Pattern Recognition 104 (2020) 107352

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Joint graph regularized dictionary learning and sparse ranking for multi-modal multi-shot person re-identification

Aihua Zheng[a], Hongchao Li[a], Bo Jiang[a,b,*], Wei-Shi Zheng[c], Bin Luo[a]

[a] Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, School of Computer Science and Technology, Anhui University, Hefei, China
[b] Institute of Physical Science and Information Technology, Anhui University, Hefei, China
[c] School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China

ARTICLE INFO

Article history:
Received 22 February 2019
Revised 24 March 2020
Accepted 29 March 2020
Available online 4 April 2020

Keywords:
Person re-identification
Sparse ranking
Joint graph regularization

ABSTRACT

The promising achievement of sparse ranking in image-based recognition gives rise to a number of development on person re-identification (Re-ID) which aims to reconstruct the probe as a linear combination of few atoms/images from an over-complete dictionary/gallery. However, most of the existing sparse ranking based Re-ID methods lack considering the geometric relationships between probe, gallery, and cross-modal images of the same person in multi-shot Re-ID. In this paper, we propose a novel joint graph regularized dictionary learning and sparse ranking method for multi-modal multi-shot person Re-ID. First, we explore the probe-based geometrical structure by enforcing the smoothness between the codings/coefficients, which refers to the multi-shot images from the same person in probe. Second, we explore the gallery-based geometrical structure among gallery images, which encourages the multi-shot images from the same person in the gallery making similar contributions while reconstructing a certain probe image. Third, we explore the cross-modal geometrical structure by enforcing the smoothness between the cross-modal images and thus extend our model for the multi-modal case. Finally, we design an APG based optimization to solve the problem. Comprehensive experiments on benchmark datasets demonstrate the superior performance of the proposed model. The code is available at https://github.com/ttaalle/Lhc.

1. Introduction

Person re-identification (Re-ID), which aims to identify person images from the gallery that shares the same person as the given probe, is an active task driven by the applications of visual surveillance and social security. Despite years of extensive efforts, It still faces various challenges due to the changes in illumination, poses, camera view and occlusions. Multi-shot Re-ID, where each person is recorded by multiple frames, is more realistic in real-life applications with more visual aspects than single-shot Re-ID. Extensive methods have been proposed for multi-shot Re-ID including appearance modeling based methods [1–5], learning-based methods [6–9] and CNN-based methods [10–16]. Moreover, a single visual sensor suffers from the intra-class difference and inter-class similarity due to the poor illumination, clothing changes and background clutters in Re-ID. On one hand, the same person appears distinctly under different cameras due to the clothing changes or illumination changes. On the other hand, the distinct persons may share similar visual appearance due to the similar clothing and low-resolution environment. Therefore, it is essential to integrate the complementary thermal infrared or depth sensors to relieve this intra-class difference and help distinguish this inter-class similarity in single visible RGB modality. We focus on multi-modal multi-shot Re-ID in this paper.

Sparse Ranking (SR) has been successfully applied to extensive image-based applications which give rise to a number of developments on Re-ID [4,17–20]. The basic idea is to reconstruct the probe image as a linear combination of a few atoms/images from an over-complete dictionary gallery. However, most existing sparse ranking based Re-ID methods encode the probe images from the same person independently and therefore fail to take advantage of the relationship among the probe, gallery and cross-modal images of the same person. We argue that these intrinsic geometric structures are crucial in multi-modal multi-shot Re-ID.

In multi-shot Re-ID, first, the multiple images of the same person (as shown in Fig. 1) under the same camera generally have a

* Corresponding author.
E-mail addresses: ahzheng214@ahu.edu.cn (A. Zheng), zeyiabc@163.com (B. Jiang), wszheng@ieee.org (W.-S. Zheng), luobin@ahu.edu.cn (B. Luo).

https://doi.org/10.1016/j.patcog.2020.107352
0031-3203/

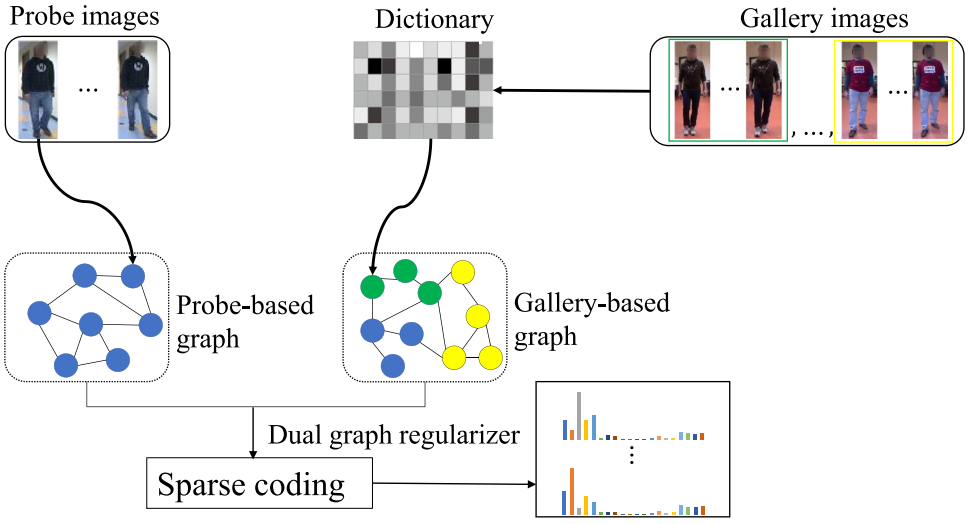

Fig. 1. Probe and gallery-based dual graph regularized sparse ranking for multi-shot Re-ID.

similar visual appearance. Therefore, we assume that the images of the same probe person produce similar codings during sparse ranking. This is known as a manifold assumption which has been commonly employed in many other computer vision problems. Second, the multiple images of a certain person in the gallery also share similar visual appearance, as shown in Fig. 1. Therefore, we further assume that the images of the same person in the gallery make a similar contribution to the reconstruction. This constraint encourages the sparse presentation of probe images more compactly. Third, although some works investigate the thermal infrared or depth information as the complementary modality in person Re-ID [21–25], most of the existing works directly utilize the thermal infrared or depth information as auxiliary information or connective feature for Re-ID while ignoring the cross-modal coherence. Therefore, we further argue to explore the cross-modal geometrical structure by enforcing the cross-modal coherence constraint into the sparse ranking model for multi-modal person Re-ID.

Based on the above discussion, we propose a novel joint graph regularized dictionary learning and sparse ranking for multi-modal multi-shot person re-identification in this paper. Our model incorporates the geometrical structures embedded in probe, gallery, and cross-modalities simultaneously. First, to capture the geometrical structure among the probe images in multi-shot Re-ID, we propose to enforce the probe-based graph regularizer into the sparse ranking model. This regularization encourages the probe images from the same person generating similar codings. Then, to capture the geometrical structure among the gallery images, we further propose to enforce another gallery-based graph regularizer. This regularization encourages the gallery images of the same person making similar contributions while reconstructing a certain probe image. At last, to handle the multi-modal case, we propose to explore the cross-modal geometrical structure among the cross-modal images by imposing the cross-modal coherence constraint, which thus extends our dual graph regularized sparse ranking model for multi-modal person Re-ID. The main contributions can be summaries as:

- We propose to enforce the probe and gallery-based dual graph regularizer into the sparse ranking formulation, by enforcing the smoothness between the images of the same person in both probe and gallery. It can better capture the intrinsic geometric structure for multi-shot Re-ID.
- We propose to introduce a cross-modal regularizer for multi-modal person Re-ID, by enforcing the same person in different modalities making similar contributions while reconstructing probe image from another modality.
- Comprehensive experiments on challenging benchmark datasets with both hand-crafted and deep features validate the superior performance of our model for multi-modal multi-shot person Re-ID.

A preliminary version of this work appeared in [26]. In this work, we further automatically learn the dictionary. Moreover, we extend our model to the multi-modal case and propose a cross-modal consistency to preserve the geometric structure between different modalities for multi-modal person Re-ID. In addition, more extensive experiments have been conducted in both single-modal and multi-modal person Re-ID.

2. Related work

2.1. Sparse ranking based person Re-ID

Recent works show that Sparse Ranking (SR) can effectively resist noise, handle partial occlusion and image corruption in Re-ID. Liu et al. [27] learned two coupled dictionaries for both probe and gallery from both labeled and unlabeled images to transfer the features of the same person from different cameras. Karanam et al. [18] learned a single dictionary for both gallery and probe images to overcome the viewpoint and associated appearance changes and then discriminatively trained the dictionary by enforcing explicit constraints on the associated sparse representations. Zhou et al. [28] proposed a joint learning framework that unifies representative feature learning and discriminative metric learning. Lisanti et al. [4] proposed to learn a discriminative sparse basis expansion of targets in terms of a labeled gallery of known individuals for multi-shot Re-ID. Recently, Jing et al. [17] proposed a semi-coupled dictionary learning method for super-resolution multi-shot Re-ID. In addition, Li et al. [19] designed a semi-coupled projective dictionary learning framework for multi-shot Re-ID with great resolution diversities. Li et al. [20] designed an effective cross-view dictionary learning model to reformulate the data encoding and reconstruction for multi-shot Re-ID. However, most of the existing sparse ranking based Re-ID methods encode the probe images from the same person independently and thus fail to take advantage of their intrinsic geometric structure information.

2.2. Multi-modal person Re-ID

In the early stage, single-modal Re-ID methods devoted to design discriminative features. Representative methods included Symmetry-Driven Accumulation of Local Features (SDALF) [3], Weighted Histograms of Overlapping Stripes (WHOS) [4], Local Maximal Occurrence (LOMO) [7] and so on. Meanwhile, metric learning was designed to bridge the gaps between the low-level feature and high-level human semantic. Roth et al. [9] proposed a Keep It Simple and Straightforward Metric Learning (KISSME) to conduct a hypothesis test on similar/dissimilar pairs. Pedagadi et al. [8] tried to minimize the intra-class divergence and maximize the inter-class diversities via Local Fisher Discriminant Analysis (LFDA). Liao et al. [7] performed the KISSME algorithm in a subspace with a reduced feature dimension by proposing Cross-view Quadratic Discriminant Analysis (XQDA). However, most of the existing metric learning methods tend to learn a projection matrix through training data, which is difficult to overcome the impact of data quality and distribution.

Person Re-ID can also benefit from multi-modal resources such as thermal or depth information. Mogelmose et al. [21] combined the thermal features into RGB appearance modeling and later on further integrated the depth information [22]. Pala et al. [23] improved the accuracy of appearance descriptors by fusing them with anthropometric measures extracted from depth data. John et al. [24] combined RGB-Height histogram and gait feature of depth information. Wu et al. [25] proposed a kernelized implicit feature transfer scheme to estimate the Eigen-depth feature from RGB images implicitly when the depth device was not available. However, they ignore the common expression ability of the multi-modal features in the metric procedure.

