Group Re-Identification via Unsupervised Transfer of Sparse Features Encoding

Giuseppe Lisanti*[,1], Niki Martinel*[,2], Alberto Del Bimbo[1] and Gian Luca Foresti[2]
[1]MICC - University of Firenze, Italy
[2]AViReS Lab - University of Udine, Italy
giuseppe.lisanti@unifi.it, niki.martinel@uniud.it
alberto.delbimbo@unifi.it, gianluca.foresti@uniud.it

Abstract

Person re-identification is best known as the problem of associating a single person that is observed from one or more disjoint cameras. The existing literature has mainly addressed such an issue, neglecting the fact that people usually move in groups, like in crowded scenarios. We believe that the additional information carried by neighboring individuals provides a relevant visual context that can be exploited to obtain a more robust match of single persons within the group. Despite this, re-identifying groups of people compound the common single person re-identification problems by introducing changes in the relative position of persons within the group and severe self-occlusions. In this paper, we propose a solution for group re-identification that grounds on transferring knowledge from single person re-identification to group re-identification by exploiting sparse dictionary learning. First, a dictionary of sparse atoms is learned using patches extracted from single person images. Then, the learned dictionary is exploited to obtain a sparsity-driven residual group representation, which is finally matched to perform the re-identification. Extensive experiments on the i-LIDS groups and two newly collected datasets show that the proposed solution outperforms state-of-the-art approaches.

1. Introduction

Person re-identification is the problem of associating a *single* person that moves across disjoint camera views. The open challenges like changes in viewing angle, background clutter, and occlusions have recently yield to a surge of effort by the community [49]. In particular, existing works have focused on seeking either the best feature representations (*e.g.*, [31, 39, 29]) or propose to learn optimal matching metrics (*e.g.*, [26, 30, 53]). Despite obtaining interesting results on benchmark datasets (*e.g.*, [13, 40, 61], such works have generally neglected the fact that in crowded public environments people often walk in *groups*.

(a) (b) (c)

Figure 1. Major group re-identification issues. Examples of: (a) position swap between person in a group; (b) different background between group images; (c) group images with partial occlusion.

We believe that being able to associate the same *group of people* can be a powerful tool to improve classic *single-person* re-identification. Indeed, the appearance of the whole group provides a rich visual context that can be extremely useful to reduce the ambiguity in retrieving those persons that are partially occluded or to understand the behavior of the group over time if a person in the group is missed for a certain period of time.

Group re-identification introduce some additional difficulties with respect to classic person re-identification (see Figure 1). First of all, the focus is no longer on a single subject, hence the visual appearance of all the persons in the group should be considered. The relative displacement of the subjects in a group can be different from camera to camera. Self-occlusions or occlusions generated by other people near by, as well as the fact that an individual in a group may be missing because he/she left the group, bring in additional challenges. Such challenges deny the direct application of existing representation descriptors and

*G. Lisanti and N. Martinel should be considered as joint first-authors.

This research was partially supported by the Social Museum and Smart Tourism, MIUR project no. CTN01_00034_23154_SMST and by the "PREscriptive Situational awareness for cooperative autoorganizing aerial sensor NETworks" project CIG68827500FB.

matching methods for single-person re-identification to the group association problem. *In this work, we investigate the problem of associating groups of people*.

Contribution: The contribution of this work is twofold: i) To handle the spatial displacement configuration of persons in a group, we introduce a visual descriptor that is invariant both to the number of subjects and to their displacement within the image. Such a task is accomplished by ii) introducing a sparse feature encoding solution that leverages on the knowledge that can be acquired from the large quantity of data that is available in the *single-person* re-identification domain and *transfer* it to the *group* re-identification domain in an unsupervised fashion.

To validate the proposed solution, we compare with existing methods on the i-LIDS group benchmark dataset. In addition, to study the behavior of the approach under different conditions, we have collected two new group re-identification datasets. Extensive evaluations demonstrate that better performances than current solutions are obtained on all datasets.

2. Related Work

While being a young field of research, the community has recently produced several works to address the re-identification problem [49]. In the following we provide a brief overview of the most relevant works to our approach.

Single Person Re-Identification: The literature on single person re-identification can be clustered into two main categories: i) direct matching and ii) metric learning-based methods. Works belonging to the first group aim to address the re-identification problem by designing –or learning– the most discriminative appearance feature descriptors. Multiple local and global feature [5, 4, 33] were combined with reference sets [2], patch matching strategies [57], saliency learning [56, 50, 37], joint attributes [43, 24, 27] and camera network-oriented schemes [36]. Among all the methods in this category, to date, the most widely used appearance descriptors are the Gaussian of Gaussian (GOG) [39], the Local Maximal Occurrence (LOMO) [29] and the Weighted Histogram of Overlapping Stripes (WHOS) [31, 22].

Approaches grouped in the second family represent the trend in person re-identification. In particular, metric learning approaches have been proposed by relaxing [19] or enforcing [30] positive semi-definite (PSD) conditions, by considering equivalence constraints [26, 47, 46] or by exploiting the null-space [53]. While most of the existing methods capture the global structure of the dissimilarity space, local solutions [28, 41, 13] have been proposed too. Sample-specific metrics were also investigated in [54]. Following the success of both approaches, methods combining them in ensembles [40, 51, 38] have been introduced. Different solutions yielding similarity measures have also been investigated by proposing to learn listwise [8] and pairwise [60] similarities.

To deal with the re-identification of a single person all such works assume that the provided images represent good detections of a single person only. This limits their application when more than a person appears in the given image.

Group Person Re-Identification: The first work concerning group association over space and time was proposed in [59]. The authors introduced a group representation and matching algorithm based on a learned dictionary. Since then, the literature on this task is limited to two works [6, 48]. Specifically, in [6], independence of persons locations within the group was captured by the covariance descriptor, while in [48], spatio-temporal group features were explored to improve single person re-identification.

Differently from our work, such approaches either assume that background/foreground segmentation masks are available or exploit training data coming from the same domain (*i.e.*, dataset) of the evaluation data.

Other works have addressed the problem of group-based verification [61] and group membership prediction [55], but both tasks still assume that the input datum represents a single person only. Group information was also explored to address visual tracking [52, 20, 3] and behavior analysis [1] among other tasks.

Object Displacement Invariant Descriptors: The most relevant problem in group re-identification is determined by the fact that people often change their positions while walking in a group. A standard approach to deal with a similar problem in image retrieval is to extract a set of local descriptors, encode them and pool them into an image-level signature which is independent from the spatial location.

