Multimed Tools Appl (2018) 77:10715–10732
DOI 10.1007/s11042-017-4893-5

Sparse coding with cross-view invariant dictionaries for person re-identification

Yunlu Xu[1] · Jie Guo[1] · Zheng Huang[1] · Weidong Qiu[1]

Received: 24 March 2017 / Revised: 29 May 2017 / Accepted: 30 May 2017/
Published online: 8 June 2017

Abstract The task of matching observations of the same person in disjoint views captured by non-overlapping cameras is known as the person re-identification problem. It is challenging owing to low-quality images, inter-object occlusions, and variations in illumination, viewpoints and poses. Unlike previous approaches that learn Mahalanobis-like distance metrics, we propose a novel approach based on dictionary learning that takes the advances of sparse coding of discriminatingly and cross-view invariantly encoding features representing different people. Firstly, we propose a robust and discriminative feature extraction method of different feature levels. The feature representations are projected to a lower computation common subspace. Secondly, we learn a single cross-view invariant dictionary for each feature level for different camera views and a fusion strategy is utilized to generate the final matching results. Experimental statistics show the superior performance of our approach by comparing with state-of-the-art methods on two publicly available benchmark datasets VIPeR and PRID 2011.

Keywords Dictionary learning · Sparse coding · Person re-identification · Intelligent surveillance

1 Introduction

The person re-identification problem is known as recognizing an individual at a different location from non-overlapping camera views. It has huge potentials for security and safety management applications and thus has received increasing attention in the past years. Despite great efforts from researchers in this field, it remains an unsolved problem. It is

✉ Jie Guo
guojie@sjtu.edu.cn

[1] Department of Information Security Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China

challenging owing to low-quality images, inter-object occlusions, and variations in illumination, viewpoints and poses as can be seen samples from Fig. 1. The procedure of re-identifying a person is generally under two main steps.

The first step is constructing distinctive feature descriptors. Common approaches apply the descriptor to character the color and textural information of human clothes [4, 6, 8, 9, 20, 31, 51]. A good descriptor is expected to only have the discriminative power to distinguish different people but also possess the robustness of intra-personal variations. Since the images from different camera views are in low resolution and have large variations, it has been proved that color is the most valuable cue for person re-identification, i.e. color histograms and color name descriptors [44]. Liu et al. [27] employs a strategy of the color feature ensemble to extend the distinct inter-personal ability and invariant intra-personal power of different color features. Because color alone can hardly sufficiently differentiate different persons of similar colors, texture descriptors such as Local Binary Pattern (LBP) [42] and filters (i.e. Gabor filters and Schimid filters [9]) are often combined with color descriptors. However, it is difficult to unify feature representation of cross-view individuals and to obtain both distinct and robust low-level features under severe condition variations.

The next step of feature extraction is metric distance computation. Commonly, the dimensions of informative feature descriptors are extremely high, direct distance computation does not work. So metric is often obtained by learning a proper similarity for images from different camera views. The widely used traditional method ground on the idea of learning, in a supervised fashion, a Mahalanobis-like distance metric:

$$d_M(y_1, y_2) = (y_1 - y_2)^T M(y_1 - y_2) \tag{1}$$

where M is a positive semi-definite matrix [5, 15, 19, 34, 42, 50].

To improve performance in both steps, we propose a sparse coding method based on the fusion of different level features. We firstly extract both the stripe-level and patch-level features to deal with misalignment of images from different camera views. In our stripe-level

Fig. 1 Example pairs of images from the VIPeR dataset [9]

features, we maximize the local occurrence of each histogram bin. And in the patch-level feature we find the most similar neighbor patch as a corresponding feature representation. In addition, to solve the high-dimensional feature problem, we use a subspace transformation to a lower computation space. With representation descriptors at different feature levels, we do not follow the existing popular state-of-the-art Mahalanobis distance approaches. Instead, we focus on the dictionary learning method and learn a single viewpoint invariant dictionary for each feature level. It strengthens the capability of holding the invariant features between different camera views.

The major contribution of this paper is twofold. Firstly, as our work belongs to the category of distance learning for person re-identification, we adopt a different idea from those popular Mahalanobis-like metric learning algorithms. Dictionary learning for person re-identification is a new direction, where only a few researchers have explored before. Among dictionary learning trials, different from existing methods learning separate sub-dictionaries for each data class or each camera view, our work straightforwardly learns a single viewpoint invariant dictionary for all camera views. Secondly, we model our input images in different representation levels, denoted by the stripe-level and the patch-level, followed by subspace learning, jointly for discrimination of each individual and robustness to pose and camera view variations.

The rest of the paper is organized as follows: In Section 2, we review the related works. In Section 3, we analyze the overall frameworks of our re-identification approach. We propose our feature extraction method and sparse coding framework with viewpoint invariant dictionaries for re-identification and describe its mathematical insight in Section 4. In Section 5, we show the experimental results of our proposed approaches and compare our performance with the state-of-the-art methods. The conclusions of the work are presented in Section 6.

2 Related work

Most existing approaches of the re-identification problem can be divided into two categories: feature representations and metric learning.

2.1 Feature extraction

Feature extraction base methods aim at the exploration of extracting a robust and distinctive representation under various conditions, including illumination, camera view and pose change [4, 9, 20, 31]. Gray and Tao [9] propose an ensemble of localized color and texture features. Shape and structural constraints has been exploited when SDALF [4, 6] takes the symmetry and asymmetry property into account to handle viewpoint changes. Recently, Liao et al. [23] propose a superior method by maximizing the horizontal occurrence of local features to make a stable representation against viewpoint changes and is reported outstanding performance for person re-identification. And Zhao et al. [49] use an unsupervised learned salience model for patch matching. The eBicov [31] method computes a covariance based bio-inspired feature. Semantic attribute representations [20] generate low-dimensional attribute descriptions, similar to descriptions provided verbally to an eyewitness, which is more straightforward to human understanding. To further improve the performance by taking advantage of the complementary of multiple feature representations, Zheng et al. [51] propose an effective feature fusion method to combine multiple feature extraction methods. Deep learning based methods have a good ability of extracting features in variant fields and tasks [21, 28–30, 43] and are also successfully applied in person

re-identification problems as proposed in [2, 41]. Wu et al. [2] use a combination of handcraft features and deep features on a singe image to obtain an enhanced representation. Lin et al. [41] deeper the weight layer to 10 layers using 3 $\times$3 convolution filters. Although the existing feature representations achieve good re-identification performance, the multiple scales and different levels haven't been extensively studied. Assuming that distinguishing level of feature descriptors hold complementary information, our approach models both the stripe-level and patch-level feature representation to achieve a comprehensive effect for better re-identification result.

