f_{w,i}(\mathbf{x}_j^p) - f_{w,i}(\mathbf{x}_j^n)\big|\right)^{-1} f_{w,i}(\cdot), \tag{25}$$

so that $f_{w,i}(\mathbf{x}_j^p) - f_{w,i}(\mathbf{x}_j^n) \in [-1, +1]$.

By learning RDC in an ensemble way, each weak model is learned on a smaller set of data and the final distance function of the ensemble model is based on the score values of each weak model. Define $N^+(\mathbf{z}_i)$ ($N^-(\mathbf{z}_i)$) as the number of relevant (irrelevant) observations for query $\mathbf{z}_i$ in the training set. Note that the space complexity (memory cost) of creating all the training samples $\mathbf{x}_i^p$ and $\mathbf{x}_i^n$ is

$$O\left(\sum_{i=1}^{N} q \cdot N^+(\mathbf{z}_i) \cdot N^-(\mathbf{z}_i)\right), \tag{26}$$

where $N^-(\mathbf{z}_i) = N - N^+(\mathbf{z}_i)$-1, q is the number of features to describe each data sample. Assuming there are $\frac{N}{L}$ images for each person, we then have $N^+(\mathbf{z}_i) = \frac{N}{L}$-1. Therefore, to generate each weak RDC model in learning an ensemble RDC, the space complexity is reduced to $O(q \cdot ((\frac{b^2}{L} - \frac{b}{L^2}) \cdot N^3 + (\frac{b}{L}\text{-}b^2) \cdot N^2))$, where b is the percentage of all training samples used for building a weak RDC.[4] After generating the weak RDCs, the ensemble learning process itself has a space complexity of $O(H \cdot ((\frac{1}{L}\text{-}\frac{1}{L^2}) \cdot N^3 + (\frac{1}{L} - 1) \cdot N^2))$, where H is the number of groups (i.e., the total number of weak RDC models). As $H \ll q$, the boosting process has much less memory usage during training.

Algorithm 2: Algorithm of Ensemble RDC

Data: Pairwise relevant difference vector set $\mathbb{O}$, a set of weak RDC models $\{f_{w,i}\}_{i=1}^{H}$, Initial distribution D
begin
  $D_1 \longleftarrow D$;
  for $t = 1, \cdots, T$ **do**
    Select the best weak RDC mdoel f_{w,k_t} by Eq. (22);
    Compute the weight α_t by Eq. (24);
    Update the distribution D_{t+1} by Eq. (23).
  end
end
Output: $f_s(\mathbf{x}) = \sum_{t=1}^{T} \alpha_t \cdot f_{w,k_t}(\mathbf{x}) = \sum_{i=1}^{H} \beta_i \cdot f_{w,i}(\mathbf{x})$

Apart from reducing the space complexity of RDC, ensemble learning also alleviates the local optimum problem of the iterative algorithm proposed to solve the RDC optimization problem in Section 3.2. Note that each RDC model we described above is weak because it is only learned on a small set of training data and it may still suffer from the local optimum problem. As the ensemble learning theory in [9] ensures the matching error is minimized, the ensemble learning introduced above thus is able to alleviate the effect of being trapped in a local optimum. Our experiments show that the Ensemble RDC can generally yield equal or better performance as compared to the proposed batch mode RDC for large scale computing and is with reduced memory usage.

5 Relations to Alternative Models

Given the RDC model and its ensemble formulation, we shall now discuss the relations between these models and alternative models, specifically ranking models and distance learning models.

Relations to existing ranking models. Our RDC model is a special ranking model, concerned with only two ranks, i.e., the true match being ranked higher than any mismatches. In our early work [28], we investigated the use of a rank support vector machine (RankSVM)-based ranking model for person reidentification. In particular, the primal

4. The value of b is always smaller than 50 percent in our experiments under the aforementioned setting of H and B.

No commercial use ^_^

RankSVM proposed by Chapelle and Keerthi [2] is adopted, which is more suitable for large-scale learning compared to a standard RankSVM. The primal RankSVM aims to solve the following ranking optimization problem:

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2+\beta\sum_{i=1}^{|\mathcal{O}|}\max\left(0,1-\mathbf{w}^T\left(\mathbf{x}_i^p-\mathbf{x}_i^n\right)\right)^2, \quad (27)$$

where β is a positive importance weight on the ranking performance, and $\mathbf{x}_i^p$ and $\mathbf{x}_i^n$ are also computed in the absolute data difference space. Comparing this optimization problem with the one our RDC model attempts to solve (6), one can note the following fundamental differences between the two models:

1. RDC is able to explore the second-order information extracted from data due to the quadratic formulation in (4), learning weights for not only each individual feature but also the combination of each pair of features, while primal RankSVM only computes the weights $\mathbf{w}$ based on the first-order information, ignoring the correlations between features. This difference is due to the distance learning formulation of RDC and the linear SVM formulation of primal RankSVM.
2. With the hinge loss function, primal RankSVM is essentially a large margin-based optimization model due to the offset 1 and minimization of $\|\mathbf{w}\|$ in (27). In contrast, our RDC model enforces a softer constraint by using logistic function modeling. This enables the RDC model to be more tolerant to large intra and interclass variations and less prone to underfitting given undersampled data.
3. Differing from RDC, there is a free parameter β in the cost function of primal RankSVM which determines the relative weighting between the margin function and the ranking error function. Determining the optimal value of β is critical and can be achieved by cross-validation. However, person reidentification based on learning to rank is typically a large scale learning problem. Using cross-validation would further increase the computational cost a lot, making the model less tractable.

Another related ranking model one can consider is RankBoost based on the boosting technique. Comparing RDC to RankBoost [9], the major difference is that RDC quantifies the joint combination of different features rather than quantifying each feature independently. This individual local selection process makes the RankBoost model computationally much more expensive than either RDC or RankSVM, as demonstrated by our experiments (see Section 6.6). It is worth pointing out that although boosting technique is also used in our ensemble version of RDC, the objective is completely different: We aim to combine a handful of weak RDC models together rather than quantifying features individually and independently.