Research in this area is quite vast, but almost all approaches inherit or extend the Bag-of-Words (BoW) [9], the Vector of Locally Aggregated Descriptors (VLAD) [21] or the Fisher Vector (FV) [44]. In person re-identification such solutions have been explored to encode first and second-order derivatives for each pixel [32] and to address large-scale applications [58]. Similar solutions exploiting an encoding scheme based on dictionary learning have also been proposed in [25, 42].

These works did not address the group re-identification problem and mainly adopted the encoding schemes to deal with extremely high dimensional image descriptors.

The closest work to our approach [59] exploited a classical BoW scheme on densely extracted features and combined them with a proposed global descriptor. In addition, authors assumed that foreground/background segmentation masks were available such that only features extracted for foreground pixels were used to construct visual words for group image representation. Our approach has three key differences with such a work: i) we propose a novel encoding scheme based on dictionary learning; ii) there is no requirement of foreground/background segmentation masks

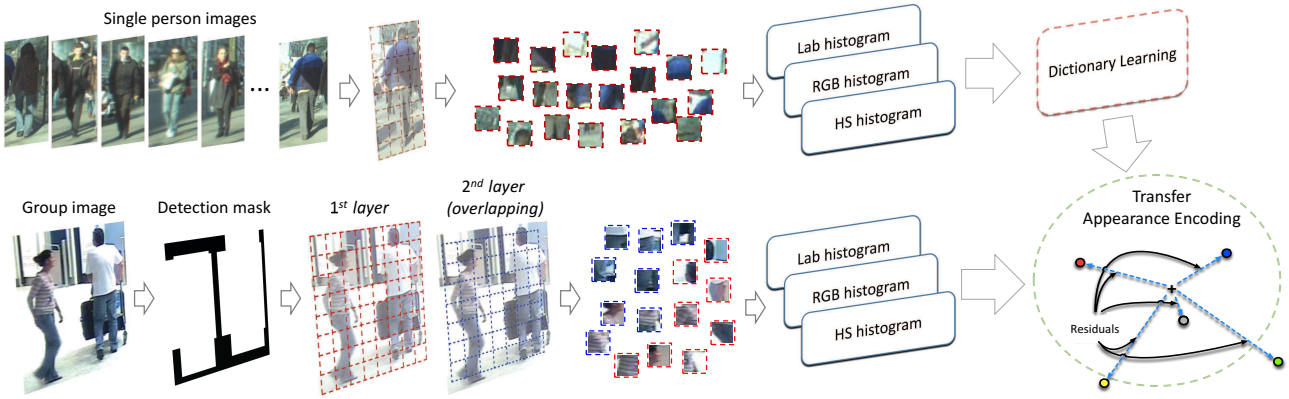


Figure 2. Proposed group re-identification pipeline. Top row shows the unsupervised single-person dictionary learning. Bottom row depicts the re-identification process with feature extraction and subsequent sparse residual encoding obtained with the transferred dictionary atoms.

which demands a substantial hand-work; iii) knowledge obtained from single person re-identification domain is transferred to tackle the group re-identification problem.

3. Proposed Approach

In the following we first define how group appearance is modeled. Then we introduce a transfer learning solution that allows us to exploit knowledge available in single-person re-identification to better tackle the group re-identification. Finally, we describe how groups matching is performed. The whole process is depicted in Figure 2.

3.1. Group Appearance Modeling

In our representation, the image of a group is resized to 128×128 pixels. Two set of patches with fixed dimension 16×16 are then extracted. The first set is obtained from the whole image, whereas the second one is chosen so as to collect information that overlaps with the first layer, see figure 2. For each patch, we compute three histograms considering the same image projected onto different color spaces, namely: HS, RGB and Lab. For the HS images we consider 8 bins for each channel, while for the RGB and Lab we use 4 bins for each channel. This results in a 64 dimensional histogram for each patch and color space (*e.g.*, 8×8 for HS or $4 \times 4 \times 4$ for RGB and Lab).

To obtain a representation that does not preserve location information and is more robust to changes in the group configuration, we separately consider each histogram extracted from each patch (*i.e.*, we do not concatenate them).

Due to the unconstrained patch image subdivision, noisy background information is captured by the feature representation. To circumvent such an issue, we first run three different person detectors, namely Deformable Part Models [12], Aggregated Channel Feature [10] and R-CNN [14]. Then, the filtering mask obtained as the combination of the responses of these three detectors is used to weight the contribution of each pixel in the histogram computation (*i.e.*, pixels belonging to the background have zero contribution).

3.2. Unsupervised Learning of Person Appearance

We propose to exploit a sparse dictionary learning framework that allows us to represent a group of persons as a combination of few human body parts (*i.e.*, patches/atoms). Since these atoms does not necessarily need to be structured accordingly to the relative person displacements, we obtain a flexible group representation. Such a solution resembles visual encoding schemes (*e.g.*, BoW [9], FV [44], VLAD [21]) that are widely adopted for image classification with local descriptors.

We first exploit the dictionary learning solution in [34] to find the basis set of patches that yields to the optimal reconstruction accuracy for single person re-identification. Then, we leverage on such a basis set to introduce a sparse residual group representation.

Problem Definition: Let $\mathcal{I}^{tr} = \{\mathbf{I}_1, \ldots, \mathbf{I}_N\}$ be a training set composed of N images belonging to a *single person* re-identification domain (*i.e.*, images in $\mathcal{I}^{tr}$ may come from the ETHZ [45], CAVIAR [7], or VIPeR [15] dataset). Also let P denote the number of patches into which each image is divided such that $\mathcal{X}^{\text{tr}} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{NP}\}$ is a training set containing NP d-dimensional vectors $\mathbf{x}$, each representing the visual features extracted from a single patch[1].

With this, we define our optimization objective as

$$\mathcal{L}(\mathbf{D}) = \frac{1}{NP} \sum_{i=1}^{NP} l(\mathbf{x}_i, \mathbf{D}) \tag{1}$$

where $\mathbf{D} = [\mathbf{d}_1^T, \ldots, \mathbf{d}_k^T]$, with $\mathbf{d} \in \mathbb{R}^d$ is the dictionary of k atoms to be learned and $l(\cdot, \cdot)$ is a suitable loss function

[1] In our current solution, $\mathbf{x}$ represents a 64-D histogram extracted either from the HS, RGB or Lab color space. These are obtained with a similar approach to the one described in Sec. 3.1 but with input images not processed by the detectors and resized to 128×64.

