A Subspace Learning Approach to Multishot Person Reidentification

Aihua Zheng, Xuehan Zhang, Bo Jiang, Bin Luo, and Chenglong Li

***Abstract*—This paper addresses the challenging problem of multishot person reidentification (Re-ID) in real world uncontrolled surveillance systems. A key issue is how to effectively represent and process the multiple data with various appearance information due to the variations of pose, occlusions, and viewpoints. To this end, this paper develops a novel subspace learning approach, which pursues regularized low-rank and sparse representation for multishot person Re-ID. For the images of a person crossing a certain camera, we assume that the appearances of those subset images with similar viewpoints against a camera draw from the same low-rank subspace, and all the images of a person under a camera lie on a union of low-rank subspaces. Based on this assumption, we propose to learn a nonnegative low-rank and sparse graph to represent the person images. Moreover, the recurring pattern prior is integrated into our model to refine the affinities among images. Extensive experiments on four public benchmark datasets yield impressive performance by improving 22.9% on imagery library for intelligent detection systems video re identification (iLIDS-VID), 42.4% on person RE-ID (PRID) dataset 2011, 39.7% and 30.6% on speech, audio, image, and video technology-SoftBio camera 3/8 and camera 5/8, respectively, and 1.6% on motion analysis and re identification set compared to the state-of-the-art methods.**

***Index Terms*—Low-rank and sparse representation, multishot reidentification (Re-ID), recurring pattern prior, subspace learning.**

I. Introduction

PERSON reidentification (Re-ID) is used to reidentify the same person crossing the cameras with nonoverlapping views in the camera networks. It plays an important role in public security [1], automatic surveillance [2], human behavior analysis [3], and vehicle navigation, and has been widely investigated during the past decades.

Generally speaking, there are two categories of the Re-ID problem: the first category is the single-shot Re-ID, where only a single frame/image is recorded for each person within each camera view. Although many remarkable methods have been proposed for single-shot Re-ID [4]–[6], the performance is restrained by the limited information contained in a single image of a person. In the real-life surveillance system, the task of Re-ID normally produces multiple frames for a single person. Thus, it is natural to improve the performance of Re-ID in the second category, multishot case, where a video sequence is recorded for each person, through taking advantage of the multiple visual aspects. Recently, more and more works focus on multishot Re-ID, including appearance-based methods [7] which focus on appearance modeling to leverage the various changes between cameras, and learning-based methods [8]–[11] which focus on mitigating the appearance gaps between the low-level features and the high-level semantics. However, few of the existing methods exploit the subspace structure of the images for a certain person.

As a subspace representation method, low-rank representation (LRR) was proposed by Liu *et al.* [12] to recover the low-rank subspace structure, which can better capture the global structure of data against the influences of outliers and occlusions and robust to illumination or pose changes for recognition [13]. LRR has been widely applied in image/video segmentation [14], [15], saliency detection [16], and background modeling [17]. Recently, some works also proposed using LRR for multishot Re-ID. Jing *et al.* [18] proposed a novel multishot Re-ID framework by jointly learning a dictionary pair and a mapping function from high-resolution gallery images and low-resolution probe images, where low-rank matrix recovery was employed in a dictionary learning procedure to separate the noises from patches. Except for the low-level features, Chi *et al.* [19] proposed to produce continuous sematic attributes by embedding the low-rank attributes into the original binary attributes, based on which, a multitask learning framework is utilized for multishot Re-ID. However, how to learn the subspace correlations during the person images remains not well studied.

In this paper, we propose to employ a subspace learning approach based on regularized low-rank and sparse representation for multishot Re-ID. LRR has been widely applied in image/video segmentation [14], [15], due to its capability of capturing the global low-rank structure among data, and thus is robust to noises and/or corruptions. Therefore, we employ the idea of LRR in the problem of multishot Re-ID as follows. For each person sequence, some of them are similar in appearance, and thus are correlated without considering the image or video noises and/ or corruptions. This observation

Manuscript received April 6, 2017; revised July 7, 2017 and September 20, 2017; accepted December 7, 2017. Date of publication January 5, 2018; date of current version December 31, 2019. This work was supported in part by the National Nature Science Foundation of China under Grant 61502006, Grant 61702002, Grant 61602001, and Grant 61671018, in part by the Natural Science Foundation of Anhui Province under Grant 1508085QF127, in part by the Natural Science Foundation of Anhui Higher Education Institutions of China under Grant KJ2017A017, and in part by the Co-Innovation Center for Information Supply and Assurance Technology, Anhui University. This paper was recommended by Associate Editor K. Huang. *(Corresponding author: Chenglong Li.)*

The authors are with the School of Computer Science and Technology, Anhui University, Hefei 230601, China (e-mail: ahzheng214@ahu.edu.cn; zhangxh1234@foxmail.com; jiangbo@ahu.edu.cn; luobin@ahu.edu.cn; lcl1314@foxmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMC.2017.2784356

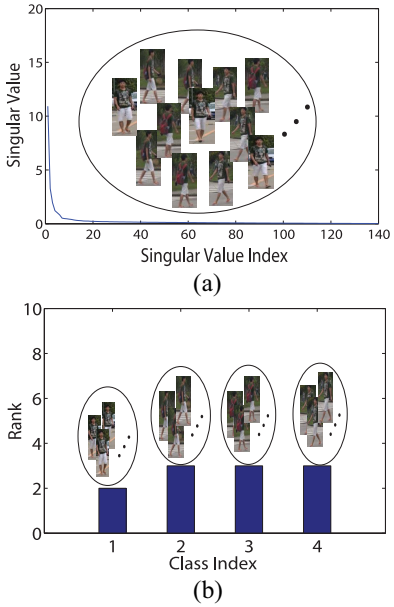


Fig. 1. Demonstration of the low-rank observation of the person images collected on the campus supervenience system. (a) Singular values of the feature matrix of the images. (b) Ranks of 95% principle components of four nature clusters, with ranks of 2, 3, 3, and 3 for corresponding clusters.

is similar with [11]. Therefore, for the sequential images of a person crossing a certain camera, we assume that the appearances of those subset images derived from similar appearance characterization draw from the same low-rank subspace, and all the images of a person under a camera lie on a union of low-rank subspaces and nature clusters lie on corresponding low-rank subspaces. As demonstrated in Fig. 1, Fig. 1(a) shows some representative person images with various appearances that sampled from a video sequence. We partition them into four clusters according to their appearance, and perform SVD on the feature matrix of each cluster. The singular value of the matrix is the results of SVD factorization of matrix, which denotes the correlation between elements. Thus, the more rapidly this value declines, the lower rank of the matrix. The results show that all feature matrices are low rank, as shown in Fig. 1(b). Specifically, we preserve 95% principle components of each feature matrix, and the rank of each cluster is less than 4, which justifies the low-rank assumption. The similar justification on the low-rank assumption is presented in [15]. Based on this assumption, we employ the LRR model to recover the subspace structures of the person images against noises and/or corruptions of low-level features.

According to the work of Wright *et al.* [20], we take each image as graph node and incorporate the sparse and nonnegative constraints on the representation coefficient matrix into our model to refine the affinities between the images of a certain person, where the nonnegativity and the sparsity ensure the convex combination of data points and the local linear relationship, respectively. To further refine the low-rank affinity matrix, the recurring pattern prior is integrated into our model, which exploits the nonlocal recurring regions [21] to refine the affinities among images.

Give the learned affinity matrix, the subspace clustering method, normalized cut (NCut) [22] is employed to generate the subspace clusters of each person. After that, each cluster is represented by the center of the features of the images in the corresponding cluster, which then fed into a modern metric learning scheme, cross-view quadratic discriminant analysis (XQDA) [6], to mitigate the cross-view gaps.

