Person Re-Identification by Iterative Re-Weighted Sparse Ranking

Giuseppe Lisanti, Iacopo Masi, Andrew D. Bagdanov, *Member, IEEE*, and Alberto Del Bimbo, *Member, IEEE*

Abstract—In this paper we introduce a method for person re-identification based on discriminative, sparse basis expansions of targets in terms of a labeled gallery of known individuals. We propose an iterative extension to sparse discriminative classifiers capable of ranking many candidate targets. The approach makes use of soft- and hard- re-weighting to redistribute energy among the most relevant contributing elements and to ensure that the best candidates are ranked at each iteration. Our approach also leverages a novel visual descriptor which we show to be discriminative while remaining robust to pose and illumination variations. An extensive comparative evaluation is given demonstrating that our approach achieves state-of-the-art performance on single- and multi-shot person re-identification scenarios on the VIPeR, i-LIDS, ETHZ, and CAVIAR4REID datasets. The combination of our descriptor and iterative sparse basis expansion improves state-of-the-art rank-1 performance by six percentage points on VIPeR and by 20 on CAVIAR4REID compared to other methods with a single gallery image per person. With multiple gallery and probe images per person our approach improves by 17 percentage points the state-of-the-art on i-LIDS and by 72 on CAVIAR4REID at rank-1. The approach is also quite efficient, capable of single-shot person re-identification over galleries containing hundreds of individuals at about 30 re-identifications per second.

Index Terms—Person re-identification, video surveillance, sparse methods

1 Introduction

Person re-identification is the task of recognizing a person, captured by one or more cameras, over a range of candidate targets represented as a gallery of already-labeled subjects. This gallery may contain imagery of known subjects from one or more sensors, and there may be no guarantee that the unknown person observed has already been imaged from the same point of view or in the same conditions. In fact, some of the main complications in person re-identification are due to the fact that the same person is usually acquired at different times, by different disjoint cameras, and this can result in large variations in target appearance because of different illumination conditions, different poses or partial occlusions.

Person re-identification is a critical component of modern surveillance systems as it is a way of maintaining identity information about targets in multiple views over potentially long periods of time. This matching across cameras is traditionally cast as a retrieval problem: given one or more images of an unknown target, the re-identification task is to *rank* all individuals in a gallery of known target images in terms of similarity to the person to be recognized.

- *G. Lisanti, I. Masi, and A. Del Bimbo are with the Media Integration and Communication Center (MICC), Università di Firenze, Firenze 50134, Italy. E-mail: {giuseppe.lisanti, iacopo.masi, alberto.delbimbo}@unifi.it.*
- *A.D. Bagdanov is with the Computer Vision Center, Barcelona and Media Integration and Communication Center, Università di Firenze, Firenze 50134, Italy. E-mail: bagdanov@cvc.uab.es.*

Manuscript received 6 Aug. 2013; revised 4 July 2014; accepted 27 Oct. 2014. Date of publication 9 Nov. 2014; date of current version 6 July 2015. Recommended for acceptance by S. Sarkar.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2014.2369055

Much of the research on person re-identification has concentrated on human appearance modeling [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11]. A number of descriptors of image content have been proposed to discriminate identities while compensating for appearance variability due to changes in pose, illumination and camera viewpoint. Re-identification has also been cast as a learning problem in which either metrics or discriminative models are learned. Metric learning approaches [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22] require labeled training data and most of them also require new training data when camera settings change. Discriminative models, on the other hand, can suffer from lack of training data in small gallery image sets and are often unsuitable for ordering large numbers of candidates due to their inability to reliably rank all but a few of the best ones.

The literature on person re-identification focuses on several different modalities or scenarios that are recognized as de facto standards for performance evaluation. These modalities are characterized in terms of how many images of each individual are known a priori to be in the gallery and probe sets, and according to whether or not it is known that multiple images in the probe set correspond to a single target. The three most common are: the single-versus-single (SvsS) modality where there is a single exemplar for each person in the gallery and at least one exemplar for each person in the probe set (multiple exemplars of the same identity are considered independently); the multi-versus-single (MvsS) modality in which there is one group of multiple exemplars for each person in the gallery and a single exemplar of each person in the probe set; and the multi-versus-multi (MvsM) modality in which there is a group of multiple exemplars for

each person in the gallery and group of multiple images of each person in the probe set.

In this article we propose a robust and efficient approach to person re-identification that is applicable to all modalities considered in the literature. Our technique builds on sparse basis expansions that have been demonstrated to be a powerful tool for face recognition [23]. The use of sparse basis expansions for recognition problems is based on the observation that, even if data is high dimensional, samples from the same class tend to lie on the same low-dimensional subspace of the original feature space. If a good basis can be found, regularization can then be used to enforce sparsity and leverage this subspace structure to discriminate new test examples. Since faces under changing illumination lie near a linear subspace of the original feature space [24], sparse linear reconstructions are able to explain both noise (through linear reconstruction) and identity (by sparseness) in the model. However, the approach of [23] does not directly generalize to the person re-identification problem because, on the one hand, re-identification imagery does not have the same benefit of controlled imaging conditions, and on the other hand ℓ_1-regularized basis expansions, by their very nature, can only support ranking of a limited number of individuals. Our intuition, nevertheless, is that this type of sparse, discriminative approach can be also applied to re-identification problems after addressing these issues.

In more detail, variations in pose, changing target appearance due to articulated motion, and illumination changes make it unlikely that linear reconstructions of unknown targets both explain noise in the model and discriminate identities well. Instead of operating directly on images, as in [23], we use a feature representation for target appearance that approximate the desired invariants so that the sparse linear reconstruction model does not have to explain both noise and target identity. We thus propose a novel descriptor of person appearance, demonstrate its robustness to pose and illumination variations, and show that its use in our sparse discriminative framework yields state-of-the-art results. Our descriptor has the additional advantage of requiring no foreground/background segmentation or body part localization.

At the same time, person re-identification is an application where recall is often important. In fact, since whole-body appearance is a less persistent biometric than faces, in many re-identification scenarios recall is important in order to maximize the likelihood of finding the correct identity in the first 10 or even 20 results. Sparse reconstructions, however, by their very nature can provide inadequate support for ranking more than a few gallery individuals. We address this problem by analyzing the reconstruction error and partially ranking the gallery in terms of similarity to the query probe. We then re-weight this initial solution in order to mute the response of vectors contributing little to the initial expansion. Through the use of this novel, iterative re-weighting algorithm, we can then proceed to rank the remaining gallery individuals through analysis of re-weighted sparse basis expansions.

In the next section we review the person re-identification literature. Our approach to describing the visual appearance of persons is given in Section 3, and in Section 4 we show how to perform re-identification with sparse basis expansions. In Section 5 we give an extensive comparative evaluation of our technique with respect to the state-of-the-art on four publicly available datasets and give a detailed analysis of each component of our approach. Finally, in Section 6 we draw some conclusions and discuss new directions for research.

2 Related Work

Many recent works have addressed the problem of person re-identification. Most focus primarily on either new descriptors for person appearance, or on learning techniques for person re-identification.

2.1 Descriptors for Person Re-Identification

Much research on person re-identification has addressed the definition of discriminative features for person appearance. The first work that considered the problem of appearance models for person recognition, reacquisition and tracking was that of Gray et al. [1]. The authors argue that, until then, these problems had been evaluated independently and that there is a need for metrics that apply to complete systems [2], [3]. They proposed a standard protocol to compare results using the Cumulative Match Curve (CMC) and introduced the VIPeR dataset for re-identification. The first work based on these guidelines was [4] in which the authors propose an algorithm that learns a domain-specific similarity function using an ensemble of local features and AdaBoost. Features are raw color channels in many color spaces and texture information captured by Schmid and Gabor filters.

Descriptors of visual appearance for person recognition can be highly susceptible to background clutter, and many approaches to person re-identification use background modeling [5], [6], [7] or part-based person appearance models [5], [8] to separate foreground from background signals. In [5] the authors use a sophisticated appearance model, the Symmetry-Driven Accumulation of Local Features (SDALF) descriptor that models human body parts by estimating the axis of symmetry of a person and obtaining the head, torso, and legs positions. Each part is then represented by weighted HSV color histograms, maximally stable color region descriptors, and recurrent highly-structured patches. This work also applies a strong, generative background prior that enhances the discriminative power of the descriptor by segmenting the person from the background [25]. In [6] and [7] a multi-shot appearance model similar to [5] is proposed in order to condense a set of frames of the same individual into a highly informative signature, which they call the Histogram Plus Epitome (HPE). In [8] the authors employ an estimate of body pose to guide feature extraction. They extend the Pictorial Structure (PS) model [26] with their Custom Pictorial Structure (CPS), which is a two-step iterative process that alternates between estimating pose and updating the appearance model.

Another state-of-the-art approach with performance similar to [8] is proposed in [9]. The authors use an appearance model that, in contrast with [5] and [8], does not rely on body parts. The approach is based on a descriptor called the Mean Riemannian Covariance Grid (MRCG), which is an

extension of Spatial Covariance Regions (SCR) [10], that is the covariance of a vector of eleven cues derived from equalized RGB colors. The MRCG descriptor is computed as a mean of gallery examples and is only applicable to multi-shot re-identification modalities. The person re-identification problem was extended to groups in [11]. The authors show that groups represent a contextual cue that can be exploited to improve person re-identification.

Re-identification problems are often characterized by poor and variable image quality on which it can be hard to fit background or part-based models without relying on scene-specific information. Our approach, on the other hand, is able to exploit multiple gallery examples and does not require sophisticated background or body part modeling.