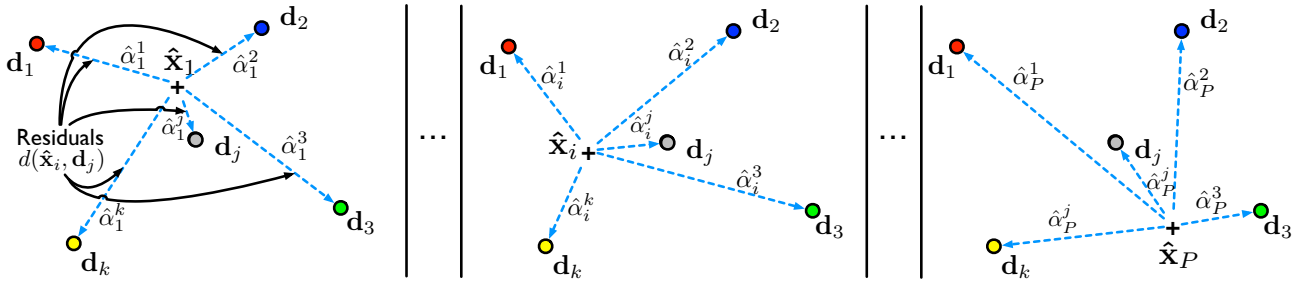


Figure 3. Proposed sparse residual encoding. Colored circles represent the learned dictionary atoms. Black crosses denote the visual features extracted from each of the $\hat{P}$ patches into which the group image is divided. Blue dashed arrows show the residual computed via $d(\cdot,\cdot)$, which are then weighted by the corresponding sparse reconstruction coefficients $\hat{\alpha}$.

such that its output should be "small" if $\mathbf{D}$ is able to provide a good representation for any training input datum $\mathbf{x}_i$.

It has been demonstrated in many fields, ranging from image compression to person re-identification itself [23], that obtaining a representation of a signal $\mathbf{x}$ using only a few elements of a dictionary $\mathbf{D}$ performs better than considering all the atoms. We let our loss function l be the optimal value of the ℓ_1-sparse coding problem, *i.e.*

$$l(\mathbf{x}, \mathbf{D}) = \min_{\boldsymbol{\alpha}} \frac{1}{2}\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1 \tag{2}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^k$ is the sparse vector of coefficients and λ is a regularization parameter that balances the trade-off between a perfect reconstruction and the sparsity of $\boldsymbol{\alpha}$.

By solving the minimization problem in eq. (2), we find the set of atoms in $\mathbf{D}$ that yields the best reconstruction for the signal $\mathbf{x}$. Despite this being compliant to our objective, it does not answer the problem of finding the set of all k atoms that minimize eq. (1).

To address such a problem the ℓ_1-sparse coding problem can be rewritten as a joint optimization whose solution should result in the best combination of dictionary atoms and sparse coefficients. Thus, eq. (2) can be rewritten as

$$l(\mathbf{x}_i, \mathbf{D}) = \min_{\mathbf{D}\in\mathcal{C},\boldsymbol{\Theta}} \frac{1}{NP}\sum_{i=1}^{NP}\left(\frac{1}{2}\|\mathbf{x}_i - \mathbf{D}\boldsymbol{\Theta}_i\|_2^2 + \lambda\|\boldsymbol{\Theta}_i\|_1\right). \tag{3}$$

where $\boldsymbol{\Theta} = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{NP}]$ contains the sparse coefficients to be found for each patch and $\mathcal{C} = \{\mathbf{D} \in \mathbb{R}^{d\times k} | \mathbf{d}_j^T\mathbf{d}_j \leq 1,\ \forall j = 1, \ldots, k\}$ is a convex set introducing an ℓ_2 norm constraint on the single atoms.

Optimization Solution: The problem in eq. (3) is not jointly convex, but convex with respect to each of the two variables when the other one is fixed. To solve the optimization problem, the solution proposed in [34] is exploited. It alternatively solves the classical sparse coding first, then it updates the learned dictionary using the so computed optimal sparse coefficient.

Specifically, let t denote the optimization iteration counter. Also let $\mathbf{D}_t$ be a randomly initialized dictionary, and $\mathbf{A}_t \in \mathbb{R}^{k\times k} = \mathbf{0}$ and $\mathbf{B}_t \in \mathbb{R}^{k\times k} = \mathbf{0}$ (with $t = 0$) be two matrices which will carry the information of all the sparse coefficients $\boldsymbol{\alpha}$'s. Then, we start the optimization by randomly drawing an image training sample from the training set and computing the visual representation of a randomly chosen patch $\mathbf{x}_t$. Such a datum is then considered by the *least angle regression* (LARS) [11] to solve the sparse coding problem in eq. (2), hence to obtain the vector of sparse coefficients for the t-th iteration $\boldsymbol{\alpha}_t$.

The computed sparse coefficient vector is then exploited to revise $\mathbf{A}_t$ and $\mathbf{B}_t$ such that these can be used in a block-coordinate descent solution to update the learned dictionary. More precisely, the two matrices carry all the information brought in by all the sparse coefficients computed so far as

$$\mathbf{A}_t = \mathbf{A}_{t-1} + \boldsymbol{\alpha}_t\boldsymbol{\alpha}_t^T \tag{4}$$

$$\mathbf{B}_t = \mathbf{B}_{t-1} + \mathbf{x}_t\boldsymbol{\alpha}_t^T. \tag{5}$$

Exploiting the block coordinate descent to update the dictionary $\mathbf{D}_t$ yields to the following solution for each dictionary atom, *i.e.* for each column $\mathbf{d}_j$ with $j = 1, \ldots, k$

$$\mathbf{v} = \frac{1}{\mathrm{Tr}(\mathbf{A}_t)_j}(\mathbf{b}_j - \mathbf{D}_t\mathbf{a}_j) + \mathbf{d}_j \tag{6}$$

$$\mathbf{d}_j = \frac{1}{\max(\|\mathbf{v}\|_2, 1)}\mathbf{v} \tag{7}$$

where $\mathrm{Tr}(\mathbf{A}_t)_j$ is the j-th element on the diagonal of $\mathbf{A}_t$, while $\mathbf{a}_j$ and $\mathbf{b}_j$ are the j-th columns of $\mathbf{A}_t$ and $\mathbf{B}_t$, respectively.

The optimization is run for T iterations. Once such a limit is reached, we let $\mathbf{D}^* = \mathbf{D}_T$ be the solution for eq. (1).

3.3. Transfer Single-to-Group Appearance

Inspired by the recent success of residual learning both for visual encoding [21] and for deep learning [16, 17], we propose to exploit the single-person learned dictionary and introduce a sparsity-driven residual representation for an unseen *group* image $\hat{\mathbf{I}}$. The process is shown in Figure 3.

