Discriminative Dictionary Learning With Ranking Metric Embedded for Person Re-Identification

De Cheng[1,2], Xiaojun Chang[2], Li Liu[3],
Alexander G. Hauptmann[2], Yihong Gong[1] *, Nanning Zheng[1].
Xi'an Jiaotong University, China[1], Carnegie Mellon University,USA[2], Malong Technologies Co. Ltd[3].
{chengde19881214}@stu.xjtu.edu.cn,{ygong,nnzheng}@xjtu.edu.cn, li.liu@malongtech.cn,
{uqxchan1,alex}@cs.cmu.edu.

Abstract

The goal of person re-identification (Re-Id) is to match pedestrians captured from multiple non-overlapping cameras. In this paper, we propose a novel dictionary learning based method with ranking metric embedded, for person Re-Id. A new and essential ranking graph Laplacian term is introduced, which minimizes the intra-personal compactness and maximizes the inter-personal dispersion in the objective. Different from the traditional dictionary learning based approaches and their extensions, which just use the same or not information, our proposed method can explore the ranking relationship among the person images, which is essential for such retrieval related tasks. Simultaneously, one distance measurement matrix has been explicitly learned in the model to further improve the performance. Since we have reformulated these ranking constraints into the graph Laplacian form, the proposed method is easy-to-implement but effective. We conduct extensive experiments on three widely used person Re-Id benchmark datasets, and achieve state-of-the-art performances.

1 Introduction

Person Re-Id aims at the maintenance of a global identity as a person moves among non-overlapping surveillance cameras. It is essential for video surveillance and has drawn great attention recently [Cheng *et al.*, 2016; Xiao *et al.*, 2016; Xiong *et al.*, 2014; Wang *et al.*, 2016]. Many algorithms have been proposed to tackle this problem, which can be mainly divided into two categories, which are the distance metric learning methods and feature learning methods. The distance learning methods usually learn distance metrics that are expected to be robust to sample variations [Jose and Fleuret, 2016; Chen *et al.*, 2016a], while feature representation learning methods aim to extract discriminative and distinct features from pedestrian images [Chen *et al.*, 2016b; Wu *et al.*, 2016b; Varior *et al.*, 2016a; Ahmed *et al.*, 2015]. However, the representation power of the learned features or metrics might be limited, and this task still remains a challenging problem due to the following two main reasons: 1) A person's appearance often changes dramatically across camera views due to occlusion, lighting conditions, illumination and pose changes, in real-world scenarios; 2) Different people in public spaces wear similar clothes (e.g. dark coats, jeans) thus having the similar visual appearance.

In order to overcome aforementioned challenges and improve the person Re-Id performances, we propose a novel ranking metric embedded dictionary learning method, which makes the traditional dictionary learning method more suitable for person Re-Id. The embedded ranking metric pulls the same person images to be close while pushes different individuals' images far apart. Thus, by embedding these ranking constraints, we can both reduce the intra-personal variations and enlarge the inter-personal variations, which is essential for the retrieval related tasks, especially for person Re-Id. Although the dictionary learning methods have also been well studied in the past several years, and there are many dictionary learning based methods specifically designed for person Re-Id [Karanam *et al.*, 2015; Jiang *et al.*, 2013; Yang *et al.*, 2016; Zhang and Li, 2010], most of them have just focused on embedding the same identity information without embedding this necessary and essential ranking information.

In this paper, our proposed dictionary learning method mainly consists of two components: one is the dictionary related part, which minimizes the reconstruction error between the original image features and the projected feature coefficients; another is the embedded ranking graph Laplacian matrix, which makes the projected feature coefficients of the same person images closer than that of different person images by a large margin. Though this intuition has been widely explored in other metric learning areas, no existing dictionary learning approach has fully explored this property. To the best of our knowledge, we are the first to reformulate all these triplet ranking constraints on all the datasets into the graph Laplacian form, and then explicitly integrate it into the dictionary learning. Moreover, one distance measurement matrix has been simultaneously learned in the dictionary learning, which further improves the Re-Id performance. Since we have learned one common dictionary to represent both the gallery and probe images, the learned dictionary is invariant to the viewpoint changes. Hence, our learned dictionary is capable of discriminatively encoding the feature vectors of different people, and can also encourage signals from the same

*Corresponding author: Yihong Gong(ygong@mail.xjtu.edu.cn)