2.2 Learning-Based Re-Identification

Among the methods that interpret re-identification as a learning problem, the authors of [12] propose a discriminative model created using Partial Least Squares (PLS) which weights features according to their discriminative power for each different gallery instance. In [13], a metric learning framework is used to obtain a robust Mahalanobis metric for Large Margin Nearest Neighbor classification with Rejection (LMNN-R). The approach in [14] is a supervised technique that uses pairs of similar and dissimilar images and a relaxed RankSVM algorithm to rank probe images. Another metric learning approach is that of [15] which learns a Mahalanobis distance from equivalence constraints derived from target labels.

The Probabilistic Distance Comparison (PRDC) approach [16] introduces a novel comparison model which aims to maximize the probability of a pair of correctly matched images having a smaller distance than that of an incorrectly matched pair. The same authors in [17] model person re-identification as a transfer ranking problem where the goal is to transfer similarity observations from a small gallery to a larger, unlabeled probe set. A set-based discriminative ranking approach was also recently proposed which alternates between optimizing a set-to-set geometric distance and a feature space projection, resulting in a discriminative set-distance-based model [18]. Camera transfer approaches have also been proposed that use images of the same person captured from different cameras to learn metrics [19], [20]. In [21] the authors apply learning in a covariance metric space using an entropy-driven criterion to select the most descriptive features for a specific class of objects. Recently saliency has been considered when matching people across views and a novel method eSDC [22] has been proposed in order to learn saliency parts of a human in a unsupervised fashion.

Learning-based approaches have recently reported higher re-identification accuracy with respect to the state-of-the-art. However, re-identification problems are often also characterized by a lack of reliably labeled data. The need to label image data for each scenario, camera configuration and parameter settings is a disadvantage of metric learning approaches. Our approach outperforms the state-of-the-art at rank-1 in most modalities without learning metrics or fitting discriminative models to gallery image sets.

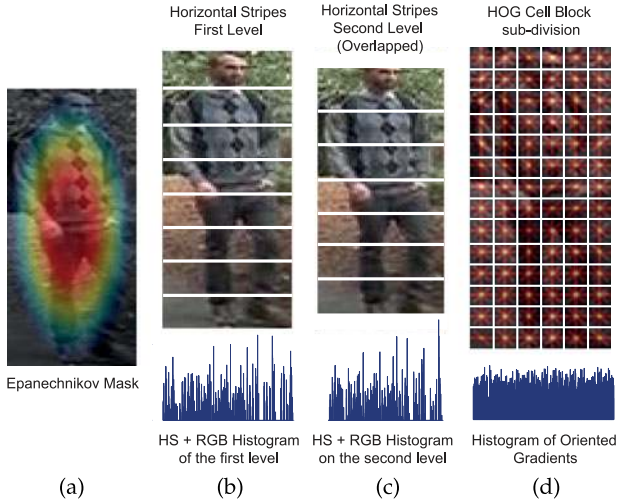


Fig. 1. Our feature descriptor. (a) An Epanechnikov kernel weights the contribution of each pixel to HS and RGB histograms computed on overlapping stripes (b) and (c). Overlapping HOG descriptors are concatenated with these (d).

3 Weighted Histograms of Overlapping Stripes (WHOS)

We have designed a discriminative and efficient descriptor of person appearance for re-identification based on coarse, striped pooling of local features. It exploits a simple yet effective center support kernel to approximately segment foreground from background. The entire descriptor construction process is illustrated in Fig. 1.

Given an input image of a target, it is scaled to a canonical size $W \times H$ (64×128 pixels in all our experiments) and a spatial pyramid is built by dividing the person image into overlapping horizontal stripes of 16 pixels in height.

From each stripe we extract Hue-Saturation (HS) and RGB histograms. Each pixel's contribution to its corresponding histogram bin is weighted using Epanechnikov kernel centered on the image:

$$K(x,y) = \begin{cases} \frac{3}{4}\left(1 - \left(\frac{x}{W}\right)^2 - \left(\frac{y}{H}\right)^2\right), & \text{if } \left|\left(\frac{x}{W}\right)^2 + \left(\frac{y}{H}\right)^2\right| \leq 1, \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$

where W and H are the image width and height, respectively, and are the only parameters of the Epanechnikov kernel. To the HS and RGB histograms we concatenate a Histogram of Oriented Gradient (HOG) descriptor computed on a grid over the image as described in [27].

The HS histograms contain 8×8 bins, while RGB is quantized to $4 \times 4 \times 4$ bins. Both the HS and RGB histograms are computed for the 15 levels of the pyramid (eight stripes for the first level plus seven for the second level of overlapping stripes). The result is a total of $1{,}920$ color histogram bins. The HOG is extracted from a sub-image obtained by removing 8 pixels from top, bottom, left and right of the original in order to remove background details. Each block of the HOG consists of a grid of 2×2 cells of 8×8 pixels. For each cell we compute the gradient histogram over only four angular bins (to capture vertical, horizontal and diagonal patterns) for each HOG block. Given the canonical image size used in our experiments, the dimension of the HOG component is $1{,}040$ bins, and the final descriptor

dimensionality is thus 2,960. As the final stage of descriptor computation, we take the square root of all descriptor bins.

The motivations for a composite descriptor such as ours are many:

- The striped pooling model grants a degree of pose invariance in the representation. Horizontal stripes capture information about vertical color distribution in the image, while overlapping stripes maintain color correlation information between adjacent stripes in the final descriptor.
- Color information is captured by HS and RGB histograms, and local texture by the HOG component. The use of HS histograms renders a portion of the descriptor invariant to illumination variations, while the RGB histograms capture more discriminative color information, especially for dark and greyish colors. We equalize RGB color channels before extracting histograms.
- The Epanechnikov kernel approximately segments the foreground by diminishing the influence of background information near the image boundary. This avoids learning a background model for each scenario, gaining in simplicity and efficiency compared to techniques that use complex background or part-based models.
- Taking the square root of descriptor bins is a well-known technique in image classification [28] that helps to reduce the "burstiness" of features by discounting the effect of small changes in bins that already have significant weight. In preliminary experiments we found this to improve robustness of euclidean distances between descriptors.

An extensive evaluation of the performance of our descriptor confirming these motivations is given in Section 5.2. Though it requires no complex segmentation or fitting of body part models, our descriptor in combination with our sparse framework performs comparably to or better than the state-of-the-art.

4 Sparse Basis Expansions for Re-Identification

In this section, we first describe basis expansions for classification and show how this basic approach does not generalize in a straightforward way to problems like re-identification due to its inability to rank all but a few confidently classified individuals. Hence, we introduce an iterative algorithm for ranking with sparse basis expansions that addresses these shortcomings and permits its effective application to re-identification.

4.1 Sparse Basis Expansion

The main idea behind the use of basis expansions for building discriminative classifiers is that, given sufficient samples $\mathbf{t}_{i,1}, \ldots, \mathbf{t}_{i,n_i}$ from some class i, a test sample $\mathbf{y}$ of the same class should approximately lie in the linear span of the training samples:

$$\mathbf{y} \approx \alpha_{i,1}\mathbf{t}_{i,1} + \alpha_{i,2}\mathbf{t}_{i,2} + \cdots + \alpha_{i,n_i}, \mathbf{t}_{i,n_i} \tag{2}$$

$$= \sum_{j=1}^{n_i} \alpha_{i,j}\mathbf{t}_{i,j} \tag{3}$$

$$= \mathbf{T}_i\boldsymbol{\alpha}_i \tag{4}$$

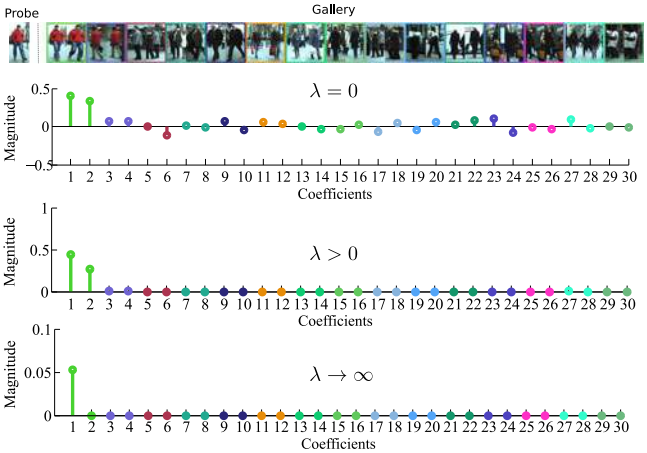


Fig. 2. Basis expansion for MvsS re-identification on ETHZ1. Top: (left) probe sample, (right) the first 15 samples in the gallery, two instances for each subject ($N = 2$). Bottom: reconstruction coefficients for least squares ($\lambda = 0$), sparse ($\lambda = 0.2$) and nearest neighbour ($\lambda = 0.6$). Each color represents a single subject which has two instances.

for some optimal choice of scalar coefficients of reconstruction $\alpha_{i,j}$, for $j = 1, \ldots, n_i$. We use $\mathbf{T}_i$ to represent the matrix of basis vectors for class i, and $\boldsymbol{\alpha}_i = [\alpha_{i,1}, \ldots, \alpha_{i,n_i}]^T$ to represent the vector of reconstruction coefficients for the same class.