Our use of two-level features for person re-identification is motivated by the recently proposed Local Maximal Occurrence [42] and the densely sampled patch descriptor [49]. The former locally constructs a histogram of pixel features and then takes its maximum values with horizontal stripes to overcome viewpoint variations while the latter densely computes the descriptor on each patch. Differently, in the case of patch-level feature extraction, we utilize the k-Nearest-Neighbor algorithm thoughts to generate the descriptors for gallery images in each pixel hierarchy. It needs to highlight that our computed descriptor for patch-level feature is asymmetric for probe and gallery (not in the same computation steps), and the k-Nearest-Neighbor algorithm is only applied to the gallery images rather than both the probe and the gallery ones. Details are demonstrated in Section 4.1.2.

2.2 Metric learning

Metric learning base methods ground on the idea of obtaining a proper similarity metric for images from different camera views by learning classifiers and metrics. Most researches in this category is to learn a Mahalanobis-like distance metric. KISSME [19] obtain the distance metric by computing the difference between the intra-class and inter-class covariance matrix. PRDC [50] derives the Mahalanobis metric by maximizing the probability of that a pair of the same individual has a smaller distance than that of different individuals. LFDA [34] maximizes the inter-class distance while preserving the multi-class modality. PCCA [15] conducts sparse pairwise constraints on the similarity metric. Regarding the performance bottleneck due to the inherent linearity of the Mahalanobis-like distance metric, Xiong et al. [42] propose kernel-based algorithms to learn non-linear distance metrics. However, these techniques are still limited by the recognition power of the learned distance metric.

Besides the Mahalanobis learning based approachds, dictionary learning has not received much attention for the person re-identification problem, although dictionary learning has achieved impressive performance in classification and recognition problems [16] in recent years. With the label information and sparse representation over the learned dictionary, the classification-oriented dictionary has strong representational power. K-SVD [1], discriminative K-SVD [46], K-SVD with label consistency constraints [13] and projective dictionary pair [10] are among the most popular dictionary learning methods in recent trends.

To the best of our knowledge, the very few applications of dictionaries in re-identification are as follows. Liu et al. [26] present an approach of semi-supervised coupled dictionaries in the assumption that a pair of patches in separate views should have the similar coding. Jing et al. [14] propose a semi-coupled low-rank discriminant dictionary learning approach to super-resolution person re-identification and apply low-rank regularization in dictionary learning procedure to characterize intrinsic feature space of high-resolution and low-resolution, but both the methods of learning coupled dictionaries mentioned above pose difficulties when the dimensionality of the training data increases. Most recently, Karanam et al. [16] learn a single dictionary instead of several dictionaries to represent both gallery

and probe images and discriminatingly apply explicit constraints on the corresponding sparse codings of image pairs in training the dictionary. However matching rate has not been improved remarkably compared to the classic metric learning methods.

We are dealing with a supervised re-identification problem with dictionary learning and sparse coding. Instead of multiple sub-dictionaries to learn, we reconstruct a single cross-view dictionary for all the camera views. Furthermore, our model is learned using the alternating directions framework by alternately fixing part of the variables and optimizing over the other variable.

3 Frameworks

We separately learn a single dictionary for stripe-level and patch-level feature representations, denoted as D_{stripe} and D_{patch} respectively. The dictionary is assumed viewpoint invariant, so it is the same one for both the probe and gallery camera views. In our work, the stripe-level and patch-level dictionaries and sparse codings are trained independently and the distance between individuals computed based on the learned dictionary are fused to output a final result. Figure 2 shows the general framework of our method.

In the *training phase*, given image sequences in both probe and gallery camera views, we first compute their two-level representative feature vectors. After transforming the feature descriptors into a subspace, stripe-level and patch-level dictionaries are independently learned through the framework of alternating directions optimization.

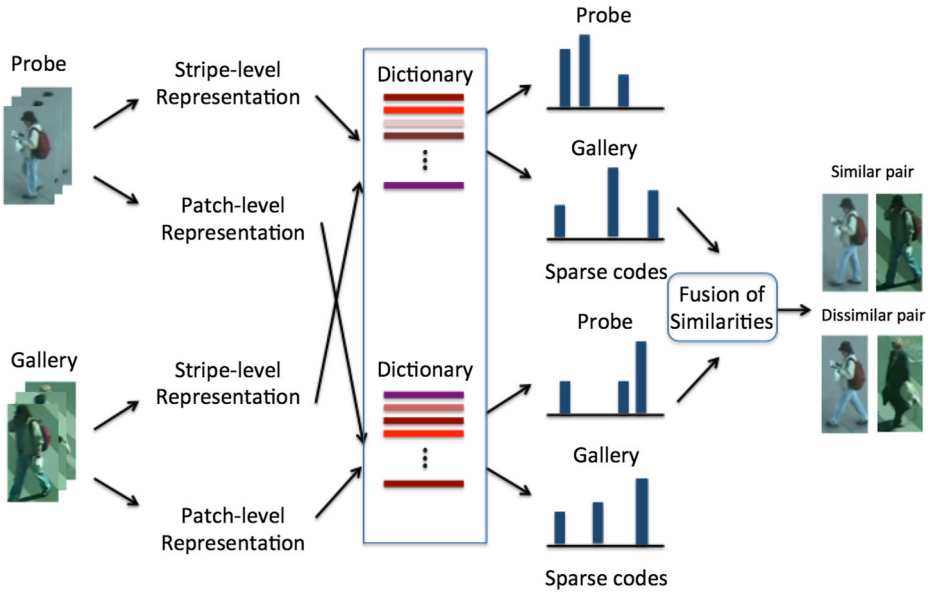

Fig. 2 A visual summary of our framework. Given images in both gallery and probe camera views for n persons, we first compute their representative feature vectors and then iteratively train a discriminative viewpoint invariant dictionary and sparse codes for two feature levels. In the test phase, we compute the corresponding sparse codes respect to the learned dictionary and conduct the fusion of similarities to obtain the final prediction

In the *test phase*, for each gallery image and a certain probe image, we generate the feature representation in the subspace and then optimize the corresponding sparse codes with respect to the learned dictionaries D_{stripe} and D_{patch} in the training phase. The distance between the sparse codes of a image pair from two camera views represent their similarities. The final distance is obtained by the fusion strategy and for each probe image, all the matching proposals are ranked by distance.