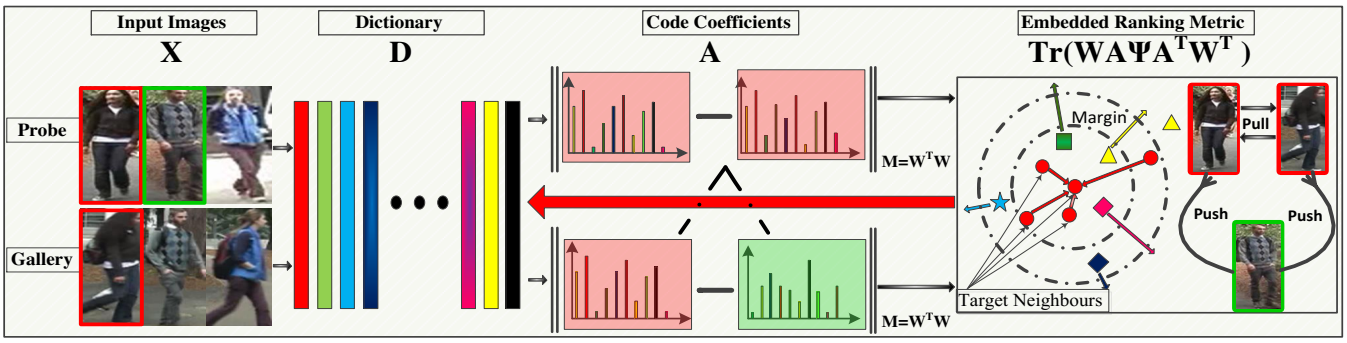

Figure 1: Illustration of the framework for our proposed novel dictionary learning method for person Re-Id. First, we formulate the traditional triplet ranking constraint into the graph Laplacian form $Tr(\mathbf{WA}\Psi\mathbf{A}^T\mathbf{W}^T)$, and then embed it into the dictionary learning process. Simultaneously, the Mahalanobis distance metric $\mathbf{M} = \mathbf{W}^T\mathbf{W}$ has been explicitly learned. Thus, our proposed method iteratively trains a discriminative viewpoint invariant dictionary, and jointly learns the dictionary $\mathbf{D}$, code coefficients $\mathbf{A}$ and the distance measurement matrix $\mathbf{M}$ by the essential ranking graph Laplacian matrix embedded dictionary learning method.

person to have more similar features, and signals from different persons to have dissimilar features. Figure 1 illustrates the framework for our proposed method. Our experimental results show that the proposed ranking metric embedded dictionary learning method is effective for improving the person Re-Id performance.

To summarize, the main contributions are as follows:

- To the best of our knowledge, we are the first to formulated the triplet ranking constraints into the graph Laplacian form, and then embed it into the dictionary learning, which makes the traditional dictionary learning method more suitable for person Re-Id task.
- Simultaneously, one distance measurement matrix has been explicitly learned in the dictionary learning objective, which can further improve the Re-Id performance.
- We conduct experiments on three widely used benchmark datasets and achieve state-of-the-art performances.

2 Related Work

In this section, we review some of the representative related works of person Re-Id and dictionary learning.

Person Re-Id. In recent years, many algorithms have been proposed for person Re-Id. Some traditional methods focus on learning effective metrics to measure the distance between two images captured from different camera views [Xiong *et al.*, 2014][Pedagadi *et al.*, 2013][Liao *et al.*, 2015]. Among them, the Mahalanobis distance function [Chen *et al.*, 2016a], triplet loss function and its extensions have been well explored [Cheng *et al.*, 2016]. Other research works focus on learning discriminative features, including the attributes, salience features, gaussian descriptors, and some other learned features [Matsukawa *et al.*, 2016; Liao *et al.*, 2015]. Nowadays, deep learning based methods have learned good feature representations and achieved promising performances on almost all the person Re-Id benchmark datasets [Xiao *et al.*, 2016; Ahmed *et al.*, 2015; Cheng *et al.*, 2016]. Our proposed method falls into the category of metric learning, which embeds the ranking distance metric into the dictionary learning.

Dictionary Learning. Recently, dictionary learning methods have been successfully applied to various recognition problems, and many extension works have also been proposed. [Jiang *et al.*, 2013] employed label consistency constraints to jointly learn a discriminative dictionary and a linear classifier. [Zhang and Li, 2010] extended this algorithm by incorporating classification error into the problem formulation and learned class-wise dictionaries. Also, many other extension works used the dictionary learning methods to learn discriminative feature encodings for person Re-Id. For example, [Yang *et al.*, 2016] embedded one metric into the dictionary learning by using the same or not information, and [Karanam *et al.*, 2015] enforces the discriminability by imposing explicit constraints on the projected sparse codes. There are also a lot of works using the dictionary learning methods for unsupervised person Re-Id [Kodirov *et al.*, 2016].

In contrast to the aforementioned approaches, our method explicitly incorporates the ranking metric into the dictionary learning, and simultaneously learns a distance measurement matrix to improve the discriminability for person Re-Id.

3 Algorithm Description

In this section, we first briefly review the basics of dictionary learning. Then we present the proposed ranking metric embedded approach to learn discriminative and viewpoint invariant dictionaries, followed by its optimization method.

3.1 Dictionary Learning Revisit

Formally, assume $\mathbf{X} \in \mathbf{R}^{M\times N}$ is an input feature matrix, with each column $\mathbf{x}_i$ corresponding to an M dimensional feature vector representing the i-th person image's appearance feature, and N represents the total number of samples in the dataset. Adopting the dictionary learning model, our goal is to learn a dictionary $\mathbf{D} \in \mathbf{R}^{M\times K}$. With this dictionary, each M dimensional feature vector is projected onto a lower K dimensional subspace $\mathbf{A}$ spanned by the K dictionary atoms (columns of $\mathbf{D}$), thus their corresponding coefficients (code vectors) can be matched by the Euclidean distance in the subspace. This can be formulated as the following objective:

$$\begin{aligned}\mathbf{D}^*, \mathbf{A}^* = &\underset{\mathbf{D},\mathbf{A}}{\operatorname{argmin}} ||\mathbf{X} - \mathbf{D}\mathbf{A}||_F^2 + \lambda||\mathbf{A}||_F^2 \\ &s.t. \quad ||\mathbf{d}_i||_2^2 \leq 1, \forall i,\end{aligned} \tag{1}$$

where $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_N]$ correspond to the coding vectors of the input signals $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$, and $\mathbf{d}_i \in \mathbf{R}^{M\times 1}$ is the i-th column of the learned dictionary $\mathbf{D}$. The constraint of $\mathbf{d}_i$ in Eq. (1) enforces the learned dictionary atoms to be compact. This problem is typically solved by alternately fixing $\mathbf{D}$ and $\mathbf{A}$, and then optimize over the other variables.