Relations to existing distance learning models. Among various existing distance learning methods, the methods in [29], [33], and [20] are the most relevant ones to our model as they also exploit the idea of relative distance comparison. However, there is a fundamental difference in their distance learning formulation; that is, in their models relative distance comparison is used as a constraint rather than as part of the cost function as in the RDC model. In some work, a common form of the constraint in these related models [29], [20] is as follows:

$$\mathbf{x}_n^T\mathbf{M}\mathbf{x}_n-\mathbf{x}_p^T\mathbf{M}\mathbf{x}_p\geq 1,$$

where $\mathbf{x}_p$ is the difference between relevant samples, $\mathbf{x}_n$ is that of the related irrelevant ones, and $\mathbf{M}$ is the distance matrix. Hence, when those models minimize the $\|\mathbf{M}\|_F$, it is equivalent to maximizing the margin $\frac{1}{\|\mathbf{M}\|_F}$ between a relevant pair and the corresponding related irrelevant one with a normalized distance matrix $\widetilde{\mathbf{M}}=\frac{\mathbf{M}}{\|\mathbf{M}\|_F}$. In [33], the model explicitly minimizes the intraclass variation and maximizes the interclass variation. As a result, these relative distance comparison models still either implicitly [29], [20] or explicitly [33] aim to learn a distance by which each class becomes more compact while being more separable from each other in an absolute sense. In contrast, RDC is only concerned with the relative distance comparison and using the comparison error itself as its cost function. This enables a distance to be learned with a softer constraint with the benefit of being more tolerant to intra and interclass variations and undersampling.

6 Experiments

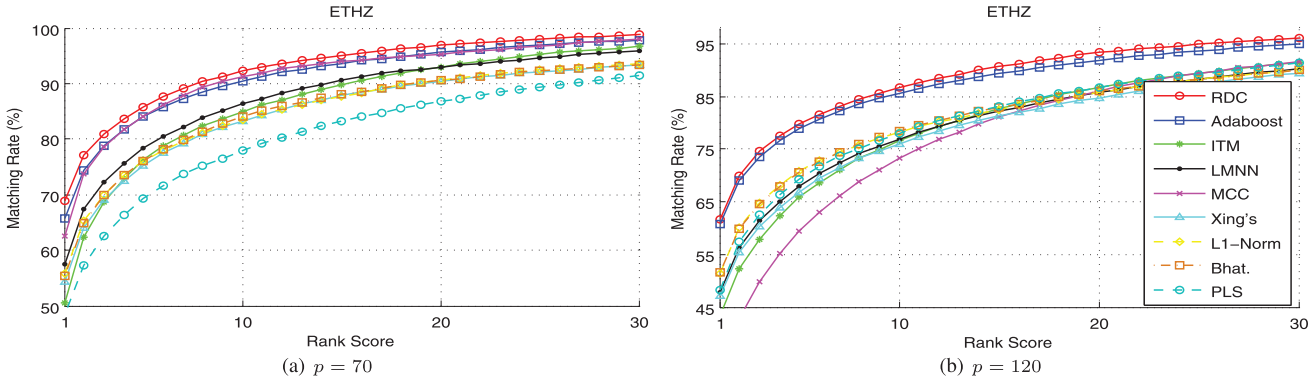
6.1 Datasets and Settings

Three publicly available person reidentification datasets, ETHZ [7], i-LIDS Multiple-Camera Tracking Scenario (MCTS) [37], [31], and VIPeR [14] were used for evaluation. The ETHZ dataset was originally designed for person detection and tracking in image sequences captured from a moving camera in a busy street scene. Schwartz and Davis [30] converted it into a person reidentification dataset by extracting images of a set of people selected from the video sequences[5] (i.e., those images of each person were assumed to have been taken from different camera views). This resulted in 146 people and 8,555 images in total. To make it more realistic to a multicamera setup, we randomly chose six images for each person for training in the dataset for our experiments. The image size is normalized to 128×64 pixels. The challenges of this dataset are the illumination changes and occlusions on people's appearance while the view angle change is small (see Fig. 5). 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 nonoverlapping cameras with an average of four images for each person. The images were normalized to a size of 128×64 pixels. Many of these images undergo large illumination change, considerable view angle change, and are subject to large occlusions (see Fig. 6). The VIPeR dataset[6] is a person reidentification dataset available consisting of 632 people captured outdoor with two images for each person with normalized size at 128×64 pixels. View angle change

5. The dataset can be downloaded at http://www.umiacs.umd.edu/~schwartz/datasets.html.

6. The dataset can be downloaded at http://vision.soe.ucsc.edu/?q=node/178.

No commercial use ^_^



(a) $p = 70$ (b) $p = 120$

Fig. 2. Performance comparison using CMC curves on ETHZ the dataset.

was 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. 7). Illumination change could also be drastic, but there was little occlusion. It is noted that these three datasets have different characteristics (e.g., outdoor/indoor, large/small variations in view angle, presence/absence of occlusion) and therefore are ideal for evaluating person reidentification algorithms given different challenges. Among them, the ETHZ dataset is considered to be the easiest one due to the fact that it was not actually captured by multiple nonoverlapping view cameras and thus lack of view angle change. Note that across the three datasets, the average number of training images of each person ranges from two (VIPeR) to six (ETHZ), highlighting the undersampled class distribution typical for the person reidentification problem.

In our experiments, we randomly selected all images of p people (classes) to set up the test set, and the rest of the people (classes) were used for training. Different values of p were used to evaluate the matching performance of models learned with different amounts of training data. 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 formed the pairwise set $\mathbb{O}$ defined in Section 3.

For evaluation, we use the average cumulative match characteristic (CMC) curves [14] 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 [14].