The main contributions of this paper can be summarized as the following three aspects.

1) We propose an effective subspace learning approach for multishot Re-ID in the LRR framework, in which the nonnegative, low-rank and sparse constraints are simultaneously employed to construct an informative graph for refining the affinities among person images.
2) We introduce the internal image statistical prior, called recurring pattern prior, to further refine the low-rank affinity matrix. This prior is originally used for image and video segmentation [15], [21], and we extend it to the multishot Re-ID task to improve its robustness.
3) We evaluate the performance of our approach on four benchmark datasets. The extensive experiments demonstrate that our approach significantly outperforms the state-of-the-art methods.

The rest of this paper is organized as follows. In Section II, the relevant existing methods are introduced. In Section III, we describe the details of our methods and the associated optimization algorithm. The experimental results on the benchmark datasets are shown in Section IV. Finally, Section V concludes this paper.

II. Related Work

Different from the single-shot Re-ID, multishot Re-ID concerns more on the additional sequential information. In the early stage, gait recognition [23] and temporal sequence matching [24] were employed on multishot Re-ID. However, the rigorous assumptions on temporal consistency restricts their performance on person Re-ID scenarios, which derives from the uncontrolled real-world camera networks. The recent commendable methods include the following two categories.

1) *Appearance-Based Methods:* They leverage the illumination, pose, and viewpoint changes in Re-ID by appearance modeling. Zhao *et al.* [25] proposed a dColorSIFT, which combines LAB color and SIFT for each patch to ensure the robustness in matching. Liao *et al.* [6] introduced another effective feature representation, which analyzes the horizontal occurrence of local features, and maximizes the occurrence to make a stable representation against viewpoint changes. Guo *et al.* [7] proposed an ambiguity removal approach on the shape feature to recognize and remove ambiguous samples. However, due to the discrimination in the inner-class images and resemblance in the interclass images, especially caused by the pose and illumination changes [26], none of the appearance models themselves can competent the challenging task of Re-ID.
2) *Learning-Based Methods:* To bridge the appearance gaps between the low-level feature and high-level human semantic, many learning-based methods have been

developed. Wang *et al.* [8], [9] presented a discriminative video ranking method based on the HOG3D spatial–temporal features of the selected video fragments. Li *et al.* [10] proposed to train a random forest based on pairwise constraints in the reduced random projection subspace. By learning a feature transformation through the adaptive Fisher discriminant analysis (FDA), Li *et al.* [11] proposed a hierarchical clustering on image sequences and followed by RankSVM as the metric learning step. You *et al.* [27] designed a top-push distance learning (TDL) model by integrating a top-push constraint during video feature matching. Meanwhile, a set of sparse coding and dictionary learning methods have been proposed for multishot Re-ID. However, as we observed in Fig. 1, the person images usually draw from the combination of several low-rank subspaces due to the appearance variations, and unfortunately, none of the above methods considered this issue, which motivates us to develop a low-rank subspace learning approach for multishot Re-ID.

III. Proposed Approach

In this section, we will detail our approach for the multishot Re-ID problem.

A. Overview

Our Re-ID method performs the following three steps. First, we propose a novel model, i.e., regularized nonnegative low-rank and sparse representation, to refine the affinities among images in an image sequence. Second, we obtain the optimal clusters of the image sequence via the NCut method on the optimized affinity matrix. Based on these clusters, we compute several representatives to represent the corresponding image sequence. Finally, we perform multishot Re-ID with XQDA to mitigate the cross-view gaps.

B. Regularized NonNegative Low-Rank and Sparse Representation

1) Formulation: For each image of a person, we extract the d-dimensional local maximal occurrence (LOMO) feature [6] to characterize the image, and the feature descriptors of the person forms the data matrix $\mathbf{X} = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d\times n}$, where n denotes the number of images of the person in a certain camera. We assume that the appearances of those subset images derived from a similar appearance characterization draw from the same low-rank subspace, and all the images of a person under a camera lie on a union of low-rank subspaces. Based on this assumption, each image descriptor can be represented as the linear combination of remaining image descriptors, and the LRR of all image descriptors can then be pursued in a joint fashion, i.e., $\mathbf{X} = \mathbf{XZ}$, where $\mathbf{Z}$ is the desired LRR coefficient matrix. Since the feature matrix is often noisy or grossly corrupted, the LRR can be solved by the following program:

$$\mathbf{X} = \mathbf{XZ} + \mathbf{E}, \ \text{s.t. rank}(\mathbf{Z}) \le r \tag{1}$$

Fig. 2. Illustrations of the recurring pattern prior. The patches with the same color indicate the recurring patterns.

where r is the desired rank, and $r \ll n$, $\mathbf{Z}_{ij}$ represented the affinity between the images i and j. In real applications, the data are often noisy and even grossly corrupted. Therefore, we add a noise term $\mathbf{E}$ to (1) for each person image. However, LRR often results in negative $\mathbf{Z}_{ij}$. In fact, the non-negativity property is more realistic for informative data, which often leads to better structure for data representation [28]. Besides that, in order to capture the local linear structure of data, we enhance the sparse property to the LRR formulation. Therefore, we seek a coefficient matrix $\mathbf{Z}$ to better capture the subspace structure by solving the following optimization problem:

$$\begin{aligned} \min_{\mathbf{Z},\mathbf{E}} \ & \text{rank}(\mathbf{Z}) + \beta\|\mathbf{Z}\|_0 + \alpha\|\mathbf{E}\|_0 \\ \text{s.t.} \ & \mathbf{X} = \mathbf{XZ} + \mathbf{E}, \mathbf{Z} \ge 0 \end{aligned} \tag{2}$$

where α and β are balance parameters. $\|\cdot\|_0$ denotes the l_0-norm of a matrix.

To further refine the low-rank coefficient matrix, we integrate the recurring pattern prior into our model based on the assumption that the local small-size patches (e.g., with size of 8×8 pixels) tend to recur frequently within the subset images with same appearance, as shown in Fig. 2, the patches with the same color cover the similar appearance and tend to reoccur in the same region of two adjacent images. The recurring pattern prior can be employed to evaluate the probability whether two images are drawn from the same subspace. Recurring pattern prior is a sort of internal image statistics which has been successfully used for image and video segmentation [21], [29] to refine the affinity between superpixels or supervoxels.

Denoting that Λ_i is the patch set covered by image i, and $\mathbf{Q} \in \mathbb{R}^{n\times n}$ is the recurring pattern prior between the n images. We have

$$\begin{aligned} Q_{ij} &= e^{-\left(\frac{1}{|\Lambda_i|}\Sigma_{p\in\Lambda_i}\varphi_\zeta(p,\Lambda_j)+\frac{1}{|\Lambda_j|}\Sigma_{q\in\Lambda_j}\varphi_\zeta(q,\Lambda_i)\right)} \\ \varphi_\zeta(p,\Lambda) &= \frac{1}{|\Lambda|}\sum_{q\in\Lambda}\delta_\zeta\left(K\left(\left\|\mathbf{f}_p - \mathbf{f}_q\right\|\right)\right) \end{aligned} \tag{3}$$

where $|\Lambda_i|$ indicates the number of patches within Λ_i, $\mathbf{f}_p$ and $\mathbf{f}_q$ are the histogram of oriented gradient (HOG) [30] features of the patches p and q, and K is a kernel function, such as Gaussian. The threshold operator $\delta_\zeta(a)$ is indicated as

$$\delta_\zeta(a) = aI(|a| > \zeta) \tag{4}$$

where $I(\cdot)$ is indicated as 1 if $|a|$ is larger than threshold ζ which is assigned as 0.6 empirically. By incorporating the

recurring pattern prior, our model is finalized as

$$\min_{\mathbf{Z},\mathbf{E}} \ \operatorname{rank}(\mathbf{Z}) + \beta\|\mathbf{Z}\|_0 + \alpha\|\mathbf{E}\|_0 + \gamma \ \operatorname{tr}\big(\mathbf{Z}^T\mathbf{Q}\big) \quad \text{s.t.} \quad \mathbf{X} = \mathbf{XZ} + \mathbf{E}, \mathbf{Z} \geq 0. \tag{5}$$