As for the person Re-Id problem, there always exists large viewpoint changes among the probe and gallery cameras. In order to learn a dictionary $\mathbf{D}$ satisfying the property of viewpoint invariant, we have learned a common dictionary to represent both the gallery and probe images.

3.2 The Proposed Dictionary Learning Algorithm With Ranking Metric Embedded

Person Re-Id aims at searching for a person of interest from a large amount of candidate gallery images captured from different cameras. Since the ranking information is essential for this task, we intuitively incorporate the widely explored triplet ranking constraints into our objective for discriminative dictionary learning.

Suppose we have a coding coefficient matrix $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_N] \in R^{K\times N}$ corresponding to the original data matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N] \in R^{M\times N}$ as illustrated in Eq. (1). Each column of A denotes a new representation of each data $\mathbf{x}_i$ in the new space. With the training data, we hope that the objective function can encourage the dictionary to find the embeddings where the distance between the same person images should be closer than that of the different person images by a large margin τ, which is inspired by the triplet loss function. Moreover, the distance between a pair is measured by the widely used Mahalanobis distance, instead of directly using the Euclidean distance between the projected feature vectors $(\mathbf{a}_i, \mathbf{a}_j)$. Thus, we simultaneously minimize the following term of Eq. (2) on all the datasets, besides Eq. (1).

$$\mathbf{\Gamma}(\mathbf{A}, \mathbf{W}) = \sum_{\substack{i,j,k=1;\\ l_i=l_j\neq l_k}}^{N} [f_{\mathbf{W}}(\mathbf{a}_i, \mathbf{a}_j) - f_{\mathbf{W}}(\mathbf{a}_i, \mathbf{a}_k) + \tau]_+, \tag{2}$$

where $[.]_+$ is the hinge loss function $max(0, .)$, l_i is the identity of the i-th training sample, $f_{\mathbf{W}}(\mathbf{a}_i, \mathbf{a}_j) = ||\mathbf{W}(\mathbf{a}_i - \mathbf{a}_j)||_2^2 = (\mathbf{a}_i - \mathbf{a}_j)^T\mathbf{M}(\mathbf{a}_i - \mathbf{a}_j)$, and $\mathbf{M} = \mathbf{W}^T\mathbf{W}$ is semi-definite, which indicates that the distance between a pair is measured by the Mahalanobis distance [Weinberger and Saul, 2009]. As illustrated in Eq. (2), we have used all the training data to generate all possible sample triplets to form the ranking cost $\mathbf{\Gamma}(\mathbf{A}, \mathbf{W})$. Thus, the ranking triplet loss is constituted by all the Mahalanobis distance of the sample pairs $(\mathbf{a}_i, \mathbf{a}_j)$ from the training set. Then, we reformulate Eq. (2) into the graph Laplacian form as Eq. (3):

$$\begin{aligned}\mathbf{\Gamma}(\mathbf{A}, \mathbf{W}) &= \sum_{i,j=1}^{N} s_{ij} f_{\mathbf{W}}(\mathbf{a}_i, \mathbf{a}_j) + C \\ &= \sum_{i,j=1}^{N} s_{ij} ||\mathbf{W}(\mathbf{a}_i - \mathbf{a}_j)||_2^2 + C \\ &= 2\,Tr(\mathbf{W}\mathbf{A}\Psi\mathbf{A}^T\mathbf{W}^T) + C,\end{aligned} \tag{3}$$

where C is a constant, s_{ij} is the adjacent weight between the sample pair $(\mathbf{a}_i,\ \mathbf{a}_j)$, $\Psi = \mathbf{G} - (\mathbf{S} + \mathbf{S}^T)/2$, $\mathbf{G} = diag(g_{11}, \ldots, g_{NN})$, $g_{ii} = \sum_{j=1, j\neq i}^{N} \frac{s_{ij}+s_{ji}}{2}, j = 1, 2, \ldots, N$, and Ψ is called the Laplacian matrix of $\mathbf{S}$, $Tr(.)$ denotes the trace of a matrix. The deduction from line 2 to 3 in Eq. (3) can refer to [Shi *et al.*, 2016]. The element s_{ij} of the adjacent matrix $\mathbf{S}$ in Eq. (3) can be deduced from Eq. (2) as follows:

$$s_{ij} = \begin{cases} \sum\limits_{\substack{k=1,\\ l_i=l_j\neq l_k}}^{N} \delta[f_{\mathbf{W}}(\mathbf{a}_i, \mathbf{a}_j) - f_{\mathbf{W}}(\mathbf{a}_i, \mathbf{a}_k) + \tau], i \neq j, \\ -\sum\limits_{\substack{k=1,\\ l_i=l_k\neq l_j.}}^{N} \delta[f_{\mathbf{W}}(\mathbf{a}_i, \mathbf{a}_k) - f_{\mathbf{W}}(\mathbf{a}_i, \mathbf{a}_j) + \tau], i \neq j, \\ 0, \quad i = j, \end{cases} \tag{4}$$

where the function $\delta[.]$ is an indicator function which takes one if the argument is bigger than zero, and zeros otherwise.

Therefore, the proposed dictionary learning algorithm with ranking metric embedded arrives at:

$$\begin{aligned}\underset{\mathbf{D},\mathbf{A},\mathbf{W}}{\operatorname{argmin}} &||\mathbf{X} - \mathbf{D}\mathbf{A}||_F^2 + \frac{\beta}{N(\tau)} Tr(\mathbf{W}\mathbf{A}\Psi\mathbf{A}^T\mathbf{W}^T) \\ &+ \lambda||\mathbf{A}||_F^2 + \alpha||\mathbf{W}||_F^2 \\ s.t. \quad &||\mathbf{d}_i||_2^2 \leq 1, \forall i,\end{aligned} \tag{5}$$

where C has been ignored from Eq. (3) as the constant has no influence on the objective, and $N(\tau)$ is the number of all the sample triplets constructed by the N training examples. The parameters λ, α and β are used to control the contributions of the corresponding terms. In Eq. (5), the first term denotes the reconstruction error. The second term is the embedded raking metric which maintains the distance of similar sample pairs to be closer than that of the dissimilar pairs by a large margin in the learned dictionary space, thus reduce the intra-personal variations. The last two terms are the regularization terms to avoid over-fitting.