6.2 Feature Representation

We apply our RDC model as well as other models to an appearance representation of people captured by a set of different basic features. We start with a mixture of color and texture histogram features similar to those used in [15], [28] 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 across different radiuses and scales, and in total, 13 Schimid filters and 8 Gabor filters were obtained. In total, 29 feature channels were constructed for each stripe and each feature channel was represented by a 16D histogram vector. The details can be referred to in [15], [28]. Each person image was thus represented by a feature vector in a 2,784D feature space $\mathcal{Z}$. Since the features computed for this representation include low-level features widely used by existing person reidentification techniques, this representation is considered generic and representative.

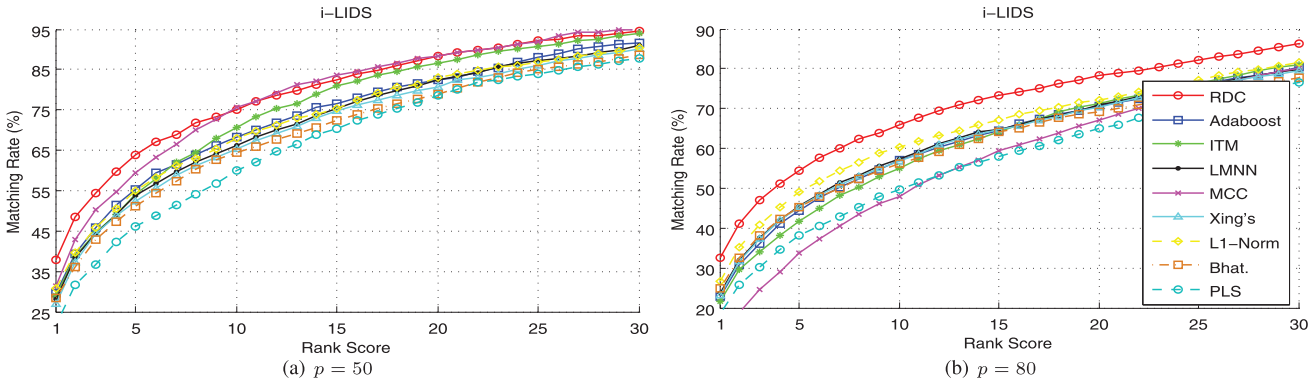
6.3 RDC versus Baseline Methods.

We first compared our RDC with baseline methods, namely nonlearning based l_1-norm distance and Bhattacharyya distance, which were used by most existing person reidentification work. Our results (Figs. 2, 3, and 4, Tables 2, 3, and 4) show clearly that with the proposed RDC, the matching performance for all three datasets is improved significantly, more so when the training set size increases. The improvement is particularly dramatic on the VIPeR dataset. In particular, Table 5 shows that a fourfold 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 RDC for the three datasets are shown in Figs. 5, 6, and 7 respectively.

6.4 RDC versus Adaboost and PLS

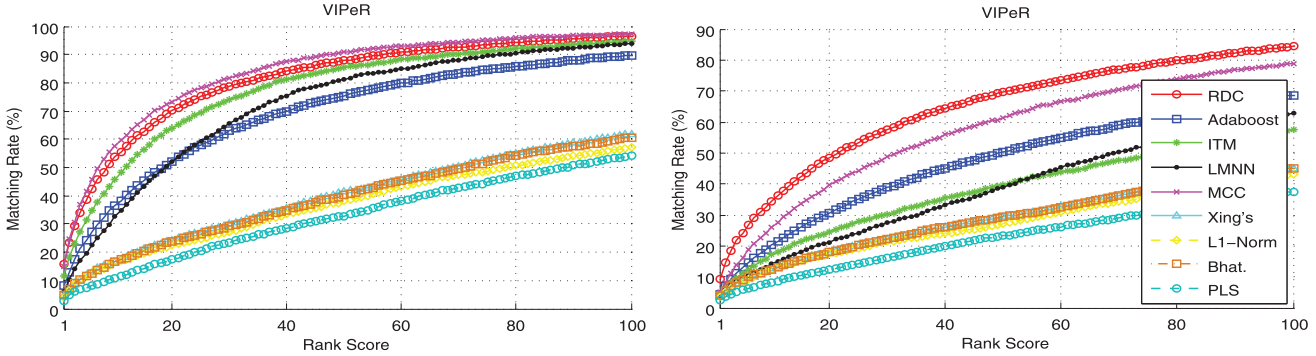
The Adaboost algorithm was formulated in [15] and the partial least squares (PLS) method was proposed in [30]. They are the only learning-based person reidentification methods we are aware of. In our experiments, the suggested settings in [15] and [30] were used. The Adaboost method in [15] is motivated by the observation that not all features are equally distinctive and reliable for matching people and aims to learn the weighting of different features. The proposed RDC algorithm also aims to compute the importance weight, but it differs in that 1) RDC performs a ranking-based soft discriminant feature selection while Adaboot in [15] performs large margin-based discriminant selection; 2) RDC is able to evaluate the importance of different combinations of features (second

No commercial use ^_^



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

Fig. 3. Performance comparison using CMC curves on the i-LIDS MCTS dataset.



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

Fig. 4. Performance comparison using CMC curves on the VIPeR dataset.