The general, multi-class basis expansion for C classes then becomes:

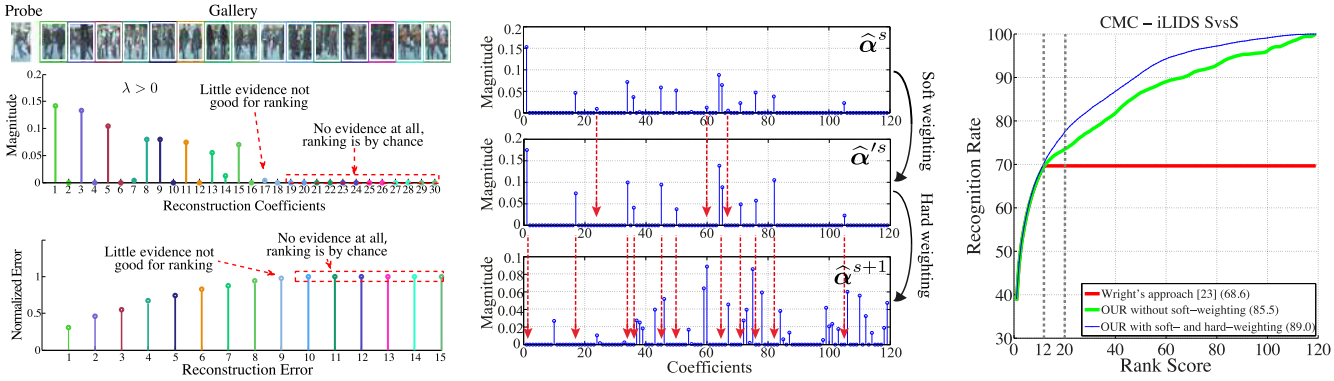
$$\begin{aligned}\mathbf{y} &\approx [\mathbf{T}_1\ \mathbf{T}_2\ \cdots\ \mathbf{T}_C][\boldsymbol{\alpha}_1\ \boldsymbol{\alpha}_2\ \cdots\ \boldsymbol{\alpha}_C]^T \\ &= \mathbf{T}\boldsymbol{\alpha}.\end{aligned} \tag{5}$$

The basis $\mathbf{T}$ can be highly overcomplete, but if $\mathbf{y}$ is an instance of a person we desire that the energy in the basis expansion be concentrated in the relatively few coefficients from the gallery examples corresponding to the identity of $\mathbf{y}$. We can impose this sparsity constraint on the solution by formulating it as an ℓ_1-regularized least squares problem:

$$\widehat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \|\mathbf{y} - \mathbf{T}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1, \tag{6}$$

where λ controls the tradeoff between minimization of the ℓ_2 reconstruction error and the ℓ_1 norm of the coefficients used to reconstruct $\mathbf{y}$. This formulation is known as Lasso Regression in the statistics literature and there exist very efficient algorithms for solving it [29].

Regularized basis expansions of this type are generally referred to as *sparse* because the ℓ_1 regularization term, depending on the sparseness factor λ, tends to cause the coefficients of reconstruction to collapse to zero except for a few important basis vectors. The form of Eq. (6) is particularly convenient because it represents a whole class of solutions to the approximate reconstruction problem of Eq. (5). When $\lambda = 0$, Eq. (6) results in a standard least squares solution. For $\lambda > 0$, we obtain solutions of increasing sparseness with increasing λ. Eventually, as $\lambda \to \infty$, only a single non-zero coefficient will be admitted in the solution of Eq. (5). We refer to this last solution, with $\lambda \to \infty$, as the *nearest neighbor* solution since only the ℓ_2-closest training sample to $\mathbf{y}$ will have a corresponding non-zero coefficient in $\widehat{\boldsymbol{\alpha}}$. In Fig. 2 we illustrate these three types of solutions for a MvsS re-identification problem. The top row of Fig. 2 illustrates the probe and gallery images for a re-identification query.



(a) Ranking with limited information (b) Soft- and hard-weighting effects (c) Soft-weighting results

Fig. 3. (a) Ranking with limited information from a single basis expansion (MvsS, N = 2). Ranking decisions must be made on the basis of little information (low coefficient energy) or no information (zero coefficient energy). In the middle are reconstruction coefficients, at the bottom the corresponding normalized reconstruction errors for each multi-shot probe. (b) Effects of soft- and hard-weighting. Top: reconstruction coefficients from the first solution $\widehat{\boldsymbol{\alpha}}^s$ at iteration s; Middle: refined reconstruction coefficients after soft-weighting $\widehat{\boldsymbol{\alpha}}'^s$; and Bottom: coefficients $\widehat{\boldsymbol{\alpha}}^{s+1}$ at iteration $s+1$ after hard-weighting. (c) The effects of soft-weighting on performance on the i-LIDS dataset.

The plot in the second row shows the coefficients of a least squares solution ($\lambda = 0$), followed by a sparse solution ($\lambda = 0.2$), and finally the nearest neighbor solution ($\lambda = 0.6$ for this example).

We can derive a decision rule for classification by analyzing the reconstruction error for solutions to Eq. (5) restricted to basis vectors corresponding to individual gallery subjects. The normalized reconstruction error corresponding to the ith subject is:

$$e_i = \frac{\|\mathbf{y} - \mathbf{T}_i \widehat{\boldsymbol{\alpha}}_{|i}\|_2}{\|\mathbf{y}\|_2}, \quad \text{for } i \in \{1, \ldots, C\}. \tag{7}$$

where $\widehat{\boldsymbol{\alpha}}_{|i}$ represents the sparse solution of Eq. (6) restricted to the coefficients corresponding to gallery examples of class i. That is, $\widehat{\boldsymbol{\alpha}}_{|i}$ is equal to $\widehat{\boldsymbol{\alpha}}$ at coefficients corresponding to gallery examples from individual i and zero elsewhere.

Our decision rule is:

$$\text{class}(\mathbf{y}) = \arg\min_i e_i. \tag{8}$$

This decision rule based on sparse discriminative basis expansion performs very well for classification problems [23]. However, as mentioned in the introduction, recall can be critical for re-identification and it is important to be able to rank gallery individuals. We can extend the decision rule of Eq. (8) in a straightforward manner to rank candidate individuals using their corresponding residual e_i.

In Fig. 3a we show two views of an MvsS re-identification problem in terms of normalized reconstruction error e_i with respect to the probe $\mathbf{y}$. In the middle are illustrated the coefficients $\widehat{\boldsymbol{\alpha}}$ of a probe reconstruction in terms of a multi-shot gallery. Below are illustrated the normalized reconstruction errors e_i corresponding to each gallery individual. Each error on the bottom thus corresponds to *two* coefficients in the middle, since the gallery is multi-shot with $N = 2$. The fundamental problem with using discriminative sparse basis expansions derived from solutions to problems like Eq. (6) is that, for many reasonable values of λ, we are deliberately forcing the majority of coefficients to zero, which limits the number of ranks the basis expansion can support. In Fig. 3a we see that after the first few individuals (ranks), the coefficient energy collapses and we have no more information upon which to base ranking decisions. The result is that beyond this point we cannot rank the remaining gallery individuals.

A more subtle problem is that in many cases we may be basing ranking decisions on inadequate evidence from the basis expansion. After the first eight individuals in Fig. 3a, even before collapsing to zero, there is very little coefficient energy upon which to base individual ranking decisions. In the next section we introduce an iterative sparse basis expansion technique that addresses these problems of lack of sufficient ranking support in sparse reconstructions.

4.2 Iterative Sparse Re-Weighting

In this section we develop an iterative technique to address the problems with applying sparse discriminative classifiers to ranking. We arrive in the process at an algorithm that is able to robustly perform re-identification up to all ranks. Our approach is an iterative extension of the weighting described in [30] which we use to first re-weight basis vectors in the sparse solutions of Eq. (6) and arrive at a more robust solution that does not rely on basis vectors contributing little to the reconstruction. A similar weighting approach is then used to proceed with ranking after damping the influence of basis vectors that have already contributed to ranking.

4.2.1 *Soft-Weighting for Ranking Robustness*

The first refinement step we perform is a sort of *soft-weighting* that is used to remove those coefficients that weakly contribute to the reconstruction of the given test sample. Assume we have computed sparse reconstruction coefficients $\widehat{\boldsymbol{\alpha}}$ for a given instance of a re-identification problem. At each iteration we define for each element in the basis a weight that is inversely proportional to its coefficient magnitude in the initial reconstruction:

$$w_{i,j} = \frac{1}{|\widehat{\alpha}_{i,j}| + \varepsilon} \quad \text{for } i \in \{1 \ldots C\} \text{ and } j \in \{1 \ldots n_i\}, \tag{9}$$

where ε is chosen to be slightly smaller than the minimum nonzero coefficient of $\widehat{\boldsymbol{\alpha}}$ to avoid division by zero and to not influence the solution with respect to the other coefficients.

We then solve a weighted Lasso problem by weighting the regularization magnitudes using the $w_{i,j}$ defined above:

$$\widehat{\boldsymbol{\alpha}}' = \arg\min_{\boldsymbol{\alpha}} \|\mathbf{y} - \mathbf{T}\boldsymbol{\alpha}\|_2^2 + \lambda \sum_{i=1}^{C} \|\mathrm{diag}(\mathbf{w}_i)\boldsymbol{\alpha}_i\|_1, \quad (10)$$

where $\mathbf{w}_i = [w_{i,1}, w_{i,2}, \ldots, w_{i,n_i}]$ are the weights from Eq. (9) corresponding to the basis vectors for individual i and $\mathrm{diag}(\mathbf{w}_i)$ denotes the diagonal matrix with vector $\mathbf{w}_i$ on its diagonal. Just like the unweighted counterpart in Eq. (6), this convex problem can also be efficiently solved as a linear program. The weights $w_{i,j}$ are free parameters in the convex relaxation, whose values can be used to penalize or favor specific basis vectors in the regularized expansion. The new solution $\widehat{\boldsymbol{\alpha}}'$ is the refined solution that is used to rank individuals with respect to probe $\mathbf{y}$.

Fig. 3b graphically illustrates this soft-weighting procedure. The initial solution contains a few dominant coefficients that contribute most to the reconstruction of the probe $\mathbf{y}$. It also contains a number of basis vectors that contribute little to the overall reconstruction, as indicated by very small coefficients in the initial solution shown in the top plot of Fig. 3b. The refined solution using the weights from Eq. (9) is shown in the middle plot of Fig. 3b. Note that this refinement eliminates small coefficients from the solution and redistributes their energy among more relevant basis vectors. It is not equivalent to a simple thresholding of coefficients.