4 Methodology

In this section, we first introduce our feature representation extraction for pedestrian in the stripe-level and the patch-level. Then we formulate person re-identification problem using the dictionary learning approach.

4.1 Feature representation

We take stripe-level and patch-level features into consideration to get detailed feature representations. Given an image of individuals from a camera view, we first normalize it to a fixed size of 128×48 and we fetch the center region of 120×40 (shaving 4 pixels for all the four directions, top, bottom, left, and right, of the original images), for ignoring the distractions from the image background. A sliding window of size 8×8 is applied to the image with a stride of 4 pixels in both the stripe-level and patch-level feature extraction method.

4.1.1 Stripe-level representation

For the extraction of distinctive features, we apply the idea of the local maximal occurrence (LOMO) feature descriptor, which has demonstrated state-of-the-art performance in re-identification [23]. The LOMO feature descriptor applies HSV color histogram to extract color features and the Scale Invariant Local Ternary Pattern (SILTP) [22] descriptor to extract texture features, which are both proven to have a nice invariant property. Within each sliding window, we compute the HSV histogram and SILTP features in multiple scales. Each histogram bin is the occurrence probability of one pattern in a sub-window. In each stripe (patches at the same horizontal location), we compute the maximization of local occurrences in each pattern. Then the features from each stripe are concatenated and log transformed to produce the final descriptor of $(8 \times 8 \times 8$ color bins $+ 81 \times 2$ texture bins$) \times (29 + 14 + 6$ stripes$) = 33{,}026$ dimensions. The resulting feature representation holds the characteristics of both invariance to viewpoint and distinctiveness for different individuals.

4.1.2 Patch-level representation

The pedestrian images in different camera views usually cause the misalignment problem. Although the former LOMO approach relieves this to some extent, it only take the maximization of the patch features in a horizontal stripe. In this way, other patch information in the same horizontal stripe are neglected thus some valuable information is missing. To remedy the neglect of detailed image feature information, we also apply a method of extracting local patches on a dense grid [49]. We compute the 32-bin color histograms and 128-dim dense SIFT features in each of LAB channels. To obtain detailed information, we compute histograms on two more down sampled scales for each patch with down sampling factor 0.5 and 0.75. All the features of one patch are normalized with l_2-norm. Finally, each

patch is represented by a feature vector and the concatenation of all the patches sequentially constitutes the probe image representation.

In this patch-level descriptor computation process, we propose an asymmetry strategy in which image representations in the probe and the gallery set are obtained in different routes. The gallery representations are generated according to the to-be-matched probe image and for each probe image, all the gallery representations are generated once newly. For image from camera A (the probe set), dense local features are denoted by $x^{A,u} = \{x_{m,n}^{A,u}\}$,where $x_{m,n}^{A,u}$ denotes the features of a patch at the m-th row and n-th column in the u-th image from camera view A (or the probe). For patch $x_{m,n}^{A,u}$, the corresponding patch in the v-th image from camera view B (or the gallery), i.e. $x^{B,v} = \{x_{i,j}^{B,v}\}$, v=1, ... , V, is represented as the constrained search set of $x_{i,j}^{A,u}$ in $x^{B,v}$

$$\begin{aligned} \aleph\left(x_{m,n}^{A,u}, x^{B,v}\right) = \{x_{i,j}^{B,v} | j = 1, ..., N, \\ i = max(0, m-h), ..., min(M, m+h)\}, \end{aligned} \tag{2}$$

where h represents the height of search regions and we follow [49] to set it to 2 in our experiment setting. For each patch $x_{m,n}^{A,u}$ in images from the probe, we do a nearest-neighbor search in its constrained search set $\aleph(x_{m,n}^{A,u}, x^{B,v})$. The nearest k neighbors in the search set is assembled to compute a mean 672-dimensional descriptor of the corresponding patches in the gallery. All the mean descriptors are concatenated to generate a final representation of image in the gallery set.

4.1.3 Fusion strategy

The final distance of a pair image is the fusion of stripe-level and patch-level distances by

$$d_{similarity} = (1-\mu) \times d_{stripe} + \mu \times d_{patch} \tag{3}$$

with a tunable parameter μ: $0 < \mu < 1$. The matching for each probe person is obtained as the index of the minimum value in the computed distances. We later show the performance of the feature fusion strategy between rank and matching rate in Section 5.2.2.

4.2 Dictionary learning

We first briefly introduce the basics of dictionary learning and notations that are used in our work. Then, we explain our approach to learn a sparse coding with viewpoint invariant dictionaries in both dictionary learning and subspace projection stage.

4.2.1 Preliminaries

Given a query image X = { $x_{ij} \mid x_{ij} \in \Re^d$, feature vector of person with index i in the view of camera j}. For the person re-identification problem, we aim to learn a dictionary $D \in \Re^{d \times N}$, that is capable of discriminating the feature vectors of the similar individual and dissimilar individuals from different camera viewpoints. We can compute s_{ij}, the sparse code of feature vectors with respect to learned dictionary D by solving the following problem:

$$s_{ij} = \min_{s} ||x_{ij} - Ds||_2^2 + \lambda ||s||_1 \tag{4}$$

where the first part is an l_2-norm term for the discrimination of the learned dictionary, and the second term ensures the sparsity of the variable s.