TABLE 2
Top Ranked Matching Rate (Percent) on ETHZ

Methods	$p = 40$				$p = 70$				$p = 120$			
	$r = 1$	$r = 5$	$r = 10$	$r = 20$	$r = 1$	$r = 5$	$r = 10$	$r = 20$	$r = 1$	$r = 5$	$r = 10$	$r = 20$
RDC	**72.65**	90.08	95.59	98.77	**68.96**	**85.82**	**92.23**	**96.85**	**61.58**	**79.70**	**86.65**	**93.33**
Adaboost	69.21	87.76	93.54	97.99	65.63	84.00	90.45	95.60	60.73	78.82	85.66	91.96
LMNN	64.88	84.23	92.04	97.11	57.58	78.37	86.29	92.94	47.87	67.90	76.96	85.78
ITM	65.38	86.81	94.06	98.63	56.26	80.74	88.64	94.06	43.09	65.95	76.55	86.75
MCC	71.92	**90.96**	**95.96**	**98.88**	62.52	84.14	91.20	95.32	31.08	59.40	73.19	86.02
Xing's	60.78	80.28	87.37	93.62	54.39	75.16	83.26	90.44	47.09	66.68	76.04	84.78
PLS	54.55	75.09	83.30	92.37	48.33	69.36	77.98	86.75	43.12	63.00	71.77	80.62
L1-norm	60.71	80.85	87.90	93.94	55.70	76.07	83.40	90.69	51.30	70.75	78.20	85.78
Bhat.	60.97	80.91	87.79	94.09	55.48	76.10	84.02	90.55	51.60	70.49	78.45	85.93

p is the size of the gallery set (larger p means smaller training set) and r is the rank.

TABLE 3
Top Ranked Matching Rate (Percent) on i-LIDS MCTS

Methods	$p = 30$				$p = 50$				$p = 80$			
	$r = 1$	$r = 5$	$r = 10$	$r = 20$	$r = 1$	$r = 5$	$r = 10$	$r = 20$	$r = 1$	$r = 5$	$r = 10$	$r = 20$
RDC	**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
PLS	25.76	57.36	73.57	90.31	22.10	46.04	59.95	78.68	18.32	38.23	49.68	64.95
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

p is the size of the gallery set and r is the rank.

order information), while Adaboost assumes different features are independent and selects them individually. As shown in Figs. 2, 3, and 4, and Tables 2, 3, and 4, our RDC model clearly outperforms the Adaboost-based method in all three datasets. The advantage is particularly significant on the more challenging i-LIDS and VIPeR datasets. For instance, for the VIPeR dataset, the rank 1 matching rate of RDC is twice of that of Adaboost for all three training/testing splits. This result highlights the importance of quantifying features globally rather than locally (individually).

Although PLS does not quantify features individually as Adaboost does, it does not perform well for person reidentification in our experiments. This is because PLS is a regression method and it can only be learned on the gallery dataset. Since there are only limited samples per

No commercial use ^_^

TABLE 4
Top Ranked Matching Rate (Percent) on VIPeR

Methods	$p = 316$				$p = 432$				$p = 532$			
	$r = 1$	$r = 5$	$r = 10$	$r = 20$	$r = 1$	$r = 5$	$r = 10$	$r = 20$	$r = 1$	$r = 5$	$r = 10$	$r = 20$
RDC	**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
PLS	2.72	7.53	10.92	17.34	2.43	6.6	9.33	13.84	2.31	5.75	8.21	12.50
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

p is the number of classes in the testing set; r is the rank.

TABLE 5
RDC versus Primal RankSVM (Percent) on ETHZ, i-LIDS, and VIPeR

DataSet	p	RDC				Primal RankSVM			
		$r = 1$	$r = 5$	$r = 10$	$r = 20$	$r = 1$	$r = 5$	$r = 10$	$r = 20$
ETHZ	$p = 40$	72.65	90.08	95.59	98.77	**73.91**	**90.44**	**96.10**	**98.85**
	$p = 70$	68.96	85.82	92.23	96.85	**69.11**	**86.19**	**92.25**	**97.18**
	$p = 120$	**61.58**	**79.70**	**86.65**	**93.33**	61.27	78.92	85.93	92.74
i-LIDS	$p = 30$	**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
VIPeR	$p = 316$	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

Fig. 5. Examples of person reidentification on ETHZ using RDC. 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.

person for training PLS and the people's appearance varies largely, PLS is sensitive to the learned data and may not generalize to new data very well. In contrast, our RDC model and the Adaboost model are learned using an independent training set consisting of different people from those in the gallery set. This not only contributes to better performance but also makes the methods more general applicable (i.e., applicable even with only a single gallery image per person).

6.5 RDC versus Related Distance Learning Methods

We also compared RDC with four alternative popular discriminant distance learning methods, namely, Xing's method [35], LMNN [33], ITM [5], and MCC [13]. Among the four methods, only LMNN exploits relative distance comparison, but it is used as an optimization constraint rather than the main objective function, and moreover a hard rather than a soft margin measure is used to quantify each relative distance comparison. MCC is based on Bayesian modeling, 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. Due to the space limitation, the standard derivations of all methods are not shown in the table. In our experiments, the standard derivations of all methods are mainly around 2-4 percent, where the proposed RDC is always around 2.5 percent and MCC is always between 3-4 percent.

The first thing we discovered in our experiments was that none of the four models were tractable due to the high dimensionality of the input data. PCA was thus performed