3.3 Optimization

Remark: *Since there are many negative elements in* $\mathbf{S}$ *(with each element* s_{ij} *computed by Eq.* (4)*), the ranking Laplacian term* $Tr(\mathbf{W}\mathbf{A}\Psi\mathbf{A}^T\mathbf{W}^T)$ *in Eq.* (5) *is not convex for* $\mathbf{A}$ *or* $\mathbf{W}$ *with other variables fixed. Hence, we optimize both* $\mathbf{A}$ *and* $\mathbf{W}$ *by the gradient decent method. While fixing* $\mathbf{A}$ *and* $\mathbf{W}$*,*

the objective function in Eq. (5) *is convex for* $\mathbf{D}$. *Thus* $\mathbf{D}$ *can be computed by its analytic solution.*

In order to efficiently optimize Eq. (5), we first initialize the parameters $\mathbf{D}$, $\mathbf{A}$ and $\mathbf{W}$ as follows, respectively. 1) $\mathbf{W}$ can be initialized as an identity matrix; 2) $\mathbf{D}$ and $\mathbf{A}$ can be initialized by solving the standard dictionary learning problem as defined in Eq. (1), the optimization method and its implementation can refer to [Peng *et al.*, 2016]. Because our work is based on the standard dictionary learning, optimizing Eq. (1) as the initialization can use its analytic solution which is much more efficient than directly using the gradient decent algorithm all the time.

Given the above initialization, we solve the problem of Eq. (5) by alternating among the following three subproblems:

Fix A and W, update D: Given $\mathbf{A}$ and $\mathbf{W}$, the objective function becomes

$$\mathbf{D}^* = \underset{\mathbf{D}}{\operatorname{argmin}} ||\mathbf{X} - \mathbf{D}\mathbf{A}||_F^2,\ s.t.\ ||\mathbf{d}_i||_2^2 \leq 1, \forall i. \tag{6}$$

To solve Eq. (6), we use the Lagrange dual method [Lee *et al.*, 2007]. The analytical solution of $\mathbf{D}$ can be computed as: $\mathbf{D}^* = \mathbf{X}\mathbf{A}^T(\mathbf{A}\mathbf{A}^T + \Lambda^*)^{-1}$. Where Λ^* is a diagonal matrix constructed by all the optimal dual variables. In practice, $\mathbf{A}\mathbf{A}^T + \Lambda^*$ is not guaranteed to be invertible, therefore pseudo inverse is used in place of computing it directly.

Fix D and W, update A: When $\mathbf{D}$ and $\mathbf{W}$ are fixed, the objective becomes minimizing the following function:

$$F(A) = ||\mathbf{X} - \mathbf{D}\mathbf{A}||_F^2 + \frac{\beta}{N(\tau)} Tr(\mathbf{W}\mathbf{A}\Psi\mathbf{A}^T\mathbf{W}^T) + \lambda||\mathbf{A}||_F^2. \tag{7}$$

Since the Laplacian matrix Ψ in Eq. (7) is based on the adjacent matrix $\mathbf{S}$, and each s_{ij} is computed according to Eq. (3) and Eq. (4), which is always changing during the iterations. Besides, since S is not positive semi-definite, we choose to use the gradient decent method to optimize Eq. (7). In order to keep the convergence of Eq.(7), we always keep Ψ fixed when optimizing $\mathbf{A}$. After that, we update $\mathbf{S}$ and Ψ according to Eq. (3) and Eq. (4). Solving the dictionary learning objective in Eq. (1) by its analytic solution as the initialization for $\mathbf{A}$ first, can greatly improve the optimization efficiency. The gradient of $F(\mathbf{A})$ in Eq. (7) can be computed as Eq. (8),

$$\frac{\partial F(\mathbf{A})}{\partial \mathbf{A}} = 2\mathbf{D}^T(\mathbf{D}\mathbf{A} - \mathbf{X}) + \frac{\beta}{N(\tau)}\mathbf{W}^T\mathbf{W}\mathbf{A}(\Psi^T + \Psi) + 2\lambda\mathbf{A}. \tag{8}$$

Then we use the gradient decent method to update $\mathbf{A}$ at step t, $\mathbf{A}^{t+1} = \mathbf{A}^t - \eta\frac{\partial F(\mathbf{A})}{\partial \mathbf{A}}$, and η is the learning rate. Detailed optimization procedure can be shown in Algorithm 1.

Fix D and A, update W: Given $\mathbf{D}$ and $\mathbf{A}$, the objective function becomes minimizing the following function:

$$\urcorner(\mathbf{W}) = \frac{\beta}{N(\tau)} Tr(\mathbf{W}\mathbf{A}\Psi\mathbf{A}^T\mathbf{W}^T) + \alpha||\mathbf{W}||_F^2. \tag{9}$$

The same reason for optimizing $\mathbf{A}$, we also need to keep Ψ fixed when optimizing $\mathbf{W}$. After getting $\mathbf{W}$, we update $\mathbf{S}$ and Ψ according to Eq. (3) and Eq. (4). The gradient of $\urcorner(\mathbf{W})$ is deduced as Eq. (10),

$$\frac{\partial \urcorner(\mathbf{W})}{\partial \mathbf{W}} = \frac{\beta}{N(\tau)}\mathbf{W}\mathbf{A}(\Psi^T + \Psi)\mathbf{A}^T + 2\alpha\mathbf{W}. \tag{10}$$

Then, we also use the gradient decent method to update $\mathbf{W}$ at step t, $\mathbf{W}^{t+1} = \mathbf{W}^t - \eta\frac{\partial \urcorner(\mathbf{W})}{\partial \mathbf{W}}$. The detailed optimization procedure is illustrated in Algorithm 1.