We start by extracting the visual features from each of the $\hat{P}$ patches as computed in Sec. 3.1. Then, for each $\hat{\mathbf{x}}_i$

with $i = 1, \ldots, \hat{P}$ we compute its *residual* $d(\hat{\mathbf{x}}_i, \mathbf{d}_j)$ with every $j = 1, \cdots, k$ atom in the learned dictionary $\mathbf{D}^*$. Notice that the residual $d(\cdot, \cdot)$ can be any suitable function that describes how much "dissimilar" the two inputs are (*e.g.*, the euclidean distance, *etc*.).

Then, we solve the ℓ_1-sparse coding problem in eq. (2) to obtain the sparse vector of coefficients $\hat{\boldsymbol{\alpha}}_i$ for each $\hat{\mathbf{x}}_i$. Since each patch is considered separately, every element in $\hat{\boldsymbol{\alpha}}_i = [\hat{\alpha}_i^1, \ldots, \hat{\alpha}_i^k]$ specifies *how important* a particular atom is in the reconstruction of $\hat{\mathbf{x}}_i$.

Armed with the aforementioned results, we want to assign more importance to the residuals computed with respect to those dictionary atoms that are relevant for the sparse reconstruction of the considered sample $\hat{\mathbf{x}}_i$. A reasonable approach to meet this objective is to weight the residuals through the corresponding sparse dictionary coefficients. This results in:

$$\hat{\mathbf{x}}_i^* = [\hat{\alpha}_1 d(\hat{\mathbf{x}}_i, \mathbf{d}_1), \ldots, \hat{\alpha}_k d(\hat{\mathbf{x}}_i, \mathbf{d}_k)]. \tag{8}$$

3.3.1 Sparse Residual Pooling

The proposed residual representation is obtained for each of the $\hat{P}$ patches of a group. To compute the final representation that can be used to match two groups of persons we should introduce a suitable combination of all the $\hat{\mathbf{x}}_i^*$'s. A classical approach would be to concatenate all such elements. However, in doing so we may lose one of the relevant features of visual encoding schemes, *i.e.*, represent any number of feature vectors as a sample in a feature space of fixed dimensionality. In addition, such a solution is likely to bring in the problem of the curse of dimensionality since the final dimension is linear with respect to both k and $N\hat{P}$.

To overcome these issues, we propose to use different pooling schemes that produce a compact representation, denoted $\hat{\mathbf{f}}$, that depends only on the number of atoms k. Specifically, we exploited the *average pooling* (*i.e.*, $\hat{\mathbf{f}}_j = \frac{1}{\hat{P}} \sum_i^{\hat{P}} \hat{\mathbf{x}}_{i,j}^*$) and *max pooling* (*i.e.*, $\hat{\mathbf{f}}_j = \max(\hat{\mathbf{x}}_{1,j}^*, \cdots, \hat{\mathbf{x}}_{\hat{P},j}^*)$), where $j = 1, \ldots, k$ indicates the j-th element of the corresponding vectors.

3.3.2 Group Representation and Matching

The final group representation is computed as $\hat{\mathbf{s}} = \Phi(\hat{\mathbf{f}})$ where $\Phi(\cdot)$ is the Principal Component Analysis (PCA) mapping function $\mathbb{R}^k \mapsto \mathbb{R}^u$ with $u \ll k$. With such a representation, the dissimilarity between two group images $\hat{\mathbf{I}}_A$ and $\hat{\mathbf{I}}_B$ is computed as $\delta(\hat{\mathbf{I}}_A, \hat{\mathbf{I}}_B) = \prod_f \Psi(\hat{\mathbf{s}}_A^f, \hat{\mathbf{s}}_B^f)$ with Ψ denoting the cosine distance and $f \in \{\mathrm{HS}, \mathrm{RGB}, \mathrm{Lab}\}$.

4. Experimental Results

In this section we report on a series of experiments to assess the performance of the proposed method. From now on we refer to our solution as: Pooling Residuals of Encoded Features (PREF).

Plenty of single-person re-identification datasets have been publicly released –each one with different characteristics– but just one of them is for group re-identification, namely the i-LIDS groups dataset [59]. In order to evaluate the proposed solution under different scenarios, we collected two additional group datasets.

i-LIDS Groups Dataset: This dataset has been obtained from the i-LIDS MCTS dataset which was captured at an airport arrival hall in the busy times under a multi-camera CCTV network. The authors of [59] extracted 274 images of 64 groups. Most of the groups have 4 images, either from different camera views or from the same camera but captured at different locations at different times. Sample images for this dataset are shown in Figure 4(a).

Museum Groups Dataset[2]**:** This dataset has been acquired in the hall of a national museum through four cameras, with small or no overlap. The cameras are installed so as to observe the artworks present in the hall and capture groups during their visits. The dataset contains 524 manually annotated images of 18 groups, composed by a variable number of persons. Each group has about 30 images distributed between each one of the four cameras. Some samples are shown in Figure 4(b).

Outdoor Groups Re-Identification Dataset (OGRE)[3]**:** This dataset contains images of 39 groups acquired by three disjoint cameras pointing at a parking lot. This results in approximatively 2,500 images acquired at different time instants and with different weather conditions. The dataset has been acquired through a weakly supervised approach in which, given a manually selected group region, subsequent detections are obtained by running the KCF tracker [18]. This results in a set of coarsely segmented group images that better resemble a real world scenario. Moreover, the dataset has severe viewpoint changes and a large number of self-occlusions (see Figure 4(c) for few samples).

4.1. Evaluation Protocol and Settings

Protocol: Tests are conducted following a single-vs-single shot scheme: for each group, one randomly selected image is included in the gallery, all the remaining images form the probe set. As commonly performed [5, 62, 29], such a process is repeated 10 times, then average results are computed.

Performance Measure: All the results are reported in terms of Cumulative Matching Characteristic (CMC) curves and normalized Area Under Curve (nAUC) values. The CMC curve represents the expectation of finding the correct match in the first r matches, whereas the nAUC gives a comprehensive measure on how well a method performs independently from the considered dataset.

[2] https://github.com/glisanti
[3] https://github.com/iN1k1

(a) i-LIDS Groups (b) Museum Groups (c) OGRE

Figure 4. Image samples from the (a) i-LIDS groups, (b) Museum groups, and (c) OGRE datasets. Each column represents a same group, while each row depicts the image acquired by a different camera.

Source Datasets: Three of the most commonly used single person re-identification datasets are employed as source domain from which to learn the dictionary of visual words. Among all the possible ones, we selected the i) ETHZ [45] dataset since it contains multiple images of a same person from a similar viewpoint, ii) CAVIAR [7] dataset because of the low-resolution and occluded multiple images of a same person, and iii) VIPeR [15] dataset due to its challenging pose and illumination variations.