4.2.2 Viewpoint invariant dictionary for re-ID

The method in [16] assumes that the images from variant camera views of the same person should possess similar sparse codings corresponding to the learned dictionary. D is expected to satisfy the following two properties: firstly, D should be invariant with large viewpoint changes. Secondly, D should be discriminative between feature representations of the same individual and those of different individuals. Taking these two properties of D into consideration, we formulate the problem as the following overall minimization problem:

$$\begin{aligned}\min_{D,s_{i1},s_{i2}} \sum_{i=1}^{n} \{ & ||x_{i1} - Ds_{i1}||_2^2 + \lambda_1 ||s_{i1}||_1 \\ & ||x_{i2} - Ds_{i2}||_2^2 + \lambda_2 ||s_{i2}||_1 \} \\ s.t.\ & ||s_{i1} - s_{i2}||_2 < ||s_{i1} - s_{j2}||_2, \forall j \neq i, \forall i \end{aligned} \tag{5}$$

For solving the optimization problem, Karanam et al. [16] introduces two constants c_1 and c_2 to reformulate the problem:

$$\begin{aligned}\min_{D,s_{i1},s_{i2}} \sum_{i=1}^{n} \{ & ||x_{i1} - Ds_{i1}||_2^2 + \lambda_1 ||s_{i1}||_1 \\ & + ||x_{i2} - Ds_{i2}||_2^2 + \lambda_2 ||s_{i2}||_1 \\ & s.t. \quad ||s_{i1} - s_{i2}||_2 < c_1, \forall i \\ & ||s_{i1} - s_{j2}||_2 < c_2, \forall j \neq i, \forall i \end{aligned} \tag{6}$$

where $c_1 << c_2$

This form is not convex in the three variables, D, s_{i1}, s_{i2} simultaneously, but the objective function is convex in one of the variables when the other two are fixed. Therefore, we utilize the technique of alternating directions to solve the minimization problem. Instead of the existing dictionary methods in the person re-ID, our appliance of the single invariant dictionary holds a better characteristic of representing robustness when camera view changes.

4.2.3 Subspace projection for re-ID

We use $\varphi(\cdot)$ to represent a nonlinear mapping of input representations to a feature subspace $\mathscr{F}$, i.e, $\varphi : x_{ij} \in \Re^m \rightarrow \varphi(x_{ij}) \in \mathscr{F}$, and $\mathscr{D}$ is the resulting dictionary matrices after mapping D and α is the sparse code corresponding to $\mathscr{D}$. So the optimal function is reformulated as:

$$\begin{aligned}\min_{\mathscr{D},\alpha_{i1},\alpha_{i2}} \sum_{i=1}^{n} \{ & ||\varphi(x_{i1}) - \mathscr{D}\alpha_{i1}||_2^2 + \lambda_1 ||\alpha_{i1}||_1 \\ & + ||\varphi(x_{i2}) - \mathscr{D}\alpha_{i2}||_2^2 + \lambda_2 ||\alpha_{i2}||_1 \} \\ & s.t. \quad ||\alpha_{i1} - \alpha_{i2}||_2 < c_1, \forall i \\ & ||\alpha_{i1} - \alpha_{j2}||_2 < c_2, \forall j \neq i, \forall i \end{aligned} \tag{7}$$

where $c_1 << c_2$

The dimensions of stripe-level and patch-level feature representations in the original space are very large, and we expect a low dimensional space $\Re^m (m < d)$ for classification.

Work [23] extended the Bayesian face [32] and KISSME [19] approaches to cross-view metric learning. The original distance function between x_{i1} and x_{j2} is

$$d(x_{i1}, x_{j2}) = (x_{i1} - x_{j2})^T \left(\Sigma_I^{-1} - \Sigma_E^{-1}\right)(x_{i1} - x_{j2}) \tag{8}$$

where Σ_I and Σ_E are the covariance matrices of intrapersonal variances Ω_I and the extrapersonal variations Ω_E.

In [23], a subspace $W = (w_1, w_2, ..., w_m) \in \Re^{d\times m}$ is learned with cross-view data. Given a training set$\{X_1, X_2\}$, where $X_1 = (x_{11}, x_{21}, ..., x_{p1}) \in \Re^{d\times p}$ contains p examples in a d-dimensional subspace from the probe and $X_2 = (x_{12}, x_{22}, ..., x_{q2}) \in \Re^{d\times q}$ contains q examples in a d-dimensional subspace from the gallery.

Then the distance function is formulated as

$$d(x_{i1}, x_{j2}) = (x_{i1} - x_{j2})^T W \left(\Sigma_I^{'-1} - \Sigma_E^{'-1}\right) W^T (x_{i1} - x_{j2}) \tag{9}$$

where $\Sigma_I^{'-1} = W^T \Sigma_I W$ and $\Sigma_E^{'-1} = W^T \Sigma_E W$. We learn the matrix $M(W) = W\left(\Sigma_I^{'-1} - \Sigma_E^{'-1}\right) W^T$ using the XQDA [23] algorithm.

4.2.4 Re-identification scheme

In the test stage, for each gallery feature vector we compute the corresponding sparse codes with respect to the learning dictionary $\mathscr{D}$ as:

$$\alpha_{i2} = \min_{\alpha_{i2}} ||\varphi(x_{i2}) - \mathscr{D}\alpha_{i2}||_2^2 + \lambda_2||\alpha_{i2}||_1, \forall i \tag{10}$$

This is the standard LASSO problem [39], we can use the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) [3] to solve the problem. Alternatively, we use l_2-constraint instead of l_1-norm on coefficients. Thus it will be transformed into a simple least squares problem, which has a closed-form solution and can be solved by derivation. Similarly, the sparse codes α_{i1} of probe images are computed. After obtaining α_{i1} and α_{i2} for all persons in the test set, their similarity is obtained by computing the cosine distance between their respective sparse code vectors.

5 Experiments

In this section, we evaluate the performance of our proposed viewpoint invariant representation based dictionary learning method. We show the assessment of different strategies in the experimental results in Section 5.2 The experimental results of our proposed approaches and the comparison with the state-of-the-art methods are shown in detail in Section 5.3.

5.1 Experimental settings

Datasets We choose two public baseline datasets, i.e. the VIPeR dataset [11] and the PRID 2011 dataset [36] for testing our proposed approach. Both are very challenging datasets for the person re-identification task because their pictures show great significant variations in illuminations, viewpoints, poses, and are of low resolutions, even with occlusions and background clutters. *The VIPeR dataset* [11] is mostly widely used and relatively large. It consists of 632 people, each of which has two images scaled to 128 × 48 pixels captured in two different outdoor views from arbitrary viewpoints under varying illumination conditions. In our experiments, the whole set of 632 image pairs are randomly divided into two

even sets of 316 pairs, one for training and the other for testing. *The PRID 2011 dataset* [40] consists of image sequences in two non-overlapping cameras. Camera A contains 385 individuals, and camera B contains 749 individuals, in which 200 individuals appearing in both cameras. The images are captured in an outdoor environment with viewpoint, pose and illumination changes. It is generally considered being even more challenging than VIPeR since it includes distractors as well as strong cross-camera changes. In our experiments we use the single shot version of the dataset as in [12, 18], i.e. one image per person for both views. In each process of data split, 100 people with one image from both views are randomly chosen for the training set from the 200 ones appearing in both cameras, while for testing, the remaining 100 of camera A are used as the probe set, and remaining 649 of camera B are used as gallery (also containing the 100 people in the probe set).

