Person Re-Identification with Block Sparse Recovery

Srikrishna Karanam*, Yang Li, Richard J. Radke

Department of Electrical, Computer, and Systems Engineering
Rensselaer Polytechnic Institute

Abstract

We consider the problem of automatically re-identifying a person of interest seen in a "probe" camera view among several candidate people in a "gallery" camera view. This problem, called person re-identification, is of fundamental importance in several video analytics applications. While extracting knowledge from high-dimensional visual representations based on the notions of sparsity and regularization has been successful for several computer vision problems, such techniques have not been fully exploited in the context of the re-identification problem. Here, we develop a principled algorithm for the re-identification problem in the general framework of learning sparse visual representations. Given a set of feature vectors for a person in one camera view (corresponding to multiple images as they are tracked), we show that a feature vector representing the same person in another view approximately lies in the linear span of this feature set. Furthermore, under certain conditions, the associated coefficient vector can be characterized as being block sparse. This key insight allows us to design an algorithm based on block sparse recovery that achieves state-of-the-art results in multi-shot person re-identification. We also revisit an older feature transformation technique, Fisher discriminant analysis, and show that, when combined with our proposed formulation, it outperforms many sophisticated methods. Additionally, we show that the proposed algorithm is flexible and can be used in conjunction with existing metric learning algorithms, resulting in improved ranking performance. We perform extensive experiments on several publicly available datasets to evaluate the proposed algorithm.

1. Introduction

Automated human re-identification, or re-id, systems play a key role in several security and surveillance applications. Given a sequence of images of a person of interest in one camera view (the probe view), the goal is to re-identify the same person among several candidate person image sequences in another camera view (the gallery view). This is a particularly challenging problem since inter-camera appearance and illumination variations are often quite pronounced. Background clutter, viewpoint variations, and occlusions can further complicate this task. Some of these commonly encountered challenges are visually summarized in Figure 1.

*Corresponding author. email: karans3@rpi.edu.

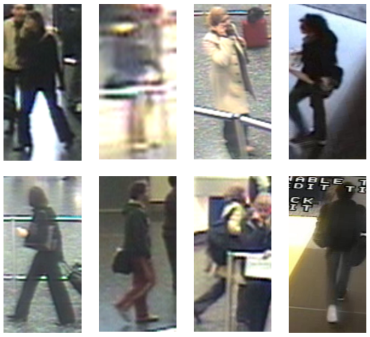


Figure 1: Person re-identification in networks of non-overlapping camera views is a challenging problem due to background clutter, illumination variations, occlusions, and viewpoint variations. Each column illustrates two images of the same person in two different camera views.

Recent advances in compressed sensing and sparse recovery have motivated several solutions for traditional computer vision problems such as face recognition [1, 2] and object tracking [3, 4]. The basic idea in such sparse recovery methods is to model high-dimensional visual data using sparse vectors. These methods are generally characterized by learning representations of high-dimensional visual data that are more discriminative for making downstream recognition and classification decisions. While it is conceivable that such methods would naturally extend to the re-id problem, they are currently not well represented in existing re-id algorithms.

In this article, we present a principled approach to address the multi-shot re-id problem. Our algorithm design is inspired by the success of learning sparse representations of visual data and based on the following key observations:

- In some discriminative feature space, the feature vector corresponding to a probe image of a person can be approximately expressed as a linear combination of the feature vectors of the corresponding gallery images of the same person.
- If we construct a matrix $\mathbf{D}$ (a dictionary) of feature vectors corresponding to all the available images for each unique person in the gallery view, the intuition above suggests that the probe feature vector can be expressed as a sparse linear combination of the columns of $\mathbf{D}$.
- Due to the manner of its construction, the dictionary $\mathbf{D}$ will have a block structure because multiple images are available for each unique person in the gallery. Furthermore, these sets of images corresponding to each person will ideally result in clustered sets of points in the feature space. This suggests that the sparse coefficient vector will also have a block structure.

Based on these observations, we formulate the task of re-identifying a person of interest as a block sparse recovery problem. We construct the feature dictionary $\mathbf{D}$ using the available gallery feature vectors and solve the associated optimization problem in the framework of alternating directions minimization. The feature space for the dictionary is learned using a well-established technique, Fisher discriminant analysis

(FDA). We show that using features learned with FDA and ranking gallery candidates based on our block sparse minimization approach outperforms many sophisticated and recently proposed algorithms for re-id.

We show that our block sparse formulation is flexible and can be used in conjunction with any existing metric learning technique to rank gallery candidates. Typically, a metric learning technique ranks candidates using the Euclidean distance metric in the learned feature space. We empirically demonstrate that using our approach to rank candidates after learning the feature space transformation can result in significant performance improvement. In this regard, our approach is complementary to existing metric learning methods and can improve their performance.

To evaluate each of our primary contributions, we perform extensive quantitative evaluations on three publicly available multi-shot re-id datasets: iLIDS-VID [5], PRID 2011 [6], and SAIVT-SoftBio[7]. Our results show a rank-1 performance improvement of about 6% for iLIDS-VID, 6% for PRID 2011 and 12% for SAIVT-SoftBio over the existing state of the art. An earlier version of a portion of this paper appeared in [8].

2. Related work

Most existing research on person re-id is focused on the single-shot version of the problem [9, 10, 11, 12, 13], i.e., the assumption that only one image per person per camera is available. However, in practice, this is not the case. In a typical surveillance application, such as the "tag and track" problem described in [14, 15], after the person of interest is identified in the probe view, s/he is tracked until the end of the current view. Similarly, candidates observed in the gallery view will also be tracked, generating a sequence of images for each candidate person. Therefore, real-world re-id necessitates the formulation of the task as a "multi-shot", or image sequence matching problem rather than the usual single-image matching problem.

A typical approach to address the re-id problem is to compute features from all the available probe and gallery images and then compare them. This has led to considerable research in two directions: (1) determining the most discriminative appearance features and subsequently ranking gallery candidate images using a traditional metric, such as the Euclidean distance, and (2) starting with basic feature vectors, such as global color histograms, and learning, in a supervised fashion, a distance metric such that feature vectors belonging to the same person are close while those belonging to different people are far apart. In the following, we summarize some key methods in these directions.

2.1. Appearance features, distance metric learning and single-shot re-id

An early appearance based re-id method was proposed by Gray and Tao [16], where color and texture histograms were extracted from locally sampled image regions. Color histograms were extracted in the RGB, YCbCr, and HSV color spaces and texture histograms were computed as filter responses to several channels of Schmid [17] and Gabor [18] filters. Following this work, several metric learning methods were proposed that started with these color and texture histograms as the basic features. Prosser *et al.* [9] used these features to learn a RankSVM [19] model by enforcing explicit constraints that the feature vectors belonging to the same person be close whereas the feature vectors of different people be far apart. Mignon and Jurie [20] formulated the problem of learning the metric as a generalized logistic loss minimization problem, while enforcing pairwise similarity and dissimilarity constraints. Zhang *et al.* [21] also started with the basic color and texture histograms and formulated re-id as a relative distance comparison problem, where a logistic objective function in a soft margin framework was employed to learn the distance metric. Pedagadi *et al.* [22] employed dense patch-based image sampling, computed color histograms in each patch, and used local Fisher discriminant analysis [23] [24] to learn a distance metric. Recently, Xiong *et al.* [12] proposed kernelized variants of several popularly used metric learning techniques.

Constructing appearance descriptors that are robust to common inter-camera variations has also been an extensive line of study. Such an approach typically uses existing distance functions to compute the similarity score between the probe and the gallery feature vector. Farenzena *et al.* [25] constructed an ensemble of localized features, including weighted color histograms and maximally stable color regions to describe the appearance of each image. Subsequently, similarity scores were accumulated across the feature channels using the Euclidean and Bhattacharya distance measures. Cheng *et al.* [26] described the appearance of each image using an ensemble of multiple feature channels. These features were computed from parts of the image that were found automatically by fitting a pictorial structure to each image. Bak *et al.* [27] also employed a part-based appearance modeling scheme, where the body parts were determined using histograms of oriented gradients [28]. Each part was then described using a covariance matrix of multiple feature cues, and a spatial pyramid matching [29] scheme was used to compute the similarity score.

Martinel *et al.* [30] hypothesized that identifying salient regions in the image of a person would lead to robust appearance descriptors. To this end, a saliency detection method based on kernelized graphs was proposed. The detected salient regions were then used as priors in constructing appearance features, following which a Mahalanobis distance metric was learned to account for inter-camera variations. Zheng *et al.* [31] approached the re-id problem from an *open-world* perspective, where it is likely that a probe person might not appear in the search gallery. Such a scenario is common in real-world re-id applications such as the tag-and-track problem described in Camps *et al.* [15]. While not directly solving this problem, Zheng *et al.* [31] proposed to verify whether each of the observed gallery persons exists in a pre-existing *watch-list* of people provided beforehand. In this way, the problem of person re-identification was converted to one of person verification, which was tackled using a distance metric learning method based on the principle of transfer learning.

However, as noted earlier, these techniques were developed specifically to address the single-shot setting of the re-id problem and do not have any mechanism to exploit the inherent discriminability available in the multi-shot setting. In the following, we summarize some key techniques that specifically address the multi-shot re-id problem.

2.2. Multi-shot re-id

An early approach that used multi-shot data was based on gait analysis [32], where gait patterns extracted from video were used to identify people. Subsequently, the advances in appearance models and distance metric learning for single-shot re-id motivated several specialized multi-shot re-id methods. Most methods follow one or more of the following lines of thought: (i) exploiting all the available data to build aggregated appearance descriptors that can then be compared using either existing distance functions or learned distance metrics, (ii) selecting images or fragments from the set of available images that are most discriminative, and (iii) representing the set of available images as a sequence of feature vectors that can then be used to perform direct sequence-to-sequence matching.

Cong *et al.* [33] used a graph embedding approach to learn a manifold and employed Euclidean distances between the centers of the points in the new space to measure similarity/dissimilarity. Bazzani *et al.* [34] constructed histogram and epitome-based features to embed discriminative information from all the available images and used the Bhattacharya distance metric to measure similarity/dissimilarity. Wang *et al.* [5] constructed a model based on a combination of optical flow, space-time features and multiple-instance learning to simultaneously select and match fragments from the available data that are most discrminative for re-id. Li *et al.* [35] learned discriminative random forests and aggregated classification scores for all the available images for each person to make a decision. Li *et al.* [36] also developed a feature transformation algorithm that combined hierarchical image sequence clustering and Fisher discriminant analysis to learn a discriminative feature space. Subsequently, a RankSVM model was used to rank the gallery candidates. Image sequences have also been used to perform direct sequence matching. Simonnet *et al.* [37] used a tracking-by-detection approach to generate a track of images for each candidate, followed by direct sequence-to-sequence appearance feature matching using the dynamic time warping algorithm.