Dictionary Learning: We set $\lambda = 0.1$ because we did not notice significant changes in the performance with other values, while for the number of atoms, we run different experiments with $k \in \{300, 500, 1000\}$ when comparing with state-of-the-art in Sec. 4.3.

4.2. Ablation Study

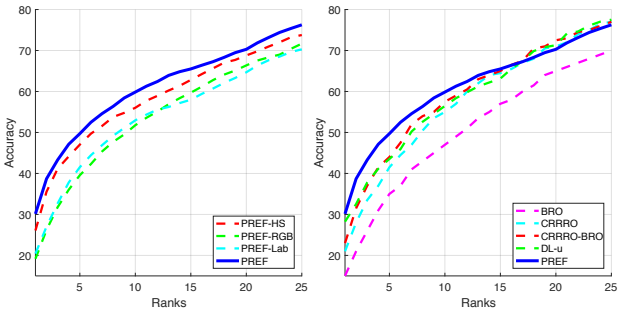
In this section, we thoroughly show how the performance of the proposed approach vary depending on the source dataset(s) considered for training, the distances used for residuals computation, the pooling method and the number of PCA components. The analysis is carried out considering the i-LIDS groups dataset. To run all the following experiments, we considered $k = 500$.

Source Datasets and Residuals: To evaluate the performance of our solution considering different combinations of single person datasets, and different distances and pooling functions, we have computed the results in Table 1.

Results demonstrate that by considering more source datasets the overall performances tend to improve. This might indicate that more discriminative atoms can be learned by considering heterogeneous visual patches together with a robust sparse reconstruction.

As regards distances and pooling, the best results are obtained if average pooling is considered along with the cosine distance. Such an outcome should be attributed to the fact that average pooling is able to better handle noisy assignments. Similarly, the cosine distance is suitable because of the nature of the learned dictionary atoms [11].

PCA Components: In Table 2 the performance of our solution are evaluated with varying number of PCA components. Results show that the overall performances improve little when increasing the value of such an hyperparameter (*i.e.*, there is an nAUC improvement of 0.5 only). Similar results are shown if the rank-1 indicator is considered. This demonstrates that our solution does not hinge on the selection of such a value and is robust to the many group re-identification challenges even if only 20 principal components are considered.



(a) (b)

Figure 5. CMC curves for the i-LIDS groups (a) obtained using the encoded features for each color histogram and their fusion; and (b) in comparison with state-of-the-art.

Features: In Figure 5(a), we show the contribution of each encoded color histogram feature and their combination. It is possible to appreciate that considering features projected onto the HS color space yields to the best results both in terms of rank-1 as well as nAUC. However, as demonstrated by the literature [51, 40], considering all the color spaces helps in improving the overall performance.

Qualitative Performance: Qualitative samples, showing the ranked gallery groups for critical probe images are reported in figures 6 for the i-LIDS groups, Museum groups and OGRE datasets, respectively. Results show that our solution is able to handle situations in which subjects are exchanging their relative displacement as well as cases with severe occlusions. However, drastic illumination variations challenge our approach since direct feature matching is not strong enough to tackle the feature transformation between

Table 1. Rank-1 accuracy for different training datasets (**E** = ETHZ, **V** = VIPeR, **C** = CAVIAR), encoding distances and pooling of the residuals. Results are obtained on the i-LIDS groups dataset using PREF with 500 atoms and 50 PCA components. In parenthesis the nAUC. Best performances are marked with bold, whereas the second bests are marked with underline. Values are in percentage.

	L1		**Cosine**		**Chi Square**		**Euclidean**	
Dataset	*Max*	*Average*	*Max*	*Average*	*Max*	*Average*	*Max*	*Average*
E	18.1 (73.3)	28.4 (77.9)	24.7 (75.8)	31.2 (77.8)	17.2 (72.7)	28.1 (77.9)	14.5 (72.1)	28.5 (77.8)
V	18.4 (72.8)	29.7 (78.3)	21.6 (75.7)	29.5 (78.2)	17.1 (71.6)	28.5 (77.7)	13.7 (70.6)	27.2 (77.9)
C	14.5 (71.7)	27.6 (77.2)	19.7 (76.8)	29.6 (77.8)	13.8 (69.9)	25.5 (77.4)	12.2 (68.8)	26.5 (77.1)
E + V	18.4 (72.5)	28.5 (77.8)	23.7 (75.7)	30.5 (78.0)	16.6 (71.0)	29.5 (78.0)	13.9 (70.8)	27.2 (77.3)
E + C	15.4 (72.5)	29.1 (78.3)	22.3 (75.5)	30.6 (78.1)	15.7 (71.6)	28.8 (77.4)	14.3 (68.9)	27.8 (77.4)
V + C	17.3 (71.9)	28.1 (**78.7**)	25.4 (77.0)	**31.4** (78.1)	15.0 (71.5)	28.9 (78.5)	13.8 (71.2)	27.4 (78.6)
E + V + C	18.2 (72.9)	30.1 (78.4)	24.6 (77.1)	31.1 (**78.7**)	18.3 (73.2)	29.5 (77.9)	17.1 (71.3)	28.5 (78.1)