No commercial use ^_^

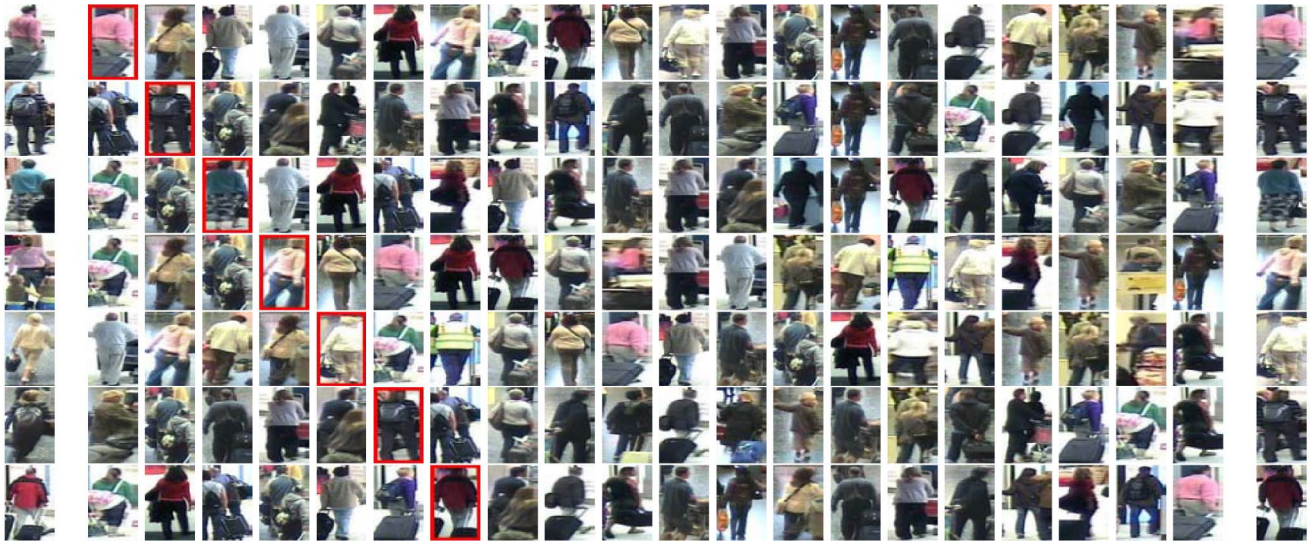


Fig. 6. Examples of person reidentification on i-LIDS MCTS using RDC.

Fig. 7. Examples of person reidentification on VIPeR using RDC.

to reduce the dimensionality while preserving 100 percent of the data. Our results (Figs. 2, 3, and 4, Tables 2, 3, and 4) clearly show that our model yields the best rank 1 matching rate and overall much superior performance compared to the compared models. The advantage of RDC 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 4 shows that on VIPeR when 100 people are used for learning and 532 people for testing ($p = 532$), the correct matching rate for RDC is almost more than doubled against any alternative distance learning methods. It is noted that, benefiting from being a Bayesian modeling, MCC gives the most comparable results to RDC when the training set is large. However, its performance degrades dramatically when the size of the training data decreases (see columns under $p = 120$ in Table 2, $p = 80$ in Table 3, and $p = 532$ in Table 4). Overall the results suggest that overfitting to undersampled training data is the main reason for the inferior performance of the compared alternative learning approaches.

6.6 RDC versus Related Ranking Methods

We first compare RDC with the primal RankSVM method used in [28]. Different from RDC, RankSVM has a free parameter β which determines the relative weights between the margin function and the ranking error function. We cross-validated the parameter β in $\{0.0001, 0.005, 0.001, 0.05, 0.1, 0.5, 1, 10, 100, 1{,}000\}$ for primal RankSVM. As shown in Table 5, the two methods all perform very well compared to non-learning-based methods and the four distance learning-based methods. Our RDC yields overall better performance, especially at lower rank matching rate and given less training data over the more challenging i-LIDS and VIPeR datasets. The better performance of RDC is mainly due to the logistic function-based modeling that enforces a softer constraint on relative distance comparison and exploiting second-order rather than first-order feature quantification. It is discovered that tuning the free parameter for primal RankSVM is not a trivial task and the performance can be sensitive to the tuning especially given

No commercial use ^_^

TABLE 6
RDC versus RankBoost (Percent) on ETHZ, i-LIDS, and VIPeR

DataSet	p	RDC				RankBoost			
		$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$
ETHZ	$p=120$	**61.58**	**79.70**	**86.65**	**93.33**	55.20	75.29	82.24	90.61
i-LIDS	$p=80$	**32.60**	**54.55**	**65.89**	**78.30**	18.25	40.09	53.01	68.86
VIPeR	$p=532$	**9.12**	**24.19**	**34.40**	**48.55**	3.01	10.06	15.60	24.89

TABLE 7
RDC versus Ensemble RDC (Precent) on ETHZ, i-LIDS, and VIPeR

ETHZ	Methods	$p=40$				$p=70$			
		$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$
	RDC	72.65	**90.08**	95.59	**98.77**	**68.96**	85.82	92.23	96.85
	Ensemble RDC	**73.51**	90.01	**95.88**	98.73	68.92	**86.11**	**92.37**	**96.94**
i-LIDS	Methods	$p=30$				$p=50$			
		$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$
	RDC	44.05	**72.74**	84.69	96.29	37.83	63.70	75.09	**88.35**
	Ensemble RDC	**45**	72.70	**85.11**	**96.44**	**39.73**	**64.93**	**75.71**	87.32
VIPeR	Methods	$p=316$				$p=432$			
		$r=1$	$r=5$	$r=10$	$r=20$	$r=1$	$r=5$	$r=10$	$r=20$
	RDC	15.66	38.42	53.86	70.09	12.64	31.97	44.28	59.95
	Ensemble RDC	**18.29**	**42.72**	**57.82**	**72.41**	**13.43**	**33.50**	**46.60**	**61.37**

p is the number of classes in the testing set; r is the rank.

undersampled data. Importantly, this results in more computational cost. The training of primal RankSVM took about 2.5 hours for each trial on i-LIDS and VIPeR, and about 8 hours for each trial on ETHZ. Hence, learning primal RankSVM is costly and could potentially be a serious problem for large-scale learning (e.g., matching in a camera network comprising hundreds of cameras). In contrast, the training of our RDC model was at least 10 times faster. (See Section 6.9 for more discussion on computational cost.) In addition, a more advanced development, namely, ensemble RDC, would achieve better performance than RDC in challenging cases.

We also compare RDC with RankBoost [9]. However, it turned out that RankBoost is intractable for our high-dimensional feature space (2,784D). Without access to special hardware, RankBoost was only tractable for the smallest training dataset setting for all three datasets. The main reason for this high computational cost is because RankBoost needs to learn an optimal weak classifier at each iteration, which has to determine a threshold parameter optimally over a large number of pairwise comparison ($O(N^3)$ with N the number of training images). Table 6 shows the results. It can clearly be seen that Rankboost performs much worse than our RDC. The possible reasons include: 1) The weak ranker in RankBoost is too weak based on a single feature, and 2) all features are treated independently.