Evaluation protocol All the quantitative results are reported in standard Cumulated Matching Characteristics (CMC) curves. The CMC curve is a plot of the re-identification rate with respect to the rank [17]. We randomly partition the dataset into two even parts, for training and testing respectively, without overlap on person identities. Im ages from a camera view are used as probe and those from a different camera view as gallery. Each probe image is matched with every image in gallery, and the rank of correct match is obtained. Rank-k matching rate is the expectation of correct match at rank k, and the cumulated values of recognition rate at all ranks is recorded as CMC result.

Parameter settings All orginal images are resized to 128×48 pixels for later feature computation. We set the subspace dimension $m = 315$ for the VIPeR dataset and $m = 150$ for the PRID dataset. The regularization parameters are set as $\lambda_1 = 0.001$ and $\lambda_2 = 0.001$. The pre-defined constant μ for similarity fusion of different feature levels is set to 0.5 by default. To make a convincing comparison, we repeat a random partition of training and test set 10 times, then evaluate and report the expectation of recognition results.

5.2 Strategy evaluation

We assess the effect of proposed subspace learning strategy in Section 5.2.1. Effects of feature level fusion is discussed in Section 5.2.2.

5.2.1 Subspace learning evaluation

The distinctive feature representations consisting of concatenating complementary feature descriptors are high-dimensional, which result in a prohibitive computational cost. An alternative way is to transform the original space to a low-dimensional subspace that maintains the computational cost much more acceptable and enforces features from the same individual to be closer than features from different individuals yet improves the results. Local Fisher Discriminant Analysis (LFDA) [34] is applied in [16] and proved as a particularly suitable approach. The image sequence for each person displays multi-modality due to occlusions, background and illumination variations. LFDA overcomes the bias of Fisher Discriminant Analysis (FDA) [7] methods and instead preserves the local data structure during the embedding process. The transformations for each pair of probe and gallery individuals are utilized for a better feature representation, and differently from [16], in our experiments, we employed a different XQDA [23] subspace learning method.

Based on our feature extraction methods in two levels, we evaluate the proposed subspace using XQDA [23] with original feature space and Local Fisher Discriminant Analysis

(LFDA) [34] representation space. Table 1 illustrates the experimental results using subspace learning XQDA of the feature representation. As can be seen from Table 1, compared to the original feature space and the transform of top linear transformation algorithm LFDA, our employed algorithm reveals greater performance in the matching result. We attribute the improvement to the implementation of the subspace learning strategy. Because of the advanced results, we use XQDA in the rest of our experiments in this paper.

5.2.2 Feature level fusion evaluation

Computation cost is the bottleneck when concatenating different features, so researches often select the most effective features. Ensemble of method both in feature and extraction [51] is a popular field. We assume that different levels of features embrace complementary useful information, so we fuse the patch-level and stripe-level feature. Of course, in our consideration, the fusion of other effective features can also improve the results. Our experimental exploration is to highlight the contributions of complementary feature fusion. Figure 3 shows the CMC curve of average matching rates on VIPeR dataset [9]. In the figure, we can conclude that the fusion of our choosing two levels of feature representations achieves better performance than the use of only one type of feature representation.

The obvious improvement of matching result shows the effect of feature fusion. The two level of features, consisted of color and texture descriptor, help each other in detailed and invariant information and improve 2.7% and 4.5% at rank-1 accuracy respectively compared to stripe-level and patch-level only.

5.3 State-of-the-art comparisons

In this section, we give a comprehensive evaluation on our proposed method. The overall results of our method and the comparison with related dictionary learning methods and state-of-the-art approaches are analyzed on the VIPeR and PRID 2011 dataset in Sections 5.3.1 and 5.3.2.

5.3.1 Comparison with dictionary learning methods

We evaluate the performance of the top dictionary learning methods for person re-identification. The recent results of learned dictionaries reported on VIPeR dataset includes SSCDL [26], CPDL [37], SLDDL [14], ISR [24], UMDL [35]. The results on the VIPeR dataset are shown in Table 2. Statistics in the table shows that the recent UMDL [35] and CPDL [37] obtain better recognition rate than the other former proposed method. It is valuable to highlight that the UMDL [35] is a novel cross-dataset transfer learning approach and it is unsupervised in the sense that the target dataset is completely unlabelled which does not rely on manually labelled data any more thus has a much larger application prospect than the supervised ones that largely rely on the manually labelled data. However, the recognition

Table 1 Re-identification accuracy (%) evaluation of subspace learning on the VIPeR dataset

The best results in comparison are in bold

Method	rank=1	rank=5	rank=10	rank=20
Original	35.6	59.8	70.4	86.1
LFDA [34]	37.5	62.4	73.6	88.2
XQDA [23]	**45.6**	**73.5**	**83.5**	**92.6**