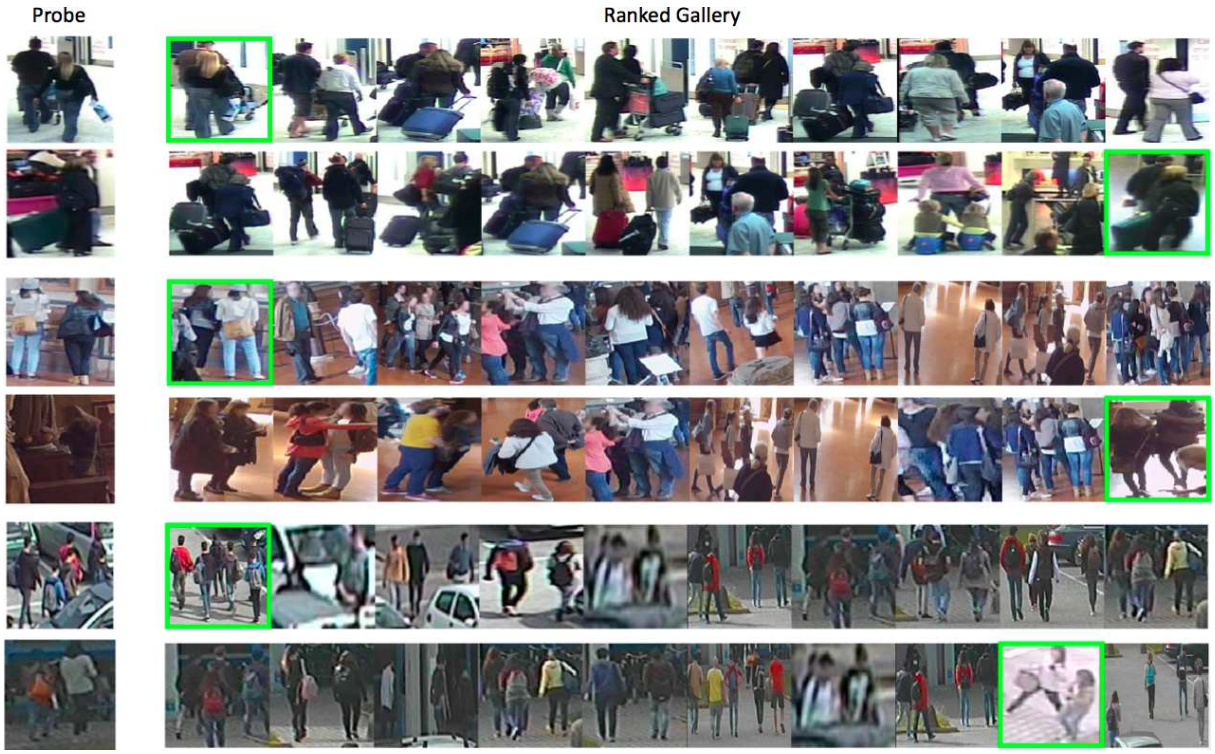


Figure 6. Qualitative samples from the i-LIDS groups (*top rows*), Museum groups (*middle rows*) and OGRE datasets (*bottom rows*). The correct match is highlighted in green and ranked galleries are sorted from left (Rank-1) to right (Rank-10).

Table 2. Re-Identification using different number of PCA components. Results are obtained on i-LIDS groups using PREF with 500 atoms, ETHZ, VIPeR and CAVIAR for training and average pooling of cosine residuals. Best performance is marked with bold, the second best is underlined. Values are in percentage.

Components	**Rank-1**	**Rank-10**	**Rank-25**	**nAUC**
20	29.4	58.5	75.1	78.2
30	30.1	59.9	**76.3**	**78.8**
40	30.2	59.6	75.6	78.6
50	**31.1**	**60.3**	75.5	78.7
60	30.7	**60.3**	76.0	78.7

cameras. Such an issue could be addressed by exploiting metric learning solutions [26, 29, 61, 35].

4.3. Comparison with State-of-the-Art

In the following, we report on the comparison with state-of-the-art in group re-identification and feature encoding.

Re-Identification: In Figure 5(b), we report on the comparison with the state-of-the-art on the i-LIDS groups dataset. Results show that the proposed solution outperforms existing approaches by about 8% at rank-1, whereas at higher ranks similar performance is achieved. It is worth noticing that the two group descriptors proposed in [59], *i.e.*, the Center Rectangular Ring Ratio-Occurrence Descriptor (CRRRO) and the Block based Ratio-Occurrence Descriptor (BRO), exploit shape features in addition to color ones. More importantly, they proposed to learn a visual represen-

Table 3. Re-Identification results on i-LIDS groups, Museum groups and OGRE datasets. Results are obtained using ETHZ, VIPeR and CAVIAR for training, average pooling of cosine residuals and 50 components for PCA. **C** = number of clusters used for the encoding; **A** = number of atoms used for the dictionary learning. Best performance in bold, the second best is underlined. Values are in percentage.

		i-LIDS groups				Museum groups				OGRE			
Method	**C/A**	**Rank-1**	**Rank-10**	**Rank-25**	**nAUC**	**Rank-1**	**Rank-5**	**Rank-15**	**nAUC**	**Rank-1**	**Rank-10**	**Rank-25**	**nAUC**
IFV [44]	64	26.3	58.6	74.4	77.2	24.1	50.0	87.6	62.7	14.6	**43.3**	**76.8**	**62.4**
IFV [44]	128	26.1	60.2	**75.8**	77.8	23.6	49.0	88.4	**62.8**	14.4	42.7	75.6	61.8
IFV [44]	256	26.7	57.4	75.7	76.8	24.8	49.2	88.3	62.6	14.1	42.4	75.9	61.8
VLAD [21]	300	23.8	55.4	74.2	76.0	22.2	47.4	87.2	61.6	13.0	41.1	74.3	60.5
VLAD [21]	500	24.6	54.0	75.6	76.5	23.0	47.6	88.2	61.8	12.6	40.4	73.8	59.9
VLAD [21]	1000	26.0	57.0	75.0	76.7	22.9	48.4	**88.7**	62.4	12.3	39.6	73.2	59.6
PREF	300	29.3	58.2	73.0	77.5	25.6	49.8	87.3	62.6	14.3	41.0	74.9	61.0
PREF	500	**31.1**	**60.3**	75.5	**78.7**	**25.8**	**50.2**	87.6	62.7	**15.1**	41.6	75.8	61.6
PREF	1000	30.1	57.8	74.5	76.9	24.5	49.7	88.0	62.3	12.9	40.3	74.8	60.4

tation considering images that are from the same i-LIDS dataset. This might indicate that our approach is able to learn a robust visual representation from a source domain that is different from the target one.

In Figure 5(b), we also report the CMC curve obtained with the unsupervised solution proposed in [25] based on dictionary learning (DL-u). This experiment has been conducted using the same features as in our solution, so as to have a fair comparison. Lower performance for [25] can be motivated by the fact that it considers all the patches as a unique descriptor, thus it hinges on the spatial displacement of persons within the image.

Feature Encoding: To have a more thorough with respect to the state-of-the-art, we performed experiments considering two encoding techniques, namely IFV [44] and VLAD [21], and our group representation. Experiments are conducted on the i-LIDS groups dataset and on the two newly introduced datasets. For IFV and VLAD, we considered $\{64, 128, 256\}$ and $\{300, 500, 1000\}$ number of clusters, respectively. For these two methods and the proposed solution we obtained the encoding model using ETHZ, VIPeR and CAVIAR datasets.

Results in Table 3 demonstrate that the proposed encoding scheme has better rank-1 performance than existing approaches on all datasets. We hypothesize that this result is due to the fact that the clustering solutions exploited to obtain the encoding models for IFV and VLAD are more sensitive to outliers (*i.e.*, noise), whereas dictionary learning with sparse coding helps in reducing this effect [34].