3. The proposed model

Given $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in R^{d\times n}$, where n denotes the number of images of a person in probe, with $\mathbf{x}_j \in R^d$, $j = \{1, \ldots, n\}$ denoting the corresponding d-dimensional feature. $\mathbf{D} = [\mathbf{D}^1, \mathbf{D}^2, \ldots, \mathbf{D}^G] \in R^{d\times M_G}$ denotes the total M_G images of G persons in gallery, where $\mathbf{D}^p = [\mathbf{d}_1^p, \mathbf{d}_2^p, \ldots, \mathbf{d}_{g_p}^p] \in R^{d\times g_p}$, $p = \{1, \ldots, G\}$ represents the matrix of g_p basis vectors for the p-th person, and g_p denotes the number of images of the p-th person in gallery. Obviously, $M_G = \sum_{p=1}^{G} g_p$. The basic idea of sparse ranking based Re-ID [4] is to reconstruct a testing probe image $\mathbf{x}_j$ with linear spanned training gallery images of G persons in the gallery:

$$\mathbf{x}_j \approx \sum_{p=1}^{G} \mathbf{D}^p \mathbf{c}_j^p = \mathbf{D}\mathbf{c}_j, \tag{1}$$

where $\mathbf{c}_j^p = [\mathbf{c}_{j,1}^p, \mathbf{c}_{j,2}^p, \ldots, \mathbf{c}_{j,g_p}^p]^T \in R^{g_p\times 1}$ represents the coding coefficients of the p-th person against the probe instance $\mathbf{x}_j$, each column $\mathbf{c}_j = [\mathbf{c}_j^{1^T}, \mathbf{c}_j^{2^T}, \ldots, \mathbf{c}_j^{G^T}]^T \in R^{M_G\times 1}$ is a sparse representation for a probe image. In other words, each probe image can be represented as a sparse linear combination of those basis vectors in the dictionary. A good representation together with dictionary should minimize the empirical loss function and impose a function to measure the sparseness [29,30], which can be represented as:

$$\min_{\mathbf{D},\mathbf{C}} \|\mathbf{X} - \mathbf{D}\mathbf{C}\|_F^2 + \lambda\|\mathbf{C}\|_1, \ s.t. \ \|\mathbf{d}_i\|_2^2 \le \epsilon, \tag{2}$$

where $\|\cdot\|_F$ denotes the Frobenius norm of the matrix, $\|\mathbf{C}\|_1$ is L_1 norm to impose the sparse constraint on related solution of $\mathbf{C}$ and λ is a penalty parameter balancing the sparseness term and the reconstruction term. Each $\mathbf{d}_i$ represents a basis vector in the dictionary and $\|\mathbf{d}_i\|_2^2 \le \epsilon$ is used to avoid the scaling problem of $\mathbf{d}_i$. In this paper ϵ is set as 1.

3.1. Dual graph regularized dictionary learning and sparse representation

Recall that sparse representation tries to find a dictionary gallery and a sparse coefficient matrix whose product can best approximate the original probe matrix. The column vectors of $\mathbf{D}$ can be regarded as the basis vectors and each column of $\mathbf{C}$ is the new representation of each probe image in this new gallery space. Graph regularized sparse coding performs great superiority in image-based applications [31–33]. In this paper, we propose a novel dual graph regularized dictionary learning and sparse representation framework to consider the intrinsic geometric structures among probe and gallery images for multi-shot person Re-ID.

3.1.1. Probe-based graph regularization

In multi-shot Re-ID, we argue that the feature vectors derived from the multiple images of the same person tend to have similar geometric distribution. We propose to enforce this constraint on probe (test) images to obtain a kind of consistent sparse representations for probe images by incorporating the geometric structure of the data. As shown in Fig. 1, we first enforce a probe-based graph regularizer over the coding coefficients to exploit the intrinsic geometric distribution among the probe images:

$$J_p = \sum_{i=1}^{n}\sum_{j=1}^{n} \|\mathbf{c}_i - \mathbf{c}_j\|_2^2 \mathbf{S}_{i,j}, \tag{3}$$

where $\{\mathbf{c}_i, \mathbf{c}_j\} \in R^{M_G\times 1}$ are the coding coefficients of probe images $\mathbf{x}_i$ and $\mathbf{x}_j$ from the same person over gallery dictionary $\mathbf{D}$ respectively. The probe-based graph matrix $\mathbf{S} \in R^{n\times n}$ is defined as:

$$\mathbf{S}_{i,j} = exp\Big(\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2\sigma_1^2}\Big), \tag{4}$$

where σ_1 is a scaling parameter and fixed as 0.1 in this paper. The probe-based regularizer in Eq. (3) encourages the probe images from the same person with higher similarity to generate closer coding coefficients during reconstruction.

3.1.2. Gallery-based graph regularization

We further argue that the multiple images of the same person in the gallery fall into similar geometry in multi-shot Re-ID, which in turn means for each pair of gallery images of the same person, they should contribute similarly while reconstructing each probe image since they are generally with similar features. Therefore, as shown in Fig. 1, we further propose to enforce this constraint on gallery data (or dictionary) to exploit their intrinsic geometry. Specifically, we enforce a gallery-based graph regularizer over the codings:

$$J_g = \sum_{p=1}^{G}\sum_{k=1}^{g_p}\sum_{l=1}^{g_p} (\mathbf{c}_k^p - \mathbf{c}_l^p)^2 \mathbf{B}_{k,l}^p, \tag{5}$$

where $\mathbf{c}_k^p = [\mathbf{c}_{1,k}^p, \mathbf{c}_{2,k}^p, \ldots, \mathbf{c}_{n,k}^p]^T \in R^{g_p\times 1}$ represents the coding coefficients to reconstruct $\mathbf{x}_j$ for the p-th person. The similarity matrix $\mathbf{B} = \mathrm{diag}\{\mathbf{B}^1, \mathbf{B}^2, \ldots, \mathbf{B}^G\} \in R^{M_G\times M_G}$, and each element $\mathbf{B}^p \in R^{g_p\times g_p}$ is defined as:

$$\mathbf{B}_{k,l}^p = exp\Big(\frac{-\|\mathbf{d}_k^p - \mathbf{d}_l^p\|_2^2}{2\sigma_2^2}\Big), \tag{6}$$

where σ_2 is a scaling parameter and fixed as 0.1 in this paper. $\mathbf{D} = [\mathbf{D}^1, \mathbf{D}^2, \ldots, \mathbf{D}^G] \in R^{d\times M_G}$ denotes the total M_G images of G persons in gallery, where $\mathbf{D}^p = [\mathbf{d}_1^p, \mathbf{d}_2^p, \ldots, \mathbf{d}_{g_p}^p] \in R^{d\times g_p}$, $p = \{1, \ldots, G\}$ represents the matrix of g_p basis feature vectors for the pth person. The gallery-based regularizer in Eq. (5) encourages the higher similarity between the gallery images from the same person, the closer contribution to the reconstruction.

Therefore, our dual graph regularized dictionary learning and sparse representation model can be summarised as:

$$J_d = \min_{\mathbf{D},\mathbf{C}} \|\mathbf{X} - \mathbf{D}\mathbf{C}\|_F^2 + \lambda\|\mathbf{C}\|_1 + \frac{1}{2}\beta J_p + \frac{1}{2}\gamma J_g,\ s.t.\ \|\mathbf{d}_i\|_2^2 \le \epsilon, \quad (7)$$

where β and γ denote the balance parameters controlling the contribution of the probe-based graph and gallery-based graph respectively.

3.2. Joint model for multi-modal problem

To capture the cross-modal coherence while handling the multi-modal problem, we further propose a joint model to consider the intrinsic geometric structures among cross-modal images of the same person and thus extend our dual graph regularized sparse ranking model in Eq. (7) to multi-modal person Re-ID.

Given $\mathbf{X}^M = [\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(M)}]$, $\mathbf{D}^M = [\mathbf{D}^{(1)}, \mathbf{D}^{(2)}, \ldots, \mathbf{D}^{(M)}]$, where $\mathbf{X}^{(v)} \subseteq \mathbf{X}^M$, $\mathbf{D}^{(v)} \subseteq \mathbf{D}^M$, $\mathbf{X}^{(v)}$ and $\mathbf{D}^{(v)}$ denote the query and corresponding dictionary of the v-th modality respectively. In order to capture the geometric structures among the multispectral data, we further propose to enforce the cross-modal coherence among M modalities by introducing a cross-modal regularizer as:

$$J_m = \sum_{v=1}^{M}\sum_{u=1}^{M} \|(\mathbf{C}^{(u)} - \mathbf{C}^{(v)})\|_F^2, \quad (8)$$

which encourages the images in different modalities to have similar codings in the relevant dictionaries. Inspired by the method on heterogeneous camera network [34], our aim here is to learn a consistent feature representation across different modalities.

Therefore, the final objective function can be summarised as:

$$J = \min_{\mathbf{D}^{(v)},\mathbf{C}^{(v)}} \sum_{v=1}^{M} \Big\{ \|\mathbf{X}^{(v)} - \mathbf{D}^{(v)}\mathbf{C}^{(v)}\|_F^2 + \lambda\|\mathbf{C}^{(v)}\|_1 + \frac{1}{2}\beta {J_p}^{(v)} + \frac{1}{2}\gamma {J_g}^{(v)} \Big\} + \mu J_m,\ s.t.\ \|\mathbf{d}_i^{(v)}\|_2^2 \le \epsilon, \quad (9)$$

where β, γ and μ are the balance parameters controlling the contribution of the probe-based graph, gallery-based graph and cross-modal regularizer respectively. ${J_p}^{(v)}$ and ${J_g}^{(v)}$ denote the probe-based graph and gallery-based graph from the vth modality. With simple algebra, Eq. (9) can be rewritten as:

$$J = \min_{\mathbf{D}^{(v)},\mathbf{C}^{(v)}} \sum_{v=1}^{M} \Big\{ \|\mathbf{X}^{(v)} - \mathbf{D}^{(v)}\mathbf{C}^{(v)}\|_F^2 + \lambda\|\mathbf{C}^{(v)}\|_1 + \beta tr(\mathbf{C}^{(v)}{\mathbf{L}_1}^{(v)}{\mathbf{C}^{(v)}}^T) + \gamma tr({\mathbf{C}^{(v)}}^T{\mathbf{L}_2}^{(v)}\mathbf{C}^{(v)}) \Big\} + \mu\sum_{v=1}^{M}\sum_{u=1}^{M}\|\mathbf{C}^{(u)} - \mathbf{C}^{(v)}\|_F^2,\ s.t.\ \|\mathbf{d}_i^{(v)}\|_2^2 \le \epsilon, \quad (10)$$

where $\mathbf{C}^{(v)} = [\mathbf{c}_1{}^{(v)}, \mathbf{c}_2{}^{(v)}, \ldots, \mathbf{c}_n{}^{(v)}] \in R^{M_G \times n}$, ${\mathbf{L}_1}^{(v)} = {\mathbf{D}_\mathbf{S}}^{(v)} - \mathbf{S}^{(v)}$ and ${\mathbf{L}_2}^{(v)} = {\mathbf{D}_\mathbf{B}}^{(v)} - \mathbf{B}^{(v)}$ denote the probe and gallery-based graph Laplacian matrix respectively for each modality, ${\mathbf{D}_\mathbf{S}}^{(v)} = \mathrm{diag}\{\sum_j {\mathbf{S}_{1,j}}^{(v)}, \sum_j {\mathbf{S}_{2,j}}^{(v)}, \cdots\}$ and ${\mathbf{D}_\mathbf{B}}^{(v)} = \mathrm{diag}\{\sum_j {\mathbf{B}_{1,j}}^{(v)}, \sum_j {\mathbf{B}_{2,j}}^{(v)}, \cdots\}$ indicate the degree matrix of $\mathbf{S}^{(v)}$ and $\mathbf{B}^{(v)}$ respectively, and $\mathrm{diag}\{\cdots\}$ indicates the diagonal operation, $tr\{\cdots\}$ indicates the trace of a matrix.