Martinel *et al.* [38] proposed a novel concept called *warp functions*, based on the observation that inter-camera feature variations lie in a non-linear function space of all the possible feature transformations between two cameras. Subsequently, a discriminative surface was learned using a random forest classifier, separating warp functions of images of the same person and the warp functions of images of different people. Chakraborty *et al.* [39] proposed a generalized technique to address the re-id problem in camera networks with more than two cameras. Their framework involved formulating a binary integer programming problem that minimizes the cost of associating pairs of target persons across the entire camera network, while enforcing network consistency constraints depending on whether a single image or multiple images were available for a person in each camera view.

2.3. Sparsity and regularization in computer vision

The notion of sparse data representations has seen a surge in interest due to recent results [40, 41] from signal processing. Since then, sparsity has gained traction in several computer vision tasks including face recognition [1, 42], object tracking [3, 43, 4] and image restoration [44]. The notion of using l_1 regularization to promote sparse solutions has, in part, motivated several algorithms that adopt some form of regularization in the problem formulation. Xiong *et al.* [12] developed a regularized variant of the popular PCCA metric learing algorithm [20], enforcing a Frobenius norm regularization on the associated feature space projection matrix. Hu *et al.* [45] reformulated the problem of finding the pair of closest points on affine hulls of data, enforcing sparsity-promoting l_1 regularization in the problem formulation. l_1 regularization is also extremely popular in so-called online object tracking algorithms, and as discussed in the benchmarking paper of [46], these methods perform well in the presence of target appearance changes, such as occlusions. A nice review of the various types of penalty and regularization methods, with specific focus on formulations that promote sparse solutions, can be found in the paper by Bach *et al.* [47].

2.3.1. Sparsity in re-identification

In spite of their demonstrated success in these areas, classification methods based on sparse representations have not received much attention in the context of the re-id problem. Harandi *et al.* [48] posed the re-id problem as a dictionary learning task and constructed the associated sparse codes on a Riemannian manifold. Kheder *et al.* [49] represented each gallery image using the SURF [50] features, constructed the appearance dictionary in a dynamic fashion by finding the closest gallery SURF feature vectors to a given test feature vector using a k-d tree, and posed re-id as a sparse vector retrieval problem. Lisanti *et al.* [51] designed a discriminative feature descriptor based on weighted histograms and used them as part of a sparse basis expansion scheme, which was iteratively re-weighted to rank the gallery candidates. Martinel *et al.* [52] constructed a localized person descriptor by randomly sampling image patches and used the notion of sparsity to only consider relevant patches during the descriptor matching process. Zheng *et al.* [53] addressed a variant of the traditional re-id problem wherein only a partial observation, likely due to inaccurate detection or occlusions, of each person was available. To solve this problem, a local patch-to-patch matching framework was proposed. A sparse modeling framework was proposed to select the most suitable patches for matching at test time.

In contrast to these approaches, we exploit the inherent block sparsity that characterizes the multi-shot re-id problem. Furthermore, our algorithm is very flexible and can be used in conjunction with any existing metric learning method.

3. Algorithm Description

3.1. Features, image clustering, and feature transformation

We begin by describing each image using the Fisher vector (LDFV) approach of Ma *et al.* [54] to compute features, which has been demonstrated to be effective for re-id. As proposed in that work, we convert the image from the RGB space to the HSV space, divide the image uniformly into 6 horizontally-striped regions, construct 17-dimensional local pixel descriptors comprising of spatial, intensity and gradient information, and estimate a Gaussian mixture model in each region using the Expectation-Maximization algorithm [55] as implemented in the VLFeat library [56]. We set the number of Gaussian components to 16. The local descriptors in each region are then encoded in the Fisher vector representation [57, 58], giving a 3264-dimensional descriptor $\mathbf{f}$ for the image $\mathbf{I}$. Since this feature extraction process involves access to training data, we use the generated data splits described in Section 4.2 to compute the training and testing features separately for each split.

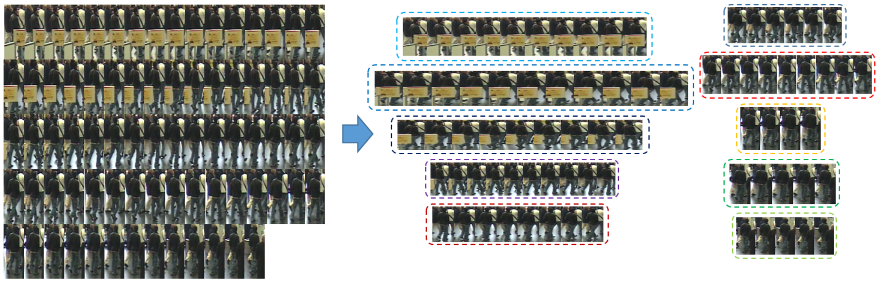


Figure 2: Clustering a tracking sequence of a person from the iLIDS-VID dataset into ten clusters.

Next, we cluster the feature vectors of a person in both the gallery and the probe views to obtain a clustered feature space. Clustering the feature vectors is motivated by several observations:

- Since people are tracked in a real-world re-id application, it is often the case that we obtain redundant image data as the output of a tracking algorithm. Clustering the feature vectors can result in a more compact and meaningful representation of the trajectory of the person.
- As we will see briefly in the next section and in more empirical detail in Section 4, the time required to retrieve the identity of a probe image is directly related to the number of gallery feature vectors used per person. Clustering the available feature vectors results in significant computational gains when compared to using all the available data.

An example of clustering a tracking sequence of a person from the iLIDS-VID dataset into ten clusters using the K-Means clustering algorithm is shown in Figure 2. The figure illustrates that considering the centers of each of the clusters as representative points of the tracking sequence is a succinct representation of the available data. Therefore, we compute the mean of the points in each cluster, resulting in n feature vectors for each person in each of the probe and gallery views, where n is the number of clusters.

Given the feature vectors $\mathbf{y}_{ij}^{g}$ in the gallery view and $\mathbf{y}_{ij}^{p}$ in the probe view in the clustered feature space, where the index j denotes the j^{th} cluster center for the person denoted by index i, we learn a new feature space using Fisher discriminant analysis (FDA). In the following, we give a brief overview of FDA. The goal of FDA is to learn a feature space transformation that maximizes the between-class data scatter while minimizing the within-class data scatter. By stacking all the N gallery and probe feature vectors in the clustered space column-wise, we construct the matrix $\mathbf{F} \in \mathbb{R}^{c \times N}$, where $c = 3264$. We can then define the associated within-class and between-class data scatter matrices as:

$$\mathbf{S}_w = \frac{1}{2} \sum_{a,b=1}^{N} \mathbf{A}_{ab}^{w} (\mathbf{F}_a - \mathbf{F}_b)(\mathbf{F}_a - \mathbf{F}_b)^\top \qquad (1)$$

$$\mathbf{S}_b = \frac{1}{2}\sum_{a,b=1}^{N} \mathbf{A}^b_{ab}(\mathbf{F}_a - \mathbf{F}_b)(\mathbf{F}_a - \mathbf{F}_b)^\top \tag{2}$$

where $\mathbf{F}_a$ denotes the a^{th} column of the matrix $\mathbf{F}$, and $\mathbf{A}^w_{ab}$ and $\mathbf{A}^b_{ab}$ are defined as

$$\mathbf{A}^w_{ab} = \begin{cases} \frac{1}{n_c} & \text{if class}(\mathbf{F}_a) = \text{class}(\mathbf{F}_b) = c \\ 0 & \text{if class}(\mathbf{F}_a) \neq \text{class}(\mathbf{F}_b) \end{cases} \tag{3}$$

$$\mathbf{A}^b_{ab} = \begin{cases} \frac{1}{N} - \frac{1}{n_c} & \text{if class}(\mathbf{F}_a) = \text{class}(\mathbf{F}_b) = c \\ \frac{1}{N} & \text{if class}(\mathbf{F}_a) \neq \text{class}(\mathbf{F}_b) \end{cases} \tag{4}$$

Here, n_c denotes the number of feature vectors available for the person indexed by c. We then take the trace of the resulting within-class and between-class data scatter matrix in the $d-$dimensional projected feature space as the scalar measure of the data variance and learn a transformation $\mathbf{T} \in \mathbb{R}^{d_1 \times d}$ from the following optimization problem:

$$\mathbf{T} = \arg\max_{\mathbf{T}} \ \text{trace}\{(\mathbf{T}^\top \mathbf{S}_w \mathbf{T})^{-1}\mathbf{T}^\top \mathbf{S}_b \mathbf{T}\} \tag{5}$$

3.2. Block Sparsity for Re-Identification

In this section, we first describe our formulation of re-id as a block sparse recovery problem. We then describe an efficient algorithm in the alternating directions framework to recover the block sparse vector given the data dictionary and an observation vector.

3.2.1. Preliminaries and Notation

Let $\mathbf{T}$ be the feature space transformation matrix learned as described above using the available training data. Let K be the number of unique people in the gallery view of the test dataset. Given the feature vectors in the LDFV space for each of the available images in the gallery and the probe views for each person in the test set, we first cluster them into n clusters and determine the cluster centers. We then project these cluster centers into the learned feature space using the transformation matrix $\mathbf{T}$. Let $\mathbf{g}_{ij} = \mathbf{T}^\top \mathbf{y}^g_{ij}$ and $\mathbf{p}_{ij} = \mathbf{T}^\top \mathbf{y}^p_{ij}$ denote these projected gallery and probe cluster centers respectively, where i represents the i^{th} test person, with $i = 1, 2, \ldots, K$ and j represents the j^{th} projected cluster center, with $j = 1, 2, \ldots, n$.

We let $\mathbf{G}_i \in \mathbb{R}^{d \times n}$ denote the dictionary specific to the person with index i, and define it as:

$$\mathbf{G}_i = \begin{bmatrix} \mathbf{g}_{i1} & \mathbf{g}_{i2} & \cdots & \mathbf{g}_{in} \end{bmatrix} \tag{6}$$

Essentially, the columns of $\mathbf{G}_i$ are the n cluster centers in the gallery view of the person indexed by i. Now, we construct the gallery feature dictionary $\mathbf{D} \in \mathbb{R}^{d \times N}$, where $N = K \times n$ is total number of available gallery feature vectors across all the people, as:

$$\mathbf{D} = \begin{bmatrix} \mathbf{G}_1 & \mathbf{G}_2 & \cdots & \mathbf{G}_K \end{bmatrix} \tag{7}$$

By construction, the dictionary has a block structure since it is a concatenation of K disparate blocks of vectors. This is further illustrated in Figure 3. This characterization is unique to the multi-shot setting of the re-id problem, which we exploit as explained next to develop a block sparsity approach to person re-id.