The larger $\mathbf{Q}_{ij}$, the higher probability that the image i and j derives from different clusters/subspaces, which will encourage smaller $\mathbf{Z}_{ij}$ by minimizing the last term. Therefore, minimizing $\operatorname{tr}(\mathbf{Z}^T\mathbf{Q})$ prefers to enforce the coefficient matrix $\mathbf{Z}$ to be block diagonal, where $\mathbf{Z}_{ij}$ is zero if the image i and j are from different cluster subspaces.

2) Optimization: Directly minimizing (5) is not trivial due to the nonconvexity of the low rank term (rank function) and the sparse term (l_0-norm). The problem of finding the sparsest solution of an underdetermined system of linear equations is NP-hard and difficult even to approximate [31]. In practice, convex relaxation is usually adopted to transfer (5) into a convex optimization problem by replacing the l_0-norm with the l_1-norm. The theory of sparse representation and compressed sensing [32] reveals that if the solution is sparse enough, the solution of the l_0-minimization problem is equal to the solution of the l_1-minimization problem. Hence, to tackle this issue, we will relax the nonconvexity by convex substituting. Specifically, we substitute the low-rank term and the l_0-norm by nuclear norm and l_1-norm, respectively. Thus, (5) can be relaxed as

$$\min_{\mathbf{Z},\mathbf{E}} \ \|\mathbf{Z}\|_* + \beta\|\mathbf{Z}\|_1 + \alpha\|\mathbf{E}\|_1 + \gamma \ \operatorname{tr}\big(\mathbf{Z}^T\mathbf{Q}\big) \quad \text{s.t.} \quad \mathbf{X} = \mathbf{XZ} + \mathbf{E}, \mathbf{Z} \geq 0 \tag{6}$$

where $\|\cdot\|_*$ and $\|\cdot\|_1$ denote the nuclear norm and the l_1-norm of a matrix, respectively.

The optimization problem of (6) can be efficient solved via the Augmented Lagrange Multiplier (ALM) method [33]. By introducing the auxiliary variables $\mathbf{J}$ and $\mathbf{P}$, (6) can be rewritten as the following countertype:

$$\min_{\mathbf{Z},\mathbf{E},\mathbf{J},\mathbf{P}} \ \|\mathbf{J}\|_* + \beta\|\mathbf{P}\|_1 + \alpha\|\mathbf{E}\|_1 + \gamma \ tr\big(\mathbf{Z}^T\mathbf{Q}\big) \quad \text{s.t.} \quad \mathbf{X} = \mathbf{XZ} + \mathbf{E}, \ \mathbf{Z} = \mathbf{J}, \ \mathbf{Z} = \mathbf{P}, \ \mathbf{P} \geq 0. \tag{7}$$

The related unconstrained problem of (7) is defined as

$$\begin{aligned}\min_{\mathbf{Z},\mathbf{E},\mathbf{J},\mathbf{P}} \ & \|\mathbf{J}\|_* + \beta\|\mathbf{P}\|_1 + \alpha\|\mathbf{E}\|_1 + \gamma \ tr\big(\mathbf{Z}^T\mathbf{Q}\big) \\ & + \ < \mathbf{Y}, \mathbf{X} - \mathbf{XZ} - \mathbf{E} > + \frac{\mu}{2}\|\mathbf{X} - \mathbf{XZ} - \mathbf{E}\|_F^2 \\ & + \ < \mathbf{V}, \mathbf{Z} - \mathbf{J} > + \frac{\mu}{2}\|\mathbf{Z} - \mathbf{J}\|_F^2 \\ & + \ < \mathbf{U}, \mathbf{Z} - \mathbf{P} > + \frac{\mu}{2}\|\mathbf{Z} - \mathbf{P}\|_F^2 \end{aligned} \tag{8}$$

where $\mu > 0$ is the penalty parameter, $\mathbf{Y}$, $\mathbf{V}$, and $\mathbf{U}$ are ALMs. As shown in [33], ALM procedure will converge with μ increasing. The entire algorithm is summarized in Algorithm 1. Noted that: 1) step 2 is solved via the singular value thresholding operator [34]; 2) steps 3(1) and 4 are convex problems which can be solved by the soft-threshold (or shrinkage) method in [33] and the operator $\max(\cdot)$ in step 3(2) ensures the non-negativity of $\mathbf{P}$; and 3) a series of increasing μ can be obtained by setting the update parameter ρ to be 1.1.

Algorithm 1 Optimization Procedure of Our Model

Input:

The data matrix $\mathbf{X}$ of person images;
The recurring pattern prior $\mathbf{Q}$;
Set parameter β, α, γ, $\rho = 1.1$ and $\mu = 10^{-5}$;
Set $\mathbf{Z} = \mathbf{J} = \mathbf{P} = 0$; $\mathbf{E} = 0$; $\mathbf{Y} = 0$; $\mathbf{V} = 0$; $\mathbf{U} = 0$; $\epsilon = 10^{-8}$; *MAXIter* $= 400$.

Output: $\mathbf{J}, \mathbf{P}, \mathbf{E}, \mathbf{Z}$;

1: **while** Not converged **do**

2: Fix the others and update $\mathbf{J}$ by solving

$$\mathbf{J} = \arg\min_{\mathbf{J}} \frac{1}{\mu}\|\mathbf{J}\|_* + \frac{1}{2}\left\|\mathbf{J} - \left(\mathbf{Z} + \frac{\mathbf{V}}{\mu}\right)\right\|_F^2$$

3: Fix the others and update $\mathbf{P}$ by solving

$$(1)\mathbf{P} = \arg\min_{\mathbf{P}} \frac{\beta}{\mu}\|\mathbf{P}\|_1 + \frac{1}{2}\left\|\mathbf{P} - \left(\mathbf{Z} + \frac{\mathbf{U}}{\mu}\right)\right\|_F^2;$$
$$(2)\mathbf{P} = \max(\mathbf{P}, 0);$$

4: Fix the others and update $\mathbf{E}$ by solving

$$\mathbf{E} = \arg\min_{\mathbf{E}} \frac{\alpha}{\mu}\|\mathbf{E}\|_1 + \frac{1}{2}\left\|\mathbf{E} - \left(\mathbf{X} - \mathbf{XZ} + \frac{\mathbf{Y}}{\mu}\right)\right\|_F^2$$

5: Fix the others and update $\mathbf{Z}$ by solving

$$\mathbf{Z} = \big(\mathbf{X}^T\mathbf{X} + \mathbf{2}\big)^{-1}(\mathbf{X}^T\mathbf{X} - \mathbf{X}^T\mathbf{E} + \mathbf{J} + \mathbf{P} + \frac{1}{\mu}(\mathbf{X}^T\mathbf{Y} - (\mathbf{V} + \mathbf{U}) - \gamma\mathbf{Q}))$$