In summary, the pipeline of the proposed joint graph regularized sparse ranking (JGRSR) model is illustrated in Fig. 2 as for RGB-D multi-modal problem. The proposed JGRSR will optimize the coding coefficients enforced by both the dual graph regularizer and the cross-modal regularizer during each iteration, which will further propagate to the dictionary learning phase in each modality. The optimized codings investigate the geometric structure among all probe, gallery and cross-modal spaces in a unified framework. The final rankings of the certain probe against given gallery are achieved by the accumulation, which will be detailed in Section 5 based on the coding coefficients and redistribution errors.

4. Optimization

As summarized in Algorithm 1. To solve Eq. (10), our aim is to solve the coefficient matrix $\mathbf{C}$ and the dictionary matrix $\mathbf{D}$. We first fix the initial dictionary matrix as the gallery feature matrix to gain the initial coefficient matrix. Then we fix the coefficient matrix to obtain the new dictionary matrix. The proposed model can be optimized by iteratively learning the coefficient matrix and the dictionary matrix until convergence based on Accelerated Proximal Gradient (APG) [35] algorithm. We will elaborate on the process in the following two subsections.

Algorithm 1 Optimization Procedure to Eq. (10).

Input: probe matrix $\mathbf{X}^{(v)}$, initial gallery dictionary matrix $\mathbf{D}^{(v)}$, Laplacian matrix ${\mathbf{L}_1}^{(v)}, {\mathbf{L}_2}^{(v)}$, parameters λ, β, γ and μ, v = 1, ..., M;
Set $\mathbf{C}^{(v)^0} = \mathbf{C}^{(v)^1} = \mathbf{0}$, $\mathbf{D}^{(v)^0} = \mathbf{D}^{(v)^1} = \mathbf{D}^{(v)}$, $\xi = 1 \times 10^3$, $\theta = 1 \times 10^{-4}$, $\rho_0 = \rho_1 = 1$, maxIter = 30;
Iterate: for t = 1, 2, ..., maxIter
Sparse Representation: for k = 1, 2, ..., maxIter
1: While not converged **do**
2: Update $\mathbf{K}^{(v)^{k+1}}$ by $\mathbf{K}^{(v)^{k+1}} = \mathbf{C}^{(v)^k} + \frac{\rho_{k-1}-1}{\rho_k}(\mathbf{C}^{(v)^k} - \mathbf{C}^{(v)^{k-1}})$;
3: Update $\mathbf{C}^{(v)^{k+1}}$ by Eq. (18);
4: Update $\rho_{k+1} = \frac{1+\sqrt{1+4\rho_k^2}}{2}$;
5: Update k by $k = k + 1$;
6: The convergence condition: the maximum element change of $\mathbf{C}^{(v)^k}$ between two consecutive iterations is less than θ.
7: **end While**
Dictionary Update: for k = 1, 2, ..., maxIter
1: While not converged **do**
2: Update $\mathbf{P}^{(v)^{k+1}}$ by $\mathbf{P}^{(v)^{k+1}} = \mathbf{D}^{(v)^k} + \frac{\rho_{k-1}-1}{\rho_k}(\mathbf{D}^{(v)^k} - \mathbf{D}^{(v)^{k-1}})$;
3: Update $\mathbf{D}^{(v)^{k+1}}$ by Eq. (22);
4: Update $\rho_{k+1} = \frac{1+\sqrt{1+4\rho_k^2}}{2}$;
5: Update k by $k = k + 1$;
6: The convergence condition: the maximum element change of $\mathbf{D}^{(v)^k}$ between two consecutive iterations is less than θ.
7: **end While**
8: $\mathbf{D}^{(v)}$ is normalized.
Output: $\mathbf{C}^{(v)}, \mathbf{D}^{(v)}$

In Algorithm 1, ξ and θ are empirically set as in APG (Accelerated Proximal Gradient) [35]. Specifically, ξ is the Lipschitz constant denoting the step size in the gradient descent algorithm. θ is an error threshold for convergence. ρ_k is a positive and incremental number with $\rho_0 = \rho_1 = 1$, which can gradually reduce the step size and speed up the convergence rate of the algorithm.

4.1. Learning the graph regularized sparse codes

In this section, we discuss how to learn the graph regularized sparse codes by all dictionary $\mathbf{D}^{(v)}, v = \{1, \ldots, M\}$. The Eq. (10) be-

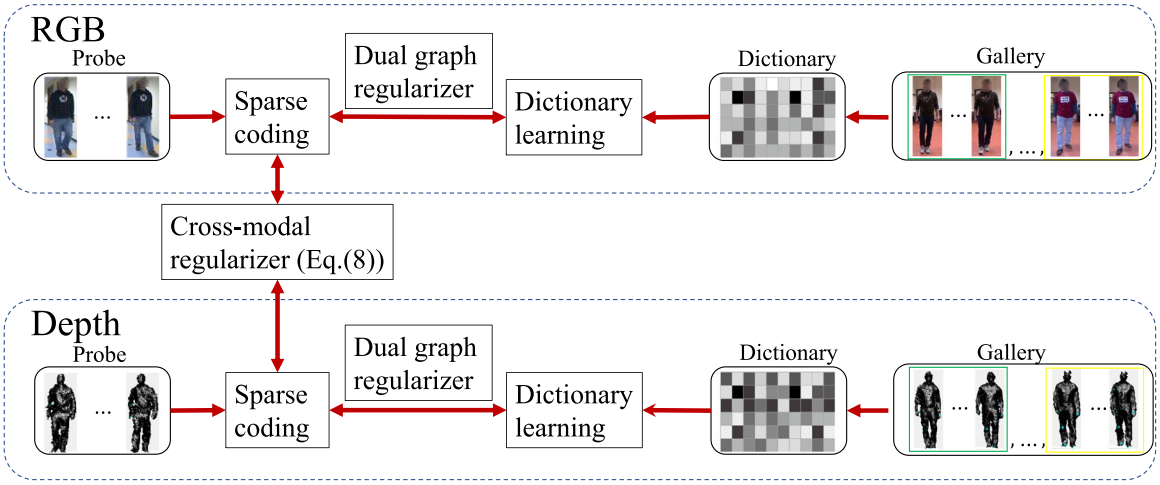

Fig. 2. Pipeline of the proposed joint graph regularized sparse ranking (JGRSR) model for multi-modal Re-ID in the RGB-D case.

comes:

$$\min_{\mathbf{C}^{(v)}} \sum_{v=1}^{M} \Big\{ \|\mathbf{X}^{(v)} - \mathbf{D}^{(v)}\mathbf{C}^{(v)}\|_F^2 + \lambda\|\mathbf{C}^{(v)}\|_1 + \beta tr(\mathbf{C}^{(v)}\mathbf{L}_1^{(v)}\mathbf{C}^{(v)^T}) + \gamma tr(\mathbf{C}^{(v)^T}\mathbf{L}_2^{(v)}\mathbf{C}^{(v)}) \Big\} + \mu \sum_{v=1}^{M}\sum_{u=1}^{M} \|\mathbf{C}^{(u)} - \mathbf{C}^{(v)}\|_F^2. \tag{11}$$

Eq. (11) with L_1-regularization is nondifferentiable when $\mathbf{c}_j^{(v)}$ contains values of 0, the standard unconstrained optimization methods can not be applied. For the sake of more meaningful interpretation [35], we further consider the non-negativeness of the codings $\mathbf{C}^{(v)}$, Eq. (11) can be written as:

$$\min_{\mathbf{C}^{(v)}} \sum_{v=1}^{M} \Big\{ \|\mathbf{X}^{(v)} - \mathbf{D}^{(v)}\mathbf{C}^{(v)}\|_F^2 + \lambda\mathbf{1}^T\mathbf{C}^{(v)}\mathbf{1} + \beta tr(\mathbf{C}^{(v)}\mathbf{L}_1^{(v)}\mathbf{C}^{(v)^T}) + \gamma tr(\mathbf{C}^{(v)^T}\mathbf{L}_2^{(v)}\mathbf{C}^{(v)}) \Big\} + \mu \sum_{v=1}^{M}\sum_{u=1}^{M} \|\mathbf{C}^{(u)} - \mathbf{C}^{(v)}\|_F^2, \ s.t. \ \mathbf{C}^{(v)} \geq \mathbf{0}, \tag{12}$$

where $\mathbf{1}$ denotes the vector that its all elements are 1. To solve Eq. (12), we convert it to an unconstrained form as:

$$\min_{\mathbf{C}^{(v)}} \sum_{v=1}^{M} \Big\{ \|\mathbf{X}^{(v)} - \mathbf{D}^{(v)}\mathbf{C}^{(v)}\|_F^2 + \lambda\mathbf{1}^T\mathbf{C}^{(v)}\mathbf{1} + \beta tr(\mathbf{C}^{(v)}\mathbf{L}_1^{(v)}\mathbf{C}^{(v)^T}) + \gamma tr(\mathbf{C}^{(v)^T}\mathbf{L}_2^{(v)}\mathbf{C}^{(v)}) \Big\} + \mu \sum_{v=1}^{M}\sum_{u=1}^{M} \|\mathbf{C}^{(u)} - \mathbf{C}^{(v)}\|_F^2 + \psi(\mathbf{C}^{(v)}), \tag{13}$$

where

$$\psi(\mathbf{C}_{i,j}^{(v)}) = \begin{cases} 0, & if \ \mathbf{C}_{i,j}^{(v)} \geq 0, \\ \infty, & otherwise \end{cases} \tag{14}$$

We utilize the accelerated proximal gradient (APG) [35] approach to optimize. We denote:

$$F(\mathbf{C}^{(v)}) = \min_{\mathbf{C}^{(v)}} \sum_{v=1}^{M} \Big\{ \|\mathbf{X}^{(v)} - \mathbf{D}^{(v)}\mathbf{C}^{(v)}\|_F^2 + \lambda\mathbf{1}^T\mathbf{C}^{(v)}\mathbf{1} + \beta tr(\mathbf{C}^{v}\mathbf{L}_1^{v}\mathbf{C}^{v^T}) + \gamma tr(\mathbf{C}^{(v)^T}\mathbf{L}_2^{(v)}\mathbf{C}^{(v)}) \Big\} + \mu \sum_{v=1}^{M}\sum_{u=1}^{M} \|\mathbf{C}^{(u)} - \mathbf{C}^{(v)}\|_F^2,$$

$$Q(\mathbf{C}^{(v)}) = \psi(\mathbf{C}^{(v)}). \tag{15}$$

Obviously, $F(\mathbf{C}^{(v)})$ and $Q(\mathbf{C}^{(v)})$ are a differentiable convex function and a non-smooth convex function, respectively. Therefore, according to the APG method, we obtain:

$$\mathbf{C}^{(v)^{k+1}} = \min_{\mathbf{C}^{(v)}} \frac{\xi}{2}\|\mathbf{C}^{(v)} - \mathbf{K}^{(v)^{k+1}} + \nabla F(\mathbf{K}^{(v)^{k+1}})/\xi\|_F^2 + Q(\mathbf{C}^{(v)}), \tag{16}$$

where k indicates the current iteration time, and ξ is the Lipschitz constant.