Spatial Encoding: To verify whether the proposed solution is robust to the spatial appearance ambiguities of group images (*e.g.*, distinguishing two groups of people with opposite appearance), we have conducted the following experiment: to each 64-D feature extracted from each patch (Sec. 3.1) we have concatenated its (x, y) position, thus producing a 66-D vector. The considered (x, y) position of the patch is calculated with respect to the detected person image size. This avoids the problem of having an absolute (x, y) information that depends on the person location within the group image. Results in Table 4, show that

Table 4. Results on i-LIDS groups dataset obtained using the same configuration adopted for Table 3 and spatial information. **C** = number of clusters used for the encoding; **A** = number of atoms used for the dictionary learning. Best performance in bold, the second best is underlined. Values are in percentage.

Method	**C/A**	**Rank-1**	**Rank-10**	**Rank-25**	**nAUC**
IFV [44]	64	22.6	51.2	71.1	73.6
IFV [44]	128	**24.5**	**53.4**	**73.4**	**75.5**
IFV [44]	256	23.3	52.1	72.0	74.2
VLAD [21]	300	18.2	48.8	72.4	73.2
VLAD [21]	500	17.2	46.1	70.2	72.3
VLAD [21]	1000	17.1	49.2	70.3	72.1
PREF	300	20.1	48.5	66.9	71.6
PREF	500	21.7	49.6	67.5	71.9
PREF	1000	21.1	48.5	66.9	71.4

IFV/VLAD/PREF performances degrade by about 7% if such spatial information is included in the feature vector. This might indicate that, spatially constraining the patches of a person may limit re-identification performance due to appearance variations caused by pose changes and the different viewpoints from which a person can be observed.

5. Conclusion

In this paper we have proposed a solution for associating group of persons across different cameras. The proposed solution grounds on the idea of transferring knowledge from single person re-identification to groups, in an unsupervised way, exploiting sparse dictionary learning. The sparse dictionary is learned from classical single person re-identification images. Then a sparsity-driven residual along with a pooling strategy have been introduced to encode features coming from the group and to obtain the final representation. An extensive evaluation shows that the proposed solution achieves state-of-the-art performance on the three datasets for group re-identification. Moreover, results show that it is worth investigating the introduction of a learning scheme to better handle cross-view re-identification issues.

References

[1] X. Alameda-Pineda, J. Staiano, R. Subramanian, L. Batrinca, E. Ricci, B. Lepri, O. Lanz, and N. Sebe. SALSA: A novel dataset for multimodal group behavior analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1707–1720, 2016.

[2] L. An, M. Kafai, S. Yang, and B. Bhanu. Reference-Based Person Re-Identification. In *Advanced Video and Signal-Based Surveillance*, 2013.

[3] S. M. Assari, H. Idrees, and M. Shah. Human Re-identification in Crowd Videos using Personal, Social and Environmental Constraints. In *European Conferece on Computer Vision*, 2016.

[4] S. Bak, E. Corvée, F. Brémond, and M. Thonnat. Boosted human re-identification using Riemannian manifolds. *Image and Vision Computing*, 30(6-7):443–452, June 2012.

[5] L. Bazzani, M. Cristani, and V. Murino. Symmetry-driven accumulation of local features for human characterization and re-identification. *Computer Vision and Image Understanding*, 117(2):130–144, Nov. 2013.

[6] Y. Cai, V. Takala, and M. Pietikäinen. Matching groups of people by covariance descriptor. *International Conference on Pattern Recognition*, pages 2744–2747, 2010.

[7] CAVIAR. CAVIAR Dataset, 2004.

[8] J. Chen, Z. Zhang, and Y. Wang. Relevance Metric Learning for Person Re-Identification by Exploiting Listwise Similarities. *IEEE Transactions on Image Processing*, 7149(c):1–1, 2015.

[9] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. *ECCV International Workshop on Statistical Learning in Computer Vision*, pages 59–74, 2004.

[10] P. Dollár, R. Appel, S. J. Belongie, and P. Perona. Fast feature pyramids for object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(8):1532–1545, 2014.

[11] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

[12] P. F. Felzenszwalb, R. B. Girshick, and D. A. McAllester. Cascade object detection with deformable part models. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, pages 2241–2248, 2010.

[13] J. García, N. Martinel, A. Gardel, I. Bravo, G. L. Foresti, and C. Micheloni. Modeling feature distances by orientation driven classifiers for person re-identification. *Journal of Visual Communication and Image Representation*, 38:115–129, jul 2016.

[14] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition*, 2014.

[15] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recongnition, reacquisition and tracking. In *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, Rio De Janeiro, Brazil, oct 2007.

[16] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *International Conference on Computer Vision and Pattern Recognition*, pages 770—-778, 2016.

[17] K. He, X. Zhang, S. Ren, and J. Sun. Identity Mappings in Deep Residual Networks. In *European Conference on Computer Vision*, pages 630–645, 2016.

[18] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista. High-Speed Tracking with Kernelized Correlation Filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):583–596, mar 2015.

[19] M. Hirzer, P. M. Roth, K. Martin, and H. Bischof. Relaxed Pairwise Learned Metric for Person Re-identification. In *European Conference Computer Vision*, volume 7577 of *Lecture Notes in Computer Science*, pages 780–793, 2012.

[20] E. Insafutdinov, M. Andriluka, L. Pishchulin, S. Tang, B. Andres, and B. Schiele. Articulated Multi-person Tracking in the Wild. *arXiv*, 2016.

[21] H. Jegou, M. Douze, C. Schmid, and P. Perez. Aggregating local descriptors into a compact image representation. In *International Conference on Computer Vision and Pattern Recognition*, 2010.

[22] S. Karaman, G. Lisanti, A. D. Bagdanov, and A. Del Bimbo. Leveraging local neighborhood topology for large scale person re-identification. *Pattern Recognition*, 47(12):3767–3778, 2014.

[23] S. Karanam, Y. Li, and R. J. Radke. Person Re-Identification with Block Sparse Recovery. *Image and Vision Computing*, pages 1–33, 2016.

[24] S. Khamis, C.-h. Kuo, V. K. Singh, V. D. Shet, and L. S. Davis. Joint Learning for Attribute-Consistent Person. In *European Conference on Computer Vision Workshops and Demonstrations*, 2014.

[25] E. Kodirov, T. Xiang, and S. Gong. Dictionary Learning with Iterative Laplacian Regularisation for Unsupervised Person Re-Identification. In *British Machine Vision Conference*, pages 44.1–44.2, 2015.

[26] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *International Conference on Computer Vision and Pattern Recognition*, pages 2288–2295, 2012.

[27] R. Layne, T. M. Hospedales, and S. Gong. Re-id : Hunting Attributes in the Wild. In *British Machine Vision Conference*, 2014.

[28] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith. Learning Locally-Adaptive Decision Functions for Person Verification. In *International Conference on Computer Vision and Pattern Recognition*, pages 3610–3617, jun 2013.