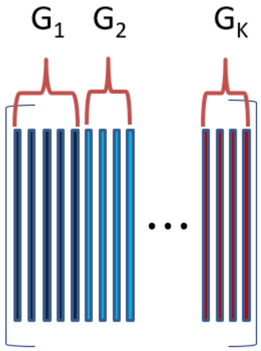


Figure 3: The dictionary $\mathbf{D}$, comprised of K disparate blocks of feature vectors, has a block structure.

3.2.2. Problem Formulation

Consider the j^{th} feature vector in the probe view $\mathbf{p}_{ij}$ of person i. If the feature space is sufficiently discriminative, $\mathbf{p}_{ij}$ should approximately lie in the feature subspace spanned by the corresponding gallery feature vectors of the same person i, $i.e.$,

$$\mathbf{p}_{ij} \approx x_{i1}\mathbf{g}_{i1} + x_{i2}\mathbf{g}_{i2} + \cdots + x_{in}\mathbf{g}_{in} \tag{8}$$

where each $x_{ik} \in \mathbb{R}$, with $k = 1, 2, \ldots, n$. This equation can be conveniently rewritten as

$$\mathbf{p}_{ij} \approx \mathbf{G}_i\mathbf{x}_i \tag{9}$$

where $\mathbf{G}_i$ is the feature dictionary corresponding to person i as defined above and $\mathbf{x}_i \in \mathbb{R}^n$.

Now, define $\mathbf{x} = \begin{bmatrix}\mathbf{x}_1^\top & \mathbf{x}_2^\top & \cdots & \mathbf{x}_K^\top\end{bmatrix}^\top$ and consider the following linear inverse problem:

$$\mathbf{p}_{ij} = \mathbf{D}\mathbf{x} \tag{10}$$

where $\mathbf{D}$ is the gallery feature dictionary as defined above. This equation can be rewritten as:

$$\mathbf{p}_{ij} = \mathbf{G}_1\mathbf{x}_1 + \mathbf{G}_2\mathbf{x}_2 + \cdots + \mathbf{G}_K\mathbf{x}_K \tag{11}$$

Since our hypothesis is that $\mathbf{p}_{ij}$ approximately lies in the span of the columns of the dictionary $\mathbf{G}_i$ of person i, we note that in the most desirable solution vector $\mathbf{x}$, the contribution from the vector block $\mathbf{x}_i$ dominates the contributions from the vector blocks $\mathbf{x}_k$, $k = 1, 2, \ldots, K, k \neq i$. Thus, we are looking for such a solution $\mathbf{x}$ to the linear inverse problem of Equation 10 that has a property of being block sparse. Furthermore, we also note that this hypothesis is stronger than the model

$$\mathbf{p}_{ij} = \mathbf{D}\mathbf{x} \tag{12}$$

with the hypothesis that $\mathbf{x}$ is sparse. Put another way, instead of looking for a solution vector that has as few non-zero entries as possible, our approach seeks a solution vector that has these non-zero entries concentrated in one of the person-specific blocks of the feature dictionary.

Following [59], we mathematically pose our problem as the following l_1/l_2 minimization task:

$$\begin{array}{ll} \min\limits_{\mathbf{x}} & \sum_{l=1}^{K} \|\mathbf{x}_l\|_2 \\ \text{s.t.} & \mathbf{p}_{ij} = \mathbf{D}\mathbf{x} \end{array} \tag{13}$$

Intuitively, this problem formulation attempts to minimize the l_2 norm, or the energy, of the blocks in the coefficient vector $\mathbf{x} = \begin{bmatrix}\mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_K\end{bmatrix}$. Once we have the solution vector $\mathbf{x}^s$, the identity of the person represented by the test feature vector $\mathbf{p}_{ij}$ is found by determining the vector block m giving the least residual value for the solution vector. Specifically, we determine the residual vector $r_l = \|\mathbf{p}_{lj} - \mathbf{G}_l\mathbf{x}_l^s\|$, $l = 1, 2, \ldots, K$. Subsequently, m is determined as the index of the minimum value in the vector $\mathbf{r}$. Figure 4 provides a visual summary of our entire multi-shot re-id pipeline.

3.2.3. Dealing with clutter and noise

In a typical surveillance application, it is likely that the images of people captured by the cameras will be cluttered with spurious background or noise. Our problem formulation, as described in the previous section, enables us to explicitly deal with such issues. Specifically, we introduce an error term $\mathbf{e} \in \mathbb{R}^d$ into the problem formulation of Equation 10. The modified hypothesis is written as:

$$\mathbf{p}_{ij} = \mathbf{D}\mathbf{x} + \mathbf{e} \tag{14}$$

The l_1/l_2 optimization problem of Equation 13 is appropriately modified to:

$$\begin{array}{ll} \min\limits_{\mathbf{x},\mathbf{e}} & \sum\limits_{l=1}^{K} \|\mathbf{x}_l\|_2 + \|\mathbf{e}\|_1 \\ \text{s.t.} & \mathbf{p}_{ij} = \mathbf{D}\mathbf{x} + \mathbf{e} \end{array} \tag{15}$$

Similarly, the procedure to determine the identity of the test feature vector $\mathbf{p}_{ij}$ given solution vector $\mathbf{x}^s$ and the error vector $\mathbf{e}^s$ is modified to include the effect of the error vector. Specifically, the residual computation becomes $r_l = \|\mathbf{p}_{lj} - \mathbf{G}_l\mathbf{x}_l^s - \mathbf{e}\|$, $l = 1, 2, \ldots, K$ and the identity is determined as before.

3.2.4. Block Sparse Recovery using Alternating Directions

In this section, we describe an iterative scheme to obtain the solution vector $\mathbf{x}^s$ given the test feature vector $\mathbf{p}_{ij}$ and the feature dictionary $\mathbf{D}$. Our approach fits into the general framework of alternating directions minimization [60].

We begin describing the approach by re-formulating Equation 15 using an auxiliary variable $\mathbf{s} \in \mathbb{R}^N$ as:

$$\begin{aligned}
&\min_{\mathbf{s},\mathbf{x},\mathbf{e}} \sum_{l=1}^{K} \|\mathbf{s}_l\|_2 + \|\mathbf{e}\|_1 \\
&\text{s.t. } \mathbf{s} = \mathbf{x} \\
&\qquad \mathbf{p}_{ij} = \mathbf{D}\mathbf{x} + \mathbf{e}
\end{aligned} \tag{16}$$

We convert this constrained minimization problem into an unconstrained one by introducing two Lagrange multipliers $\boldsymbol{\alpha} \in \mathbb{R}^N$ and $\boldsymbol{\beta} \in \mathbb{R}^d$. The resulting minimization problem is:

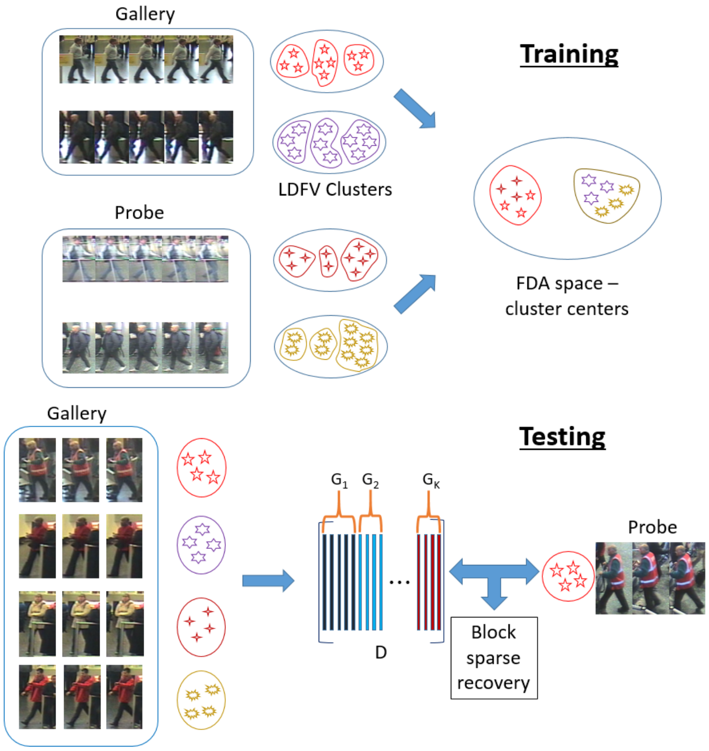


Figure 4: A visual summary of our approach to re-identify a test probe image. In the training stage, given the gallery and probe images for each person, we cluster the feature vectors in the LDFV feature space, giving n cluster centers in both the gallery and probe views for that person. We then learn a feature space transformation using Fisher discriminant analysis and project the cluster centers into the learned feature space. In the testing stage, we first cluster the gallery and probe images and project the resulting cluster centers into the learned feature space. We then construct the feature dictionary $\mathbf{D}$ using the projected gallery feature vectors. Given a probe feature vector, we formulate a block-sparse recovery problem to retrieve the associated identity.

$$\min_{\mathbf{s},\mathbf{x},\mathbf{e}} \sum_{l=1}^{K} \|\mathbf{s}_l\|_2 + \|\mathbf{e}\|_1 - \boldsymbol{\alpha}^\top(\mathbf{s}-\mathbf{x}) - \boldsymbol{\beta}^\top(\mathbf{D}\mathbf{x}+\mathbf{e}-\mathbf{p}_{ij}) \tag{17}$$

We add two quadratic penalty terms $\frac{\eta_1}{2}\|\mathbf{s}-\mathbf{x}\|^2$ and $\frac{\eta_2}{2}\|\mathbf{D}\mathbf{x}+\mathbf{e}-\mathbf{p}_{ij}\|^2$ to the cost function, resulting in a smooth minimization objective. The overall unconstrained minimization problem we now work with is:

$$\begin{aligned} \min_{\mathbf{s},\mathbf{x},\mathbf{e}} \sum_{l=1}^{K} \|\mathbf{s}_l\|_2 + \|\mathbf{e}\|_1 - \boldsymbol{\alpha}^\top(\mathbf{s}-\mathbf{x}) - \boldsymbol{\beta}^\top(\mathbf{D}\mathbf{x}+\mathbf{e}-\mathbf{p}_{ij}) + \\ \frac{\eta_1}{2}\|\mathbf{s}-\mathbf{x}\|^2 + \frac{\eta_2}{2}\|\mathbf{D}\mathbf{x}+\mathbf{e}-\mathbf{p}_{ij}\|^2 \end{aligned} \tag{18}$$