$\nabla F(\mathbf{K}^{(v)^{k+1}})$ is:

$$\nabla F(\mathbf{K}^{(v)^{k+1}}) = 2(\mathbf{D}^{(v)^T}\mathbf{D}^{(v)}\mathbf{K}^{(v)^{k+1}} - \mathbf{D}^{(v)^T}\mathbf{X}^{(v)}) + \lambda\mathbf{E} + 2\beta\mathbf{K}^{(v)^{k+1}}\mathbf{L}_1^{(v)} + 2\gamma\mathbf{L}_2^{(v)}\mathbf{K}^{(v)^{k+1}} + 2\mu\sum_{u=1}^{M}(\mathbf{K}^{(u)^{k+1}} - \mathbf{K}^{(v)^{k+1}}), \tag{17}$$

where $\mathbf{K}^{(v)^{k+1}} = \mathbf{C}^{(v)^k} + \frac{\rho_{k-1}-1}{\rho_k}(\mathbf{C}^{(v)^k} - \mathbf{C}^{(v)^{k-1}})$, ρ_k is a positive sequence with $\rho_0 = \rho_1 = 1$. Eq. (16) can be solved by:

$$\mathbf{C}^{(v)^{k+1}} = \max(0, \mathbf{K}^{(v)^{k+1}} - \nabla F(\mathbf{K}^{(v)^{k+1}})/\xi). \tag{18}$$

4.2. Learning dictionary

In this section, we describe the method of learning the dictionary $\mathbf{D}^{(v)}$, while fixing the coefficient matrix $\mathbf{C}^{(v)}$. The problem becomes a least squares problem with quadratic constraints.

$$F(\mathbf{D}^{(v)}) = \min_{\mathbf{D}^{(v)}} \|\mathbf{X}^{(v)} - \mathbf{D}^{(v)}\mathbf{C}^{(v)}\|_F^2, \ s.t. \ \|\mathbf{d}_i^{(v)}\|_2^2 \leq \epsilon. \tag{19}$$

In the process of dictionary learning, we continue to use the accelerated proximal gradient (APG) [35] approach. We denote:

$$\mathbf{D}^{(v)^{k+1}} = \min_{\mathbf{D}^{(v)}} \frac{\xi}{2}\|\mathbf{D}^{(v)} - \mathbf{P}^{(v)^{k+1}} + \nabla F(\mathbf{P}^{(v)^{k+1}})/\xi\|_F^2, \tag{20}$$

where k indicates the current iteration time, and ξ is the Lipschitz constant.

$\nabla F(\mathbf{P}^{(v)^{k+1}})$ is:

$$\nabla F(\mathbf{P}^{(v)^{k+1}}) = 2(\mathbf{P}^{(v)^{k+1}}\mathbf{C}^{(v)}\mathbf{C}^{(v)^T} - \mathbf{X}^{(v)}\mathbf{C}^{(v)^T}), \tag{21}$$

where $\mathbf{P}^{(v)^{k+1}} = \mathbf{D}^{(v)^k} + \frac{\rho_{k-1}-1}{\rho_k}(\mathbf{D}^{(v)^k} - \mathbf{D}^{(v)^{k-1}})$, ρ_k is a positive sequence with $\rho_0 = \rho_1 = 1$. Eq. (20) can be solved by:

$$\mathbf{D}^{(v)^{k+1}} = \mathbf{P}^{(v)^{k+1}} - \nabla F(\mathbf{P}^{(v)^{k+1}})/\xi. \tag{22}$$

The specific process can be referred to as **Alg. 1**. It's worth noting that we end up with the dictionary as normalized.

5. Ranking implementation for multi-shot Re-ID

The basic idea of sparse ranking based Re-ID is to encode a testing probe image $\mathbf{x}_j^{(v)}$ with linear spanned training dictionary gallery $\mathbf{D}^{(v)}$ in the v-th modality. $\mathbf{c}_j^{(v)}$ is a sparse code for probe image $\mathbf{x}_j^{(v)}$ in dictionary $\mathbf{D}^{(v)}$. Each item $\mathbf{c}_{i,j}^{(v)}$ represents the contribution of each image in dictionary space to encode the probe image $\mathbf{x}_j^{(v)}$. The larger the contribution, the more likely they are the same person.

Due to the sparsity of the coding coefficients, the majority of which collapse to zero after a few higher coding coefficients. Therefore, we can not support ranking for all individuals in the gallery. Lisanti et al. [4] propose a soft and hard re-weighting technique to deal with this issue. To avoid re-running the whole sparse ranking algorithm after each re-weighting, we provide an Error Distribution measurement. First, we obtain the normalized coding error $e_j^{(v)}$ for current probe $\mathbf{x}_j^{(v)}$ using:

$$e_j^{(v)} = \frac{\|\mathbf{x}_j^{(v)} - \mathbf{D}^{(v)}\mathbf{c}_j^{(v)}\|_2}{\|\mathbf{x}_j^{(v)}\|_2}. \tag{23}$$

Then, we re-distribute the coding error into the dictionary gallery individuals according to their similarity to the current probe image $\mathbf{x}_j^{(v)}$:

$$\mathbf{W}_{j,k}^{p\ (v)} = \frac{1/dis(\mathbf{x}_j^{(v)}, \mathbf{d}_k^{p(v)})}{\sum_{p=1}^{G}\sum_{k=1}^{g_p}(1/dis(\mathbf{x}_j^{(v)}, \mathbf{d}_k^{p(v)}))}, \ k \in \{1, \ldots, g_p\}, \tag{24}$$

where $\mathbf{d}_k^{p(v)}$ represents the feature of the kth image from the pth person in dictionary gallery $\mathbf{D}^v$ under the vth modality, $dis(\mathbf{x}_j^{(v)}, \mathbf{d}_k^{p(v)})$ denotes the Euclidean distance between probe $\mathbf{x}_j^{(v)}$ and each element $\mathbf{d}_k^{p(v)}$ in gallery. $\mathbf{W}_{j,k}^{p\ (v)}$ indicates the similarity/weight of $\mathbf{d}_k^{p(v)}$ relative to $\mathbf{x}_j^{(v)}$.

In this paper, we employ the coding coefficients as the similarity measures and define the accumulated coding coefficients from the pth person as a part of the ranking value of the probe person with n images against the pth person. Moreover, we use the coding residues to assign the pth category whose coding coefficients are all zeros with the ranking score. Therefore, the final ranking score of the probe person with n images against the pth person in the gallery under the vth modality is defined as follows:

$$\mathbf{r}^{p(v)} = \sum_{j=1}^{n}\sum_{k=1}^{g_p}(\mathbf{c}_{j,k}^{p\ (v)} + \mathbf{W}_{j,k}^{p\ (v)} e_j^{(v)}), \ \ p \in \{1, \ldots, G\}. \tag{25}$$

The higher similarity of $\mathbf{d}_k^{p(v)}$ relative to $\mathbf{x}_j^{(v)}$, the higher value distributed to $\mathbf{c}_{j,k}^{p\ (v)}$. Since $e_j^{(v)}$ is usually a small value, the value distributed to $\mathbf{c}_{j,k}^{p\ (v)}$ is also very small, which will not change the ranks of the non-zero coding coefficients but will reorder the zero coding coefficients according to Euclidean distance.

Our final decision rule in the vth modality is:

$$class(\mathbf{X}^{(v)}) = \arg\max_p \mathbf{r}^{p(v)}. \tag{26}$$

The rankings from different modalities can be fused as:

$$\mathbf{r}^p{}_{fusion} = \sum_{v=1}^{M}\eta_v \mathbf{r}^{p(v)}, \sum_{v=1}^{M}\eta_v = 1, \tag{27}$$

where η_v balances the contributions of different modalities.

6. Experimental results

We evaluate our method on eight multi-shot person Re-ID benchmark datasets including: (1) three single-modal datasets, i-LIDS [36], CAVIAR4REID [2] and MARS [37]; (2) three multi-modal RGB-D datasets, PAVIS [38], BIWI [39] and IAS-Lab [40]; (3) two RGB transferred depth datasets, transferred 3DPeS [2] and transferred CAVIAR4REID [2]. We use the standard measurement named Cumulated Match Characteristic (CMC) curve [41] to figure out the matching results, where the matching rate at rank-n indicates the percentage of correct matchings in top n candidates according to the learned ranking function Eq. (26). Our previous non-negative dual graph regularized sparse ranking [26] is referred to as NNDGSR in the following content.

6.1. Evaluation on single-modal benchmarks

The single modal representation, referred as *JGRSR_single*, can be regarded as the special case of our model in Eq. (9) with $M = 1$. We evaluate the proposed method on both hand-crafted and deep features. Followed by the protocol in [4], we use WHOS [4] as hand-crafted feature. As for deep feature, we generate APR [11] features pre-trained on large Re-ID dataset Market-1501 [41] for i-LIDS [36] and CAVIAR4REID [2], while utilize IDE feature [37] for MARS [37] as provided.

6.1.1. Comparison on i-LIDS

i-LIDS [36] dataset is composed by 479 images of 119 people, which was captured at an airport arrival hall under two non-overlapping camera views with almost two images each person per camera views. This dataset consists of challenging scenarios with heavy occlusions and pose variance.

Evaluation results on i-LIDS [36] dataset are shown in Table 1. From which we can see, our approach significantly outperforms the state-of-the-art. The Rank-1 accuracies of our approach achieve 84.9% and 79.6% on hand-crafted and deep features respectively, which improve 22% and 2.4% than the second best method ISR [4]. It is worth noting that: (1) The limited number of samples in i-LIDS [36] compromises the performance of deep learning. (2) Our NNDGSR [26] significantly improves the ranking results on both hand-crafted and deep features. (3) By introducing the dictionary learning to our NNDGSR [26], our *JGRSR_single* can further boost the performance.

6.1.2. Comparison on CAVIAR4REID

CAVIAR4REID [2] dataset contains 72 unique individuals with averagely 11.2 images per person extracted from two non-overlapping cameras in a shopping center, 50 of which with both the camera views and the remaining 22 with only one camera view. The images for each camera view have variations with respect to resolution changes, light conditions, occlusions and pose changes.

Evaluation results on CAVIAR4REID [2] are shown in Table 1. We evaluate our method with APR [11] deep features in the same manner as on i-LIDS [36] and adopt the same experimental protocols as ISR [4] by 50 random trials. Clearly, our approach significantly outperforms the state-of-the-art algorithms on both hand-crafted and deep features. Specifically, the Rank-1 accuracies with $N = 5$ achieve 93.7% and 89.5% on hand-crafted features and deep features respectively. Together with the results on i-LIDS [36] and CAVIAR4REID [2], it suggests that the proposed method achieves impressive performance on small size datasets.

6.1.3. Comparison on MARS

MARS [37] dataset is the largest and newly collected dataset for video-based Re-ID. It is collected from six near-synchronized cameras in the campus of Tsinghua University. MARS [37] consists of 1261 pedestrians each of which appears at least two cameras. It contains 625 identities with 8298 tracklets for training and 636 identities with 12,180 tracklets for testing. Different from the other

Table 1
Comparison results at Rank-1 on i-LIDS and CAVIAR4REID (in %).