6: Update the multipliers and parameter

$$\mathbf{Y} = \mathbf{Y} + \mu(\mathbf{X} - \mathbf{XZ} - \mathbf{E});$$
$$\mathbf{V} = \mathbf{V} + \mu(\mathbf{Z} - \mathbf{J});$$
$$\mathbf{U} = \mathbf{U} + \mu(\mathbf{Z} - \mathbf{P});$$
$$\mu = \rho\mu;$$

7: Check the convergence conditions

$(1)\|\mathbf{X} - \mathbf{XZ} - \mathbf{E}\| < \epsilon$ and $\|\mathbf{Z} - \mathbf{J}\| < \epsilon$ and $\|\mathbf{Z} - \mathbf{P}\| < \epsilon$; Or
(2)iterations reaches *MAXIter*;

8: **end while**

C. Subspace Clustering via NCut

In this section, we will deploy the optimized low-rank and sparse representation in the previous sections to obtain the representatives for each image sequence. Given the optimized coefficient matrix $\mathbf{Z}$ and feature matrix $\mathbf{X}$, where the $\mathbf{Z}$ is coefficient matrix which is not a symmetric matrix. For applying NCut algorithm, we need a symmetric matrix. So we first

define the affinity matrix as

$$\mathbf{Z}_{ij} = \begin{cases} \mathbf{Z}_{ij}, & \text{if } \mathbf{Z}_{ij} > \theta \\ 0, & \text{otherwise.} \end{cases}$$
$$\mathbf{W}_{ij} = \omega e^{\frac{-(\mathbf{Z}_{ij}+\mathbf{Z}_{ji})/2}{2\tau_1^2}} + (1-\omega)e^{\frac{-\mathbf{D}_{ij}}{2\tau_2^2}} \quad (9)$$

where, $\theta = 0$, $\mathbf{D}_{ij}$ means the Euclidean distances between the feature vectors of x_i and x_j. Then, for each person on a different camera, we employ the NCut algorithm [22] to achieve the subspace clustering person by person, which can be formulated as

$$\max_{\mathbf{G}} \ \frac{1}{k} \operatorname{tr}\left(\mathbf{G}^T\mathbf{W}\mathbf{G}\right) \quad \text{s.t. } \mathbf{G}^T\mathbf{B}\mathbf{G} = \mathbf{I}_k \quad (10)$$

where k is the number of clusters, $\mathbf{B} = \mathbf{W}\mathbf{1}_n$ is the degree matrix, and $\mathbf{G} = \mathbf{M}(\mathbf{M}^T\mathbf{B}\mathbf{M})^{-(1/2)}$ is the scaled partition matrix. n is the total number of images. $\mathbf{1}$ and $\mathbf{I}$ denote all ones vector and identity matrix, respectively. $\mathbf{M} \in \{0,1\}^{n\times k}$ is the partition matrix. The optimization of (10) has been addressed in [22]. After NCut clustering, we can reduce the number of person images by using the mean features of the k clusters, which can also improve the efficiency of the forthcoming metric learning. For instance, it can speed up 1500 times during the training phase and 200 times during the testing phase for iLIDS-VID. The improvement will be more significant when the number of person images increases.

D. Cross-View Quadratic Discriminant Analysis

After the subspace clustering, the images of each person are clustered into k clusters. The center of each cluster, $c_{i,j}^a$ which is the mean of the image features of the ith person in the jth cluster, is used as the representative of the cluster. Then, we employ the XQDA method [6] as the metric learning step to mitigate the cross-view gap. The cross-view training set $\{\mathbf{C}_i^a, \mathbf{C}_j^b | i,j = 1,\ldots,m\}$ is formed by m person from camera a and camera b. Noted that $\mathbf{C}_i^a = \{c_{i,1}^a, c_{i,2}^a, \ldots, c_{i,k}^a\} \in \mathbb{R}^{d\times k}$ contains d-dimensional feature representatives of k clusters of person i in the camera a. Different from Bayesian Face [35] and keep it simple and straightforward metric [36] learning a distance function in d-dimension, XQDA learned a subspace $\mathbf{S} = (\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_{l_r}) \in \mathbb{R}^{d\times l_r}$ for the cross-view data, and simultaneously learned a distance function in l_r-dimensional ($l_r \ll d$) subspace as

$$d_S\left(c_{i,h}^a, c_{j,t}^b\right) = \left(c_{i,h}^a - c_{j,t}^b\right)^T \mathbf{S}\left(\Sigma_{In}'^{-1} - \Sigma_{Ex}'^{-1}\right)\mathbf{S}^T\left(c_{i,h}^a - c_{j,t}^b\right) \quad (11)$$

where $\sum_{In}'^{-1} = \mathbf{S}^T\Sigma_{In}\mathbf{S}$ and $\Sigma_{Ex}'^{-1} = \mathbf{S}^T\Sigma_{Ex}\mathbf{S}$, for $i,j \in \{1,\ldots,m\}$, $h,t \in \{1,\ldots,k\}$. $l_r = 125$ in this paper. Noted that Σ_{In} and Σ_{Ex} are the covariance matrices of the interpersonal variations *In* (the same person under different cameras) and the extrapersonal variations *Ex* (the different person under different cameras). Due to the inverse matrices in (11) and zero-mean properties of *In* and *Ex*, the projection direction $\mathbf{s}$ can be optimized by maximizing $\sigma_{Ex}(\mathbf{s})/\sigma_{In}(\mathbf{s})$

$$\max_{\mathbf{s}} \ \sigma_{Ex}(\mathbf{s})/\sigma_{In}(\mathbf{s}) = \frac{\mathbf{s}^T\Sigma_{Ex}\mathbf{s}}{\mathbf{s}^T\Sigma_{In}\mathbf{s}} \quad (12)$$

which equals to

$$\max_{\mathbf{s}} \ \mathbf{s}^T\Sigma_{Ex}\mathbf{s}, \ \text{s.t. } \mathbf{s}^T\Sigma_{In}\mathbf{s} = 1. \quad (13)$$

This can be solved by the generalized eigenvalue decomposition problem as similar in LDA. The final distance between the person i in camera a and the person j in camera b is obtained by

$$D_S\left(\mathbf{C}_i^a, \mathbf{C}_j^b\right) = \min_h\left\{\min_t\left\{d_S\left(c_{i,h}^a, c_{j,t}^b\right)\right\}\right\} \quad (14)$$

where $i,j \in \{1,\ldots,m\}$, $h,t \in \{1,\ldots,k\}$.

IV. Experimental Results and Analysis

We evaluate our approach on four benchmark datasets, iLIDS-VID [8], PRID 2011 [37], speech, audio, image, and video technology (SAIVT)-SoftBio [38], and motion analysis and re identification set (MARS) [39], and compare the performance with state-of-the-art methods symmetry-driven accumulation of local features [5], Salience [25], RankSVM [40], random-projection-based random forest [10], local fisher discriminant analysis (LFDA) [41], sparse Re-ID [42], discriminative viewpoint dictionaries learning [43], adaptive fisher discriminant analysis (AFDA) [11], discriminative selection in video ranking (DVR) [9], pairwise feature dissimilarities space (PFDS) [44], and Fused [38].

A. Benchmark Datasets

1) iLIDS-VID [8]: This dataset includes 600 image sequences for 300 indoor pedestrians recorded by two nonoverlapping nonadjacent cameras at an airport arrival hall. The length of each image sequence varies from 23 to 192 frames, with an average length of 73. The dataset is quite challenging due to large occlusions, and big viewpoint changes across the cameras.