4.2.2 Hard Re-Weighting for Improved Recall

A more serious problem related to using a sparse discriminative classifier for re-identification is the lack of sufficient non-zero support in sparse solutions. It is often the case that only a few gallery individuals can be ranked by analyzing the initial, refined sparse solution. To address this problem a set of hard weights are maintained that are used to exclude those elements that have already contributed to ranking an individual against the probe $\mathbf{y}$:

$$w_{i,j}^h \leftarrow \begin{cases} \infty, & \text{if} \quad \widehat{\alpha}'_{i,j} > 0, \\ 1, & \text{otherwise}, \end{cases} \quad (11)$$

where $\widehat{\alpha}'_{i,j}$ are the coefficients from the soft-weighted solution $\widehat{\boldsymbol{\alpha}}'$. The hard weights vector $\mathbf{w}^h$ is used in the next step of an iterative process that repeats the soft-weighting and ranking procedure. In the bottom plot of Fig. 3b we show the solution to the weighted Lasso problem using these weights. The difference in the distribution of coefficients between the hard-weighted solution and the original solution in the top plot of Fig. 3b is quite noticeable.

In Fig. 3c we give a comparison of our approach, including soft- and hard-weighting to rank the entire gallery, with our approach without soft-weighting and the technique of [23]. On the iLIDS dataset, the basic sparse reconstruction approach of [23] can rank, on average, only about 12 gallery persons per probe with a recognition rate of 70 percent. At rank 20, our approach *without* soft-weighting reaches a recognition rate of about 73 percent, while our final approach *with* soft-weighting gains another 5 percent in accuracy and reaches 78 percent recall in just the first twenty ranks. Note also how the gain due to soft-weighting improves for higher ranks and how soft-weighting provides a slightly different recall at first rank. This is because soft-weighting effectively defers the decision to rank an individual with low coefficient energy to future iterations. Without soft-weighting, persons can be ranked on the basis of very little evidence.

As we will see in more detail in the experimental results, the problem of lack of ranking support does not only have an impact on very low or very high ranks. Moreover, though low ranks are clearly the most important, there will always be applications where recall is also crucial.

4.3 Iterative Sparse Person Re-Identification

Putting it all together, our approach to person re-identification up to arbitrary ranks is an iterative process of both soft- and hard-weighting of ℓ_1 regularized probes reconstructions using gallery examples. The steps are as follows:

1) Reconstruct probe(s) using Eq. (10) with hard weights defined as in Eq. (11) to eliminate already ranked persons (if any).
2) Use soft-weighting from Eq. (9) in a weighted reconstruction using Eq. (10) to eliminate coefficients contributing little to the reconstruction of the probe and distribute their energy among more relevant basis vectors.
3) Rank gallery individuals who have non-zero coefficient energy (i.e. those individuals who have normalized reconstruction error $e_i < 1$, where e_i is defined as in Eq. (7)).
4) Update hard weights as in Eq. (11) to eliminate from subsequent iterations those basis vectors contributing to ranking in the current round.
5) Repeat until all gallery individuals are ranked.

Algorithm 1 formalizes each of these steps and how they fit together to rank all individuals in the gallery. Note that gallery individuals are not ranked by normalized reconstruction error alone. The normalized reconstruction error is used for ranking *within* an iteration, but those individuals ranked in an iteration will always be ranked higher than those in subsequent ones. Algorithm 1 is designed to work with single- and multi-shot gallery and probe sets. For this reason, the optimizations in lines 5 and 7 differ slightly from Eq. (10) in that they simultaneously optimize over coefficients $\boldsymbol{\alpha}$ and all probes $\mathbf{y} \in Y$. In this way, Algorithm 1 can be used for each of the re-identification modalities considered:

- *SvsS re-identification.* For SvsS re-identification problems, the ranked list of triples returned by Algorithm 1 uniquely ranks each gallery individual. As soon as a triple (i, s, e) is added to the ranked list R, hard-weighting of the *single* template for individual i in the gallery prevents it from further consideration. The ordered list R returned by Algorithm 1 thus represents the ordering of the gallery with respect to the probe $\mathbf{y}$.
- *MvsS re-identification.* For multi-shot galleries like MvsS in which more than one example of each person is present, hard-weighting does not necessarily eliminate a ranked individual from consideration in subsequent iterations. Due to the sparseness constraint, not all examples corresponding to one individual are necessarily used in the regularized basis expansion. Since hard-weighting by Eq. (11) only

eliminates those basis vectors already used for ranking, it is possible that some templates corresponding to already ranked individuals remain. The guard in line 9 of Algorithm 1, however, guarantees that a person only occurs once in the ordered list R and thus the rank of each gallery individual is unambiguously defined in the output.

- *MvsM re-identification.* When both the gallery and the probe sets are multi-shot the minimization over *both* $\boldsymbol{\alpha}$ and all probe templates $\mathbf{y} \in Y$ in lines 5 and 7 of Algorithm 1 uniquely defines the reconstruction error e_i in terms of the minimum error over all probe templates. Thus, re-identification with multi-shot probes is similar to running multiple MvsS re-identifications (one for each probe image of each person) and using the minimum error to represent the overall reconstruction error for the multi-shot probe.

Algorithm 1. $R(\mathbf{T}, Y, \lambda)$ Iterative Sparse Ranking

Input: $\mathbf{T} = [\mathbf{T}_1\ \mathbf{T}_2\ \cdots\ \mathbf{T}_C]$, the gallery templates;
$Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$, the probe templates; and
λ, the regularization factor.
Output: R, the ranked list of gallery individuals.

1 Initialize hard weights: $w^h_{i,j} \leftarrow 1$.
2 Initialize iteration count: $s \leftarrow 1$.
3 Initialize list of gallery individuals: $R \leftarrow \emptyset$.
4 **while** $|R| < C$ **do**
5 Hard-weighting:
$\widehat{\boldsymbol{\alpha}} \leftarrow \arg\min_{\boldsymbol{\alpha}, \mathbf{y} \in Y} \|\mathbf{y} - \mathbf{T}\boldsymbol{\alpha}\|_2^2 + \lambda \|\mathrm{diag}(\mathbf{w}^h)\boldsymbol{\alpha}\|_1$
6 Soft-weighting:
$w_{i,j} \leftarrow \frac{1}{|\widehat{\boldsymbol{\alpha}}_{i,j}| + \varepsilon}$ for $i \in \{1 \ldots C\}$ and $j \in \{1 \ldots n_i\}$
7 $\widehat{\boldsymbol{\alpha}}' \leftarrow \arg\min_{\boldsymbol{\alpha}, \mathbf{y} \in Y} \|\mathbf{y} - \mathbf{T}\boldsymbol{\alpha}\|_2^2 + \lambda \|\mathrm{diag}(\mathbf{w})\boldsymbol{\alpha}\|_1$
8 **for** $\widehat{\boldsymbol{\alpha}}'_i \neq \mathbf{0}$ **do**
9 **if** person i has not yet been ranked **then**
10 $R \leftarrow R \cup \{(i, s, e_i)\}$ (e_i from Eq. (7))
11 **end**
12 **end**
13 $s \leftarrow s + 1$
14 $w^h_{i,j} \leftarrow \begin{cases} \infty & \text{if } \widehat{\alpha}'_{i,j} > 0 \\ w^h_{i,j} & \text{otherwise} \end{cases}$
15 **end**
16 **return**
17 R ordered by $(i, s, e) \leq (i', s', e') \Leftrightarrow s < s' \vee (s = s' \wedge e \leq e')$

5 Experimental Results

In this section we report on an extensive set of experiments performed to compare our method with the state-of-the-art and to evaluate in detail each component of our approach. All experiments were conducted on standard, publicly available datasets (ETHZ, VIPeR, i-LIDS and CAVIAR4REID), and we compare our results with the following state-of-the-art approaches: SDALF [5], HPE [6], AHPE [7], SCR [10], ELF [4], CPS [8], MRCG [9], ContextB [11], PRDC [16], PRSVM [14], SBDR [18], EIML [31], COSMATI [21], RPLM [20] and eSDC [22]. Note that not all techniques report results on all four datasets or on all three modalities (SvsS, MvsS and MvsM). For example, with the exception of COSMATI, metric learning approaches report results only for SvsS scenarios. To provide the most comprehensive comparison possible, we test our method on all modalities and include all reported results from the above methods, when available.

The principal metric used for evaluating person re-identification is the Cumulative Match Characteristic (CMC) curve which summarizes overall performance by reporting recall over a range of cutoff points. A CMC curve represents the expectation of finding the correct match in the top r matches, where r is the rank considered in the final ranking result. Unless otherwise noted, all results were computed by averaging over 50 random, independent splits of dataset into gallery and probe sets, except for VIPeR where we use the 10 splits from [5]. We also report, and compare with the state-of-the-art when available, the normalized Area Under the Curve (nAUC). The nAUC is calculated as the total area under a CMC divided by $100 \times N$, where N is the total number of gallery individuals. It gives an overall score of how well a method performs over all ranks. For many applications, the most important cutoff rank is one. We thus also report a comparison of our rank-1 performance with respect to the state-of-the-art on all datasets.

We refer to our approach as Iterative Sparse Ranking (ISR) in all tables and figures that follow. Unless otherwise qualified, we use our full descriptor as described in Section 3. In our experiments reported in Section 5.3 we found $\lambda = 0.2$ to be a good trade-off between recognition rate and number of iterations. Accordingly, we fixed $\lambda = 0.2$ for all experiments reported here. In practice, the optimal λ will be descriptor- and dataset-dependent and could be cross-validated given a labeled validation set.