Features	Methods	i-LIDS	CAVIAR4REID		References
		N=2	N=3	N=5	
Hand-craft features	HPE [42]	18.5	-	-	ICPR2010
	AHPE [43]	32	7.5	7.5	PRL2012
	SCR [44]	36	-	-	ICAVSS2010
	MRCG [45]	46	-	-	ICAVSS2011
	SDALF [3]	39	8.5	8.3	CVPR2010
	CPS [2]	44	13	17.5	BMVC2011
	COSMATI [46]	44	-	-	ECCV2012
	WHOS + ISR [4]	62.9	75.1	90.1	PAMI2015
	WHOS [4] + **NNDGSR** [26]	**84.3**	**78.7**	**93.2**	**Ours**
	WHOS [4] + **JGRSR_single**	**84.9**	**79.6**	**93.7**	**Ours**
	APR [11] + EU [11]	67.7	44.3	53.8	PR2019
Deep features	APR [11] + ISR [4]	77.2	65.7	80.7	PR2019+PAMI2015
	APR [11] + **NNDGSR** [26]	**78.4**	**70.4**	**89.0**	**Ours**
	APR [11] + **JGRSR_single**	**79.6**	**71.5**	**89.5**	**Ours**

Table 2
Comparison with baselines on MARS dataset (in %).

Features	Methods	Rank-1	Rank-5	Rank-20	References
Hand-craft Features	HOG3D [47]+ KISSME [9]	2.6	6.4	12.4	BMVC2010+CVPR2012
	GEI [48]+ KISSME [9]	1.2	2.8	7.4	PAMI2005+CVPR2012
	HistLBP [49]+ XQDA [7]	18.6	33.0	45.9	ECCV2014+CVPR2015
	BoW [41]+KISSME [9]	30.6	46.2	59.2	ICCV2015+CVPR2012
	LOMO+ XQDA [7]	30.7	46.6	60.9	CVPR2015
Deep features	ASTPN [50]	44	70	81	ICCV2017
	LCAR [15]	55.5	70.2	80.2	TCSVT2018
	SFT [16]	70.6	90	97.6	CVPR2017
	MSCAN [10]	71.8	86.6	93.1	CVPR2017
	EUG [51]	62.6	74.9	82.6	CVPR2018
	BUC [52]	61.1	75.1	-	AAAI2019
	IDE+EU [37]	58.7	77.1	86.8	ECCV2016
	IDE [37] + ISR [4]	63	77.1	85.6	ECCV2016+PAMI2015
	IDE [37] + **NNDGSR** [26]	**72.50**	**88.0**	**93.30**	**Ours**
	IDE [37] + **JGRSR_single**	**75.76**	**92.68**	**97.27**	**Ours**

datasets, it also consists of 23,380 *junk* bounding boxes and 147,743 distractors bounding boxes in the testing samples.

In this dataset, the query trackelets are automatically generated from the testing samples. For each query trackelet, we construct two feature vectors via max pooling and average pooling respectively on the provide deep features, IDE [37]. For the remaining testing trackelets, since there are multiple trackelets for each person under a certain camera, we conduct the max pooling for each trackelet to construct the multiple feature vectors followed by the state-of-the-art methods on MARS [37]. Note that, our method doesn't require any training therefore only the testing set containing the query set is utilized. The performance of our method against different metrics is reported in Table 2. As we can see: (1) CNN based methods generally outperform the traditional metric learning methods on hand-crafted features. (2) The sparse ranking based method outperforms on the powerful deep features comparing with the traditional Euclidian distance. (3) By introducing the non-negative dual graph regularized into the sparse ranking framework, our NNDGSR [26] can significantly boost the performance by increasing 12.76% at Rank-1 accuracy. (4) Introducing the dictionary learning can better improve the performance on the large dataset.

6.2. Evaluation on multi-modal benchmarks

The multi-modal Re-ID in Eq. (9) is referred as *JGRSR_multi* in the following content. We evaluate the multi-modal Re-ID on three RGB-D datasets PAVIS [38], BIWI [39] and IAS-Lab [40]. Followed by the protocol in [25], we use ELF18 [1] and LBP [5] as RGB features, and DVCov [25] as depth feature. We adopt the same protocol as DVCov [25] by randomly selecting five images of each person in either probe or gallery for multi-shot evaluation for all the three multi-modal datasets. All the experimental results are based on 10 random trials.

For the single-modal case, we use LDA to learn the distance metric for contrastive features, except for the Skeleton [39] and DVCov [25], which were matched by Euclidean distance and geodesic distance [25] respectively. To fuse the multi-modal features as the comparison, we first evaluate the traditional neural network based (NN-based) method by directly feeding the RGB and depth features into the fully-connected (FC) layer and reduce the features to 1000-dim based on the softmax loss in the training phases. Then, we evaluate the prevalent metric learning methods by weighting the ranking results achieved in different modalities followed by [25].

6.2.1. Comparison on PAVIS

PAVIS [38] dataset consists of two groups denoted by *Walking*1 and *Walking*2. Images of *Walking* and *Walking*2 were obtained by recording the same 79 people with a frontal view, walking slowly in an indoor scenario, where 60 people in *Walking*2 were dressed in different clothes from *Walking*1. Following the common train-test policy [38], we randomly sampled half of the group *Walking*1, i.e., images of 40 persons for training, and the remaining 39 persons for testing. These 39 testing persons in *Walking*1 were used as gallery data and the corresponding images of these 39 persons in *Walking*2 were used as probe data.

Table 3
Comparison with baselines on PAVIS dataset (in %).

Modality	Methods	Rank-1	Rank-5	References
RGB	LOMO [7]	19.74	44.36	CVPR2015
	ELF18 [1]	52.62	**78.26**	TCSVT2017
	Color Hist [53]	48.92	74.82	ECCV2008
	HOG [54]	45.33	73.92	CVPR2010
	LBP [5]	45.64	72.36	ICIG2011
	ELF18 [1]+ISR [4]	54.62	64.62	TCSVT2017+PAMI2015
	ELF18 [1]+**NNDGSR** [26]	**58.46**	73.85	**Ours**
	ELF18 [1]+**JGRSR_single**	**58.72**	74.62	**Ours**
Depth	RIFT2M [55]	8.77	27.69	ISOP2007
	Fehrs [56]	30.56	58.67	ICRA2012
	Skeleton [39]	37.33	71.13	Springer2014
	4D RAM [57]	43.00	-	CVPR2016
	RTA [58]	52.40	-	ECCV2018
	DVCov [25]	66.00	82.92	TIP2017
	DVCov [25]+ISR [4]	64.36	86.44	TIP2017+PAMI2015
	DVCov [25]+**NNDGSR** [26]	65.13	**87.18**	**Ours**
	DVCov [25]+**JGRSR_single**	**66.67**	**87.18**	**Ours**
RGB-D	ELF18 [1]+DVCov [25] + NN	47.74	80.36	TCSVT17+TIP17
	ELF18 [1]+DVCov [25]+LFDA [8]	47.08	68.05	TCSVT17+TIP17+CVPR13
	ELF18 [1]+DVCov [25]+KISSME [9]	61.69	79.23	TCSVT17+TIP17+CVPR12
	ELF18 [1]+DVCov [25]+XQDA [7]	57.49	83.33	TCSVT17+TIP17+CVPR15
	ELF18 [1]+**DVCov** [25]+**JGRSR_multi**	**68.97**	**87.44**	**Ours**

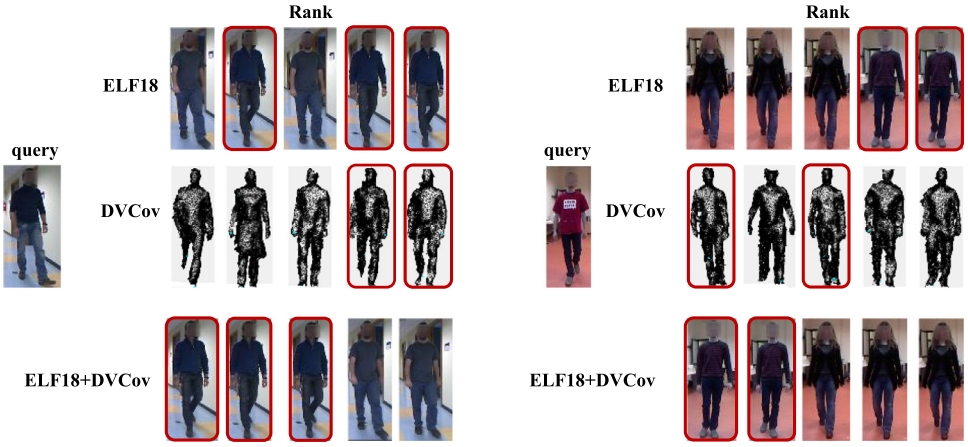

Fig. 3. Top 5 ranking results on PAVIS. In each group of images, the query image is on the left. The first, second and third rows are the ranking results of our method with LEF18, DVCov and LEF18+DVCov feature (s) respectively. The bounding boxes indicate the correct matchings.

Evaluation results on PAVIS [38] are shown in Table 3. Clearly, our approach significantly outperforms the state-of-the-art algorithms on both RGB feature, ELF18 [1] and depth feature DVCov [25]. After combing the RGB and depth features, our multi-modal representation achieves 68.97% at Rank-1 accuracy, which significantly beats the traditional NN-based method and the state-of-the-art metric learning methods. Together with the demonstration in Fig. 3 we can see, our model can relieve the problem of the person's clothing changes and the lighting influence by leveraging the RGB and depth representation.

6.2.2. Comparison on BIWI

BIWI [39] dataset contains three groups of sequences *Training, Still* and *Walk – ing* captured from 50 different people, with 300 frames of depth images and skeletons for each person. In *Training*, people performed motions, such as walking and rotating. Only 28 people presented in *Training* were recorded in *Still* and *Walking*, which were collected in a different day and a different scene so that most persons were dressed differently. In *Still*, people slightly moved, while in *Walking*, every person walked in different view angles.

For this dataset, we use images of those 28 commonly appeared persons in *Training* as gallery data and their images in *Still* and *Walking* as probe data. Table 4 reports the evaluation results on BIWI [39] dataset. From which we can see, (1) Our *JGRSR_single* significantly outperforms the state-of-the-art on RGB features. (2) The performance of sparse ranking based methods (including ISR [4] and Ours) is restricted on depth features. This might be caused by the dramatical posture changes of the pedestrian in this dataset which destroys the robustness of the features. Even though it is still clear that. (3) by introducing the probe and gallery-based graph regularizers and the dictionary learning, our NNDGSR [26] and *JGRSR_single* outperform the conventional sparse ranking method ISR [4] with competitive results. (4) Our *JGRSR_multi* achieves the best performance comparing to the NN-based method and metric learning methods.

6.2.3. Comparison on IAS-Lab

IAS-Lab [40] dataset consists of three groups of sequences *Training, TestingA* and *TestingB*. Each person is with about 500 frames of depth images and skeletons rotated on himself and walked during the recording. *TestingA* and *TestingB* were collected

Table 4
Comparison with baselines on BIWI dataset (in %).