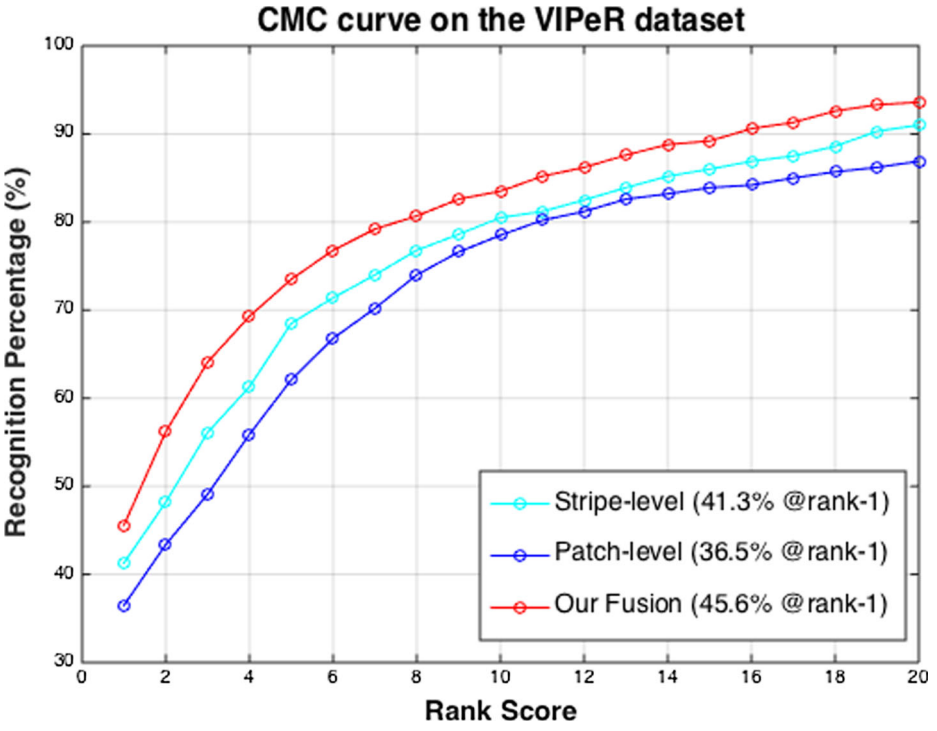

Fig. 3 Re-identification accuracy comparison of stripe-level only, patch-level only and the fusion model on VIPeR dataset

accuracy of the unsupervised method UMDL has not achieved the best. We can see from the results that our dictionary learning method presented the highest matching rate when compared to these dictionary learning methods on the VIPeR dataset.

When it comes to the PRID dataset, similarly to the VIPeR dataset, we collect nearly all the reported results of dictionary learning methods for the person re-identification problem (to the best of our knowledge). Top reported results consists of SSCDL [26], SLD^2L [14], ISR [24], UMDL [35], Kordirov's [18]. According to the results on the PRID dataset shown in Table 3 and similarly to the results on the VIPeR dataset, we achieve improved performance when compared to these dictionary learning methods. As matter of fact, we improved the best reported rank-1 accuracy by 1.7% and reached the best rank-1 matching rate at 26.9% reported in literature.

Table 2 Re-identification accuracy (%) evaluation of the learned dictionaries on the VIPeR dataset

Method	rank=1	rank=5	rank=10	rank=20
SSCDL [26]	25.6	53.7	68.1	83.6
CPDL [37]	34.0	64.2	77.5	88.6
SLD^2L [14]	16.9	41.2	58.1	79.0
ISR [24]	27.0	49.8	61.2	73.0
UMDL [35]	31.2	–	–	–
Ours	**45.6**	**73.5**	**83.5**	**93.6**

The best results in comparison are in bold

Table 3 Re-identification accuracy (%) evaluation of the learned dictionaries on the PRID dataset

Method	rank=1	rank=5	rank=10	rank=20
SSCDL [26]	4.8	16.0	32.6	48.4
SLD^2L [14]	8.8	22.8	39.2	58.6
ISR [24]	17.0	34.4	42.0	54.3
UMDL [35]	24.2	–	–	–
Kordirov et al. [18]	25.2	51.9	62.9	71.6
Ours	**26.9**	**53.7**	**63.7**	**73.5**

The best results in comparison are in bold

5.3.2 Comparison with the state-of-the-art

We compare our proposed approach with a large number of state-of-the-art method including the best reported approaches on the exploration of metric learning and feature extraction on both the VIPeR and PRID dataset.

VIPeR dataset Table 4 shows the matching rate for different methods reported on the VIPeR dataset including approaches based on metric learning [19, 23, 25, 25, 33, 37, 42, 42], and feature methods [23, 27, 38, 44, 50]. It is worth noticing that the ECM [18, 27, 47, 48] uses an ensemble of color features and achieves 38.9% of rank-1 accuracy better than other method on the exploration of noval feature extractions. We also apply a strategy of multiple features and have a broader range both in feature type and selection of extraction region. In Null Space [47] method, images of the same person are collapsed into a single point thus minimizing the within-class scatter to the extreme and maximizing the relative between-class separation simultaneously and reaching 42.3% of rank-1 accuracy, but differently, we try to transfer the feature space into a common subspace and than learn a dictionary for sparse coding. According to the results, the proposed method reaches accuracy of 45.6% at rank-1 and best results in low ranks, i.e,rank-1 to rank-10, only surpassed by [38] at rank-20 by the recognition rate of 94.9% more than ours 93.6%. It needs to highlight that it

Table 4 Comparison of state-of-the-art results on the VIPeR dataset

Method	rank=1	rank=5	rank=10	rank=20
KISSME [19]	25.4	53.3	67.7	82.1
kLFDA [42]	40.7	70.0	81.2	90.8
kCCA [25]	36.8	–	84.5	92.3
ECM [27]	38.9	–	78.4	88.9
SCNCD [44]	37.8	68.5	81.2	90.4
Mid-level [50]	29.1	52.5	67.1	80.0
Shi et al. [38]	31.1	68.6	82.8	**94.9**
CPDL [37]	34.0	64.2	77.5	88.6
Liu et al. [23]	40.0	–	80.5	91.1
Zhang et al. [48]	42.7	–	84.3	91.9
Null Space [47]	42.3	71.5	82.5	92.1
Ours	**45.6**	**73.5**	**83.5**	93.6

The best results in comparison are in bold

Table 5 Comparison of state-of-the-art results on the PRID dataset

Method	rank=1	rank=5	rank=10	rank=20
KISSME [19]	10.2	26.1	37.4	53.2
kLFDA [42]	19.7	44.9	56.4	65.9
kCCA [25]	14.5	34.3	46.6	59.1
Kodirov et al. [18]	25.2	51.9	62.9	71.6
Ensemble [33]	17.9	–	50.0	62.0
KCCA [25]	15.0	–	38.0	50.0
Ours	**26.9**	**53.7**	**63.7**	**73.5**

The best results in comparison are in bold

needs more attention in lower ranking results since lower ranks makes more sense in real surveillance environment. And in this case, our approach impressively outperforms all the other former existing approaches.