We note that this minimization problem involves three variables $\mathbf{s}$, $\mathbf{x}$, and $\mathbf{e}$. We minimize the objective iteratively with respect to only one variable at a time, while keeping the other two fixed. First, we fix $\mathbf{s}$ and $\mathbf{e}$, and minimize the cost function with respect to $\mathbf{x}$. In this case, the cost function reduces to:

$$\begin{aligned} \min_{\mathbf{x}} \quad & -\boldsymbol{\alpha}^\top(\mathbf{s}-\mathbf{x}) - \boldsymbol{\beta}^\top(\mathbf{D}\mathbf{x}+\mathbf{e}-\mathbf{p}_{ij}) \\ & + \frac{\eta_1}{2}\|\mathbf{s}-\mathbf{x}\|^2 + \frac{\eta_2}{2}\|\mathbf{D}\mathbf{x}+\mathbf{e}-\mathbf{p}_{ij}\|^2 \end{aligned} \tag{19}$$

This $\mathbf{x}$ sub-problem involves optimizing a straightforward quadratic objective, and has a closed-form solution, given by:

$$\mathbf{x}^* = (\eta_1 I + \eta_2 \mathbf{D}^\top\mathbf{D})^{-1}(\eta_2\mathbf{D}^\top(\mathbf{p}_{ij}-\mathbf{e}) + \eta_1\mathbf{s} + \boldsymbol{\beta}^\top\mathbf{D} - \boldsymbol{\alpha}) \tag{20}$$

Next, we fix $\mathbf{s}$ and $\mathbf{x}$, resulting in the following minimization problem with respect to $\mathbf{e}$:

$$\min_{\mathbf{e}} \|\mathbf{e}\|_1 - \boldsymbol{\beta}^\top(\mathbf{D}\mathbf{x}^*+\mathbf{e}-\mathbf{p}_{ij}) + \frac{\eta_2}{2}\|\mathbf{D}\mathbf{x}^*+\mathbf{e}-\mathbf{p}_{ij}\|^2 \tag{21}$$

where $\mathbf{x}^*$ is the $\mathbf{x}$ sub-problem optimal solution. This minimization problem also results in a closed-form solution, given by:

$$\mathbf{e}^* = \text{shrink}\left(\frac{\boldsymbol{\beta}}{\eta_2} - \mathbf{D}\mathbf{x}^* - \mathbf{p}_{ij}, \frac{1}{\eta_2}\right) \tag{22}$$

where $\text{shrink}(\mathbf{t},\alpha) = \text{sgn}(\mathbf{t}) \odot \max\{|\mathbf{t}|-\alpha, 0\}$, and $\odot$ denotes element-by-element multiplication.

Finally, by fixing $\mathbf{x}$ and $\mathbf{e}$, we get the following minimization problem with respect to $\mathbf{s}$:

$$\min_{\mathbf{s}} \sum_{l=1}^{K} \|\mathbf{s}_l\|_2 - \boldsymbol{\alpha}^\top(\mathbf{s}-\mathbf{x}^*) + \frac{\eta_1}{2}\|\mathbf{s}-\mathbf{x}^*\|^2 \tag{23}$$

This minimization problem also admits a closed-form solution, with the elements in each block $l = 1, 2, \ldots, K$ given by the following block shrink [61] scheme:

$$\mathbf{s}_l^* = \max\left(\left\|\mathbf{x}_l^* + \frac{\boldsymbol{\alpha}_l}{\eta_1}\right\| - \frac{1}{\eta_1}, 0\right) \frac{\mathbf{x}_l^* + \frac{\boldsymbol{\alpha}_l}{\eta_1}}{\|\mathbf{x}_l^* + \frac{\boldsymbol{\alpha}_l}{\eta_1}\|^2} \tag{24}$$

We then update the Lagrange multipliers as $\boldsymbol{\alpha} = \boldsymbol{\alpha} - \eta_1(\mathbf{s}^* - \mathbf{x}^*)$ and $\boldsymbol{\beta} = \boldsymbol{\beta} - \eta_1(\mathbf{D}\mathbf{x}^* + \mathbf{e}^* - \mathbf{p}_{ij})$. We summarize the entire iterative block sparse recovery scheme in Algorithm 1.

Algorithm 1: An alternating directions algorithm to solve the minimization problem of Equation 15

```
Input  : $\mathbf{p}_{ij}, \mathbf{D} \in \mathbb{R}^{d \times N}$
Output: $\mathbf{x}^s$, $\mathbf{e}^s$
Initialize $\mathbf{s} = \mathbf{0}$, $\mathbf{e} = \mathbf{0}$, $\boldsymbol{\alpha} = \mathbf{0}$, $\boldsymbol{\beta} = \mathbf{0}$;
$\eta_1 = \frac{2d}{\|\mathbf{p}_{ij}\|_1}, \eta_2 = \eta_1;$
for $t \leftarrow 1, 2, \dots$ do
    $\mathbf{x}^t = (\eta_1 I + \eta_2 \mathbf{D}^\top \mathbf{D})^{-1}(\eta_2 \mathbf{D}^\top(\mathbf{p}_{ij} - \mathbf{e}^{t-1}) + \eta_1 \mathbf{s}^{t-1} + \boldsymbol{\beta}^\top \mathbf{D} - \boldsymbol{\alpha});$
    $\mathbf{e}^t = \text{shrink}(\frac{\boldsymbol{\beta}}{\eta_2} - \mathbf{D}\mathbf{x}^t - \mathbf{p}_{ij}, \frac{1}{\eta_2});$
    $\mathbf{s}_l^t = \max\left(\|\mathbf{x}_l^t + \frac{\boldsymbol{\alpha}_l}{\eta_1}\| - \frac{1}{\eta_1}, 0\right) \frac{\mathbf{x}_l^t + \frac{\boldsymbol{\alpha}_l}{\eta_1}}{\|\mathbf{x}_l^t + \frac{\boldsymbol{\alpha}_l}{\eta_1}\|^2}, l = 1, 2, \dots, K;$
    $\boldsymbol{\alpha} = \boldsymbol{\alpha} - \eta_1(\mathbf{s}^t - \mathbf{x}^t);$
    $\boldsymbol{\beta} = \boldsymbol{\beta} - \eta_1(\mathbf{D}\mathbf{x}^t + \mathbf{e}^t - \mathbf{p}_{ij})$
end
$\mathbf{x}^s = \mathbf{x}^t;$
$\mathbf{e}^s = \mathbf{e}^t;$
```

3.2.5. Re-identification

Given the n cluster centers $\mathbf{p}_{ij}$, $j = 1, 2, \dots, n$ for the person with index i, we solve the minimization problem of Equation 15 for each cluster center. We compute the residual vector $r_l^j = \|\mathbf{p}_{lj} - \mathbf{G}_l \mathbf{x}_l^s - \mathbf{e}\|$, $l = 1, 2, \dots, K$ associated with each cluster center and then determine the net residual vector $\mathbf{R} = \sum_{j=1}^n \mathbf{r}^j$. Subsequently, the identity of the person represented by the n cluster centers is found as the index of the least element in $\mathbf{R}$. Our overall multi-shot re-id framework is summarized in Algorithm 2.

4. Experiments and Results

4.1. Datasets

We experimentally validate the proposed multi-shot person re-identification algorithm on three publicly available multi-shot image sequence based datasets: iLIDS-VID [5], PRID 2011 [6] and SAIVT-SoftBio [7].

4.1.1. iLIDS-VID

The iLIDS-VID dataset was created from person images obtained from two cameras with non-overlapping fields of view. The cameras were located in an airport arrival hall. For each camera view, image sequences of variable length for 300 distinct individuals are available. The number of images in each sequence varies from 23 frames

Algorithm 2: The proposed block sparse algorithm framework for multi-shot person re-identification

```
Input  : Feature vectors $\mathbf{p}_{ij} \in \mathbb{R}^d$, $j = 1, 2, \ldots, n$, of the person $P_i$ in the probe
         view, person-specific gallery dictionaries $\mathbf{G}_i$, $i = 1, 2, \ldots, K$
Output: Class $c$ of person $P_i$
$\mathbf{R} = \mathbf{0}$;
for $j \leftarrow 1, 2, \ldots, n$ do
    Solve Equation 15 for $\mathbf{p}_{ij}$ and obtain $\mathbf{x}^s$ and $\mathbf{e}^s$;
    $r_l^j = \|\mathbf{p}_{lj} - \mathbf{G}_l \mathbf{x}_l^s - \mathbf{e}^s\|$, $l = 1, 2, \ldots, K$;
    $\mathbf{R} = \mathbf{R} + \mathbf{r}^j$;
end
$c =$ index of the least element in $\mathbf{R}$;
```

to 192 frames, with an average of 73 frames. The images in each sequence across both views suffer from extreme lighting and viewpoint variations, occlusions and cluttered background.

4.1.2. PRID 2011

The PRID 2011 dataset was created from person images obtained from two cameras with non-overlapping adjacent fields of view. The cameras were located in an outdoor environment. For the first camera view, image sequences of variable length for 385 distinct individuals are available, whereas for the second camera view, image sequences of variable length for 749 distinct individuals are available. The images in each sequence across both views involve viewpoint, illumination, and background variations. However, in this dataset, occlusion is less severe than in the iLIDS-VID dataset.

In our experiments, to ensure consistency with the evaluation protocol presented in [5], we only consider image sequences corresponding to the 178 distinct individuals that appear in both the camera views and that have more than 21 frames in each sequence. The average number of frames available in image sequences for these individuals is 100.

4.1.3. SAIVT-SoftBio

The SAIVT-SoftBio dataset was created from a multi-camera surveillance network installed in an indoor environment. It consists of image sequences of variable length for 150 distinct individuals passing through the fields of view of the eight cameras in the surveillance network. However, since the dataset was created in an uncontrolled setting, not all of these people appear in each of the 8 camera views.

To ensure consistency with the evaluation protocol presented in [7], we only consider two camera pairs: cameras 3 and 8 (hereby referred to as SAIVT-38), and cameras 5 and 8 (hereby referred to as SAIVT-58). SAIVT-38 consists of image sequences corresponding to 99 distinct individuals, whereas SAIVT-58 consists of image sequences corresponding to 103 distinct individuals. The images in each sequence across both pairs of views suffer from illumination and background variations. Furthermore, the images in SAIVT-58 suffer from extreme viewpoint variations.