Modality	Methods	Still		Walking		References
		Rank-1	Rank-5	Rank-1	Rank-5	
RGB	LOMO [7]	**18.17**	35.47	**10.31**	21.39	CVPR2015
	ELF18 [1]	4.11	19.13	1.50	16.77	TCSVT2017
	Color Hist [53]	10.61	31.92	5.86	21.70	ECCV2008
	HOG [54]	12.35	30.39	6.94	23.29	ECCV2008
	LBP [5]	10.87	**35.57**	5.34	**23.31**	CVPR2015
	ELF18 [1]+ISR [4]	4.41	21.43	3.56	17.86	TCSVT2017+PAMI2015
	ELF18 [1]+ **NNDGSR** [26]	8.57	21.43	5.36	19.65	**Ours**
	ELF18 [1]+**JGRSR_single**	14.71	21.78	10.14	21.01	**Ours**
Depth	RIFT2M [55]	4.34	20.78	3.75	18.31	ISOP2007
	Fehrs [56]	14.06	43.78	12.09	39.60	ICRA2012
	Skeleton [39]	**26.55**	**62.73**	16.94	47.18	Springer2014
	DVCov [25]	23.07	58.89	**21.40**	**54.12**	TIP2017
	DVCov [25]+ISR [4]	5.7	21.01	4.06	19.31	TIP2017+PAMI2015
	DVCov [25]+**NNDGSR** [26]	17.29	25.56	14.28	20.46	**Ours**
	DVCov [25]+**JGRSR_single**	21.65	42.14	16.96	23.35	**Ours**
RGB-D	ELF18 [1]+DVCov [25] + NN	4.29	27.86	7.14	21.68	TCSVT17+TIP17
	ELF18 [1]+DVCov [25]+LFDA [8]	16.43	40.00	7.86	17.14	TCSVT17+TIP17+CVPR13
	ELF18 [1]+DVCov [25]+KISSME [9]	22.86	32.86	12.14	34.29	TCSVT17+TIP17+CVPR12
	ELF18 [1]+DVCov [25]+XQDA [7]	15.71	37.86	13.57	35	TCSVT17+TIP17+CVPR15
	ELF18 [1]+**DVCov** [25]+**JGRSR_multi**	**28.18**	**64.29**	**21.79**	46.43	**Ours**

Table 5
Comparison with baselines on IAS-Lab dataset (in %).

Modality	Methods	TestingA		TestingB		References
		Rank-1	Rank-3	Rank-1	Rank-3	
RGB	LOMO[7]	25.79	66.28	30.06	79.90	CVPR2015
	ELF18 [1]	21.81	67.77	23.01	67.81	TCSVT2017
	Color Hist [53]	24.42	66.48	23.89	60.93	ECCV2008
	HOG [54]	38.89	72.67	49.62	86.79	ECCV2008
	LBP [5]	32.81	68.22	52.88	**89.81**	CVPR2015
	LBP [5]+ISR [4]	41.65	73.61	56.67	85.41	CVPR2015+PAMI2015
	LBP [5]+ **NNDGSR** [26]	**43.78**	**73.86**	**60.42**	86.38	**Ours**
	LBP [5]+**JGRSR_single**	**47.74**	**75.28**	**62.56**	88.85	**Ours**
Depth	RIFT2M [55]	20.94	60.87	19.88	60.02	ISOP2007
	Fehrs [56]	24.05	64.95	20.46	62.65	ICRA2012
	Skeleton [39]	49.83	**91.49**	60.25	**93.58**	Springer2014
	DVCov [25]	35.56	72.53	36.14	71.45	TIP2017
	DVCov [25]+ISR [4]	41.65	72.22	36.11	66.70	TIP2017+PAMI2015
	DVCov [25]+**NNDGSR** [26]	**50.00**	69.47	41.67	68.87	**Ours**
	DVCov [25]+**JGRSR_single**	**52.80**	80.56	44.44	73.89	**Ours**
RGB-D	LBP [5]+DVCov [25] + NN	48.89	74.44	34.44	54.44	CVPR2015+TIP2017
	LBP [5]+DVCov [25]+LFDA [8]	23.33	64.00	20.00	61.33	CVPR15+TIP17+CVPR13
	LBP [5]+DVCov [25]+KISSME [9]	24.67	56.67	30.67	61.33	CVPR15+TIP17+CVPR12
	LBP [5]+DVCov [25]+XQDA [7]	38.67	62.00	38.00	58.00	CVPR15+TIP17+CVPR15
	LBP [5]+**DVCov** [25]+ **JGRSR_multi**	**55.06**	89.37	**65.28**	91.56	**Ours**

with different clothes and in different environments respectively for the same person in *Training*.

Following the evaluation settings on PAVIS [38], half of *Training* sequences were randomly selected as gallery data, while the samples in *TestingA* and *TestingB* of the same persons were selected as probe data. Table 5 reports the evaluation results on IAS-Lab [40]. As we can see, (1) our approach significantly outperforms the state-of-the-art methods on both RGB and depth features. (2) Although our methods work overshadowed to LBP [5] and Skeleton [39] on Rank-3, they achieve much higher Rank-1 which is the most important metric in real-life application. (3) By combining both RGB and depth features, our *JGRSR_multi* can further boost the performance.

6.3. Evaluation on depth transferred multi-modal benchmarks

Although it is effective to incorporate depth information into the RGB space for multi-modal Re-ID, we cannot guarantee the depth information in most of the surveillance. Therefore, we generate the transferred Eigen-depth feature (TED) [25] as depth feature for two RGB datasets 3DPeS [59] and CAVIAR4REID [2] where the depth information is not available. For RGB features, we apply two favorable hand-crafted features, WHOS [4] and ELF18 [1], and one deep feature APR [11]. Following the protocol in [25], we select the BIWI [39] as the auxiliary dataset, then estimate depth information for 3DPeS [59] and CAVIAR4REID [2].

3DPeS [2] dataset contains hundreds of video sequences of 200 people taken from a multi-camera distributed surveillance system over several days, with different light conditions and different points of view. Two images were randomly selected as gallery data and another two as probe data in 3DPeS [2]. As for **CAVIAR4REID** [2] dataset, we follow the experiment settings in Sect 6.1 for single-modal Re-ID, and randomly select five images as gallery data or probe data for the multi-modal case.

We evaluate our method on both single-modal and multi-modal cases as shown in Table 6 and Table 7 respectively. From

Table 6
Comparison of single-modal cases on 3DPeS and CAVIAR4REID (in %).

Dataset	3DPeS					CAVIAR4REID				
Rank	1	2	3	4	5	1	2	3	4	5
TED feature (+ metric)										
+ LFDA [8]	28.23	33.54	37.19	40.73	42.29	60.83	70.83	81.11	65.56	87.78
+ KISSME [9]	16.77	19.38	21.35	24.17	25.94	57.50	67.78	73.61	76.39	80.28
+ XQDA [7]	30.31	35.73	39.17	42.69	44.48	57.94	69.94	76.83	81.00	85.28
+ ISR [4]	35.23	40.07	43.10	45.41	47.54	64.47	75.31	80.11	82.81	84.81
+ NNDGSR [26]	**36.33**	**42.59**	**46.20**	**48.97**	**51.39**	**70.08**	**80.53**	**84.00**	**85.92**	**87.72**
+ JGRSR_single	**36.98**	**41.15**	**43.49**	**46.35**	**49.48**	**73.72**	**81.75**	**85.78**	**87.72**	**89.44**
WHOS feature [4] (+ metric)										
+ LFDA [8]	41.98	48.75	52.71	55.42	58.23	71.67	82.22	87.78	90.56	91.94
+ KISSME [9]	32.92	40.10	44.06	45.83	48.02	63.44	75.28	81.67	86.94	90.28
+ XQDA [7]	49.38	57.29	61.15	64.17	66.35	66.67	78.33	86.67	90.28	91.67
+ ISR [4]	67.67	75.22	79.19	81.84	83.78	90.10	93.89	95.28	96.81	97.08
+ NNDGSR [26]	**70.56**	**77.50**	**81.15**	**83.66**	**85.43**	**93.19**	**96.28**	**97.39**	**98.08**	**98.47**
+ JGRSR_single	**70.72**	**77.97**	**81.61**	**84.10**	**86.09**	**93.72**	**96.78**	**97.86**	**98.36**	**98.81**
ELF18 feature [1] (+ metric)										
+ LFDA [8]	28.65	35.42	39.27	41.56	45.00	67.78	79.72	85.83	90.83	93.33
+ KISSME [9]	21.88	26.04	32.29	33.33	35.42	63.06	77.22	82.5	85.56	89.72
+ XQDA [7]	30.94	37.50	43.75	48.96	48.96	68.89	77.78	85.28	89.72	91.94
+ ISR [4]	45.02	53.91	58.69	62.03	64.57	81.06	88.89	92.28	93.75	94.89
+ NNDGSR [26]	**46.11**	**55.67**	**60.89**	**64.54**	**67.21**	**86.22**	**91.22**	**94.06**	**95.33**	**96.19**
+JGRSR_single	**47.60**	**56.81**	**61.96**	**65.58**	**68.09**	**86.71**	**92.67**	**94.31**	**95.56**	**96.67**
APR feature [11] (+metric)										
+ LFDA [8]	38.54	44.58	48.75	51.77	55.00	63.61	77.78	84.44	88.06	91.67
+ KISSME [9]	27.60	34.37	38.85	41.56	44.06	65.00	78.61	83.61	85.56	88.06
+ XQDA [7]	36.46	45.94	52.40	57.19	60.73	69.50	83.56	88.67	91.67	93.83
+ ISR [4]	51.09	60.23	65.11	68.25	70.79	80.69	89.03	92.08	93.75	95.83
+ NNDGSR [26]	**52.80**	**62.32**	**67.29**	**69.36**	**72.46**	**89.00**	**93.47**	**95.33**	**96.17**	**96.78**
+ JGRSR_single	**53.45**	**63.58**	**68.23**	**71.48**	**73.82**	**89.50**	**94.44**	**96.11**	**96.11**	**96.57**

Table 7
Comparison of multi-modal cases on transferred 3DPeS and CAVIAR4REID (in %).