Algorithm 1: The Ranking Metric Embedded Discriminative Dictionary Learning Method

Input: Training Data matrix $\mathbf{X}$, parameters α, β, λ and τ, iteration number T.
Output: The learned dictionary $\mathbf{D}$ and the explicitly learned projection matrix $\mathbf{W}$.

```
Initialize D, A and W following the initialization description in Section 3.3;
Compute the Laplacian matrix Ψ according to the examples labels of X, Eq. (3) and Eq. (4);
for t = 1, 2, ..., T do
    Update the dictionary D according to Eq. (6);
    Update the code coefficients A as follows;
    Compute the adjacent matrix S according to Eq. (3) and Eq. (4), based on current A and W;
    Compute the Laplacian matrix Ψ based on current S as described in Section 3.2;
    while Not converged do
        Update A^{t+1} = A^t − η ∂F(A)/∂A based on Eq. (8);
    Update the projection matrix W as follows;
    Compute the adjacent matrix S according to Eq. (3) and Eq. (4), based on current W and A;
    Compute the Laplacian matrix Ψ based on current S as described in Section 3.2;
    while Not converged do
        Update W^{t+1} = W^t − η ∂⌝(W)/∂W based on Eq. (10);
```

The computational complexity of the proposed algorithm is $\mathcal{O}(K^3)$, which is mainly caused by calculating the inverse of the K-by-K matrix for solving Eq. (6), and K is the dictionary size. Since the objective in Eq. (5) is not convex, the proposed algorithm can converge to the local minimum by alternately optimizing the variables, and we have set the iteration number $T = 30$ in the experiments for Algorithm 1.

3.4 Application to Person Re-Id

Given the gallery person image feature vectors $\mathbf{x}_{gi}, i = 1, 2, \ldots, N$, we propose the following steps to re-identify a person represented by the probe feature vector $\mathbf{x}_p$.

1. For each gallery feature $\mathbf{x}_{gi}$, compute its corresponding code coefficients $\mathbf{a}_{gi}$ with respect to the dictionary $\mathbf{D}$ as:

$$\mathbf{a}_{gi} = \underset{\mathbf{a}}{\operatorname{argmin}} ||\mathbf{x}_{gi} - \mathbf{D}\mathbf{a}||_F^2 + \lambda||\mathbf{a}||_F^2. \tag{11}$$

Detailed optimization of Eq(11) refers to [Peng *et al.*, 2016] .

2. Similarly, compute the code coefficients $\mathbf{a}_p$ for the unknown probe feature vector $\mathbf{x}_p$ with respect to $\mathbf{D}$ by Eq. (11).

3. Now compute the Mahalanobis distance between $\mathbf{a}_p$ and each $\mathbf{a}_{gi}$ to form the distance vector $\mathbf{f}$, which can be computed by $\mathbf{f}(i) = ||\mathbf{W}(\mathbf{a}_p - \mathbf{a}_{gi})||_2, \forall i$.

4. Finally, the index of the probe person is obtained as the same index of the minimum value in $\mathbf{f}$.

4 Experiments

In this section, we use three widely used person Re-Id benchmark datasets, namely VIPeR, 3DPES and CUHK03, for performance evaluations. All the datasets contain a set of persons, each of whom has several images captured by different cameras. The following is their brief descriptions:

VIPeR dataset [Gray *et al.*, 2007] contains two views of 632 persons with total 1,264 images. Each pair for a person is captured by different cameras with different viewpoints, poses, and lighting conditions.

3DPES dataset [Baltieri *et al.*, 2011] includes 1,011 images of 192 persons captured from 8 outdoor cameras with significantly different viewpoints. The number of images for each person varies from 2 to 26.

CUHK03 dataset [Li *et al.*, 2014] is one of the largest person Re-Id benchmark datasets recently. It contains 13,164 images of 1,360 identities, and the images were captured by five different pairs of camera views in the campus.

4.1 Experimental Setup

Feature Representation: We have used two kinds of features in our experiments: One is the traditional handcraft features [Chen *et al.*, 2016a], which is extracted both in the whole image and the image subregions. Details about the 7538-D handcrafted feature representations can refer to [Peng *et al.*, 2016; Chen *et al.*, 2016a]. Another is the 2048-D deep residual network features(ResNet152) [He *et al.*, 2016].

Parameter Setting: We empirically set the dictionary size for $\mathbf{D}$ in Eq. (5) as $K = 200$. The parameters τ, α, γ and β are set to 1.0, 0.25, 0.1 and 0.7, respectively. The learning rate starts with $\eta = 0.01$, then at each iteration, we increase η by a factor of 1.2 if the loss function decreased and decrease η by a factor of 0.8 if the loss increased.

Evaluation protocol: Our experiments follow the evaluation protocol in [Peng *et al.*, 2016]. The dataset is separated into the training and test set, where images of the same person can only appear in either set. The test set is further divided into the probe and gallery set, and two sets contain the different images of a same person. In the VIPeR and 3DPES datasets, half of the identities are used as training or test set, while in the CUHK03 dataset, 100 pedestrians are used as the test set, and the rest are used as the training set. We match each probe image with every image in the gallery set, and rank the gallery images according to their distance.