4.2. Evaluation protocol and implementation details

For the iLIDS-VID and PRID 2011 datasets, we randomly split the available image sequences into equal-sized training and testing sets. For the SAIVT-38 dataset, we consider image sequences corresponding to 31 people for training and 68 people for testing. For the SAIVT-58 dataset, we consider image sequences corresponding to 33 people for training and 70 people for testing. We generate 10 such train-test splits[1] and report the overall average performance across all 10 splits.

For each split, we cluster the available training images for each person and learn the feature space projection matrix using Fisher discriminant analysis, as described in Section 3.1.

4.2.1. Parameters

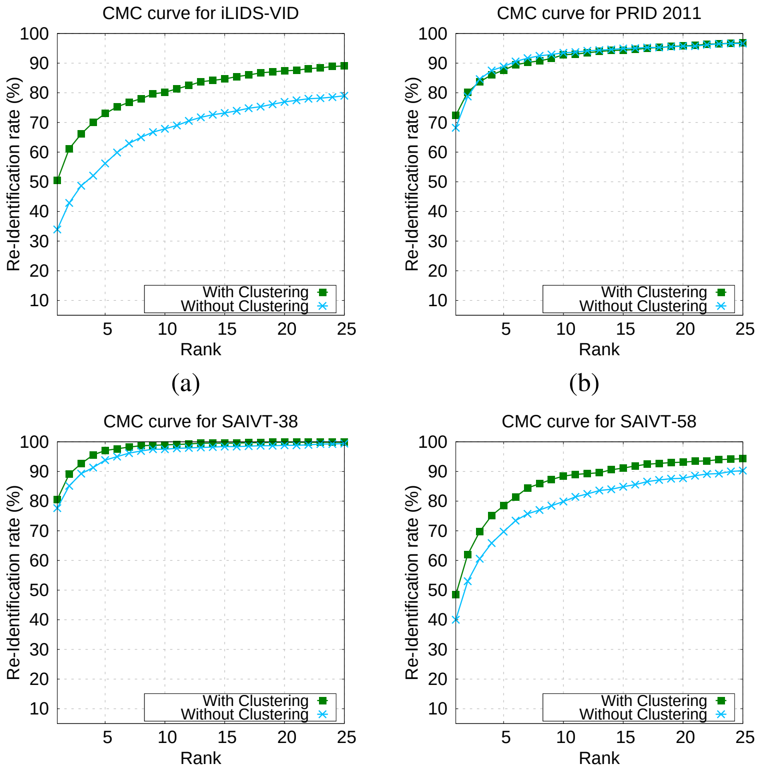
We set the number of clusters n to 20. The value of the feature space transformation parameter d was set to the same dimensions as the original feature space, 3264. Sections 4.3 and 4.5 present experiments to justify these parameter choices. The update parameters η_1 and η_2 in Equation 18 were both set to 0.1. The number of iterations in Algorithm 1 was set to 5.

4.3. Evaluating image clustering

We begin the evaluation of the proposed algorithm by studying the impact of image clustering. In this experiment, we perform re-id tests on each of the four datasets twice — first with all available images in each sequence for each person in the gallery set, and second using the clustered feature space for each person in the gallery set. In both these tests, we learn the feature space projection matrix $\mathbf{T}$ and project the available feature vectors prior to constructing the gallery dictionary. We repeat this experimental procedure for each of the 10 train-test splits and report the average performance in the cumulative match characteristic (CMC) curves shown in Figure 5. As can be seen from these plots, we have not lost any performance accuracy by clustering the available images prior to re-id. In fact, counter-intuitively, we observe that the use of clustered feature vectors results in a rank-1 performance improvement of 16.5%, 4.2%, 2.9%, and 8.4% on the iLIDS-VID, PRID 2011, SAIVT-38 and SAIVT-58 datasets respectively.

The impact of image clustering is further evident when we consider the average time needed to recover the identity of a particular test person. In this experiment, we compute the time taken by our algorithm to return the identity of a test person, given the corresponding n feature vectors. As before, we perform this twice, first using all the available images for each person, and second in the clustered feature space. We compute the average time over all the 10 train-test splits and report the results in

[1] Note that in the case of the iLIDS-VID and PRID 2011 datasets, we use the splits available at `http://www.eecs.qmul.ac.uk/~xz303/project_video_ranking/index.html`. Code, data splits, and features to reproduce our results will be made available online.



(a) (b) (c) (d)

Figure 5: Evaluating the impact of image clustering. Working in the clustered feature space results in improved re-identification performance when compared to using all the available feature vectors.

Table 1. Specifically, in both clustered and non-clustered feature spaces, we determine the time required to recover the block sparse coefficient vector for each of the available feature vectors for each probe person and compute the average of all these values. All the running times are obtained in MATLAB on an Alienware PC running on an Intel Core i7 CPU with 16.0 GB of RAM. Clearly, we see a significant reduction in the average probe recovery time. The average size (over all the 10 train-test splits) of the gallery dictionary when no clustering is employed is as follows for each of the datasets: $d \times 9957$ for iLIDS-VID, $d \times 10426$ for PRID 2011, $d \times 2804$ for SAIVT-38 and $d \times 2906$ for SAIVT-58. Here, d is the dimension of the projected feature space. The corresponding dictionary sizes after clustering are $d \times 2777$ for iLIDS-VID, $d \times 1667$ for PRID 2011, $d \times 1083$ for SAIVT-38 and $d \times 1114$ for SAIVT-58. Clearly, the numbers of feature vectors after clustering are significantly lower than those when all the available images are used. Therefore, we should expect a significant run-time gain in the clustered feature space, as reflected in the numbers in Table 1.

Table 1: Average probe recovery time (in seconds) for the PRID 2011, iLIDS-VID, SAIVT-38 and SAIVT-58 datasets. The clustered feature space offers substantial computational advantages over the non-clustered feature space.

Dataset	PRID 2011	iLIDS-VID	SAIVT-38	SAIVT-58
Without clustering	24.57	19.96	0.89	0.79
With clustering	0.1	0.2	0.06	0.04

4.3.1. Discussion

We conclude this section with additional empirical observations.

Number of clusters: We repeated experiments in the clustered feature space using four different values for the number of clusters parameter: 5, 10, 15, and 20. The CMC curves of the performance obtained in each case is shown in Figure 6. We do not observe any consistent trend in the performance as we vary the number of clusters. While the performance on SAIVT-38 and SAIVT-58 is essentially the same as we vary the number of clusters, we notice a non-negligible difference on the PRID-2011 dataset, where 5 clusters seem to be giving the best performance. In the case of iLIDS-VID, while there seems to be a non-negligible rank-1 performance difference between 5 and 20 clusters, the performance at later ranks is similar. Intuitively, we should expect to use a higher value for the number of clusters parameter in the case of datasets with a high degree of variability in the available images for each person. However, as can be seen from the results, this does not seem to be the case, with no consistent trends emerging. In such cases, we can resort to cross-validation techniques to pick the best parameter value.

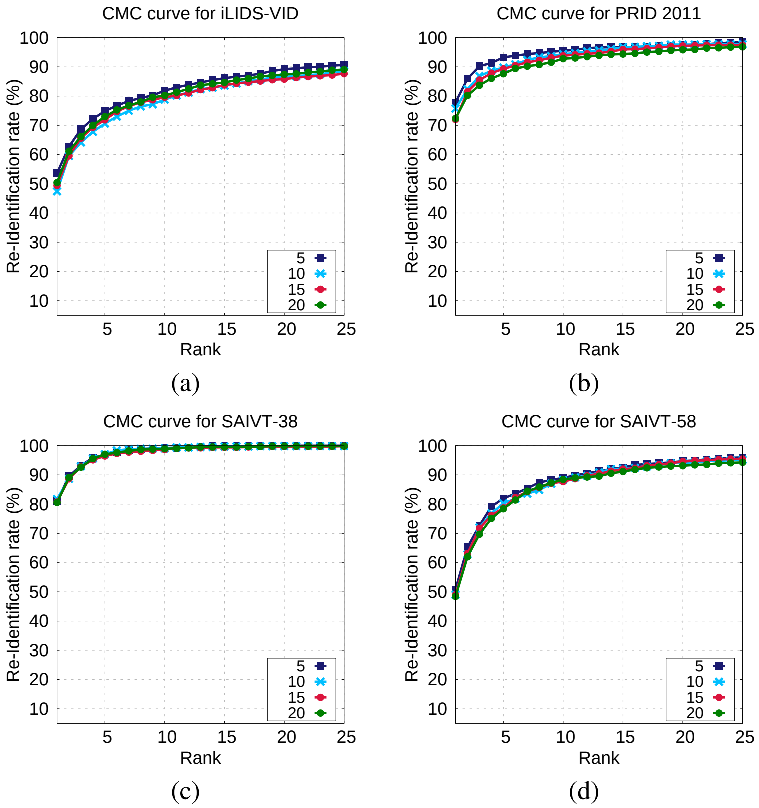
4.4. Evaluating block sparsity

We next study the impact of formulating re-id as a block sparse recovery problem as opposed to a more traditional sparse recovery problem. Let us first revisit the problem of Equation 15. Here, our hypothesis is that the coefficient vector $\mathbf{x}$ is block sparse. In this section, we validate this hypothesis by means of empirical experimental comparison with the hypothesis that $\mathbf{x}$ is sparse. To this end, we conduct experiments twice. First, we consider the problem formulation of Equation 15. Next, we change this formulation as follows:

$$\begin{aligned} \min_{\mathbf{x},\mathbf{e}} \; & \|\mathbf{x}\|_1 + \|\mathbf{e}\|_1 \\ \text{s.t.} \; & \mathbf{p}_{ij} = \mathbf{D}\mathbf{x} + \mathbf{e} \end{aligned} \tag{25}$$

i.e., we now solve a traditional sparse recovery problem. To ensure consistency with the way we solved the block sparse recovery problem in Section 3.2.4, we reformulate the above problem in the Lagrangian framework as follows:

$$\min_{\mathbf{x},\mathbf{e}} \|\mathbf{x}\|_1 + \|\mathbf{e}\|_1 - \mathbf{m}^\top(\mathbf{D}\mathbf{x} + \mathbf{e} - \mathbf{p}_{ij}) + \frac{\eta}{2}\|\mathbf{D}\mathbf{x} + \mathbf{e} - \mathbf{p}_{ij}\|^2 \tag{26}$$



(a)

(b)

(c)

(d)

Figure 6: Impact of the number of clusters parameter on the average performance.