2) PRID 2011 [37]: This dataset consists of 400 image sequences for 200 outdoor persons from two adjacent cameras. The length of each image sequence varies from 5 to 675 frames with an average number of 100. We follow the same protocol as [9] and only 178 persons with a length > 21 frames is evaluated. It is less challenging with relatively clean backgrounds and rare occlusions.

3) SAIVT-SoftBio [38]: This dataset consists of 150 people from eight nonoverlapping uncontrolled real-life indoor surveillance networks. Since not every person appears in each camera view, following the works in [11], [38], and [44], we select cameras 3/8 including 99 person pairs with similar viewpoints and cameras 5/8 including 103 person pairs with large viewpoint changes.

4) MARS [39]: MARS is the largest and newly collected dataset for person Re-ID. It is an extension of the Market-1501 dataset [45] that collected from six near-synchronized cameras in the campus of Tsinghua University. MARS consists of 1261 pedestrians each of which appears at least two cameras. It contains 625 identities with 8298 tracklets for training, which having in total 509 914 bounding boxes that automatically extracted by the deformable part model as pedestrian detector and the generalized maximum multiclique problem

TABLE I
EVALUATED THE PARAMETERS ON ILIDS-VID DATASET (IN %)

Param	Setting	Rank-1	Param	Setting	Rank-1
	0.01	61.1		0.001	62.2
β	0.1	62.4	α	0.01	62.4
	1.1	61.4		0.1	61.9
	0.001	61.8		0.04	62.0
γ	0.01	62.4	ω	0.4	62.4
	0.1	62.3		0.9	61.4
	0.03	61.8		0.06	62.1
τ_1	0.3	62.4	τ_2	0.6	62.4
	0.9	61.6		0.9	61.6
	2	62.1			
k	4	62.4			
	8	62.2			

tracker. For testing, 681 089 bounding boxes are generated in the same manner, containing 636 identities with 12 180 tracklets. The query trackelets are automatically generated from the testing samples. Different from the other dataset, it also consists 23 380 "junk" bounding boxes and 147 743 "distractors" bounding boxes in the testing samples.

B. Experiment Setup

For the parameters in our model: we empirically set $\{\beta, \alpha, \gamma\} = \{0.1, 0.01, 0.01\}$ in optimization. In affinity definition, we set $\{\omega, \tau_1, \tau_2\} = \{0.4, 0.3, 0.6\}$. For cluster number, we set $k = 4$. We evaluate these parameters on iLIDS-VID and report the ranking results in Table I. It is worth noting that our approach is incentive to parameters. The variation of parameters do not have too much effect on the final performance of our approach. We randomly and identically separate the dataset into training and testing sets for iLIDS-VID and PRID 2011. For SAIVT-SoftBio, we follow the protocol in [44] and randomly select one third sequences for training and the rest two thirds for testing. In the testing phase, the sequences from one camera are used as probe while those from the other camera are gallery. For MARS, we follow the protocol of [39] by selecting the query/probe set from the testing/gallary set and use the provided training set for training. By computing the ranking of each probe sequence with all the gallery sequences, the results are reported by cumulative match characteristic (CMC) curves for the first three datasets iLIDS-VID, PRID 2011, and SAIVT-SoftBio, where the matching rate at rank-n indicate the percentage of correct matchings in top n candidates. And all the experimental results are reported based on the average of ten trials while the splitting of training and testing is fixed for each trial. For MARS, following by [45], we use the ranking results together with the mean average precision (mAP) for accuracy evaluation, which is more comprehensive than CMC curves for person Re-ID when the number of cameras more than two. Noted that the feature of our method in Tables II–V is LOMO.

C. Evaluation on Benchmark Datasets

1) iLIDS-VID: The results for iLIDS-VID [8] dataset are shown in Table II with corresponding CMC curves in Fig. 3. As can be seen, the proposed algorithm achieves the best performance. Specifically, the Rank-1 and Rank-5 matching rates of OUR subspace learning method (OURS) are 62.4% and 88.4%, outperforming the second best DVR [9] by 22.9% and 27.3%, respectively, and consistently higher for other ranks.

TABLE II
MATCHING RATE COMPARISON ON ILIDS-VID DATASET (IN %)

Dataset	iLIDS-VID				
Ranks	Rank-1	Rank-5	Rank-10	Rank-20	Reference
SDALF	6.3	18.8	27.1	37.3	2010 CVPR [5]
Salience	10.2	24.8	35.5	52.9	2013 CVPR [25]
RankSVM	18.6	43.3	57.1	71.2	2002 SIGKDD [40]
RPRF	14.5	29.8	40.7	58.1	2015 WACV [10]
LFDA	21.1	34.8	41.3	48.7	2006 ICML [41]
SRID	24.9	44.5	55.6	66.2	2015 CVPRW [42]
DVDL	25.9	48.2	57.3	68.9	2015 ICCV [43]
AFDA	37.5	62.7	73.0	81.8	2015 BMVC [11]
DVR	39.5	61.1	71.7	81.0	2016 TPAMI [9]
OURS	**62.4**	**88.4**	**94.3**	**98.8**	Proposed

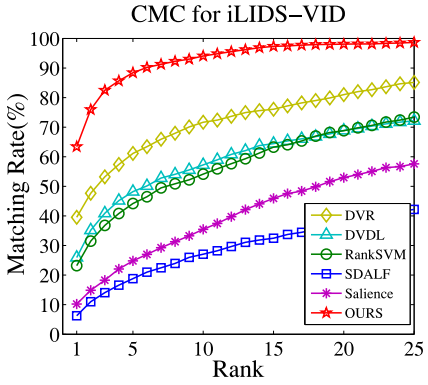


Fig. 3. Comparison results of CMC curves on iLIDS-VID.

TABLE III
MATCHING RATE COMPARISON ON PRID 2011 DATASET (IN %)

Dataset	PRID 2011				
Ranks	Rank-1	Rank-5	Rank-10	Rank-20	Reference
SDALF	5.2	20.7	32.0	47.9	2010 CVPR [5]
RPRF	19.3	38.4	51.6	68.1	2015 WACV [10]
LFDA	22.3	41.7	51.6	62.0	2006 ICML [41]
RankSVM	22.4	51.9	66.8	80.7	2002 SIGKDD [40]
Salience	25.8	43.6	52.6	62.0	2013 CVPR [25]
SRID	35.1	59.4	69.8	79.7	2015 CVPRW [42]
DVR	40.0	71.7	84.5	92.2	2016 TPAMI [9]
DVDL	40.6	69.7	77.8	85.6	2015 ICCV [43]
AFDA	43.0	72.7	84.6	91.9	2015 BMVC [11]
OURS	**85.4**	**96.6**	**98.5**	**99.8**	Proposed

2) PRID 2011: The comparison results for PRID 2011 [37] dataset are shown in Table III and Fig. 4. Compared with iLIDS-VID dataset, PRID 2011 dataset is less challenging with relatively clean backgrounds and rare occlusions, therefore the performance is generally much better than iLIDS-VID dataset for most of the existing methods. Our method significantly outperforms the state-of-the-art methods with 85.4% of Rank-1 which is better than the second-best AFDA [11] method and with promising performance for other ranks.

3) SAIVT-SoftBio: The result for SAIVT-SoftBio [38] dataset is shown in Tables IV and V. Since this dataset is not widely evaluated, we only compared our method with the reported methods, including LFDA [41], RankSVM [40], PFDS [44], Fused [38], and AFDA [11]. The proposed method has significant improvement on both subsets. For Cameras 3/8

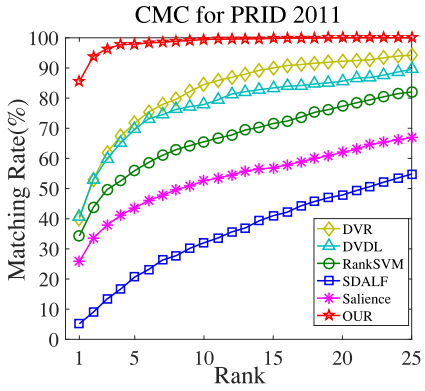


Fig. 4. Comparison results of CMC curves on PRID 2011.