Dataset	3DPeS					CAVIAR4REID				
Rank	1	2	3	4	5	1	2	3	4	5
WHOS [4] + TED [25] (+ metric)										
+ LFDA [8]	42.71	49.06	53.02	56.25	58.33	72.78	82.78	88.06	90.56	91.94
+ KISSME [9]	34.06	40.42	44.06	46.56	48.54	64.33	76.11	82.50	85.00	89.17
+ XQDA [7]	49.90	58.02	61.35	63.75	65.94	67.50	78.06	85.83	90.00	91.94
+ ISR [4]	67.85	75.27	79.52	81.82	83.85	89.25	93.97	95.97	97.06	97.69
+ NNDGSR [26]	**69.01**	**76.04**	9.43	81.51	83.59	**93.33**	**96.53**	**98.06**	**98.61**	**98.75**
+ JGRSR_multi	**71.55**	**77.60**	**81.32**	**84.05**	**85.74**	**94.17**	**97.14**	**98.31**	**98.83**	**99.08**
ELF18 [1] + TED [25] (+ metric)										
+ LFDA [8]	28.65	35.42	39.27	41.56	45.00	68.33	80.00	85.83	91.39	93.33
+ KISSME [9]	21.88	26.46	31.15	33.96	37.19	64.72	75.83	83.89	86.67	89.44
+ XQDA [7]	31.15	38.54	42.71	45.83	48.96	69.72	77.78	86.39	89.94	91.04
+ ISR [4]	46.70	56.79	62.10	65.91	68.82	84.14	90.61	93.36	94.97	95.75
+ NNDGSR [26]	**47.92**	**58.44**	**63.54**	**67.29**	**69.58**	**87.50**	**93.06**	**94.44**	**95.83**	**96.67**
+ JGRSR_multi	**48.75**	**57.60**	**62.29**	**66.25**	**69.06**	**88.42**	**93.89**	**95.5**	**96.81**	**97.50**
APR [11] + TED [25] (+ metric)										
+ LFDA [8]	38.54	44.58	48.75	51.77	55.00	63.61	77.78	84.44	88.06	91.67
+ KISSME [9]	28.23	34.17	39.17	41.56	43.33	66.67	77.50	83.06	86.39	88.89
+ XQDA [7]	36.88	46.98	53.44	57.92	61.15	70.11	83.39	88.72	91.67	93.50
+ ISR [4]	52.51	62.32	68.03	71.68	74.07	86.50	92.39	94.56	95.86	96.39
+ NNDGSR [26]	**52.86**	62.11	66.8	70.25	72.79	**87.64**	**93.61**	**95.14**	**95.83**	**96.39**
+ JGRSR_multi	**54.69**	**64.58**	**69.79**	**72.40**	**76.04**	**89.70**	**94.08**	**95.89**	**96.61**	**97.42**

Table 6 we can see, the sparse ranking based methods (ISR [4], NNDGSR [26] and *JGRSR_single*) significantly beat the conventional metric learning methods (KISSME [9] and XQDA [7]) on both hand-crafted features and deep features. The transferred depth feature TED performs overshadowed by the other three RGB features (WHOS [4], ELF18 [1] and APR [11]) since the single depth feature is not sufficient to describe the appearance of the person. However, the augmentation of TED [25] feature into the RGB features can effectively improve top-rank matching accuracies, as shown in Table 7. Furthermore, Table 7 demonstrates the promising results of our *JGRSR_multi* comparing to the state-of-the-art methods.

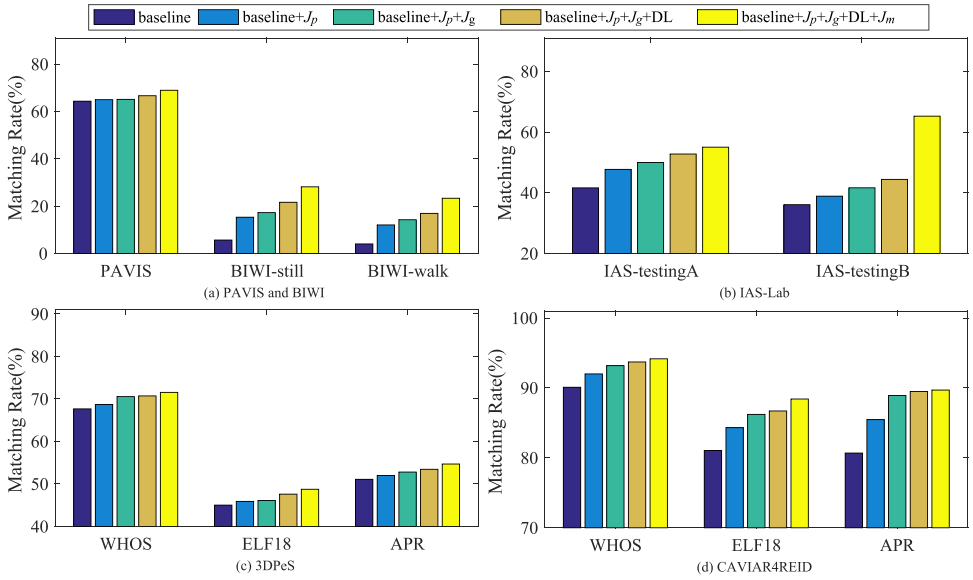

Fig. 4. The component analysis on PAVIS, BIWI, IAS-Lab, 3DPeS and CAVIRA4REID datasets.

Table 8
The probe and gallery sizes on PAVIS, BIWI, IAS-Lab, 3DPeS and CAVIRA4REID datasets (in {number of person} × {number of images per person}).

Dataset	PAVIS	BIWI	IAS-Lab	3DPeS	CAVIAR4REID
probe-size	39 × 5	28 × 5	6 × 5	100 × 2	36 × 5
gallery-size	39 × 5	28 × 5	6 × 5	100 × 2	36 × 5

6.4. Ablation study

To verify the contribution of the components in our model, we implement the ablation study of several variants of our method on PAVIS [38], BIWI [39], IAS-Lab [40], 3DPeS [59] and CAVIAR4REID [2]. Fig. 4 reports the results. From which we can see: (1) Probe and gallery-based graph regularizers play important roles in sparse ranking based Re-ID (comparing *baseline* $+J_p+J_g$ to *baseline*). (2) By enforcing the dictionary learning ($+DL$), it can improve the performance to some content. (3) By integrating other modality resources, it can further boost the performance ($+J_m$). Fig. 4 consistently demonstrates the contribution of each component in the proposed joint graph regularized dictionary learning and sparse ranking model. It should be noted that *baseline* is the initial sparse representation method (ISR [4]), *baseline* $+$ J_p+J_g denotes ours previous work NNDGSR [26], *baseline* $+J_p+$ J_g+DL denotes *JGRSR_single* and *baseline* $+J_p+J_g+DL+J_m$ denotes *JGRSR_multi*. Table 8 indicates the gallery and probe sizes of each dataset in Fig. 4.

6.5. Parameter analysis

There are five important parameters in our model: $\{\lambda, \beta, \gamma, \mu, \eta_v\}$. The first four parameters in Eq. (10) control the sparsity of the codings, the probe-based regularizer in probe images, the gallery-based regularizer in gallery images and the cross-modal coherence respectively, while the last parameter η_v from Eq. (26) balances the contribution of corresponding modality. In this paper, η_1 and η_2 indicate the contribution of RGB and depth modality respectively and $\eta_1 = 1 - \eta_2$. We empirically set: $\{\lambda, \beta, \gamma, \mu, \eta_1\} = \{0.1, 0.2, 0.5, 0.3, 0.7\}$. The results with different λ, β, γ, μ and η_1 on PAVIS are shown in Table 9, which demonstrates that our model is not sensitive to the parameters.

7. Conclusion

In this paper, we have proposed a novel joint graph regularized dictionary learning and sparse ranking method for multi-modal multi-shot person Re-ID. First, it can simultaneously capture the intrinsic geometric structures in both probe and gallery. In addition, it preserves the cross-modal consistency while handling the multi-modal (RGB-depth) Re-ID task. Then we provide a fast optimization for the proposed unified sparse ranking framework. Extensive experiments on challenging multi-modal multi-shot person Re-ID datasets demonstrate the promising performance of the proposed method. Although there are many fast algorithms to optimize sparse ranking framework, our method still faces the computational complexity problem due to the graph construction, especially on large-scale datasets. In the future, we will leverage the strong feature learning capability of convolutional neural networks (CNN) and the noise resistance ability of sparse ranking, and design

Table 9
Parameter analysis at rank-1 on PAVIS dataset (in %).

Parameter	Setting	Rank-1	Parameter	Setting	Rank-1	Parameter	Setting	Rank-1
	0.05	68.73		0.1	68.73		0.8	68.21
λ	0.1	68.97	β	0.2	68.97	η_1	0.7	68.97
	0.2	68.48		0.3	67.95		0.6	67.95
	0.2	67.68		0.2	67.69			
γ	0.5	68.97	μ	0.3	68.97			
	0.8	67.43		0.4	67.96			

more robust deep learning based Re-ID models for diverse challenging scenarios.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was partially funded by the National Key Research and Development Program of China (2016YFB1001001), the National Natural Science Foundation of China (61976002 and 61860206004), the National Laboratory of Pattern Recognition (NLPR) (201900046), and Open fund for Discipline Construction, Institute of Physical Science and Information Technology, Anhui University.