4.2 Experimental Evaluations

As illustrated in Eq. (5), our proposed Re-Id method contains mainly two novel ingredients: 1)we formulate the original triplet loss into the ranking graph Laplacian matrix as shown in Eq. 3, and then learn the dictionary with this ranking metric embedded; 2) an explicit projection matrix $\mathbf{W}$ was simultaneously learned to measure the distance between the projected image features. To reveal how each ingredient contributes to the performance improvement, we implemented the following four variants of the proposed method, and compared them with many representative works in the literature:

Table 1: Experimental results on VIPeR dataset(p=316).

Method	r=1	r=5	r=10	r=20	r=30
[Prates *et al.*, 2016]	35.8	69.1	80.8	89.9	93.8
[Chen *et al.*, 2016b]	38.4	69.2	81.3	90.4	94.1
[Xiong *et al.*, 2014]	39.2	71.8	81.3	92.4	94.9
[Lisanti *et al.*, 2014]	37.0	−−	85.0	93.0	−−
[Liao *et al.*, 2015]	40.0	68.0	80.5	91.1	95.5
[Jose and Fleuret, 2016]	40.2	68.2	80.7	91.1	−−
[Yang *et al.*, 2016]	41.1	71.7	83.2	91.7	−−
[Zhang *et al.*, 2016b]	42.3	71.5	82.9	92.1	−−
[Chen *et al.*, 2015]	43.0	75.8	87.3	94.8	−−
[Ahmed *et al.*, 2015]	45.9	77.5	88.9	95.8	−−
[Matsukawa *et al.*, 2016]	49.7	79.7	88.7	94.5	−−
[Chen *et al.*, 2016a]	53.5	82.6	91.5	96.6	−−
Dict(baseline)	51.9	76.5	84.8	90.8	94.6
DictL	52.6	77.5	85.9	91.8	94.6
DictR	55.0	82.5	90.7	95.8	97.1
Ours(DictRW)	**55.7**	**82.9**	**91.5**	**96.7**	**97.2**

Table 2: Experimental results on 3DPES dataset(p=92).

Method	r=1	r=5	r=10	r=20	r=30
[Koestinger *et al.*, 2012]	34.2	58.7	69.6	80.2	−−
[Mignon and Jurie, 2012]	43.5	71.6	81.8	91.0	−−
[Pedagadi *et al.*, 2013]	45.5	69.2	70.1	82.1	88.2
[Xiong *et al.*, 2014]	54.0	77.7	85.9	92.4	−−
[Paisitkriangkrai *et al.*, 2015]	53.3	76.8	85.7	91.4	−−
[Xiao *et al.*, 2016]	55.2	76.4	84.9	91.9	94.1
[Chen *et al.*, 2016a]	57.3	78.6	86.5	93.6	95.2
Dict(baseline)	54.3	73.2	80.8	89.7	92.4
DictL	55.3	76.3	83.7	91.4	94.2
DictR	59.0	80.7	87.6	94.1	95.8
Ours(DictRW)	**59.6**	**81.4**	**89.7**	**95.0**	**96.1**

Variant 1(denoted as Dict): We just implement the original dictionary learning method as illustrated in Eq. (1). This is our baseline method.

Variant 2(denoted as DictL): We implement the dictionary learning method with the previously used Laplacian matrix embedding, which just used the same identity information, and the matrix is constructed in the following way: $s_{ij} = 1$ only if $l_i = l_j, i \neq j$, otherwise $s_{ij} = 0$.

Variant 3(denoted as DictR): We implement the dictionary learning method as illustrated in Eq. (5), but with the projection matrix $\mathbf{W}$ removed (equal to set $\mathbf{W} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix).

Variant 4(denoted as Ours(DictRW)): This is our proposed final dictionary model as illustrated in Eq. (5).

Table 1, 2 and 3 show the evaluation results on VIPeR, 3DPES and CUHK03 datasets, respectively, using the rank 1, 5, 10, 20, 30 accuracies. Each table includes the recently

Table 3: Experimental results on CUHK03 labeled dataset(p=100).

Method	r=1	r=5	r=10	r=20	r=30
[Zhao *et al.*, 2013]	8.8	24.1	38.3	53.4	−−
[Koestinger *et al.*, 2012]	14.2	48.5	52.6	−−	−−
[Li *et al.*, 2014]	20.6	51.5	66.5	80.0	−−
[Liao *et al.*, 2015]	52.2	82.2	92.1	96.2	−−
[Xiong *et al.*, 2014]	48.2	59.3	66.4	−−	−−
[Wang *et al.*, 2016]	52.2	83.7	89.5	94.3	96.5
[Ahmed *et al.*, 2015]	54.7	86.5	94.0	96.1	**98.0**
[Varior *et al.*, 2016b]	57.3	80.1	88.3	−−	−−
[Paisitkriangkrai *et al.*, 2015]	62.1	89.1	94.3	97.8	−−
[Zhang *et al.*, 2016a]	58.9	85.6	92.5	96.3	−−
[Wu *et al.*, 2016a]	63.2	90.0	92.7	97.6	−−
[Varior *et al.*, 2016a]	68.1	88.1	94.6	−−	−
Dict	64.1	81.6	87.9	92.8	94.1
DictL	65.2	83.5	88.7	93.6	96.0
DictR	70.2	89.3	92.3	96.8	**98.0**
Ours(DictRW)	**71.1**	**91.7**	**94.7**	**98.0**	**98.0**