PRID dataset As evidenced by the results Table 5 and similarly to ons on the VIPeR dataset, among the reported experiment results of the state-of-the-arts, KISSME [19], kLFDA [42], kCCA [25], Kodirov's [18], Ensemble [33] and EIML [25] on the single-shot scenery of PRID 2011 dataset, our method of using dictionary learning with viewpoint invariant representations produces favorable performance with the best results of 26.9% at rank-1 recognition rate for the person re-identification problem. Furthermore, the similar performances on the public challenging datasets, the VIPeR and the PRID 2011 dataset, demonstrate the advantage of our overall learning strategy.

5.3.3 Discussion

Notice that our proposed viewpoint invariant dictionary with the fusion of stripe-level and patch-level cross-view variation suppression representations improves re-identification accuracy with respect to other state-of-the-art methods. The proposed method increases from 42.7% to 45.6% and 25.2% to 26.9% at rank-1 on VIPeR and PRID dataset respectively. The important thing we want to highlight is the complementation of different feature levels improve the representative ability of feature descriptors. Subspace learning and dictionary learning methods both effectively obtain the more proper distance metric of individual pairs, which contribute to our experimental improvements in the matching rate.

6 Conclusion

In this paper, we have presented an effective way to re-identify persons in non-overlapping cross-view cameras. We take different levels of feature representations into consideration and fuse them instead of only using one level of feature descriptor. After transferring the high-dimensional feature space into a common subspace, we learn a single viewpoint invariant dictionary and sparse codes for both camera views. With cross-view invariant and distinctive dictionaries, we show our method outperforming other state-of-the-art methods for the person re-identification problem in our experiments on publicly challenging datasets. Due to the time cost of our training process, we plan to extend our work for optimizing our training and testing process to make it more computationally efficient. In addition, due to

the small number of labeled data, we would like to investigate the technique to extend our supervised dictionary learning method to a semi-supervised one with only a few manually labelled data to solve the re-identification problem in our future works.

Acknowledgements This work has been supported by New Century Excellent Talents in University of Ministry of Education under Grant NCET-12-0358, and Program of Shanghai Technology Research Leader under Grant 16XD1424400.

References

1. Aharon M, Elad M, Bruckstein A (2006) K-svd: an algorithm for designing overcomplete dictionaries for sparse representation. Proc IEEE Trans Signal Process 54(11):4311–4322
2. Ahmed E, Jones M, Marks TK (2015) An improved deep learning architecture for person re-identification. In: Proceedings IEEE computer vision and pattern recognition, pp 3908–3916
3. Beck A, Teboulle M (2009) A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J Imag Sci 2(1):183–202
4. Dong SC, Cristani M, Stoppa M, Bazzani L, Murino V (2011) Custom pictorial structures for re-identification. In: Proceedings brit mach vis conf, BMVC, pp 68.1–68.11
5. Engel C, Baumgartner P, Holzmann M, Nutzel JF (2010) Person re-identification by support vector ranking. In: Proceedings brit mach vis conf, BMVC, pp 21.1–21.11
6. Farenzena M, Bazzani L, Perina A, Murino V (2010) Person re-identification by symmetry-driven accumulation of local features. In: Proceedings IEEE conference on computer vision and pattern recognition, pp 2360–2367
7. Fisher BR (2012) The use of multiple measurements in taxonomic problems. Ann Eugen 7(2):179–188
8. Gong S, Cristani M, Yan S et al (2014) Person re-identification, 1st edn
9. Gray D, Tao H (2008) Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: Proceedings eur conf computer vision, pp 262–275. ECCV
10. Gu S, Zhang L, Zuo W, Feng X (2014) Projective dictionary pair learning for pattern classification. In: Proceedings advances in neural information processing systems, pp 793–801
11. Hirzer M, Beleznai C, Roth PM, Bischof H (2011) Person re-identification by descriptive and discriminative classification. In: Proceedings Scandinavian conference on image analysis, vol 6688, pp 91–102
12. Hirzer M, Roth PM, Stinger M, Bischof H (2012) Relaxed pairwise learned metric for person re-identification. In: Proceedings ECCV, vol 7577, pp 780–793
13. Jiang Z, Lin Z, Davis LS (2003) Label consistent K-SVD: learning a discriminative dictionary for recognition. Proc IEEE Trans Pattern Anal Mach Intell 35(11):2651–2664
14. Jing XY, Zhu X, Wu F, You X, Liu Q et al (2015) Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning. In: Proceedings computer vision and pattern recognition, pp 695–704
15. Jurie F, Mignon A (2012) PCCA: a new approach for distance learning from sparse pairwise constraints. In: Proceedings IEEE conf comput vis pattern recog, pp 2666–2672
16. Karanam S, Li Y, Radke RJ (2015) Person re-identification with discriminatively trained viewpoint invariant dictionaries. In: Proceedings IEEE international conference on computer vision, pp 4516–4524
17. Karanam S, Gou M, Wu Z, Rates-Borras A, Camps O, Radke RJ (2016) A comprehensive evaluation and benchmark for person re-identification: features, metrics, and datasets
18. Kodirov E, Xiang T, Gong S (2015) Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification. In: Proceedings British machine vision conference, pp 44.1–44.12
19. Kostinger M, Hirzer M, Wohlhart P, Roth PM, Bischof H (2012) Large scale metric learning from equivalence constraints. In: Proceedings IEEE Conference on computer vision and pattern recognition, pp 2288–2295
20. Layne R, Hospedales TM, Gong S (2014) Attributes-Based Re-identification. In: Proceedings advances in computer vision & pattern recognition, pp 93–117
21. Li Y, Lu H, Li J et al (2016) Underwater image de-scattering and classification by deep neural network. In: Computers and electrical engineering, pp 68–77
22. Liao S, Zhao G, Kellokumpu V, Pietikainen M, Li SZ (2010) Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes. In: Proceedings IEEE computer society conference on computer vision and pattern recognition, pp 1301–1306