6.7 Evaluation of Ensemble RDC

Ensemble RDC is proposed as an extension to RDC in order to alleviate the large scale computation problem in RDC. Table 7 shows that the ensemble RDC yields similar matching performance to RDC on ETHZ. But on the two more challenging datasets, ensemble RDC outperforms RDC. As expected, the ensemble RDC has much less space complexity than the batch model RDC. For instance, in the case of $p = 316$ for VIPeR, ensemble RDC took at most 2G RAM for learning the weak classifier while RDC required at least 10.4G RAM in our experiments. The better performance of ensemble RDC is likely due to the fact that the ensemble learning process can effectively alleviate the local optimum of the iterative algorithm for optimizing RDC. As we explained earlier, the formulated iterative algorithm in Section 3.2 may be trapped in a local optimum. With the boosting-based learning, an RDC that is particularly weak because of being trapped in a local optimum will be given a smaller weight. It thus alleviates the local optimum problem.

6.8 Further Evaluations of RDC

In this section, we further evaluate the proposed RDC methods in the following three aspects.

Effect of using logistic function. We first evaluate the usefulness of the logistic function based modeling. Without a logistic function, Criterion (6) becomes

$$\begin{aligned} &\min_{\mathbf{W}} r'(\mathbf{W}, \mathbb{O}),\ s.t.\ \ \mathbf{w}_i^T \mathbf{w}_j = 0,\ \forall i \neq j, \\ &\text{where } r'(\mathbf{W}, \mathbb{O}) = \sum_{\mathbb{O}_i} \|\mathbf{W}^T \mathbf{x}_i^p\|^2 - \|\mathbf{W}^T \mathbf{x}_i^n\|^2. \end{aligned} \tag{28}$$

This is similar to the maximum margin criterion (MMC) for feature extraction [21], which we call RDC-MMC in our experiments. The performance of RDC-MMC is compared with RDC in Table 8. The results show that without the logistic modeling for differentiating the margin in the difference information from different types, the RDC-MMC model performs much worse for person reidentification. This highlights the importance of using a logistic function for learning a person reidentification model.

Effect of learning in an absolute data difference space. We have shown in Section 3.4 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 RDC with RDC_{raw} which learns in the normal data difference space $\mathcal{DZ}$ (see Section 3.4). The result in Table 9 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

No commercial use ^_^

TABLE 8
RDC versus RDC-MMC (Percent) on ETHZ, i-LIDS, and VIPeR

ETHZ	Methods	$p = 40$				$p = 70$				$p = 120$			
		$r = 1$	$r = 5$	$r = 10$	$r = 20$	$r = 1$	$r = 5$	$r = 10$	$r = 20$	$r = 1$	$r = 5$	$r = 10$	$r = 20$
	RDC	**72.65**	**90.08**	**95.59**	**98.77**	**68.96**	**85.82**	**92.23**	**96.85**	**61.58**	**79.70**	**86.65**	**93.33**
	RDC-MMC	63.32	82.50	89.05	95.65	57.84	78.17	85.85	91.93	53.3	72.66	80.31	87.92
i-LIDS	Methods	$p = 30$				$p = 50$				$p = 80$			
		$r = 1$	$r = 5$	$r = 10$	$r = 20$	$r = 1$	$r = 5$	$r = 10$	$r = 20$	$r = 1$	$r = 5$	$r = 10$	$r = 20$
	RDC	**44.05**	**72.74**	**84.69**	**96.29**	**37.83**	**63.70**	**75.09**	**88.35**	**32.60**	**54.55**	**65.89**	**78.30**
	RDC-MMC	37.42	67.34	79.81	93.37	32.05	58.02	69.95	84.55	28.19	51.16	62.59	74.57
VIPeR	Methods	$p = 316$				$p = 432$				$p = 532$			
		$r = 1$	$r = 5$	$r = 10$	$r = 20$	$r = 1$	$r = 5$	$r = 10$	$r = 20$	$r = 1$	$r = 5$	$r = 10$	$r = 20$
	RDC	**15.66**	**38.42**	**53.86**	**70.09**	**12.64**	**31.97**	**44.28**	**59.95**	**9.12**	**24.19**	**34.40**	**48.55**
	RDC-MMC	6.90	17.94	24.56	36.42	5.76	14.56	21.02	30.05	4.92	12.31	17.89	25.85

p is the number of classes in the testing set; r is the rank.

TABLE 9
Effect of Learning (Percent) in an Absolute Data Difference Space

Methods	ETHZ ($p = 70$)				i-LIDS, ($p = 50$)				VIPeR ($p = 316$)			
	$r = 1$	$r = 5$	$r = 10$	$r = 20$	$r = 1$	$r = 5$	$r = 10$	$r = 20$	$r = 1$	$r = 5$	$r = 10$	$r = 20$
RDC	**68.96**	**85.82**	**92.23**	**96.85**	**37.83**	**63.70**	**75.09**	**88.35**	**15.66**	**38.42**	**53.86**	70.09
RDC_{raw}	10.45	30.75	44.61	63.05	19.92	50.19	68.29	86.40	12.28	37.28	53.83	**71.77**
ITM_{abs}	43.82	66.03	76.21	85.26	29.16	53.01	66.75	82.53	5.44	14.43	22.53	33.35
MCC_{abs}	23.73	52.91	67.89	81.82	5.59	23.01	43.59	70.47	1.20	3.51	5.6	9.68

TABLE 10
Average Rank of $\mathbf{W}$ Learned by RDC

Methods	ETHZ			i-LIDS MCTS			VIPeR		
	$p = 40$	$p = 70$	$p = 120$	$p = 30$	$p = 50$	$p = 80$	$p = 316$	$p = 432$	$p = 532$
$rank(\mathbf{W})$	1.9	2	4.4	3.2	2.4	2.3	2.9	3.2	3.7

difference space $\mathcal{DZ}$. It is possible to reformulate some of them in order to learn in an absolute data difference space. In Table 9, we show that when ITM and MCC are learned in the absolute data difference space $|\mathcal{DZ}|$, termed ITM_{abs} and MCC_{abs}, respectively, their performances become worse as compared to their results in Tables 2, 3, and 4. This indicates that the absolute different space is more suitable for our relative comparison distance learning, which makes the distance comparison more consistently.