References

[1] Y.C. Chen, W.S. Zheng, J.H. Lai, P. Yuen, An asymmetric distance model for cross-view feature mapping in person re-identification, IEEE Trans. Circ. Syst. Video Technol. 27 (8) (2017) 1661–1675.
[2] S.C. Dong, M. Cristani, M. Stoppa, L. Bazzani, V. Murino, Custom pictorial structures for re-identification, in: British Machine Vision Conference, 2011, pp. 68.1–68.11.
[3] M. Farenzena, L. Bazzani, A. Perina, V. Murino, M. Cristani, Person re-identification by symmetry-driven accumulation of local features, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 2360–2367.
[4] G. Lisanti, I. Masi, A.D. Bagdanov, A.D. Bimbo, Person re-identification by iterative re-weighted sparse ranking, IEEE Trans. Pattern Anal. Mach.Intell. 37 (8) (2015) 1629–1642.
[5] Y. Zhang, S.T. Li, Gabor-lbp based region covariance descriptor for person re-identification, in: International Conference on Image and Graphics, 2011, pp. 368–371.
[6] S.C. Liao, S.Z. Li, Efficient psd constrained asymmetric metric learning for person re-identification, in: IEEE International Conference on Computer Vision, 2015, pp. 3685–3693.
[7] S.C. Liao, Y. Hu, X.Y. Zhu, S.Z. Li, Person re-identification by local maximal occurrence representation and metric learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2197–2206.
[8] S. Pedagadi, J. Orwell, S. Velastin, B. Boghossian, Local fisher discriminant analysis for pedestrian re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3318–3325.
[9] M. Koestinger, M. Hirzer, P. Wohlhart, P.M. Roth, H. Bischof, Large scale metric learning from equivalence constraints, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2288–2295.
[10] D.W. Li, X.T. Chen, Z. Zhang, K.Q. Huang, Learning deep context-aware features over body and latent parts for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 384–393.
[11] Y.T. Lin, L. Zheng, Z.D. Zheng, Y. Wu, Z.L. Hu, C.G. Yan, Y. Yang, Improving person re-identification by attribute and identity learning, Pattern Recogn. 95 (2019) 151–161.
[12] J.X. Liu, B.B. Ni, Y.C. Yan, P. Zhou, S. Cheng, J.G. Hu, Pose transferrable person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4099–4108.
[13] C. Su, S.L. Zhang, J.L. Xing, W. Gao, Q. Tian, Multi-type attributes driven multi-camera person re-identification, Pattern Recogn. 75 (2018) 77–89.
[14] T. Xiao, H.S. Li, W.L. Ouyang, X.G. Wang, Learning deep feature representations with domain guided dropout for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1249–1258.
[15] W. Zhang, S. Hu, K. Liu, Z. Zha, Learning compact appearance representation for video-based person re-identification, IEEE Trans. Circ. Syst. Video Technol. 29 (8) (2018) 2442–2452.
[16] Z. Zhou, Y. Huang, W. Wang, L. Wang, T. Tan, See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6776–6785.
[17] X.Y. Jing, X.K. Zhu, F. Wu, R.M. Hu, X.G. You, Y.H. Wang, H. Feng, J.Y. Yang, Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning, IEEE Trans. Image Process. 26 (3) (2017) 1363–1378.
[18] S. Karanam, Y. Li, R.J. Radke, Person re-identification with discriminatively trained viewpoint invariant dictionaries, in: IEEE International Conference on Computer Vision, 2015, pp. 4516–4524.
[19] K. Li, Z.M. Ding, S. Li, Y. Fu, Discriminative semi-coupled projective dictionary learning for low-resolution person re-identification, in: AAAI Conference on Artificial Intelligence, 2018, pp. 2331–2338.
[20] S. Li, M. Shao, Y. Fu, Person re-identification by cross-view multi-level dictionary learning, IEEE Trans. Pattern Anal. Mach.Intell. 40 (12) (2018) 2963–2977.
[21] A. Møgelmose, T.B. Moeslund, K. Nasrollahi, Multimodal person re-identification using rgb-d sensors and a transient identification database, in: International Workshop on Biometrics and Forensics, 2013, pp. 1–4.
[22] A. Møgelmose, C. Bahnsen, T. Moeslund, A. Clapes, S. Escalera, Tri-modal person re-identification with rgb, depth and thermal features, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2013, pp. 301–307.
[23] F. Pala, R. Satta, G. Fumera, F. Roli, Multimodal person reidentification using rgb-d cameras, IEEE Trans. Circ. Syst. Video Technol. 26 (4) (2016) 788–799.
[24] V. John, G. Englebienne, B. Krose, Person re-identification using height-based gait in colour depth camera, in: IEEE International Conference on Image Processing, 2013, pp. 3345–3349.
[25] A.C. Wu, W.S. Zheng, J.H. Lai, Robust depth-based person re-identification, IEEE Trans. Image Process. 26 (6) (2017) 2588–2603.
[26] A.H. Zheng, H.C. Li, B. Jiang, C.L. Li, J. Tang, B. Luo, Non-negative dual graph regularized sparse ranking for multi-shot person re-identification, in: Chinese Conference on Pattern Recognition and Computer Vision, 2018, pp. 108–120.
[27] X. Liu, M.L. Song, D.C. Tao, X.C. Zhou, C. Chen, J.J. Bu, Semi-supervised coupled dictionary learning for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 3550–3557.
[28] Q. Zhou, S.B. Zheng, H.B. Ling, H. Su, S. Wu, Joint dictionary and metric learning for person re-identification, Pattern Recognition. 72 (2017) 196–206.
[29] M. Aharon, M. Elad, A. Bruckstein, K-Svd: an algorithm for designing overcomplete dictionaries for sparse representation, IEEE Trans. Signal Process. 54 (11) (2012) 4311–4322.
[30] H. Lee, A. Battle, R. Raina, A.Y. Ng, Efficient sparse coding algorithms, in: Advances in Neural Information Processing Systems, 2007, pp. 801–808.
[31] M. Zheng, J.J. Bu, C. Chen, C. Wang, L.J. Zhang, G. Qiu, D. Cai, Graph regularized sparse coding for image representation, IEEE Trans. Image Process. 20 (5) (2010) 1327–1336.
[32] L.D. Sha, D. Schonfeld, Dual graph regularized sparse coding for image representation, in: IEEE Visual Communications and Image Processing, 2017, pp. 1–4.
[33] H.X. Wang, Y. Kawahara, C.Q. Weng, J.S. Yuan, Representative selection with structured sparsity, Pattern Recogn. 63 (2017) 268–278.
[34] J.X. Zhuo, J.Y. Zhu, J.H. Lai, X.H. Xie, Person re-identification on heterogeneous camera network, in: CCF Chinese Conference on Computer Vision, 2017, pp. 280–291.
[35] N. Parikh, S. Boyd, et al., Proximal algorithms, Found. Trends® Optim. 1 (3) (2014) 127–239.
[36] W.S. Zheng, S.G. Gong, T. Xiang, Associating groups of people, in: British Machine Vision Conference, 2009, pp. 23.1–23.11.
[37] L. Zheng, Z. Bie, Y.F. Sun, J.D. Wang, C. Su, S.J. Wang, Q. Tian, Mars: A video benchmark for large-scale person re-identification, in: European Conference on Computer Vision, 2016, pp. 868–884.
[38] I.B. Barbosa, M. Cristani, A. Del Bue, L. Bazzani, V. Murino, Re-identification with rgb-d sensors, in: European Conference on Computer Vision Workshops, 2012, pp. 433–442.
[39] M. Munaro, A. Fossati, A. Basso, E. Menegatti, L. Van Gool, One-shot Person Re-identification with a Consumer Depth Camera, in: Person Re-Identification, Springer, 2014, pp. 161–181.
[40] M. Munaro, A. Basso, A. Fossati, L. Van Gool, E. Menegatti, 3d reconstruction of freely moving persons for re-identification with a depth sensor, in: IEEE Conference on Robotics and Automation, 2014, pp. 4512–4519.
[41] L. Zheng, L.Y. Shen, L. Tian, S.J. Wang, J.D. Wang, Q. Tian, Scalable person re-identification: A benchmark, in: IEEE International Conference on Computer Vision, 2015, pp. 1116–1124.
[42] L. Bazzani, M. Cristani, A. Perina, M. Farenzena, V. Murino, Multiple-shot person re-identification by hpe signature, in: IEEE International Conference on Pattern Recognition, 2010, pp. 1413–1416.
[43] L. Bazzani, M. Cristani, A. Perina, V. Murino, Multiple-shot person re-identification by chromatic and epitomic analyses, Pattern Recogn. Lett. 33 (7) (2012) 898–903.
[44] E. Corvee, F. Bremond, M. Thonnat, et al., Person re-identification using spatial covariance regions of human body parts, in: IEEE International Conference on Advanced Video and Signal Based Surveillance, 2010, pp. 435–440.
[45] S. Bak, E. Corvee, F. Bremond, M. Thonnat, Multiple-shot human re-identification by mean riemannian covariance grid, in: IEEE International Conference on Advanced Video and Signal-Based Surveillance, 2011, pp. 179–184.
[46] G. Charpiat, M. Thonnat, Learning to match appearances by correlations in a covariance metric space, in: European Conference on Computer Vision, 2012, pp. 806–820.
[47] A. Klaser, M. Marszałek, C. Schmid, A spatiotemporal descriptor based on 3d–gradients, in: British Machine Vision Conference, 2008, pp. 995–1004.
[48] J. Han, B. Bhanu, Individual recognition using gait energy image, IEEE Trans. Pattern Anal.Mach.Intell. 28 (2) (2006) 316–322.
[49] F. Xiong, M. Gou, O. Camps, M. Sznaier, Person re-identification using kernel-based metric learning methods, in: European Conference on Computer Vision, 2014, pp. 1–16.
[50] S.J. Xu, Y. Cheng, K. Gu, Y. Yang, S.Y. Chang, P. Zhou, Jointly attentive spatial-temporal pooling networks for video-based person re-identification, in: IEEE International Conference on Computer Vision, 2017, pp. 4733–4742.
[51] Y. Wu, Y.T. Lin, X.Y. Dong, Y. Yan, W.L. Ouyang, Y. Yang, Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5177–5186.

[52] Y.T. Lin, X.Y. Dong, L. Zheng, Y. Yan, Y. Yang, A bottom-up clustering approach to unsupervised person re-identification, in: AAAI Conference on Artificial Intelligence, 33, 2019, pp. 8738–8745.
[53] D. Gray, H. Tao, Viewpoint invariant pedestrian recognition with an ensemble of localized features, in: European Conference on Computer Vision, 2008, pp. 262–275.
[54] O. Oreifej, R. Mehran, M. Shah, Human identity recognition in aerial images, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 709–716.
[55] L.J. Skelly, S. Sclaroff, Improved feature descriptors for 3d surface matching, in: Proceedings of SPIE, 6762, 2007.
[56] D. Fehr, A. Cherian, R. Sivalingam, S. Nickolay, V. Morellas, N. Papanikolopoulos, Compact covariance descriptors in 3d point clouds for object recognition, in: IEEE International Conference on Robotics and Automation, 2012, pp. 1793–1798.
[57] A. Haque, A. Alahi, F.F. Li, Recurrent attention models for depth-based person identification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1229–1238.
[58] N. Karianakis, Z.C. Liu, Y.P. Chen, S. Soatto, Reinforced temporal attention and split-rate transfer for depth-based person re-identification, in: European Conference on Computer Vision, 2018, pp. 715–733.
[59] D. Baltieri, R. Vezzani, R. Cucchiara, 3dpes: 3d people dataset for surveillance and forensics, in: ACM Workshop on Human Gesture and Behavior Understanding, 2011, pp. 59–64.

Aihua Zheng received her B.Eng. degrees and finished her Master-Doctor combined program in computer science and technology from Anhui University of China in 2006 and 2008 respectively. And received her Ph.D degree in computer science from the University of Greenwich of UK in 2012. She is currently an associated professor in computer science at Anhui University. Her main research interests include computer vision and artificial intelligent, especially on person/vehicle re-identificaiton, audio-visual learning and multi-modal and cross-modal lerning.

Hongchao Li received his B.Eng. degree in software.engineering in 2017 from Anhui University, Hefei, China. He is currently pursuing the PhD degree in computer science and technology at Anhui University. His current research is person re-identification.

Bo Jiang received the B.S. degrees in mathematics and applied mathematics and the M.Eng. and Ph.D. degrees in computer science from Anhui University of China in 2009, 2012 and 2015, respectively. He is currently an associated professor in computer science at Anhui University. His current research interests include image feature extraction and matching, data representation and learning.

Wei-Shi Zheng received the PhD degree in applied mathematics from Sun Yat-sen University in 2008. He is now a full Professor at Sun Yat-Sen University. He has now published more than 120 papers, including more than 100 publications in main journals (TPAMI, TNN/TNNLS, TIP, TSMC-B, PR) and top conferences (ICCV, CVPR, IJCAI, AAAI). His research interests include person/object association and activity understanding in visual surveillance, and the related large-scale machine learning algorithm. Especially, Dr. Zheng has active research on person re-identification in the last five years. He serves a lot for many journals and conference, and he was announced to perform outstanding review in recent top conferences (ECCV 2016 & CVPR 2017). He has ever joined Microsoft Research Asia Young Faculty Visiting Programme. He has ever served as a senior PC/area chairs many conferences (such as CVPR, BMVC, IJCAI and AAAI). He is an IEEE MSA TC member. He is an associate editor of Pattern Recognition. He is a recipient of the Excellent Young Scientists Fund of the National Natural Science Foundation of China, and a recipient of Royal Society-Newton Advanced Fellowship of United Kingdom.. Homepage: http://isee.sysu.edu.cn/hwshi.

Bin Luo received the B.Eng. degree in electronics and M.Eng. degree in computer science from Anhui University, Hefei, China, in 1984 and 1991, respectively, and the Ph.D. degree in computer science from the University of York, York, U.K., in 2002. From 2000 to 2004, he was a Research Associate with the University of York. He is currently a Professor with Anhui University. His current research interests include graph spectral analysis, large image database retrieval, image and graph matching, statistical pattern recognition, digital watermarking, and information security.