23. Liao S, Hu Y, Zhu X, Li SZ (2015) Person re-identification by local maximal occurrence representation and metric learning. Proc IEEE Conf Comput Vis Pattern Recog 8(4):2197–2206
24. Lisanti G, Masi I, Bagdanov AD, Bimbo AD (2013) Person re-identification by iterative re-weighted sparse ranking. Proc IEEE Trans Pattern Anal Mach Intell 37(8):1629–42
25. Lisanti G, Masi I, Bimbo AD (2014) Matching people across camera views using kernel canonical correlation analysis. In: Proceedings of the international conference on distributed smart cameras, ICDSC '14. ACM, New York, pp 10:1–10:6
26. Liu X, Song M, Tao D, Zhou X et al (2014) Semi-supervised coupled dictionary learning for person re-identification. In: Proceedings IEEE conference computer vision and pattern recognition, pp 3550–3557
27. Liu X, Wang H, Wu Y et al (2015) An ensemble color model for human re-identification. In: Applications of computer vision, pp 868–875
28. Lu H, Li B, Zhu J et al (2016) Wound intensity correction and segmentation with convolutional neural networks. In: Concurrency and computation practice and experience
29. Lu HM, Li YJ, Uemura T, Ge ZY, Xu X, He L, Serikawa S, Kim H (2017) FDCNet: filtering deep convolutional network for marine organism classification. In: Multimedia tools and applications, pp 1–14
30. Lu HM, Li YJ, Zhang YD, Chen M, Serikawa S, Kim H (2017) Underwater optical image processing: a comprehensive review. In: Mobile networks and applications, pp 1–12
31. Ma B, Yu S, Jurie F (2012) BiCov: a novel image representation for person re-identification and face verification. In: Proceedings brit. mach. vis. conf., BMVC, pp 57.1–57.11
32. Moghaddam B, Jebara T, Pentland A (2000) Bayesian face recognition. Pattern Recogn 33(11):1771–1782
33. Paisitkriangkrai S, Shen C, Anton VDH (2015) Learning to rank in person re-identification with metric ensembles. In: Proceedings computer vision and pattern recognition, pp 1846–1855
34. Pedagadi S, Orwell J, Velastin S, Boghossian B (2013) Local fisher discriminant analysis for pedestrian re-identification. In: Proceedings IEEE conf. comput. vis. pattern recog., pp 3318–3325
35. Peng P, Xiang T, Wang Y, Pontil M, Gong S, Huang T et al (2016) Unsupervised cross-dataset transfer learning for person re-identification. In: Proceedings IEEE conference on computer vision and pattern recognition, pp 1306–1315
36. Roth PM, Hirzer M, Koestinger M, Beleznai C, Bischof H (2014) Mahalanobis distance learning for person re-identification. In Person re-identification. Springer, pp 247–267
37. Sheng L, Ming S, Yun F (2015) Cross-view projective dictionary learning for person re-identification. In: Proceedings international joint conference on artificial intelligence. IJCAI
38. Shi Z, Hospedales TM, Xiang T (2015) Transferring a semantic representation for person re-identification and search. In: Proceedings computer vision and pattern recognition, pp 4184–4193
39. Tibshirani R (2011) Regression shrinkage and selection via the lasso. J R Stat Soc 73(3):273–282
40. Wang T, Gong S, Zhu X, Wang S (2014) Person re-identification by video ranking. Proc Eur Conf Comput Vis 8692:688–703
41. Wu L, Shen C, Hengel A V D (2016) PersonNet: person Re-identification with deep convolutional neural networks
42. Xiong F, Gou M, Camps O, Sznaier M (2014) Person re-identification using kernel-based metric learning methods. In: Proceedings ECCA, vol 8695, pp 1–16
43. Xu X, He L, Shimada A, Taniguchi RI, Lu H (2016) Learning unified binary codes for cross-modal retrieval via latent semantic hashing. In: Neurocomputing, pp 191–203
44. Yang Y, Yang J, Yan J et al (2014) Salient color names for person re-identification. In: Proceedings European conference on computer vision, pp 536–551
45. Zeng M, Wu Z, Tian C, Hu L (2015) Efficient person re-identification by hybrid spatiogram and covariance descriptor. In: Proceedings IEEE conference computer vision and pattern recognition, pp 48–56
46. Zhang Q, Li B (2010) Discriminative k-svd for dictionary learning in face recognition. In: Proceedings IEEE conference computer vision and pattern recognition, pp 2691–2698
47. Zhang L, Xiang T, Gong S (2016) Learning a discriminative null space for person re-identification, pp 1239–1248
48. Zhang Y, Li B, Lu H, Irie A, Ruan X (2016) Sample-specific svm learning for person re-identification. In: Proceedings IEEE conference on computer vision and pattern recognition, pp 1278–1287

49. Zhao R, Ouyang W, Wang X (2013) Unsupervised salience learning for person re-identification. In: Proceedings IEEE conference on computer vision and pattern recognition, vol 9. IEEE Computer Society, pp 3586–3593
50. Zheng WS, Gong S, Xiang T (2013) Re-identification by relative distance comparison. Proc IEEE Trans Pattern Anal Mach Intell 35(3):653-668
51. Zheng L, Wang S, Tian L, He F, Liu Z, Tian Q (2015) Query-adaptive late fusion for image search and person re-identification. In: Proceedings IEEE conference on computer vision and pattern recognition, CVPR, pp 1741–1750

Yunlu Xu received the B.Eng. degree in information security from Shanghai Jiao Tong University in 2015. She is now a Master candidate at the Computer Science and Technology of Shanghai Jiao Tong University. Her interests include computer vision, pattern recognition and machining learning.

Jie Guo received her PhD at the institute of Image Processing and Pattern Recognition from Shanghai Jiao Tong University in 2003. Since then, she worked as a teacher in the school of Information Security Engineering, Shanghai Jiao Tong University. Her research interests are in pattern recognition and multimedia security, especially in human behavior analysis.

Zheng Huang received the Ph.D. degree in computer science and technology from Shanghai Jiao Tong University, Shanghai, China in 2003. He is currently a Associated Professor in the School of Information Security and Engineering, Shanghai Jiao Tong University. His main research areas include cryptography and machine learning.

Weidong Qiu received the M.S. degree in cryptography from Xidian University, Xi'an, China, in 1998 and Ph.D. degree in computer software theory from Shanghai Jiao Tong University, Shanghai, China, in 2001. He is currently a Professor in the School of Information Security and Engineering, Shanghai Jiao Tong University. He has published more than twenty academic papers on cryptology. His main research areas include cryptography and computer forensics.