TABLE IV
MATCHING RATE COMPARISON ON SAIVT-SOFTBIO (CAMERAS 3/8) DATASET (IN %)

Dataset	SAIVT-SoftBio(Cameras 3/8)				
Ranks	Rank-1	Rank-5	Rank-10	Rank-20	Reference
LFDA	12.2	36.8	54.6	74.9	2006 ICML [41]
RankSVM	32.4	68.4	82.0	92.9	2002 SIGKDD [40]
PFDS	33.2	60.5	74.0	87.2	2014 ICPR [44]
Fused	36.4	60.3	76.0	87.6	2012 DICTA [38]
AFDA	43.0	72.7	84.6	91.9	2015 BMVC [11]
OURS	**82.7**	**96.4**	**97.6**	**99.8**	Proposed

TABLE V
MATCHING RATE COMPARISON ON SAIVT-SOFTBIO (CAMERAS 5/8) DATASET (IN %)

Dataset	SAIVT-SoftBio(Cameras 5/8)				
Ranks	Rank-1	Rank-5	Rank-10	Rank-20	Reference
LFDA	9.3	27.1	41.2	60.6	2006 ICML [41]
RankSVM	14.9	40.5	57.9	75.0	2002 SIGKDD [40]
PFDS	18.6	32.9	53.0	85.3	2014 ICPR [44]
Fused	20.0	33.0	50.4	67.8	2012 DICTA [38]
AFDA	30.9	61.6	77.3	91.1	2015 BMVC [11]
OURS	**61.5**	**86.8**	**93.5**	**97.8**	Proposed

subset, Rank 1 matching rate of our method achieves 82.7% which outperforms the second-best AFDA [11] by 39.7%, while for the more challenging Cameras 5/8 subset with larger viewpoint changes, the results of Rank-1 can still reach 61.5%, which also significantly beats the state-of-the-art methods.

4) MARS: The ranking results and mAP on MARS [39] dataset is reported in Table VI on provided CNN features, where the Max/Avg means the Max/Avg pooling on a set of feature vectors by maximizing/averaging each dimension to generate a single feature vector. Specifically, we implement our subspace learning on the images of each person under each possible camera. To reduce the huge number of samples, we simplify the samples for clustering by the following protocol: let $|c_{i,t}^{a_j}|$ be the number of the images in the tth tracklet of the ith person under the jth camera, if $|c_{i,t}^{a_j}| > 10\,000$, which exists in the junk and distractors samples, we conduct the max pooling on each 100 sequential images before clustering, while max pooling on ten sequential images if $160 < |c_{i,t}^{a_j}| \leq 10\,000$; if $21 < |c_{i,t}^{a_j}| \leq 160$, we cluster all the images; if $|c_{i,t}^{a_j}| \leq 21$, we directly employ max/average pooling instead of clustering.

TABLE VI
MATCHING RATE COMPARISON ON MARS DATASET (IN %)

Methods	MARS			
Ranks	Rank-1	Rank-5	Rank-20	mAP
HistLBP+XQDA	18.6	33.0	45.9	8.0
gBiCov+XQDA	9.2	19.8	33.5	3.7
BoW+Kissme	30.6	46.2	59.2	15.5
SDALF+DVR	4.1	12.3	25.1	1.8
LOMO+XQDA	30.7	46.6	60.9	16.4
CNN+XQDA(Max)	65.3	**82.0**	**89.0**	47.6
CNN+OURS+XQDA(Max)	**66.8**	80.0	88.2	**47.7**
CNN+XQDA(Avg)	64.6	**81.4**	**89.1**	47.5
CNN+OURS+XQDA(Avg)	**66.2**	77.9	87.3	**49.2**

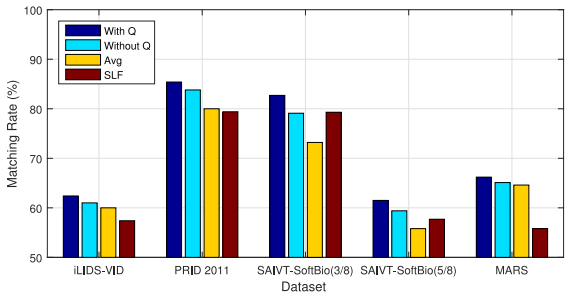


Fig. 5. Component analysis of recurring pattern prior component and nonnegative sparse subspace learning on four benchmark datasets.

In order to verify the performance of our subspace learning method comparing to the baselines, we evaluate the max or averaging pooling, respectively, on each cluster of the feature vectors of the person images for our methods. From the table we can see: CNN feature outperforms the state-of-the-art features on MARS and our subspace clustering can further improve the Rank-1 accuracy. It is also noted that our method performs a slightly better Rank-1 accuracy but inferior performance at other ranks. It may be because our method is evaluated on the down-sampled data on video frame level as mentioned above, while the baseline results are reported on the original data. Despite of this, we still achieved competitive performance.

D. Component Analysis

To justify the contribution of the components of our method, we evaluate the recurring pattern prior Q and the nonnegative sparse subspace learning in this section.

1) Recurring Pattern Prior: To evaluate the contribution of recurring pattern prior (with Q in Fig. 5) in our model, we compare our model to the model without recurring pattern prior (the bars "without Q" in Fig. 5) by setting $\gamma = 0$ in (5). The comparison results are reported in Fig. 5. The recurring pattern prior can improve the matching rates by 3.60% in maximum and 2.17% in average, which justifies the contribution of the recurring pattern prior component.

2) Nonnegative Sparse Subspace Learning: In order to evaluate the contribution of the proposed nonnegative sparse subspace learning method for multishot Re-ID, we evaluated the XQDA without subspace learning directly on the feature spaces (LOMO feature for iLIDS-VID, PRID 2011, and SAIVT-SoftBio and CNN feature for MARS) as the baseline method. Specifically, the average pooling (the bars "Avg"

TABLE VII
EVALUATION ON CNN FEATURES ON DATASETS ILIDS-VID, PRID 2011, AND SAIVT-SOFTBIO

Dataset	Methods	Rank-1	Rank-5	Rank-20
iLIDS-VID	CNN_d+OURS+XQDA	**30.5**	**55.2**	**78.2**
	CNN_d+XQDA	24.9	44.1	64.7
PRID 2011	CNN_d+OURS+XQDA	**58.8**	**83.2**	**96.7**
	CNN_d+XQDA	45.1	67.8	87.7
SAIVT-SoftBio (Cameras 3/8)	CNN_d+OURS+XQDA	**81.7**	**95.2**	**99.1**
	CNN_d+XQDA	69.1	87.1	96.3
SAIVT-SoftBio (Cameras 5/8)	CNN_d+OURS+XQDA	**42.7**	**75.1**	**92.7**
	CNN_d+XQDA	28.5	57.8	85.4

in Fig. 5) and the score level fusion (the bars "SLF" in Fig. 5) are implemented for the images of a person. Noted that, Avg means generating a single feature vector from a set of feature vectors by averaging each dimension as we stated in Section IV-C4. SLF means conducting the metric learning and distance calculation on the frame level, while fusing all obtained distance values between two tracklets by a simple averaging fusion. Particularly, due to the large quantity number of frames in each tracklet in MARS, we implement the score level fusion on the simplified/down-sampled tracklets as described in Section IV-C4. As the result, the subspace learning and clustering can improve averagely 4.92% comparing with average pooling, especially on iLIDS-VID and SAIVT-SoftBio datasets with more complex environment, such as occlusions and large viewpoint changes, which suggests that our approach is better at dealing with challenging image sequences than other methods. Meanwhile, it can improve averagely 5.72% comparing with the score level fusion, especially on MARS with large amount of images for each tracklet.