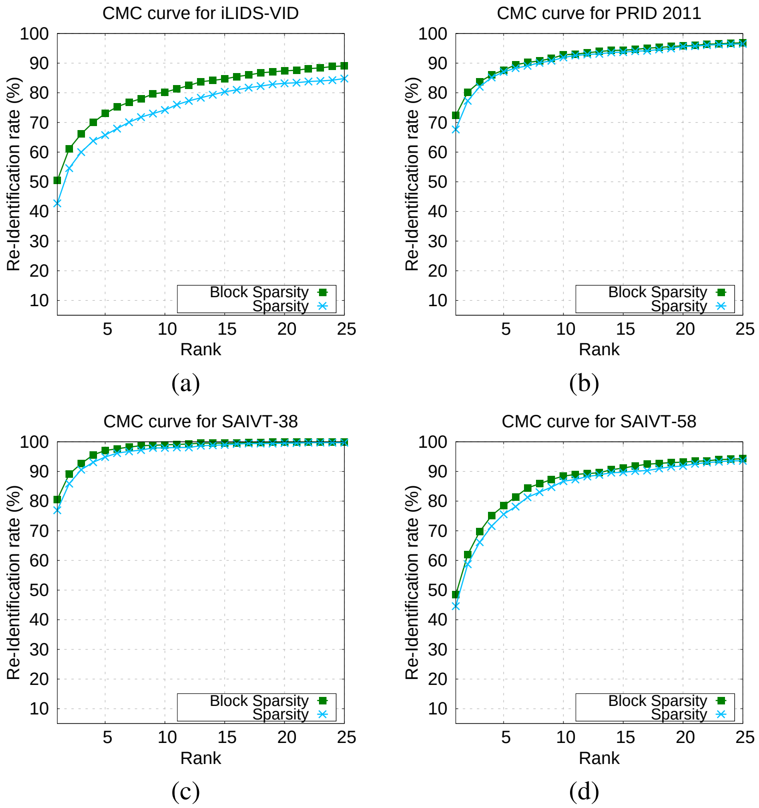
We minimize the above objective function using the primal augmented Lagrangian algorithm [2]. The re-identification protocol for a test person with n available feature vectors in this case will be similar to that described in Algorithm 2. The only difference is that now we solve for a sparse $\mathbf{x}^s$ instead of a block sparse $\mathbf{x}^s$. As before, we repeat this experiment for all 10 train-test splits and report the overall average performance in the CMC curves shown in Figure 7.

From these results, it is evident that the block sparse formulation of Equation 15 gives significantly better results than the sparse recovery formulation of Equation 25. In particular, we note that block sparse recovery results in a rank-1 performance improvement of 7.7%, 4.7%, 3.7%, and 3.9% on iLIDS-VID, PRID 2011, SAIVT-38 and SAIVT-58 datasets respectively when compared with the sparse recovery formulation. These results validate our hypothesis of formulating the person re-id problem as a block sparse recovery problem instead of a sparse recovery problem.

4.4.1. Discussion

We conclude this section with additional empirical observations.

Error term: Let us revisit the problem formulation of Equation 15. While it is



(a) (b) (c) (d)

Figure 7: Evaluating the impact of block sparsity. The block sparse problem formulation offers improved performance when compared to the sparse recovery formulation.

quite intuitive to see why we might need the error term $\mathbf{e}$, the goal of this section is to empirically verify if it is indeed necessary to achieve good results. To this end, we conducted experiments with both these problem formulations twice - first with the error term, and second without the error term. A bar plot of the average rank-1 performance in both the cases for each of the four datasets is shown in Figure 8. As can be seen from the Figure, the error term does seem to improve, albeit marginally, the rank-1 performance of the block sparse recovery formulations. Specifically, the improvement is about 1.1%, 1.4%, and 0.4% on iLIDS-VID, PRID 2011, and SAIVT-58 datasets respectively. On the SAIVT-38 dataset, we notice a rank-1 performance drop of about 0.2%.

4.5. Evaluating feature space projection

We next study the impact of the choice of the feature space on the performance of our block sparsity formulation. To this end, we perform experiments twice: first in the original feature space, and next in the feature space learned using FDA. Note that in each case, we work with the cluster centers as before. The average re-id performance

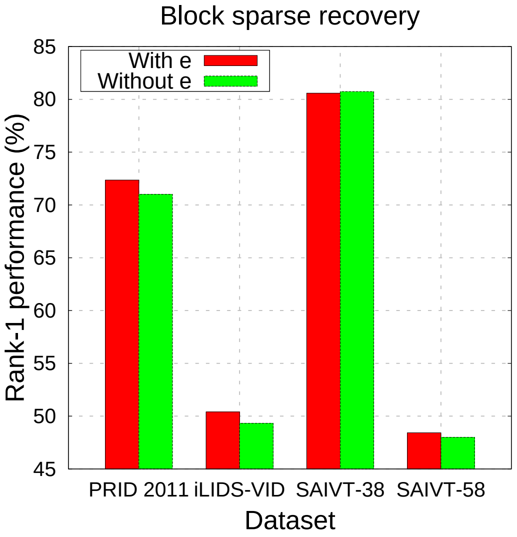


Figure 8: Impact of the error term $\mathbf{e}$ on the average rank-1 performance of the block sparse recovery formulation of Equation 15.

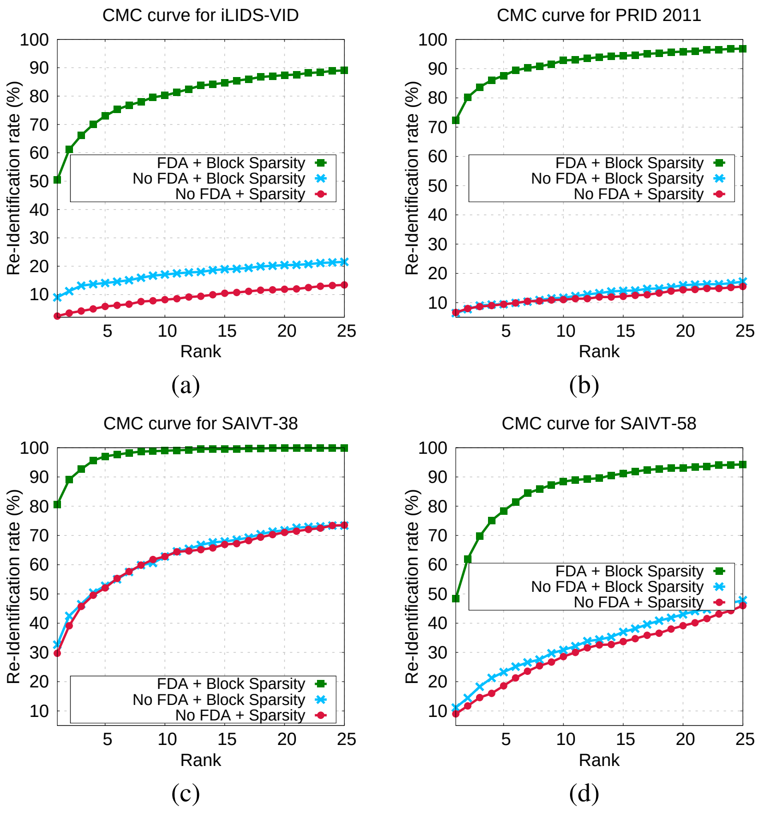
obtained in each case is shown in the CMC curves in Figure 9. It is evident from these figures that formulating the problem in the feature space learned using FDA offers substantially improved performance when compared to the original feature space. Specifically, we note that the rank-1 performance improvement is about 27.5%, 66%, 47.9%, and 37.3% for the iLIDS-VID, PRID 2011, SAIVT-38 and SAIVT-58 datasets respectively.

4.5.1. Discussion

We conclude this section with additional empirical observations.

Block sparsity vs. sparsity in the original feature space: To further evaluate the impact of block sparsity, we compared its performance with that of sparse recovery in the original feature space. Specifically, we performed experiments using the clustered feature vectors in the original feature space and the results obtained are shown in Figure 9. We see that block sparsity offers improved performance even in the original feature space. Specifically, the rank-1 performance improvement is about 4%, 2.9%, and 2.1% on iLIDS-VID, SAIVT-38 and SAIVT-58 datasets respectively. On PRID-2011, however, we notice a rank-1 performance drop of about 0.23%.

Dimension of the projected feature space: In this experiment, we study the impact of the feature space transformation parameter. To this end, we uniformly sampled 3 values for d, in addition to $d = 3264$. Specifically, we projected the feature space to $d/20$, $d/10$, and $d/5$ dimensions and performed experiments using the clustered feature vectors. The results obtained are shown in Figure 10, which plots the average rank-1 performance versus d for all the four test datasets. We observe that the rank-1 performance improves as the number of dimensions increases, and this supports our general



(a) (b)

(c) (d)

Figure 9: Evaluating the impact of feature space projection. Formulating our approach in a feature space learned using Fisher discriminant analysis results in substantial performance improvement over the original feature space.

intuition that we retain more information in a higher-dimensional feature space.

4.6. Comparison with the state of the art

In this section, we compare the results of our approach with several recently proposed approaches that report state-of-the-art re-id performance. Specifically, we consider the following algorithms: SDALF [25], Fusion Model [7], Salience [10], DVR [5], DVDL [62], ISR [51], AFDA [36], and STFV3D [63]. Furthermore, as evaluated in [5], we also consider a combination of color histograms and local binary patterns (LBP) [64] in conjunction with both RankSVM and DVR as the metric. We abbreviate our algorithm as **SRID**. In addition to these methods, we also implemented two baseline methods that serve as useful reference points with which to compare our results. These are described next.

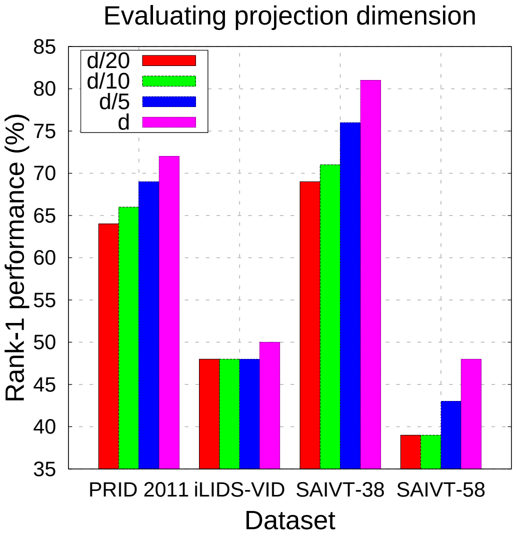


Figure 10: Evaluating the impact of the feature space projection dimension on the average rank-1 performance of the block sparse recovery formulation of Equation 15. $d = 3264$ in the legend.