5.1 Comparison with the State-of-the-Art

In this section we compare ISR with the state-of-the-art on the ETHZ, VIPeR, i-LIDS, and CAVIAR4REID datasets.

5.1.1 Performance on the ETH Zurich Datasets

The ETH Zurich dataset consists of three sequences used for tracking, from which Schwartz and Davis [12] extracted a set of samples of each person in the videos. We performed SvsS and MvsM experiments, varying the number of elements in both the probe and gallery.

In Table 1 we report rank-1 results for each sequence of the ETHZ dataset for the SvsS and MvsM ($N \in \{5, 10\}$) modalities. ISR outperforms current methods for MvsM, and performs comparably to others for SvsS. More extensive results and CMC curves comparing ISR with the state-of-the-art on ETHZ can be found in the supplementary material accompanying this article.

5.1.2 Performance on the VIPeR Dataset

The VIPeR dataset consists of 632 people imaged by two non-overlapping cameras. Image pairs exhibit viewpoint changes of up to 180 degrees and illumination changes that result in large intra-class variations. The dataset has only two samples of each person (one from each view), and thus can only be used for SvsS re-identification.

On VIPeR we use the publicly available splits into gallery and probe sets provided by the authors of SDALF [5]. Table 2 compares the rank-1 performance of ISR and the

TABLE 1
Performance at Rank-1 with Respect to the State-of-the-Art on ETHZ

Modality:	ETHZ1			ETHZ2			ETHZ3		
	SvsS	*MvsM* N = 5	*MvsM* N = 10	*SvsS*	*MvsM* N = 5	*MvsM* N = 10	*SvsS*	*MvsM* N = 5	*MvsM* N = 10
HPE [6]	–	84	85	–	81.5	79.3	–	87.3	82.6
AHPE [7]	–	91	–	–	90.6	–	–	94	–
MRCG [9]	–	–	96	–	–	97	–	–	98.3
PLS [12]	79	–	–	74.5	–	–	77.5	–	–
SDALF [5]	64.8	90.2	89.6	64.4	91.6	91.5	77	93.7	94.1
CPS [8]	–	97.7	–	–	97.3	–	–	98	–
EIML* [31]	78	–	–	74	–	–	**91**	–	–
RPLM* [20]	77	–	–	65	–	–	83	–	–
eSDC* [22]	**80**	–	–	**80**	–	–	89	–	–
ISR	79.5	**99.8**	**99.9**	76.1	**99.7**	**100**	86.2	**99.9**	**99.9**

Techniques indicated by "" set aside a portion of the data for metric learning. Recognition rates in percent.*

state-of-the-art on VIPeR. From this table we see that ISR improves by about six percentage points on the state-of-the-art performance on VIPeR, except for learning-based methods like RPLM [20] and eSDC [22], which perform similarly to us at rank-1.

Fig. 4a gives the CMC curves up to rank 50 comparing ISR with the state-of-the-art. We outperform all state-of-the-art techniques not based on metric learning up to all but the highest ranks. After about rank-5, techniques that learn on a part of the data like EIML [31], RPLM [20], and eSDC [22] begin to outperform us. Note that such techniques are not strictly comparable with ours since they set aside a portion (up to half) of the dataset on which to learn metrics. The gallery and probe sets are drawn from the remaining data and thus the standard splits cannot be used.

5.1.3 Performance on the i-LIDS Dataset

The i-LIDS dataset contains images from multiple camera views in a busy airport arrival hall. As shown in Fig. 4b, ISR outperforms the state-of-the-art at low ranks. After about rank-4, however, techniques based on metric learning begin to outperform us. Note that, due to having to use a portion of available data for learning, the SBDR [18] and PRSVM [14] methods only consider, respectively, 80 and 108 out of the 119 people in the dataset.

TABLE 2
Performance at Rank-1 with Respect to the State-of-the-Art on VIPeR, iLIDS and CAVIAR4REID

Modality:	VIPeR	iLIDS		CAVIAR4REID		
	SvsS	*SvsS*	*MvsM* N = 2	*SvsS*	*MvsM* N = 3	*MvsM* N = 5
HPE [6]	–	–	18.5	–	–	–
AHPE [7]	–	21	32	7.5	7.5	7.5
SCR [10]	–	34.5	36	–	–	–
MRCG [9]	–	–	46	–	–	–
ContextB [11]	–	24	–	–	–	–
SDALF [5]	19.9	28	39	7	8.5	8.3
ELF [4]	12	16	–	–	–	–
CPS [8]	21.8	29.5	44	8.5	13	17.5
PRSVM* [14]	15	32	–	–	–	–
PRDC* [16]	15.7	32.6	–	–	–	–
SBDR* [18]	–	37.75	–	–	–	–
EIML* [31]	22.0	–	–	–	–	–
COSMATI* [21]	–	–	44	–	–	–
eSCD (ocsvm)* [22]	26.7	–	–	–	–	–
RPLM* [20]	**27.0**	–	–	–	–	–
ISR	**27.0**	**39.5**	**62.9**	**29**	**75.1**	**90.1**

Techniques indicated with a "" set aside a portion of the dataset for learning. Recognition rates in percent.*

Table 2 summarizes the rank-1 performance of ISR and the state-of-the-art for SvsS and MvsM on i-LIDS. From this table we see that we slightly outperform other approaches on SvsS, while we significantly outperform competing methods by about 17 percentage points for MvsM.

For MvsS and MvsM, where we are able to exploit multiple images of each gallery individual, our improvement over the state-of-the-art is dramatic. As seen in Fig. 4c for MvsS $(N = 2)$ we exceed the state-of-the-art at rank-1 by nearly 19 percentage points. We similarly improve for MvsS $(N = 3)$ where we outperform SDALF by nearly 11 percentage points at rank-1. For MvsM we report results for $N \in \{2, 3\}$ in Fig. 4d along with results of other methods tested on this dataset. We outperform the state-of-the-art at all ranks for the MvsM modality.

5.1.4 Performance on the CAVIAR4REID Dataset

The CAVIAR4REID dataset contains 72 unique individuals captured in a shopping center scenario. This dataset was designed to maximize variability with respect to resolution changes, illumination conditions, occlusions, and pose changes.

We compare the rank-1 recognition rate of ISR and the state-of-the-art on CAVIAR4REID in Table 2. We significantly outperform competing methods at rank-1 in all modalities on this dataset. For MvsM we improve on the state-of-the-art by nearly 62 percentage points for MvsM $(N = 3)$ and by 72.5 for MvsM $(N = 5)$.

In Fig. 5a we report the CMC curves for ISR and the state-of-the-art for SvsS on CAVIAR4REID. Our approach outperforms current methods up to about rank-20. The improvement over the state-of-the-art at first rank is particularly noticeable: there is a difference of 20.5 percentage points at rank-1 between our performance and competing methods. In the legend we also report the nAUC for each method, which gives an idea of the trend of the curve across all ranks. Fig. 5b gives CMC curves for MvsM on CAVIAR4REID for

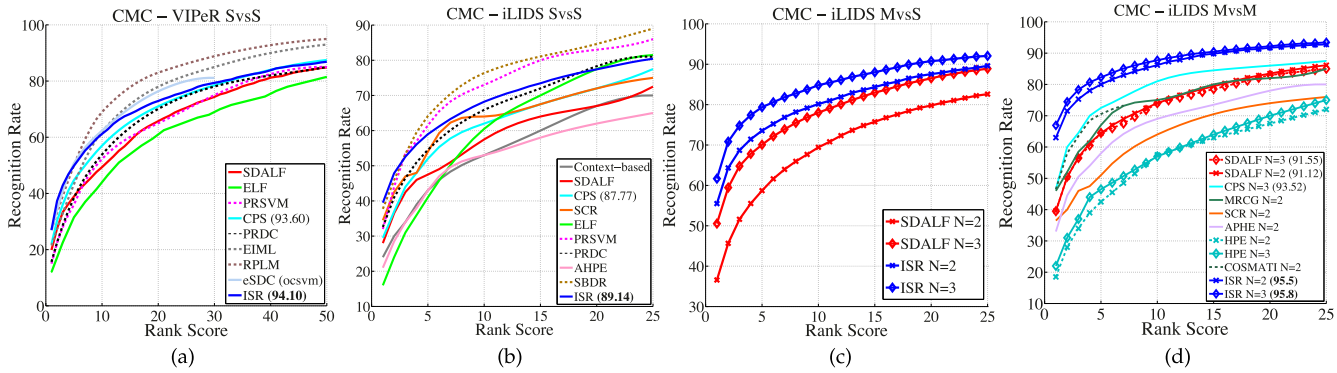


Fig. 4. Comparative performance evaluation on VIPeR and i-LIDS. (a) SvsS on VIPeR. (b) SvsS on i-LIDS. (c) MvsS on i-LIDS ($N \in \{2,3\}$). (d) MvsM on i-LIDS ($N \in \{2,3\}$). Dashed curves distinguish techniques that set aside a portion of the dataset for learning. In the legends we report the normalized area under the CMC curve (nAUC), when available.

$N \in \{3,5\}$. In the MvsM modality, as for i-LIDS, we significantly outperform the state-of-the-art at all ranks.

5.2 Evaluation of Our Person Descriptor

In this section, we first give a comparison between our descriptor detailed in Section 3 and some of the most used descriptors from the re-identification literature. We also quantify the contribution of our background separation and pooling models with respect to other models used in the literature. Then we analyze the contribution of each component of our descriptor to the overall performance of the method. Finally we provide an extensive evaluation of the sensitivity of our descriptor to illumination variations, viewpoint changes and misalignments of the person window.