3) Evaluation on CNN Features: It is noted that the results of CNN+OURS+XQDA are worse than CNN+XQDA on Rank-5 and Rank-20 although slightly better in mAP. We further evaluate our method on CNN feature (CNN_d in the following) on the other three datasets iLIDS-VID, PRID 2011, and SAIVT-SoftBio, where the down-sampling is not necessary. Specifically, CNN_d features for these three datasets are directly extracted from the FC7 layer after RELU in AlexNet, where the model weights are pretrained on ImageNet classification task. The final CNN_d feature is 4096-D for each person image. Noted that more sophisticated networks may yield higher accuracy. We provide the comparison between CNN_d+OURS+XQDA and CNN_d+XQDA on Table VII. For CNN_d+XQDA, the averaging pooling is conducted in the same manner as on MARS. It is clear to see that, our subspace learning and clustering method can improve the performance on all the ranks on the three datasets. Compared with the performance on MARS, it suggests that the worse results than CNN+XQDA on some ranks come from down-sampling. Even though, we still achieve slightly higher mAP on MARS.

E. Other Discussion

We have further assessed the proposed method with different metric learning algorithms as well as different feature descriptors. First, we fix the feature representation (LOMO) and evaluate three state-of-the-art metric learning algorithms, including TDL [27], MFA [46], [47], and LFDA [41], [47]

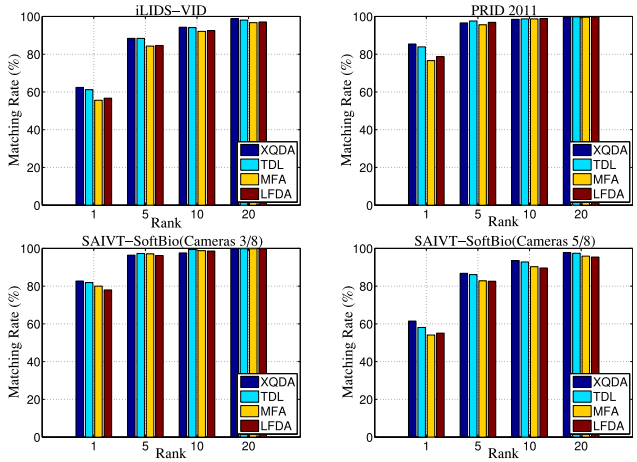


Fig. 6. Comparison results of four metric learning algorithms on three benchmark datasets.

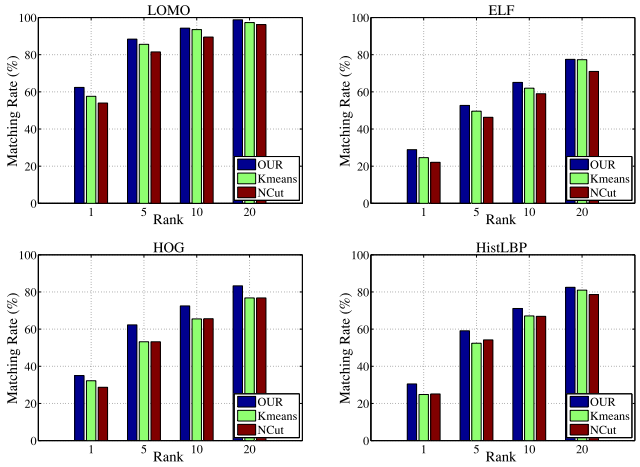


Fig. 7. Comparison results of four features with different clustering methods on iLIDS-VID.

on the first three datasets. The comparison results in Fig. 6 demonstrate the following.

1) Based on the proposed subspace learning method, the performances with different metric learning methods significantly beat the state-of-the-art methods we compared in Tables II–V.
2) All four metric learning methods generally work competitively to each other, which demonstrates the robustness of the proposed subspace learning method.

Second, we evaluate our method with three additional feature descriptors, including ensemble of localized features [4], [48], HOG [30], and hist local binary pattern [47] on iLIDS-VID while fixing the XQDA as the metric learning. It is noted that, as the appearance representation method, the feature descriptor does significantly affect the performance of the Re-ID. In order to fairly demonstrate the contribution of the proposed subspace learning method, we evaluate the Re-ID with same feature descriptor by three different clustering methods, including OURS (nonnegative sparse subspace learning and followed by NCut clustering), *K*-means (directly utilizing *K*-means clustering on the features), and NCut (where

affinity matrix is directly defined according to the Euclidean distance matrix of the features). The comparison results are reported in Fig. 7. From Fig. 7, we can see: 1) LOMO feature descriptor significantly outperforms the other three feature descriptors; 2) for any feature descriptor, our subspace learning method can achieve considerable improvement; and 3) for each fixed feature descriptor, the matching rates significantly decrease in K-means and NCut which indicates that our subspace learning approach plays an important role for multishot Re-ID. It is noted that the conventional clustering methods cannot achieve satisfactory performance. The fact that our subspace learning and clustering method outperforms either the conventional K-means and NCut or the LOMO+XQDA baseline validates its effectiveness for person Re-ID.

V. Conclusion

In this paper, we have presented a novel subspace learning method for multishot person Re-ID. We propose to construct the person data by a nonnegative sparse LRR, which can better capture the global structure (by low-rankness) and local linear structure (by sparseness) of the data simultaneously, and ensures the nonnegative weights of the graph for future clustering. Furthermore, we employed the internal image statistical prior to the representation to refine the low-rank affinity matrix. Experiments on four challenging multishot person Re-ID datasets demonstrate the promising performance of the proposed method. In future work, we shall further explore the robust spatial–temporal features for person sequence and more intelligent scheme to remove the outliers.

References

[1] Q. Wu, Z. Wang, F. Deng, Z. Chi, and D. D. Feng, "Realistic human action recognition with multimodal feature selection and fusion," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 43, no. 4, pp. 875–885, Jul. 2013.

[2] B.-W. Chen, C.-Y. Chen, and J.-F. Wang, "Smart homecare surveillance system: Behavior identification based on state-transition support vector machines and sound directivity pattern analysis," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 43, no. 6, pp. 1279–1289, Nov. 2013.

[3] A. Aztiria, J. C. Augusto, R. Basagoiti, A. Izaguirre, and D. J. Cook, "Learning frequent behaviors of the users in intelligent environments," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 43, no. 6, pp. 1265–1278, Nov. 2013.

[4] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Marseille, France, 2008, pp. 262–275.

[5] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, San Francisco, CA, USA, 2010, pp. 2360–2367.

[6] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, 2015, pp. 2197–2206.

[7] C.-C. Guo, S.-Z. Chen, J.-H. Lai, X.-J. Hu, and S.-C. Shi, "Multi-shot person re-identification with automatic ambiguity inference and removal," in *Proc. Int. Conf. Pattern Recognit. (ICPR)*, Stockholm, Sweden, 2014, pp. 3540–3545.

[8] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 688–703.

[9] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by discriminative selection in video ranking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 12, pp. 2501–2514, Dec. 2016.