[29] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person Re-identification by Local Maximal Occurrence Representation and Metric Learning. In *International Conference on Computer Vision and Pattern Recognition*, 2015.

[30] S. Liao and S. Z. Li. Efficient PSD Constrained Asymmetric Metric Learning for Person Re-identification. In *International Conference on Computer Vision*, pages 3685–3693, 2015.

[31] G. Lisanti, I. Masi, A. D. Bagdanov, and A. D. Bimbo. Person Re-Identification by Iterative Re-Weighted Sparse Ranking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(8):1629–1642, aug 2015.

[32] B. Ma, Y. Su, and F. Jurie. Local Descriptors Encoded by Fisher Vectors for Person Re-identification. In *European Conference on Computer Vision, Workshops and Demonstrations*, pages 413–422, Florence, Italy, 2012.

[33] B. Ma, Y. Su, and F. Jurie. Covariance Descriptor based on Bio-inspired Features for Person Re-identification and Face Verification. *Image and Vision Computing*, 32:379–390, Apr. 2014.

[34] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online Dictionary Learning for Sparse Coding. In *International Conference on Machine Learning*, pages 689–696, Monterey,Canada, 2009.

[35] N. Martinel, A. Das, C. Micheloni, and A. Roy-Chowdhury. Re-Identification in the Function Space of Feature Warps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2015.

[36] N. Martinel, G. L. Foresti, and C. Micheloni. Person Reidentification in a Distributed Camera Network Framework. *IEEE Transactions on Cybernetics*, pages 1–12, 2016.

[37] N. Martinel, C. Micheloni, and G. L. Foresti. Kernelized Saliency-Based Person Re-Identification Through Multiple Metric Learning. *IEEE Transactions on Image Processing*, 24(12):5645–5658, dec 2015.

[38] N. Martinel, C. Micheloni, and G. L. Foresti. A Pool of Multiple Person Re-Identification Experts. *Pattern Recognition Letters*, 71:23–30, 2016.

[39] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato. Hierarchical Gaussian Descriptor for Person Re-Identification. In *International Conference on Computer Vision and Pattern Recognition*, pages 1363–1372, 2016.

[40] S. Paisitkriangkrai, C. Shen, and A. V. D. Hengel. Learning to rank in person re-identification with metric ensembles. In *International Conference on Computer Vision and Pattern Recognition*, 2015.

[41] S. Pedagadi, J. Orwell, and S. Velastin. Local Fisher Discriminant Analysis for Pedestrian Re-identification. In *International Conference on Computer Vision and Pattern Recognition*, pages 3318–3325, 2013.

[42] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, and Y. Tian. Unsupervised cross-dataset transfer learning for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[43] J. Roth and X. Liu. On the Exploration of Joint Attribute Learning for Person Re-identification. In *Asian conference on Computer Vision*, pages 1–16, 2014.

[44] J. Sanchez, F. Perronnin, T. Mensink, and J. Verbeek. Image Classification with the Fisher Vector: Theory and Practice. *International Journal of Computer Vision*, 105(3):222–245, 2013.

[45] W. R. Schwartz and L. S. Davis. Learning Discriminative Appearance-Based Models Using Partial Least Squares. In *Brazilian Symposium on Computer Graphics and Image Processing*, pages 322–329, Rio De Janeiro, Brazil, oct 2009.

[46] D. Tao, L. Jin, Y. Wang, and X. Li. Person Reidentification by Minimum Classification Error-Based KISS Metric Learning. *IEEE transactions on Cybernetics*, pages 1–11, 2014.

[47] D. Tao, L. Jin, Y. Wang, Y. Yuan, and X. Li. Person Re-Identification by Regularized Smoothing KISS Metric Learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(10):1675–1685, Oct. 2013.

[48] N. Ukita, Y. Moriguchi, and N. Hagita. People re-identification across non-overlapping cameras using group features. *Computer Vision and Image Understanding*, 144:228–236, 2016.

[49] R. Vezzani, D. Baltieri, and R. Cucchiara. People reidentification in surveillance and forensics. *ACM Computing Surveys*, 46(2):1–37, Nov. 2013.

[50] Z. Wang, R. Hu, C. Liang, Q. Leng, and K. Sun. Region-Based Interactive Ranking Optimization for Person Re-identification. In W. T. Ooi, C. G. M. Snoek, H. K. Tan, C.-K. Ho, B. Huet, and C.-W. Ngo, editors, *Advances in Multimedia Information Processing*, volume 8879 of *Lecture Notes in Computer Science*, pages 1–10, Cham, 2014. Springer International Publishing.

[51] F. Xiong, M. Gou, O. Camps, and M. Sznaier. Using Kernel-Based Metric Learning Methods. In *European Conference Computer Vision*, pages 1–16, 2014.

[52] Y. Xu, X. Liu, L. Qin, and S.-c. Zhu. Cross-view People Tracking by Scene-centered Spatio-temporal Parsing. In *AAAI National Conf. on AI*, 2016.

[53] L. Zhang, T. Xiang, and S. Gong. Learning a Discriminative Null Space for Person Re-identification. In *International Conference on Computer Vision and Pattern Recognition*, pages 1239–1248, 2016.

[54] Y. Zhang, B. Li, H. Lu, A. Irie, and X. Ruan. Sample-Specific SVM Learning for Person Re-identification. In *International Conference on Computer Vision and Pattern Recognition*, pages 1278–1287, 2016.

[55] Z. Zhang, Y. Chen, and V. Saligrama. Group Membership Prediction. In *International Conference on Computer Vision*, 2015.

[56] R. Zhao, W. Ouyang, and X. Wang. Person Re-identification by Salience Matching. In *International Conference on Computer Vision*, pages 2528–2535, Dec. 2013.

[57] R. Zhao, W. Ouyang, and X. Wang. Unsupervised Salience Learning for Person Re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3586–3593, June 2013.

[58] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable Person Re-identification: A Benchmark. In *International Conference on Computer Vision*, pages 1116–1124, dec 2015.

[59] W.-S. Zheng, S. Gong, and T. Xiang. Associating Groups of People. In *British Machine Vision Conference*, pages 1–11, 2009.

[60] W.-S. Zheng, S. Gong, and T. Xiang. Re-identification by Relative Distance Comparison. *IEEE transactions on pattern analysis and machine intelligence*, 35(3):653–668, June 2013.

[61] W.-s. Zheng, S. Gong, and T. Xiang. Towards Open-World Person Re-Identification by One-Shot Group-Based Verification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):591–606, mar 2016.

[62] W.-s. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, and S. Gong. Partial Person Re-identification. In *International Conference on Computer Vision*, pages 4678–4686, 2015.