reported evaluation results. The compared methods include the approaches based on metric learning [Jose and Fleuret, 2016; Chen *et al.*, 2016a], common subspace based methods [Chen *et al.*, 2015; Prates *et al.*, ; Liao *et al.*, 2015; Lisanti *et al.*, 2014; Prates *et al.*, 2016; Zhang *et al.*, 2016b], and the deep learning based methods [Chen *et al.*, 2016b; Varior *et al.*, 2016a; Ahmed *et al.*, 2015; Wang *et al.*, 2016]. Compared with all the aforementioned recently representative works, our model(DictRW) has achieved the top performances on the three datasets, with all the five ranking measurements. We achieve the rank-1 accuracy to 55.7%, 59.6% and 71.1% on VIPeR, 3DPES and CUHK03 datasets, respectively. The evaluation results shown in Table 1,2 and 3 can be summarized as follows,

- Compared with many recently reported representative works, our method(DictRW) outperforms all the compared methods on the three datasets by a margin of 2.5%.
- With the novel ranking Laplacian matrix embedded, the performance accuracies can get up to $3.8\% - 7\%$ improvement compared with the baseline dictionary learning method. Also, comparing methods DictR with DictL, we can clearly see that the ranking information is better than only using the same identity information.
- By explicitly embedding the projection matrix $\mathbf{W}$ into the ranking dictionary objective, another $0.6\% - 0.9\%$ performance improvement can be obtained, compared to the method DictR on the above three datasets.

Since we have used two kinds of features in our experiments (the handcraft and the ResNet152 features), we also did experiments to reveal their performances in Table 4, respectively. We clearly see that combining the traditional handcraft features with the deep learning based features can further improve the Re-Id performance.

Table 4: Experiments comparison with handcraft(HC), ResNet152 features and its combination on CUHK03 datasets, respectively.

Method	r=1	r=5	r=10	r=20	r=30
DictRW(ResNet152)	42.7	73.8	84.7	94.1	96.2
DictRW(HC)	68.3	89.7	91.9	96.1	97.0
DictRW(HC+ResNet152)	**71.1**	**91.7**	**94.7**	**98.0**	**98.0**

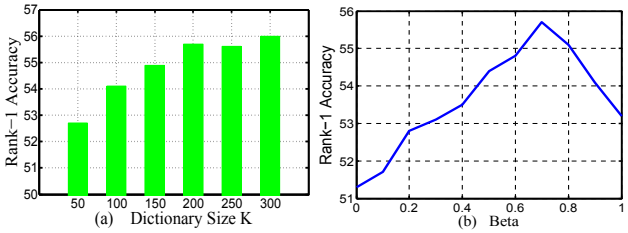

Figure 2: Parameter Analysis: We report how the rank-1 accuracy changes with (a) the dictionary size K, (b) the parameter β.

4.3 Parameter Analysis of the Method

As defined in Eq. (5), there are two important parameters in our proposed method, one is the dictionary size K of $\mathbf{D}$, and the other is the parameter β which controls the balance between the construction loss and the ranking graph Laplacian cost. To investigate the effect of the dictionary size K and the parameter β on the rank-1 accuracy, we conduct experiments on VIPeR dataset, and the rank-1 results are shown in Fig. 2.

Figure 2(a) illustrates the rank-1 accuracy with different dictionary size K from 50 to 300. We can see that firstly as the dictionary size becomes larger, the performance increases continuously. After the dictionary size K larger than 200, the performance becomes almost constant. Although higher performance can also be obtained with larger dictionary size, we choose $K = 200$ in all our experiments, because larger dictionary size requires more training and testing time.

Figure 2(b) shows the rank-1 accuracy with different parameter β from 0 to 1.0. We can clearly see that our proposed method yields the best rank-1 performance when $\beta = 0.7$. Thus, we set β to 0.7 in all our experimental evaluations.

5 Conclusion

In this paper, we present a novel dictionary learning method with ranking Laplacian matrix embedded, for person Re-Id. We formulate the triplet loss into the graph Laplacian form, and then embedded it into the dictionary learning. Overall, our proposed method has made the traditional dictionary learning methods more suitable for the retrieval related tasks. In the future, we will deploy our approach to other tasks.

Acknowledgement

This work was supported by the National Basic Research Program of China (Grant No.2015CB351705), the State Key Program of National Natural Science Foundation of China (Grant No.61332018), and the program of introducing talents of discipline to university (Grant No.B13043).

References

[Ahmed *et al.*, 2015] Ejaz Ahmed, Michael Jones, and Tim K Marks. An improved deep learning architecture for person re-identification. In *CVPR*, pages 3908–3916, 2015.

[Baltieri *et al.*, 2011] Davide Baltieri, Roberto Vezzani, and Rita Cucchiara. 3dpes: 3d people dataset for surveillance and forensics. In *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*, pages 59–64. ACM, 2011.

[Chen *et al.*, 2015] Ying-Cong Chen, Wei-Shi Zheng, and Jianhuang Lai. Mirror representation for modeling view-specific transform in person re-identification. In *IJCAI*, 2015.

[Chen *et al.*, 2016a] Dapeng Chen, Zejian Yuan, Badong Chen, and Nanning Zheng. Similarity learning with spatial constraints for person re-identification. In *CVPR*, pages 1268–1277, 2016.

[Chen *et al.*, 2016b] Shi-Zhe Chen, Chun-Chao Guo, and Jian-Huang Lai. Deep ranking for person re-identification via joint representation learning. *IEEE TIP*, 2016.

[Cheng *et al.*, 2016] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *CVPR*, pages 1335–1344, 2016.