5.2.1 Comparison of Person Descriptors

We compare our descriptor with the SDALF, PS and PRDC descriptors. Direct, side-by-side comparison is difficult because for some methods there is no publicly available code. Also, SDALF and PS use HSV histograms with an MSCR descriptor of variable-length, and is thus not suitable for reconstruction by basis expansion. In order to embed these descriptors into our framework, and considering that MSCR contributes little to re-identification with respect to the HSV histogram [8], we used only the HSV component computed on the symmetry parts for SDALF, and the HSV component computed on the person parts for PS. For fairness of comparison, we report performance of our full descriptor and the HS histogram component only. In Fig. 6a we report performance on i-LIDS for both SvsS and MvsM ($N = 2$), averaged over ten trials. Our descriptor outperforms the others, although it uses neither part-based modeling nor data-driven foreground segmentation.

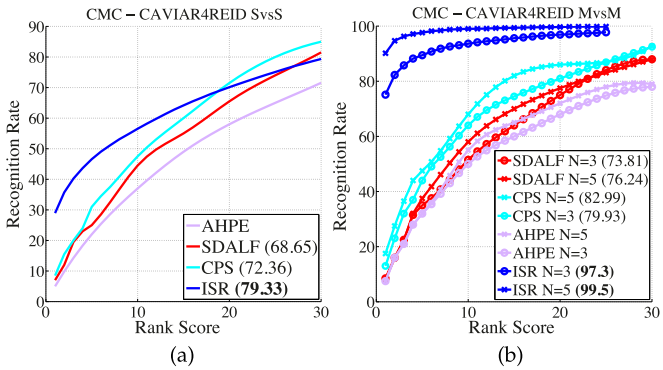


Fig. 5. Performance on CAVIAR4REID with respect to the state-of-the-art. (a) SvsS. (b) MvsM for $N \in \{3,5\}$. In the legends we report the nAUC, when available.

5.2.2 Comparison of Background Separation Models

We compare the contribution of the Epanechnikov kernel with respect to a Gaussian center-support background model with varying diagonal covariance and the STEL component analysis background model, with and without Gaussian weighting [25]. The performance figures were obtained with our ISR re-identification approach and the full descriptor with the corresponding background model.

In Fig. 6b we report performance on VIPeR, averaged over ten random trials. We see that the Epanechnikov kernel consistently outperforms Gaussian weighting, likely due to the difficulty of tuning σ_x and σ_y to balance the flatness (to include the subject) and peakedness (to exclude background). The STEL component analysis model performs similarly at high ranks. However, STEL relies on a previously learned person model and requires inference at re-identification time to segment the person from the background.

5.2.3 Comparison of Pooling Models

We compare our overlapping striped pooling model with a non-overlapping version of the same striped model, the symmetry-driven SDALF model [5], and the part-based Pictorial Structures model [8]. In all cases, we pool the local HS and RGB histograms for each region and then concatenate the HOG of the full image. We report performance on VIPeR averaged over ten trials in Fig. 6c. We see that our striped pooling strategy, together with the Epanechnikov kernel as background model, outperforms the more complex part-based approaches.

5.2.4 Contribution of Each Descriptor Component

In Fig. 6d we show the contribution of each component of our descriptor to overall performance: the HS histogram, the RGB histogram, the HOG, the Epanechnikov kernel and the application of square root to each bin of the descriptor. The experiments were performed on the VIPeR dataset, and results were averaged over ten trials. The plots show that the addition of each component improves performance.

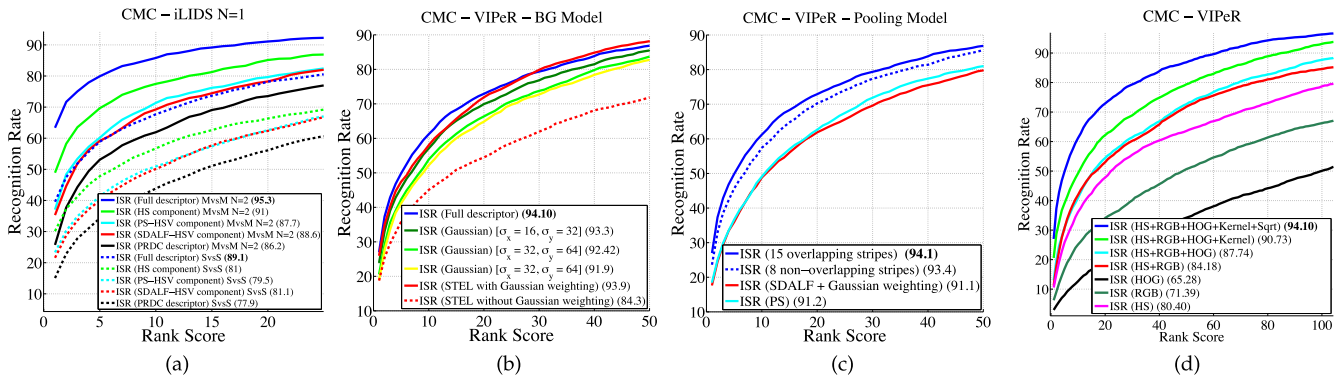


Fig. 6. Comparison with state-of-the-art descriptors and analysis of the contribution of each descriptor component. (a) State-of-the-art descriptors in our ISR framework on i-LIDS. Solid lines represent MvsM (N = 2) and dashed lines represent SvsS. (b) ISR and our descriptor with different background models on VIPeR. (c) ISR and our descriptor with different pooling models on VIPeR. (d) The contribution of each descriptor component on the VIPeR dataset.

5.2.5 Sensitivity to Illumination Changes

We performed a series of experiments on VIPeR to quantify the sensitivity of our descriptor to illumination changes. For each pair of images in the VIPeR dataset we estimate the difference in illumination by applying a Gaussian smoothing kernel to the value channel in the HSV color space for each image, weighting the filtered intensity images with the Epanechnikov kernel, and then computing the difference in average intensity between the two images. The full set of intensity variations between corresponding images in the dataset was quantized in 16 bins. We then performed leave-one-out cross validation to estimate sensitivity to illumination changes: each image was used as a probe and compared against all other images. Results are shown in Fig. 7a. The rank at which the correct gallery image was returned is recorded as a function of illumination change. All possible ranks were also quantized into 16 bins and represented with different colors from dark blue (bin 1) to red (bin 16). Looking at the first bin of illumination variations, where all the image pairs have almost equal illumination, we see that the correct image always appears in the first ranking bin. On average, re-identification in the first ranking bin is unaffected by changes in illumination in about 40 percent of cases.

5.2.6 Sensitivity to Viewpoint Changes

We also performed experiments to evaluate the sensitivity to viewpoint changes, exploiting the fact that VIPeR contains ground truth viewpoint annotations for all subjects (it is in fact the only publicly available dataset with such annotations). Four different changes of viewpoints were considered: 45, 90, 135, and 180 degrees. Also in this case, the experiments were performed using leave-one-out cross validation. All possible ranks were quantized into 16 bins and represented with different colors. From Fig. 7b, we see that the full descriptor is robust to viewpoint variations and that re-identification in the first ranking bin is approximately pose invariant in 45 percent of the cases.

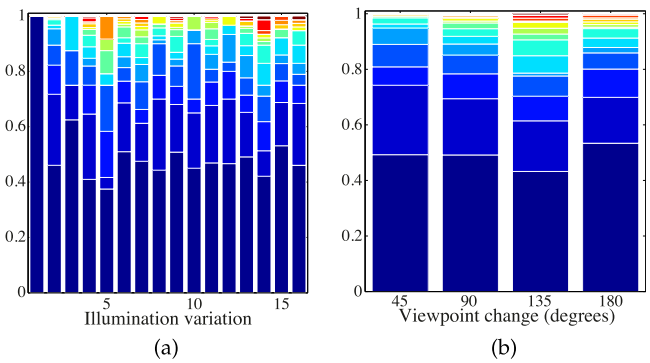


Fig. 7. Sensitivity to (a) illumination and (b) viewpoint changes.

5.2.7 Sensitivity to Misaligned Detection Windows

Finally, we analyzed the sensitivity of the full descriptor to conditions where the detection window is not well-centered on the person. To this end we used the ETHZ datasets because they also provide the original video frames from which each person image was extracted. This allowed us to artificially generate detection misalignments by shifting the location of the person bounding box in the original frame. Fig. 8 shows the SvsS recognition rate at rank-1 for different degrees of misalignment applied to the probe images. Results were averaged over ten trials, over all person images in all three ETHZ datasets. In the same figure we also report the performance of our descriptor with the PS pooling model in the presence of misalignment. We see that the ISR method with our descriptor (blue curves)

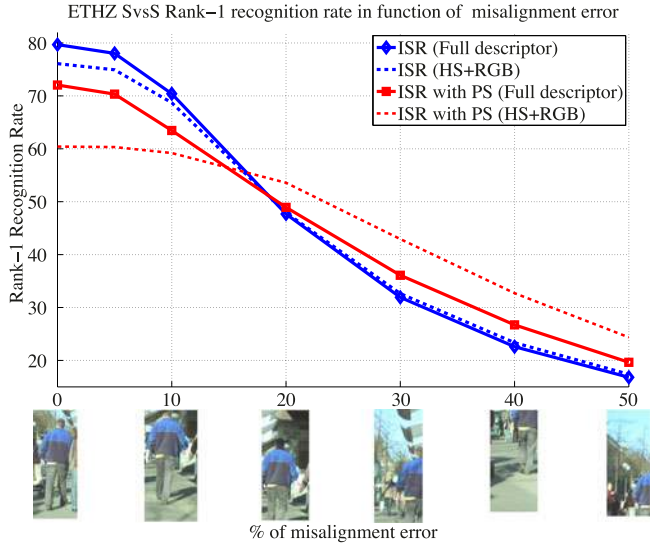


Fig. 8. Rank-1 accuracy of ISR with our descriptor and PS over a range of misalignment error. Images are random samples of misaligned imagery.