4.6.1. Clustering+FDA+L_2

In this approach, we use the same clustered and projected feature space as in SRID, but instead of using block sparse recovery, we use the Euclidean, or L_2 distance to rank candidates. Prior to ranking using the L_2 distance, we average all the feature vectors available for each person in both the gallery and probe sets.

4.6.2. Clustering+FDA+RankSVM

In this approach, we use the same clustered and projected feature space as in SRID, but instead of using block sparse recovery, we use the RankSVM [9] formulation to rank candidates. If $\mathbf{f}^g$ and $\mathbf{f}^p$ correspond to the average feature vector of a gallery candidate g and a probe candidate p, the basic idea of the RankSVM formulation is to learn a weight vector $\mathbf{w}$ using which a similarity score

$$s_p = \mathbf{w}^\top |\mathbf{f}^g - \mathbf{f}^p| \tag{27}$$

can be computed. The gallery candidates can then be ranked according to the similarity scores. To learn the weight vector $\mathbf{w}$, we minimize its norm subject to the following ranking relationship:

$$\begin{gathered} \mathbf{w}^\top (|\mathbf{f}_i^g - \mathbf{f}_i^p| - |\mathbf{f}_i^g - \mathbf{f}_j^p|) > 0 \\ i, j = 1, 2, \ldots, K, i \neq j \end{gathered} \tag{28}$$

where K is the number of people in the training set. The RankSVM method learns $\mathbf{w}$ by solving the following minimization problem:

$$
\begin{aligned}
\min_{\mathbf{x},\xi} \quad & \left(\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{K}\xi_i\right) \\
\text{s.t.} \quad & \mathbf{w}^\top(|\mathbf{f}_i^g - \mathbf{f}_i^p| - |\mathbf{f}_i^g - \mathbf{f}_j^p|) \geq 1 - \xi_i \qquad (29) \\
& \xi_i \geq 0
\end{aligned}
$$

where C is a margin trade-off parameter and ξ_i is a slack variable.

4.6.3. Results

The average performance across all the train-test splits is plotted in the CMC curves in Figure 11 and summarized in Tables 2 and 3. A point to note from these results is that the baseline methods we implemented are already very strong, offering superior performance when compared to several state-of-the-art techniques. Furthermore, we note that our algorithm results in state-of-the-art performance on all four datasets. Specifically, the rank-1 performance improvement over the best performing approach is about 6.1%, 6.2%, 12.2%, and 12.1% for the iLIDS-VID, PRID 2011, SAIVT-38 and SAIVT-58 datasets respectively.

Table 2: Comparison with the state of the art: Results on the PRID 2011 and iLIDS-VID datasets.

Dataset	PRID 2011			iLIDS-VID		
Rank	1	5	10	1	5	10
SDALF [25]	5.2	20.7	32	6.3	18.8	27.1
Salience [10]	25.8	43.6	52.6	10.2	24.8	35.5
DVR [5]	28.9	55.3	65.5	23.3	42.4	55.3
Color & LBP [64] + RankSVM [9]	34.3	56	65.5	23.2	44.2	54.1
Color & LBP [64] + DVR [5]	37.6	63.9	75.3	34.5	56.7	67.5
Color + DVR [5]	41.8	63.8	76.7	32.7	56.5	67.0
AFDA [36]	51.8	79.7	89.2	28.3	53.1	66.5
Clustering+FDA+RankSVM [9]	56.0	82.0	89.3	37.7	62.4	72.5
ISR [51]	59.3	72.8	76.7	14.1	22.3	28.7
STFV3D+KISSME [63]	64.1	87.3	89.9	44.3	71.7	**83.7**
Clustering+FDA+L_2	65.6	**89.2**	94.7	16.1	28.7	36.1
DVDL [62]	66.2	88	**95.4**	42.4	66.8	76.9
SRID	**72.4**	87.6	92.8	**50.4**	**73.0**	80.2

4.7. Improving metric learning methods

In our work, we used Fisher discriminant analysis to perform feature space projection. We could, in principle, use any metric learning algorithm prior to ranking gallery candidates. Typically, a metric learning algorithm ranks candidates by computing the distance $D(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^\top \mathbf{T}^\top \mathbf{T}(\mathbf{x} - \mathbf{y})$, where $\mathbf{T}$ is the learned projection matrix.

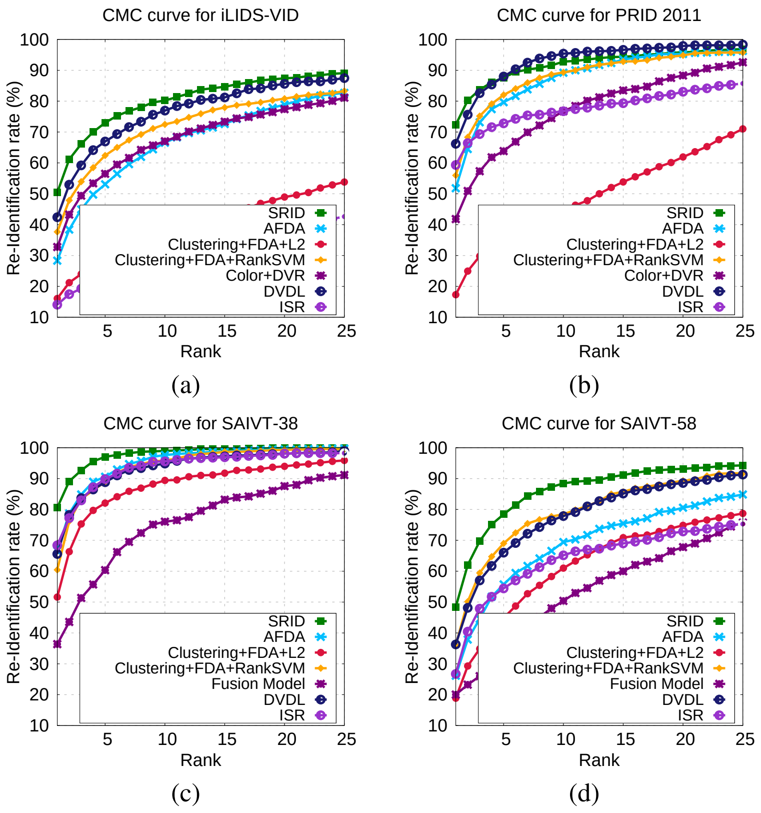


Figure 11: A comparison of the performance of the proposed algorithm with the current state of the art. We improve the rank-1 state of the art by about 6.1%, 6.2%, 12.2%, and 12.1% for the iLIDS-VID, PRID 2011, SAIVT-38 and SAIVT-58 datasets respectively.

Table 3: Comparison with the state of the art: Results on the SAIVT-38 and SAIVT-58 datasets.

Dataset	SAIVT-38			SAIVT-58		
Rank	1	5	10	1	5	10
Fusion Model [7]	36.37	60.35	76.04	20.04	33.02	50.39
DVDL [62]	65.6	88.8	94.9	36.3	66	77.9
Clustering+FDA+L_2	51.6	82.1	89.4	18.9	44.9	61.0
AFDA [36]	65.6	88.8	94.9	36.3	66	77.9
Clustering+FDA+RankSVM [9]	60.4	89.0	95.9	36.0	69.0	78.4
ISR [51]	68.4	89.9	95.6	26.7	54.4	65.1
SRID	**80.6**	**97.1**	**99.0**	**48.4**	**78.4**	**88.4**

This is equivalent to computing the L_2 distance between the feature vectors $\mathbf{x}$ and $\mathbf{y}$ in the projected feature space. In addition to the Euclidean distance, we can also use the RankSVM formulation, described in Section 4.6.2 to rank gallery candidates. In this section, we show that we can improve upon the results obtained using these methods with our block sparse recovery formulation. To this end, we consider several popularly used metric learning techniques from the re-id literature: marginal Fisher analysis (MFA) [65], local Fisher discriminant analysis (LFDA) [22], and keep-it-simple-and-straightforward (KISSME) [66].

The experimental setup is as follows: for each metric learning method, we learn the feature space projection matrix and rank candidates in the projected feature space using the L_2 distance and the RankSVM formulation. We then rank the candidates in the projected feature space using our block sparse recovery formulation of Equation 15. We repeat this for all 10 train-test splits and report the average performance in the CMC curves in Figure 12. In particular, the results for MFA, LFDA and KISSME are shown in Figures 12(a)-(d), 12(e)-(f), and 12(i)-(l) respectively. Clearly, we can see that the block sparse recovery formulation gives consistently better results when compared to both the L2 distance and RankSVM formulations. It is worth noting that, as seen from the results in Section 4.6, these baselines already give strong performance. Therefore, the performance improvement achieved by our proposed formulation in significant. Specifically, our approach improves the rank-1 performance of MFA by about 4.9%, 11.5%, 12.8%, and 5.4% on iLIDS-VID, PRID 2011, SAIVT-38 and SAIVT-58 datasets respectively. The corresponding improvements in the case of KISSME are 1.7%, 39.32%, 7.35%, and 6.7%. In the case of LFDA, the corresponding improvements are 7.4%, 12.1%, 19.4%, and 10.6% on the iLIDS-VID, PRID 2011, SAIVT-38 and SAIVT-58 datasets respectively.

5. Conclusions and future work

We presented an algorithm based on block sparse recovery for the multi-shot person re-identification problem. The formulation was motivated by the observation that the available feature vectors for each person form disparate feature sets, and that a dictionary constructed using these features will exhibit a block structure. Consequently, a test feature vector is expressed as a linear combination of the columns of this dictionary, with the resulting coefficient vector characterized as being block sparse. This insight enabled us to develop a principled approach to exploit the availability of multi-shot data, and subsequently, the identity of a test feature vector was determined using an optimization approach in the alternating directions framework. We extensively evaluated the proposed algorithm on three publicly available multi-shot re-identification datasets, and demonstrated new state-of-the-art results.

We conclude with a brief discussion on factors that are crucial to achieve good performance with our approach while working in the general framework of re-id based on image sequences. Here, we also provide avenues for promising directions for future research.