[10] Y. Li, Z. Wu, and R. J. Radke, "Multi-shot re-identification with random-projection-based random forests," in *Proc. Win. Conf. Appl. Comput. Vis. (WACV)*, 2015, pp. 373–380.

[11] Y. Li, Z. Wu, S. Karanam, and R. J. Radke, "Multi-shot human re-identification using adaptive fisher discriminant analysis," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2015, pp. 1–12.

[12] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Haifa, Israel, 2010, pp. 663–670.

[13] W. Jia, R.-X. Hu, Y.-K. Lei, Y. Zhao, and J. Gui, "Histogram of oriented lines for palmprint recognition," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 44, no. 3, pp. 385–395, Mar. 2014.

[14] X. Liu, L. Lin, and A. L. Yuille, "Robust region grouping via internal patch statistics," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Portland, OR, USA, 2013, pp. 1931–1938.

[15] C. Li, L. Lin, W. Zuo, S. Yan, and J. Tang, "SOLD: Sub-optimal low-rank decomposition for efficient video segmentation," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, 2015, pp. 5519–5527.

[16] C. Lang, G. Liu, J. Yu, and S. Yan, "Saliency detection by multitask sparsity pursuit," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 1327–1338, Mar. 2012.

[17] X. Zhou, C. Yang, and W. Yu, "Moving object detection by detecting contiguous outliers in the low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 597–610, Mar. 2013.

[18] X.-Y. Jing *et al.*, "Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, 2015, pp. 695–704.

[19] S. Chi *et al.*, "Multi-task learning with low rank attribute embedding for person re-identification," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, 2015, pp. 3739–3747.

[20] J. Wright *et al.*, "Sparse representation for computer vision and pattern recognition," *Proc. IEEE*, vol. 98, no. 6, pp. 1031–1044, Jun. 2010.

[21] A. Faktor and M. Irani, "Video segmentation by non-local consensus voting," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2014, pp. 1–12.

[22] S. X. Yu and J. Shi, "Multiclass spectral clustering," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Nice, France, 2003, pp. 313–319.

[23] M. Goffredo, I. Bouchrika, J. N. Carter, and M. S. Nixon, "Self-calibrating view-invariant gait biometrics," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 4, pp. 997–1008, Aug. 2010.

[24] D. Simonnet, M. Lewandowski, S. A. Velastin, J. Orwell, and E. Turkbeyler, "Re-identification of pedestrians in crowds using dynamic time warping," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Florence, Italy, 2012, pp. 423–432.

[25] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Portland, OR, USA, 2013, pp. 3586–3593.

[26] M. De Marsico, M. Nappi, D. Riccio, and H. Wechsler, "Robust face recognition for uncontrolled pose and illumination changes," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 43, no. 1, pp. 149–163, Jan. 2013.

[27] J. You, A. Wu, X. Li, and W.-S. Zheng, "Top-push video-based person re-identification," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 1345–1353.

[28] L. Zhuang *et al.*, "Constructing a nonnegative low-rank and sparse graph with data-adaptive features," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3717–3728, Nov. 2015.

[29] C. Li, L. Lin, W. Zuo, W. Wang, and J. Tang, "An approach to streaming video segmentation with sub-optimal low-rank decomposition," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 1947–1960, May 2016.

[30] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, San Diego, CA, USA, 2005, pp. 886–893.

[31] E. Amaldi and V. Kann, "On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems," *Theor. Comput. Sci.*, vol. 209, nos. 1–2, pp. 237–260, 1998.

[32] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution," *Commun. Pure Appl. Math.*, vol. 59, no. 6, pp. 797–829, 2006.

[33] M. Chen *et al.*, "Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix," *J. Marine Biol. Assoc. United Kingdom*, vol. 56, no. 3, pp. 707–722, 2009.

[34] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, 2010.

[35] B. Moghaddam, T. Jebara, and A. Pentland, "Bayesian face recognition," *Pattern Recognit.*, vol. 33, no. 11, pp. 1771–1782, 2000.

[36] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Providence, RI, USA, 2012, pp. 2288–2295.
[37] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Proc. Scandinavian Conf. Image Anal. (SCIA)*, 2011, pp. 91–102.
[38] A. Bialkowski, S. Denman, S. Sridharan, C. Fookes, and P. Lucey, "A database for person re-identification in multi-camera surveillance networks," in *Proc. Digit. Image Comput. Techn. Appl. (DICTA)*, Fremantle, WA, Australia, 2012, pp. 1–8.
[39] L. Zheng *et al.*, "MARS: A video benchmark for large-scale person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands, 2016, pp. 868–884.
[40] T. Joachims, "Optimizing search engines using clickthrough data," in *Proc. Int. Conf. Knowl. Disc. Data Min.*, Edmonton, AB, Canada, 2002, pp. 133–142.
[41] M. Sugiyama, "Local fisher discriminant analysis for supervised dimensionality reduction," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Pittsburgh, PA, USA, 2006, pp. 905–912.
[42] S. Karanam, Y. Li, and R. J. Radke, "Sparse re-id: Block sparsity for person re-identification," in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Boston, MA, USA, 2015, pp. 33–40.
[43] S. Karanam, Y. Li, and R. J. Radke, "Person re-identification with discriminatively trained viewpoint invariant dictionaries," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, 2015, pp. 4516–4524.
[44] J. Garcia, N. Martinel, G. L. Foresti, A. Gardel, and C. Micheloni, "Person orientation and feature distances boost re-identification," in *Proc. Int. Conf. Pattern Recognit. (ICPR)*, Stockholm, Sweden, 2014, pp. 4618–4623.
[45] L. Zheng *et al.*, "Scalable person re-identification: A benchmark," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, 2015, pp. 1116–1124.
[46] S. Yan *et al.*, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.
[47] F. Xiong, M. Gou, O. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Zürich, Switzerland, 2014, pp. 1–16.
[48] R. Layne, T. M. Hospedales, S. Gong, and Q. Mary, "Person re-identification by attributes," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2012, pp. 1–11.

Aihua Zheng received the B.Eng. and Master-Doctor degrees combined program in computer science and technology from Anhui University, Hefei, China, in 2006 and 2008, respectively, and the Ph.D. degree in computer science from the University of Greenwich, London, U.K., in 2012.

She is currently a Lecturer with Anhui University. Her current research interests include visual-based signal processing and pattern recognition.

Xuehan Zhang received the B.E. degree in information and computing science from Hefei Normal University, Hefei, China, in 2015. He is currently pursuing the M.S. degree in computer science and technology with Anhui University, Hefei.

His current research interest includes person reidentification.

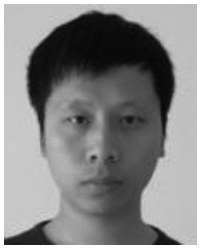
Bo Jiang received the B.S. degree in mathematics and applied mathematics and the M.Eng. and Ph.D. degrees in computer science from Anhui University, Hefei, China, in 2009, 2012, and 2015, respectively.

He is currently an Associate Professor in Computer Science with Anhui University. His current research interests include image feature extraction and matching, and data representation and learning.

Dr. Jiang is currently an Associate Editor of *Cognitive Computation* journal and the Associate Chair of IAPR-TC15.

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.



Chenglong Li received the M.S. and Ph.D. degrees from the School of Computer Science and Technology, Anhui University, Hefei, China, in 2013 and 2016, respectively.

From 2014 to 2015, he was a visiting student with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. He is currently a Lecturer with the School of Computer Science and Technology, Anhui University, and also a Postdoctoral Research Fellow with the Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China.

Dr. Li was a recipient of the ACM Hefei Doctoral Dissertation Award in 2016.