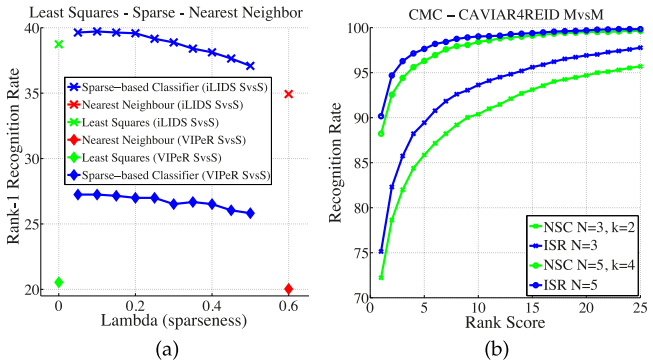


Fig. 9. (a) Rank-1 accuracy on VIPeR and i-LIDS for SvsS. Accuracy is plotted for varying sparseness (λ), including least squares ($\lambda = 0$) and the nearest neighbor ($\lambda \approx 0.6$) solutions. (b) Comparison of ISR with the Nearest Subspace Classifier on CAVIAR4REID. In the legend we report the number of instances per person (N) and number of learned subspaces (k).

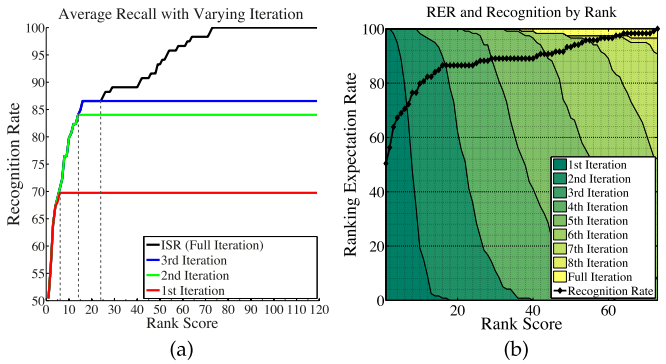


Fig. 10. Iterative ranking and its effect on recall. (a) Average recognition rates with cutoffs for each iteration. (b) RER and recognition rate as function of rank.

outperforms those with PS pooling (red curves) for misalignments up to 17 percent. For higher misalignments of the detection window, the adaptive PS pooling model performs better and the additional complexity of fitting body-part models may be warranted. Note that the HOG component (solid lines) does not negatively affect the overall performance at any degree of misalignment. In contrast, the HOG component appears to improve performance of our descriptor with PS pooling at small misalignments and affect it negatively at higher ones. From the sample thumbnails in Fig. 8, it is anyway clear that for misalignments beyond 30 percent the person image has insufficient visual content for accurate re-identification.

5.3 Evaluation of Iterative Sparse Reconstruction

In this section we investigate how ℓ_1-regularized sparse basis expansion aids in re-identification in comparison to nearest-neighbor, least-squares, and nearest-subspace classification. We also demonstrate how iteration contributes to improve recall, and thus to high recognition rates, across all ranks.

5.3.1 Contribution of Sparse Reconstruction

In Fig. 9a we show the average rank-1 re-identification accuracy for a range of solutions to the regularized least squares problem of Eq. (6) over 10 trials on the VIPeR and i-LIDS datasets. Shown are the least squares ($\lambda = 0$) solution, sparse solutions for a range of $\lambda > 0$, and the nearest neighbor solution when λ is sufficiently high to constrain the solution to a single non-zero coefficient. The sparse approach, for appropriate λ, outperforms the nearest neighbor and least squares solutions on both datasets.

In Fig. 9b we give a comparison of ISR with Nearest Subspace Classifier (NSC) [32] for MvsM on CAVIAR4REID. We chose NSC as a baseline since it is representative of linear, non-sparse approaches to recognition. A drawback of NSC is that it cannot effectively learn a subspace if the number of instances per person is low, while ISR is robust with as few as two or three gallery examples per person. This is especially evident in Fig. 9b for $N = 3$ gallery images per person. We used the CAVIAR4REID dataset for these experiments because of the need for more than four images per person to learn subspaces.

5.3.2 Contribution of Iteration to Recall

Iteration of our sparse ranking algorithm is effective not only for high ranks, but at middle and low ranks as well. Fig. 10a shows the average cutoff ranks for each iteration on the i-LIDS dataset and recognition rate (recall) across ranks. These experiments were performed for MvsS ($N = 2$) and averaged over 50 random splits. We see that with a single iteration we can rank seven gallery persons, on average, and achieve an average recognition rate of less than 70 percent. After this point, the red curve (corresponding to the first iteration) levels off since no more gallery persons can be ranked, on average. Note how the first three iterations yield a steep increase in average recognition rate at low and middle ranks, leveling off on average at rank 14 and 24, respectively. The remaining iterations contribute more slowly to recall at higher ranks.

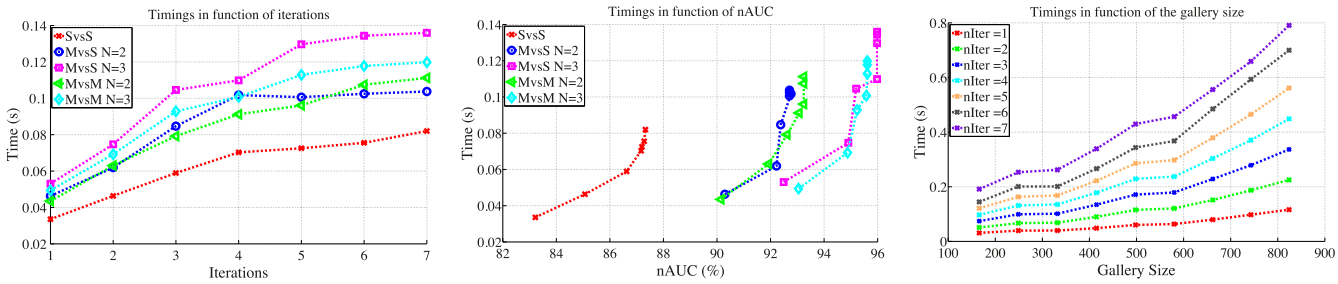
Fig. 10b shines more light on the contribution of each iteration to recall at all ranks. This plot makes use of a metric we introduce to quantify the expected number of ranks Algorithm 1 returns in each iteration, on average. The *Ranking Expectation Rate* $\mathrm{RER}(r, s)$ is the expectation of ranking at least r elements in s iterations or less. Using the notation of Algorithm 1 we define the RER for a single-shot probe set $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$ as:

$$\mathrm{RER}(r, s) = \frac{1}{m} |\{\mathbf{y} \in Y \ : \ |R_s(\mathbf{T}, \{\mathbf{y}\}, \lambda)| \geq r\}|, \qquad (12)$$

where $R_s(\mathbf{T}, \{\mathbf{y}\}, \lambda)$ denotes the restriction of the ranking of Algorithm 1 to iteration s or before:

$$R_s(\mathbf{T}, \{\mathbf{y}\}, \lambda) = \{(i', s', e') \in R(\mathbf{T}, \{\mathbf{y}\}, \lambda) \ : \ s' \leq s\}. \qquad (13)$$

Fig. 10b illustrates the RER as a function of rank for varying numbers of iterations in Algorithm 1. Superimposed on this plot is also the corresponding recognition rate at each cutoff rank. Note that iteration contributes to increased RER, and consequently recognition rate, not only at high ranks but at all ranks. For many probes only a few gallery persons are ranked by a single iteration, and thus iteration contributes significantly to increasing recall also at low ranks. We plot recall until rank 73 in this plot, at which point it saturates after eight iterations.



(a) Time as a function of iterations. (b) Time as a function of desired nAUC. (c) Time as a function of gallery size.

Fig. 11. Time required for re-identification. (a) Time as a function of the number of iterations of sparse re-weighted ranking on the i-LIDS dataset. (b) Time for a desired nAUC on the i-LIDS dataset. (c) Time as a function of the total number of gallery images on the ETHZ1 dataset for the MvsS $N = 10$ modality. Timings are averaged over all probes in fifty random splits.

5.4 Discussion

In this section we summarize our contribution terms of performance with respect to the state-of-the-art and in terms of computational efficiency.

5.4.1 General Performance Considerations

The trend that emerges from the experimental evaluation is that, with the exception of some learning-based approaches discussed below, ISR exceeds the state-of-the-art at rank-1. This can be seen in the SvsS modality on all datasets, but the increase in performance is dramatic on MvsS and MvsM on i-LIDS, ETHZ and CAVIAR4REID. On these multi-shot datasets ISR exceeds the state-of-the-art at all ranks.

Ranking based on sparse, ℓ_1-regularized basis expansions allows ISR to exploit multiple aspects of person appearance in the multi-image galleries of the MvsS and MvsM modalities. This is noticeable from the trend of the curve on the i-LIDS dataset in Fig. 4d where we quickly reach a 90 percent recognition rate at around rank-15 and on CAVIAR4REID in Fig. 5b where we reach 100 percent accuracy around rank-20. In a simple experiment on i-LIDS (MvsM with $N = 3$) to explore the effect of number of gallery exemplars per person, we found that reducing each gallery set to two exemplars had almost no effect on recognition rate at all ranks, while reducing to only one noticeably reduced performance. With one exemplar per gallery person (and a multi-shot probe) performance was still better than for SvsS, but the trend seems to be that reducing the number of exemplars in the gallery converges towards SvsS performance.