6.9 Computational Cost

Though RDC 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 2.93 GHz CPU and 48 GB RAM server to learn RDC for each trial. We observed that the low cost of RDC 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 10. In comparison among the compared other methods, Adaboost was one of the most costly which took over 7 hours for each trial. The primal RankSVM took more than 2.5 hours.

7 Conclusions

We have formulated the person reidentification as a relative distance comparison problem. In particular, we proposed a relative distance comparison model, which aims to maximize the likelihood that a pair of true match has a smaller distance than that of a wrong match pair under a soft discriminant modeling. An ensemble strategy is also introduced to develop ensemble RDC in order to overcome limitations in RDC on both space complexity and local minimum. We have demonstrated that the proposed person reidentification models can alleviate the bias of large variations during optimization of learning similarity measurement. Our experiments validate that the proposed approach outperforms the related popular person reidentification techniques and related methods in terms of matching performance and tractability.

It would be interesting to investigate how information on groups of people can assist person reidentification as contextual information. This is motivated by the observation that humans often rely on the people surrounding the target person for identification if the target is occluded or has undistinguishable appearance. This contextual information is useful in certain public spaces such as the i-LIDS airport arrival scene where people typically walk with the same group of people even when they do not know each other, as demonstrated in our previous work [37]. However, how to automatically detect a group of people in practical scenarios is still an open problem which needs to be solved in order to utilize information of group of people as contextual information for person reidentification. Also, groups of people may merge, split, or undergo occlusion, and all these issues may affect the use of group information for helping person reidentification on target people. Hence, we consider that the key problem is on exploring the most reliable and robust features for group representation based on techniques such as context quantification [39].

It is worth pointing out although our RDC model is formulated specifically for addressing the person reidentification, it can be applied to solve other pattern recognition problems. In particular, there are other vision problems that share similar characteristics as person reidentification, i.e., large intra and interclass variations, large number of classes with few samples per class. Such problems include gait

No commercial use ^_^

recognition and large scale object recognition where there exists a large number of rare classes, each containing only a handful of samples. Extending RDC to address other vision problems is part of our ongoing work. Finally, in the current work, no attempt has been made to remove the background information from a person image which could typically have an negative effect on the performance of person reidentification. The idea was to rely on the proposed feature quantification technique to select the best features in order to eliminate the negative effect of background information. Nevertheless, it will be interesting to integrate an explicit background segmentation step into the proposed framework in the future.

ACKNOWLEDGMENTS

This research was partially funded by the EU FP7 project SAMURAI with grant no. 217899. Wei-Shi Zheng was additionally supported by the National Natural Science of Foundation of China (No. 61102111), the NSFC-GuangDong (No. U0835005, U1135001), Specialized Research Fund for the Doctoral Program of Higher Education (No. 20110171120051), the Natural Science Foundation of Guangdong Province (No. S2012010009926), and the Fundamental Research Funds for the Central Universities (No. 12lgpy28, 2012350003161455) for this work.

REFERENCES

[1] S. Bak, E. Corvee, F. Brémond, and M. Thonnat, "Person Re-Identification Using Spatial Covariance Regions of Human Body Parts," *Proc. IEEE Int'l Conf. Advanced Video and Signal Based Surveillance,* pp. 435-440, 2010.

[2] O. Chapelle and S.S. Keerthi, "Efficient Algorithms for Ranking with SVMS," *Information Retrieval,* vol. 13, pp. 201-215, June 2010.

[3] K. Chen, C. Lai, Y. Hung, and C. Chen, "An Adaptive Learning Method for Target Tracking across Multiple Cameras," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2008.

[4] D. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom Pictorial Structures for Re-Identification," *Proc. British Machine Vision Conf.,* 2011.

[5] J. Davis, B. Kulis, P. Jain, S. Sra, and I. Dhillon, "Information-Theoretic Metric Learning," *Proc. Int'l Conf. Machine Learning,* 2007.

[6] P. Dollar, Z. Tu, H. Tao, and S. Belongie, "Feature Mining for Image Classification," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2007.

[7] A. Ess, B. Leibe, and L. Van Gool, "Depth and Appearance for Mobile Scene Analysis," *Proc. IEEE Int'l Conf. Computer Vision,* 2007.

[8] M. Farenzena, L. Bazzani, A. Perina, M. Cristani, and V. Murino, "Person Re-Identification by Symmetry-Driven Accumulation of Local Features," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2010.

[9] Y. Freund, R. Iyer, R.E. Schapire, and Y. Singer, "An Efficient Boosting Algorithm for Combining Preferences," *J. Machine Learning Research,* no. 4, pp. 933-969, 2003.

[10] N. Gheissari, T. Sebastian, and R. Hartley, "Person Reidentification Using Spatiotemporal Appearance," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2006.

[11] A. Gilbert and R. Bowden, "Tracking Objects Across Cameras by Incrementally Learning Inter-Camera Colour Calibration and Patterns of Activity," *Proc. European Conf. Computer Vision,* 2006.

[12] A. Gilbert and R. Bowden, "Incremental, Scalable Tracking of Objects Inter Camera," *Computer Vision and Image Understanding,* vol. 111, no. 1, pp. 43-58, 2008.

[13] A. Globerson and S. Roweis, "Metric Learning by Collapsing Classes," *Proc. Advances in Neural Information Processing Systems,* 2005.