[Gray *et al.*, 2007] Douglas Gray, Shane Brennan, and Hai Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *ECCV*, 2007.

[He *et al.*, 2016] Kaiming He, Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[Jiang *et al.*, 2013] Zhuolin Jiang, Zhe Lin, and Larry S Davis. Label consistent k-svd: Learning a discriminative dictionary for recognition. *IEEE TPAMI*, 35(11):2651–2664, 2013.

[Jose and Fleuret, 2016] Cijo Jose and Franois Fleuret. Scalable metric learning via weighted approximate rank component analysis. In *ECCV*, pages 875–890, 2016.

[Karanam *et al.*, 2015] Srikrishna Karanam, Yang Li, and Richard J Radke. Person re-identification with discriminatively trained viewpoint invariant dictionaries. In *ICCV*, 2015.

[Kodirov *et al.*, 2016] Elyor Kodirov, Tao Xiang, Zhenyong Fu, and Shaogang Gong. Person re-identification by unsupervised\ ell _1 graph learning. In *ECCV*, pages 178–195, 2016.

[Koestinger *et al.*, 2012] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012.

[Lee *et al.*, 2007] Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Y Ng. Efficient sparse coding algorithms. *NIPS*, 2007.

[Li *et al.*, 2014] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, pages 152–159, 2014.

[Liao *et al.*, 2015] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015.

[Lisanti *et al.*, 2014] Giuseppe Lisanti, Iacopo Masi, and Alberto Del Bimbo. Matching people across camera views using kernel canonical correlation analysis. In *ICDSC*, page 10, 2014.

[Matsukawa *et al.*, 2016] Tetsu Matsukawa, Takahiro Okabe, Einoshin Suzuki, and Yoichi Sato. Hierarchical gaussian descriptor for person re-identification. In *CVPR*, 2016.

[Mignon and Jurie, 2012] Alexis Mignon and Frédéric Jurie. Pcca: A new approach for distance learning from sparse pairwise constraints. In *CVPR*, pages 2666–2672, 2012.

[Paisitkriangkrai *et al.*, 2015] Sakrapee Paisitkriangkrai, Chunhua Shen, and Anton van den Hengel. Learning to rank in person re-identification with metric ensembles. In *CVPR*, 2015.

[Pedagadi *et al.*, 2013] Sateesh Pedagadi, James Orwell, Sergio Velastin, and Boghos Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *CVPR*, 2013.

[Peng *et al.*, 2016] Peixi Peng, Tao Xiang, Yaowei Wang, Massimiliano Pontil, Shaogang Gong, Tiejun Huang, and Yonghong Tian. Unsupervised cross-dataset transfer learning for person re-identification. In *CVPR*, pages 1306–1315, 2016.

[Prates *et al.*,] Raphael Prates, Felipe, and William Robson Schwartz. Appearance-based person re-identification by intra-camera discriminative models and rank aggregation. In *Biometrics (ICB), 2015 International Conference on*, pages 65–72.

[Prates *et al.*, 2016] Raphael Prates, Marina Oliveira, and William Robson Schwartz. Kernel partial least squares for person re-identification. In *AVSS*, 2016.

[Shi *et al.*, 2016] Weiwei Shi, Yihong Gong, and Jinjun. Improving cnn performance with min-max objective. In *IJCAI*, 2016.

[Varior *et al.*, 2016a] Rahul Rama Varior, Mrinal Haloi, and Gang Wang. Gated siamese convolutional neural network architecture for human re-identification. In *ECCV*, pages 791–808, 2016.

[Varior *et al.*, 2016b] Rahul Rama Varior, Bing Shuai, Jiwen Lu, Dong Xu, and Gang Wang. A siamese long short-term memory architecture for human re-identification. In *ECCV*, 2016.

[Wang *et al.*, 2016] Faqiang Wang, Wangmeng Zuo, Liang Lin, David Zhang, and Lei Zhang. Joint learning of single-image and cross-image representations for person re-identification. In *CVPR*, pages 1288–1296, 2016.

[Weinberger and Saul, 2009] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *JMLR*, 10(Feb):207–244, 2009.

[Wu *et al.*, 2016a] Lin Wu, Chunhua Shen, and Anton van den Hengel. Deep linear discriminant analysis on fisher networks: A hybrid architecture for person re-identification. *Pattern Recognition*, 2016.

[Wu *et al.*, 2016b] Shangxuan Wu, Ying-Cong Chen, Xiang Li, Jin-Jie You, and Wei-Shi Zheng. An enhanced deep feature representation for person re-identification. In *WACV*, 2016.

[Xiao *et al.*, 2016] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. *CVPR*, 2016.

[Xiong *et al.*, 2014] Fei Xiong, Mengran Gou, Octavia Camps, and Mario Sznaier. Person re-identification using kernel-based metric learning methods. In *ECCV*, pages 1–16, 2014.

[Yang *et al.*, 2016] Yang Yang, Zhen Lei, Shifeng Zhang, Hailin Shi, and Stan Z Li. Metric embedded discriminative vocabulary learning for high-level person representation. In *AAAI*, 2016.

[Zhang and Li, 2010] Qiang Zhang and Baoxin Li. Discriminative k-svd for dictionary learning in face recognition. In *CVPR*, 2010.

[Zhang *et al.*, 2016a] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a discriminative null space for person re-identification. In *CVPR*, pages 1239–1248, 2016.

[Zhang *et al.*, 2016b] Ying Zhang, Baohua Li, Huchuan Lu, Atshushi Irie, and Xiang Ruan. Sample-specific svm learning for person re-identification. In *CVPR*, pages 1278–1287, 2016.

[Zhao *et al.*, 2013] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013.