5.4.2 Comparison with Metric Learning

Some metric learning approaches outperform the ISR method at higher ranks for SvsS on VIPeR and i-LIDS. By setting aside a portion of the labeled data they are able to learn a metric that better captures the intrinsic properties of the scene, of the cameras used, and of the camera positioning and imaging conditions. This increase in performance at high rank comes at a cost. On VIPeR, for example, as much as half of all available labeled data is used for metric learning and this limits the availability of data for actual testing. Not only does this render experimental results not strictly comparable, it is also a severe limitation in real application scenarios where no labeled data may be available a priori. An important advantage of ISR with respect to learning-based ones is that learned distance metrics cannot be easily updated when camera settings or positions change, while we can easily integrate new instances per person and discard old gallery images.

5.4.3 Computational Efficiency

Our approach is implemented in MATLAB using the optimized SPAMS library for sparse modeling [33]. All tests were performed on an Intel Xeon@2.67 GHz (8-core) with 12 GB RAM.[1] Descriptor extraction in MATLAB requires about 0.016 s per person image and is included in all timing numbers reported here.

In Fig. 11 we report three views of the computational requirements of ISR. In Fig. 11a we vary the number of iterations of sparse re-weighted ranking we perform in order to quantify how computational requirements change with increasing iterations (and increasing accuracy). Fig. 11b, on the other hand, quantifies the relationship between the time required for performing a single re-identification and the area under the curve. From these curves we see that, if we are interested only in first rank, we can perform re-identification of a single probe in about 0.036 s. In real application scenarios ISR can thus perform rank-1 SvsS person re-identification at about 30 re-identifications per second.

If we are interested in higher ranks, for example in an interactive application in which a human operator will sift through re-identification results, ISR might require more than one iteration. From Fig. 11b we see that after seven iterations we arrive at a nAUC of about 88 percent, requiring 0.08 s to compute this result (which works out to about 12 re-identifications per second). In the MvsM modality ISR requires about 0.14 s (seven re-identifications per second), but yields a nAUC of more than 94 percent. These first two tests were carried out on the i-LIDS dataset.

In Fig. 11c we show how ISR scales as a function of the gallery size. The time for a single re-identification increases approximately linearly when increasing the number of images in the basis up to 600 images; then the trend becomes superlinear from 600 to 900. It is interesting that this non-linearity is more pronounced with increasing number of iterations. This test was carried out on the ETHZ1 dataset which contains the most images, and all measurements obtained by averaging over 50 random splits of gallery/probe image sets.

1. Source code available at: http://www.micc.unifi.it/lisanti/source-code/re-id/.

6 Conclusions

In this paper we described an approach to person re-identification that is based on sparse, ℓ_1-regularized basis expansions of probes in terms of a set of gallery examples used as basis vectors. We showed how to extend, through iteration and re-weighting, the concept of a Sparse Discriminative Classifier to problems requiring ranked output. Our algorithm is efficient and obtains state-of-the-art performance on both multi- and single-shot person re-identification modalities. Our results demonstrate how sparse reconstruction generally leads to higher performance at first rank, while also yielding higher nAUC using the proposed iterative ranking. It is feature agnostic and it can be applied to any feature that is encoded as a fixed-length vector. ISR is also competitive with respect to metric learning-based methods which set aside data for training.

Our approach makes use of a simple, yet discriminative descriptor of person appearance. It requires no foreground/background separation or body part segmentation. It is simple and extremely efficient to calculate, and the performance of our approach demonstrates that simple descriptors can be successfully applied in re-identification scenarios.

Iterative sparse ranking is a general approach and can be applied to retrieval problems beyond person re-identification. Our use of ℓ_1-regularized basis expansions for ranking shares some similarities with iterative algorithms such as LARS used to solve weighted Lasso problems like Eq. (10). An interesting line of research would investigate the possibility of directly incorporating soft- and hard-weighting of coefficients into a single regularization path capable of robustly ranking many candidates in a single iterative pass over basis vectors.

Acknowledgments

This work was partially supported by Thales Italia. G. Lisanti acknowledges the support of the AQUIS-CH Fellowship (POR-CRO-FSE 2007-2013/UNIFI_FSE2012), and A. D. Bagdanov the support of a Ramon y Cajal Fellowship (RYC-2012-11776).

References

[1] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proc. 10th IEEE Workshop Perform. Eval. Tracking Surveillance*, 2007, pp. 41–48.

[2] T. Gandhi and M. Trivedi, "Panoramic appearance map (PAM) for multi-camera based person re-identification," in *Proc. IEEE Int. Conf. Video Signal Based Surveillance*, 2006, p. 78.

[3] N. Gheissari, T. Sebastian, and R. Hartley, "Person reidentification using spatiotemporal appearance," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 1528–1535.

[4] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proc. 10th Eur. Conf. Comput. Vis.*, 2008, pp. 262–275.

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

[6] L. Bazzani, M. Cristani, A. Perina, M. Farenzena, and V. Murino, "Multiple-shot person re-identification by HPE signature," in *Proc. 20th Int. Conf. Pattern Recognit.*, 2010, pp. 1413–1416.

[7] L. Bazzani, M. Cristani, A. Perina, and V. Murino, "Multiple-shot person re-identification by chromatic and epitomic analyses," *Pattern Recognit. Lett.*, vol. 33, no. 7, pp. 898–903, 2011.

[8] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification," in *Proc. British Mach. Vis. Conf.*, 2011, pp. 1–11.

[9] S. Bak, E. Corvee, F. Bremond, and T. Monique, "Multiple-shot human re-identification by mean Riemannian covariance grid," in *Proc. 8th IEEE Int. Conf. Adv. Video Signal-Based Surveillance*, 2011, pp. 179–184.

[10] S. Bak, E. Corvee, F. Bremond, and M. Thonnat, "Person re-identification using spatial covariance regions of human body parts," in *Proc. 7th IEEE Int. Conf. Adv. Video Signal-Based Surveillance*, 2010, pp. 179–184.

[11] W.-S. Zheng, S. Gong, and T. Xiang, "Associating groups of people," in *Proc. British Mach. Vis. Conf.*, 2009, 23.1–23.11.

[12] W. R. Schwartz and L. S. Davis, "Learning discriminative appearance-based models using partial least squares," in *Proc. XXII Brazilian Symp. Comput. Graph. Image Process.*, 2009, 322–329.

[13] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja, "Pedestrian recognition with a learned metric," in *Proc. 10th Asian Conf. Comput. Vis.*, 2011, pp. 501–512.

[14] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by support vector ranking," in *Proc. British Mach. Vis. Conf.*, 2010, pp. 1–11.

[15] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2288–2295.

[16] W. Zheng, S. Gong, and T. Xiang, "Re-identification by relative distance comparison," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 653–668, Jun. 2012.

[17] S. G. W. Zheng and T. Xiang, "Transfer re-identification: From person to set-based verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2650–2657.

[18] Y. Wu, M. Minoh, M. Mukunoki, and S. Lao, "Set based discriminative ranking for recognition," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 497–510.

[19] T. Avraham, I. Gurvich, M. Lindenbaum, and S. Markovitch, "Learning implicit transfer for person re-identification," in *Proc. 12th Int. Conf. Comput. Vis.: Workshops Demonstrations*, 2012, pp. 381–390.

[20] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof, "Relaxed pairwise learned metric for person re-identification," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 780–793.

[21] S. Bak, G. Charpiat, E. Corvee, F. Bremond, and M. Thonnat, "Learning to match appearances by correlations in a covariance metric space," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 806–820.

[22] X. Wang, R. Zhao, and W. Ouyang, "Unsupervised salience learning for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3586–3593.

[23] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[24] R. Basri and D. W. Jacobs, "Lambertian reflectance and linear subspaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 2, pp. 218–233, Feb. 2003.

[25] N. Jojic, A. Perina, M. Cristani, V. Murino, and B. Frey, "Stel component analysis: Modeling spatial correlations in image class structure," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 2044–2051.

[26] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *Int. J. Comput. Vis.*, vol. 61, pp. 55–79, 2005.

[27] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 886–893.

[28] R. G. Cinbis, J. Verbeek, and C. Schmid, "Image categorization using Fisher kernels of non-iid image models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2184–2191.

[29] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Stat.*, vol. 32, pp. 407–499, 2004.

[30] E. Cands, M. Wakin, and S. Boyd, "Enhancing sparsity by reweighted l1 minimization," *J. Fourier Anal. Appl.*, vol. 14, pp. 877–905, 2008.

[31] M. Hirzer, P. Roth, and H. Bischof, "Person re-identification by efficient impostor-based metric learning," in *Proc. IEEE 9th Int. Conf. Adv. Video Signal-Based Surveillance*, 2012, pp. 203–208.

[32] Y. Liu, S. Ge, C. Li, and Z. You, "k-NS: A classifier by the distance to the nearest subspace," *IEEE Trans. Neural Netw.*, vol. 22, no. 8, pp. 1256–1268, Aug. 2011.

[33] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res.*, vol. 11, pp. 19–60, 2010.

Giuseppe Lisanti received the PhD degree in computer science from the Università di Firenze. He is a postdoc at the Media Integration and Communication Center and his main research interests focus on computer vision, pattern recognition, and machine learning.

Iacopo Masi received the master's degree in computer science from the Università di Firenze, Italy. He is currently working toward the PhD degree at the Media Integration and Communication Center. His research interests include pattern recognition and computer vision, specifically the subjects of tracking with PTZ cameras, person re-identification, and 2D/3D face modeling.

Andrew D. Bagdanov received the PhD degree in computer science from the University of Amsterdam. He is the head of Research Unit at the Media Integration and Communication Center, at the Università di Firenze. His research spans a broad spectrum of computer vision, image processing and machine learning. He is a member of the IEEE.

Alberto Del Bimbo is a full professor of computer engineering at the Università di Firenze, Italy, where he is the director of the Media Integration and Communication Center. His research interests include multimedia processing and computer vision. He is a member of the IEEE.