[14] D. Gray, S. Brennan, and H. Tao, "Evaluating Appearance Models for Recognition Reacquisition, and Tracking," *Proc. IEEE Int'l Workshop Performance Evaluation of Tracking and Surveillance,* 2007.

[15] D. Gray and H. Tao, "Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features," *Proc. European Conf. Computer Vision,* 2008.

[16] R. Herbrich, T. Graepel, and K. Obermayer, "Large Margin Rank Boundaries for Ordinal Regression," *Proc. Advances in Neural Information Processing Systems,* pp. 115-132, 1999.

[17] W. Hu, M. Hu, X. Zhou, J. Lou, T. Tan, and S. Maybank, "Principal Axis-Based Correspondence between Multiple Cameras for People Tracking," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 28, no. 4, pp. 663-671, Apr. 2006.

[18] O. Javed, K. Shafique, Z. Rasheed, and M. Shah, "Modeling Inter-Camera Space-Time and Appearance Relationships for Tracking across Non-Overlapping Views," *Computer Vision and Image Understanding,* vol. 109, no. 2, pp. 146-162, 2008.

[19] O. Javed, K. Shafique, and M. Shah, "Appearance Modeling for Tracking in Multiple Non-Overlapping Cameras," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2005.

[20] J. Lee, R. Jin, and A. Jain, "Rank-Based Distance Metric Learning: An Application to Image Retrieval," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2008.

[21] H. Li, T. Jiang, and K. Zhang, "Efficient and Robust Feature Extraction by Maximum Margin Criterion," *IEEE Trans. Neural Networks,* vol. 17, no. 1, pp. 157-165, Jan. 2006.

[22] G. Lian, J. Lai, and W.-S. Zheng, "Spatial-Temporal Consistent Labeling of Tracked Pedestrians across Non-Overlapping Camera Views," *Pattern Recognition,* vol. 44, no. 5, pp. 1121-1136, 2011.

[23] C.-J. Lin, "Projected Gradient Methods for Nonnegative Matrix Factorization," *Neural Computation,* vol. 19, no. 10, pp. 2756-2779, 2007.

[24] C. Loy, T. Xiang, and S. Gong, "Time-Delayed Correlation Analysis for Multi-Camera Activity Understanding," *Int'l J. Computer Vision,* vol. 90, no. 1, pp. 106-129, 2010.

[25] D. Makris, T. Ellis, and J. Black, "Bridging the Gaps between Cameras," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2004.

[26] U. Park, A. Jain, I. Kitahara, K. Kogure, and N. Hagita, "Vise: Visual Search Engine Using Multiple Networked Cameras," *Proc. Int'l Conf. Pattern Recognition,* 2006.

[27] B. Prosser, S. Gong, and T. Xiang, "Multi-Camera Matching under Illumination Change over Time," *Proc. ECCV Workshop Multi-Camera and Multi-Modal Sensor Fusion Algorithms and Applications,* 2008.

[28] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang, "Person Re-Identification by Support Vector Ranking," *Proc. British Machine Vision Conf.,* 2010.

[29] M. Schultz and T. Joachims, "Learning a Distance Metric from Relative Comparisons," *Proc. Advances in Neural Information Processing Systems,* 2004.

[30] W. Schwartz and L. Davis, "Learning Discriminative Appearance-Based Models Using Partial Least Squares," *Proc. Brazilian Symp. Computer Graphics and Image Processing,* 2009.

[31] UK, "Home Office i-LIDS Multiple Camera Tracking Scenario Definition," 2008.

[32] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu, "Shape and Appearance Context Modeling," *Proc. IEEE Int'l Conf. Computer Vision,* 2007.

[33] K. Weinberger, J. Blitzer, and L. Saul, "Distance Metric Learning for Large Margin Nearest Neighbor Classification," *Proc. Advances in Neural Information Processing Systems,* 2006.

[34] S. Xiang, F. Nie, and C. Zhang, "Learning a Mahalanobis Distance Metric for Data Clustering and Classification," *Pattern Recognition,* vol. 41, no. 12, pp. 3600-3612, 2008.

[35] E. Xing, A. Ng, M. Jordan, and S. Russell, "Distance Metric Learning with Application to Clustering with Side-Information," *Proc. Advances in Neural Information Processing Systems,* 2002.

[36] L. Yang, R. Jin, R. Sukthankar, and Y. Liu, "An Efficient Algorithm for Local Distance Metric Learning," *Proc. 21st Nat'l Conf. Artificial Intelligence,* pp. 543-548, 2006.

[37] W.-S. Zheng, S. Gong, and T. Xiang, "Associating Groups of People," *Proc. British Machine Vision Conf.,* 2009.

[38] W.-S. Zheng, S. Gong, and T. Xiang, "Person Re-Identification by Probabilistic Relative Distance Comparison," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2011.

No commercial use ^_^

[39] W.-S. Zheng, S. Gong, and T. Xiang, "Quantifying and Transferring Contextual Information in Object Detection," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 34, no. 4, pp. 762-777, Apr. 2012.

Wei-Shi Zheng received the PhD degree in applied mathematics from Sun Yat-Sen University in 2008, and has been a postdoctoral researcher on the EU FP7 SAMURAI Project at Queen Mary University London. He joined Sun Yat-sen University under the one-hundred-people program in 2011. His research interests include object association and categorization in visual surveillance. He is a member of the IEEE.

Shaogang Gong received the DPhil degree in computer vision from Keble College, Oxford University in 1989. He is a professor of visual computation at Queen Mary University London and is a fellow of the Institution of Electrical Engineers and the British Computer Society. His work focuses on motion and video analysis; object detection, tracking, and recognition; face and expression recognition; gesture and action recognition; visual behavior recognition.

Tao Xiang received the PhD degree in electrical and computer engineering from the National University of Singapore in 2002. He is a senior lecturer (associate professor) at Queen Mary University London. His research interests include computer vision, statistical learning, video processing, and machine learning, with focus on interpreting and understanding human behavior.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.