5.1. Features and discriminability

As our results in Section 4.5 demonstrated, formulating the block sparse recovery problem in a feature space learned using FDA offered substantial performance improvements over the original feature space. While this suggests that feature space

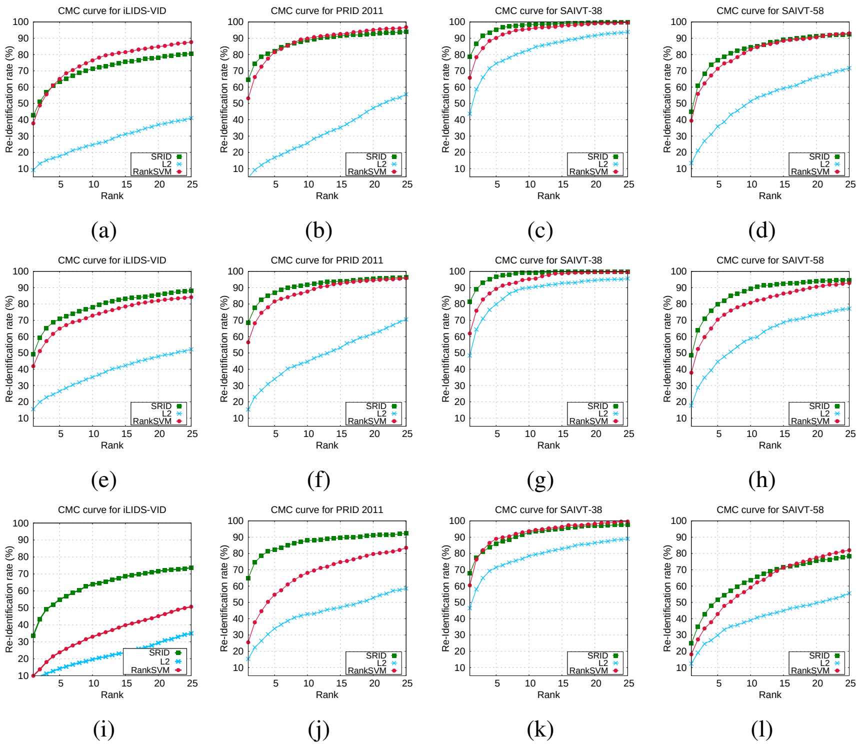


Figure 12: Improving the ranking performance of metric learning methods using SRID. (a)-(d): MFA. (e)-(f): LFDA. (i)-(l): KISSME. In each case, ranking candidates using the proposed algorithm after learning the metric offers improved re-id performance.

discriminability is key to achieving good performance with our approach, it also indicates that using learned features rather than hand-crafted features offers superior performance. A natural extension in this direction, in line with recent advances, is to employ data-hungry feature learning frameworks such as convolutional neural networks [67] prior to using our approach.

5.2. Ranking methodology

As our experiments in Section 4.7 demonstrated, ranking candidates using our block sparse recovery formulation offers superior performance when compared to the traditionally used Euclidean distance and RankSVM based approaches. These results are also intuitively satisfying since the block sparse recovery approach is a principled methodology to exploit the availability of multiple images per person.

In this work, we constructed the dictionary $\mathbf{D}$ manually using the available training data. A natural extension could be to learn the dictionary instead. In our recent work [62], we proposed a technique to learn discriminative dictionaries capable of sparsely encoding feature vectors corresponding to different people. However, this method does not fully exploit the availability of multiple images per person as part of its training and testing procedure. One possible way to use multi-shot data is to model it as an affine hull or convex hull. In fact, there has been increased recent interest in learning distance metrics for set-based recognition [68]. Learning discriminative dictionaries in conjunction with such set-based data modeling schemes could be a worthwhile research direction to pursue.

5.3. Representative data selection

As our experiments in Section 4.3 demonstrated, clustering the available image data prior to learning the feature space projection matrix can offer significant computational advantages in addition to possibly resulting in a more discriminative feature space.

Since the basic idea behind the use of image clustering is to find representative segments from the available track of images for the person of interest, we could use more advanced segment selection schemes. A closely related area is video summarization [69, 70]. Such sample selection schemes can determine the most representative parts in each image sequence, potentially leading to a more discriminative feature space in which to formulate the ranking methods discussed above.

5.4. Dealing with similar appearances

We conclude with a discussion on a specific scenario where our problem formulation of Equation 15 might fail. Consider the following case: all the observed people in the gallery camera wear similar clothes, for instance, blue jeans and a black shirt. In such a scenario, the computed appearance features will all look very similar in the feature space, giving feature dictionaries $\mathbf{G}_i$ that have similar entries for all the people. Solving the linear inverse problem of Equation 15 in this case would lead to a coefficient vector that will have similar, if not the same, entries in each block, thereby leading to an ambiguity in retrieving the identity of the probe feature vector. While such a scenario is unlikely in real-world re-id problems such as an airport [15], it is possible in environments that have a certain dress code. In such cases, clearly, we cannot solely rely on appearance features to retrieve the identity of the person of interest. A possible solution would be to integrate additional features into the problem formulation. For instance, we can use a face detector to detect person faces and compute face-specific features. We could also use person-specific gait information to construct dynamics-based features [71]. Such additional information can be integrated into our feature dictionary construction mechanism to deal with scenarios where pure appearance features might fail.

Acknowledgments

This material is based upon work supported by the U.S. Department of Homeland Security, Science and Technology Directorate, Office of University Programs, under Award 2013-ST-061-ED0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

References

[1] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2) (2009) 210–227.

[2] A. Y. Yang, Z. Zhou, A. G. Balasubramanian, S. S. Sastry, Y. Ma, Fast l_1-minimization algorithms for robust face recognition, IEEE Trans. Image Process. 22 (8) (2013) 3234–3246.

[3] X. Mei, H. Ling, Robust visual tracking using l_1 minimization, in: IEEE Int. Conf. Comput. Vision, Kyoto, Japan, 2009, pp. 1436–1443.

[4] S. Karanam, Y. Li, R. J. Radke, Particle dynamics and multi-channel feature dictionaries for robust visual tracking, in: Proc. Brit. Mach. Vision Conf., Swansea, UK, 2015, pp. 183.1–183.12.

[5] T. Wang, S. Gong, X. Zhu, S. Wang, Person re-identification by video ranking, in: Eur. Conf. Comput. Vision, Zurich, Switzerland, 2014, pp. 688–703.

[6] M. Hirzer, C. Beleznai, P. M. Roth, H. Bischof, Person re-identification by descriptive and discriminative classification, in: Image Analysis, 2011, pp. 91–102.

[7] A. Bialkowski, S. Denman, S. Sridharan, C. Fookes, P. Lucey, A database for person re-identification in multi-camera surveillance networks, in: Int. Conf. Digital Image Computing Techniques and Applicat., Fremantle, Australia, 2012, pp. 1–8.

[8] S. Karanam, Y. Li, R. J. Radke, Sparse Re-Id: Block sparsity for person re-identification, in: IEEE/ISPRS 2nd Joint Workshop on Multi-Sensor Fusion for Dynamic Scene Understanding, 2015.

[9] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, Person re-identification by support vector ranking., in: Proc. Brit. Mach. Vision Conf., Vol. 2, Aberystwyth, UK, 2010, pp. 21.1–21.11.

[10] R. Zhao, W. Ouyang, X. Wang, Unsupervised salience learning for person re-identification, in: IEEE Conf. Comput. Vision and Pattern Recognition, Portland, OR, 2013, pp. 3586–3593.

[11] E. Ahmed, M. Jones, T. K. Marks, An improved deep learning architecture for person re-identification, in: IEEE Conf. Comput. Vision and Pattern Recognition, Boston, MA, 2015, pp. 3908–3916.

[12] F. Xiong, M. Gou, O. Camps, M. Sznaier, Person re-identification using kernel-based metric learning methods, in: Eur. Conf. Comput. Vision, Zurich, Switzerland, 2014, pp. 1–16.

[13] S. Paisitkriangkrai, C. Shen, A. van den Hengel, Learning to rank in person re-identification with metric ensembles, in: IEEE Conf. Comput. Vision and Pattern Recognition, Boston, MA, 2015, pp. 1846–1855.

[14] Y. Li, Z. Wu, S. Karanam, R. J. Radke, Real-world re-identification in an airport camera network, in: Proc. Int. Conf. Distributed Smart Cameras, Venice, Italy, 2014, pp. 35:1–35:6.

[15] O. Camps, M. Gou, T. Hebble, S. Karanam, O. Lehmann, Y. Li, R. Radke, Z. Wu, F. Xiong, From the lab to the real world: Re-identification in an airport camera network, IEEE Trans. Circuits Syst. Video Technology PP (99).

[16] D. Gray, H. Tao, Viewpoint invariant pedestrian recognition with an ensemble of localized features, in: Eur. Conf. Comput. Vision, Marseille, France, 2008, pp. 262–275.

[17] C. Schmid, Constructing models for content-based image retrieval, in: IEEE Conf. Comput. Vision and Pattern Recognition, Kauai, HI, 2001, pp. II–39 – II–45.

[18] I. Fogel, D. Sagi, Gabor filters as texture discriminator, Biological cybernetics 61 (2) (1989) 103–113.

[19] O. Chapelle, S. S. Keerthi, Efficient algorithms for ranking with SVMs, Information Retrieval 13 (3) (2010) 201–215.

[20] A. Mignon, F. Jurie, PCCA: A new approach for distance learning from sparse pairwise constraints, in: IEEE Conf. Comput. Vision and Pattern Recognition, Providence, RI, 2012, pp. 2666–2672.

[21] W.-S. Zheng, S. Gong, T. Xiang, Reidentification by relative distance comparison, IEEE Trans. Pattern Anal. Mach. Intell. 35 (3) (2013) 653–668.

[22] S. Pedagadi, J. Orwell, S. Velastin, B. Boghossian, Local Fisher discriminant analysis for pedestrian re-identification, in: IEEE Conf. Comput. Vision and Pattern Recognition, Portland, OR, 2013, pp. 3318–3325.

[23] M. Sugiyama, Local Fisher discriminant analysis for supervised dimensionality reduction, in: Int. Conf. Mach. Learning, Pittsburgh, PA, 2006, pp. 905–912.

[24] X. He, P. Niyogi, Locality preserving projections, in: Annu. Conf. Neural Inform. Process. Syst., Vol. 16, Vancouver, Canada, 2004, pp. 153–160.

[25] M. Farenzena, L. Bazzani, A. Perina, V. Murino, M. Cristani, Person re-identification by symmetry-driven accumulation of local features, in: IEEE Conf. Comput. Vision and Pattern Recognition, San Francisco, CA, 2010, pp. 2360–2367.

[26] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, V. Murino, Custom pictorial structures for re-identification, in: Proc. Brit. Mach. Vision Conf., Dundee, UK, 2011, pp. 68.1–68.11.

[27] S. Bak, E. Corvee, F. Brémond, M. Thonnat, Person re-identification using spatial covariance regions of human body parts, in: IEEE Int. Conf. Advanced Video and Signal based Surveillance, Boston, MA, 2010, pp. 435–440.

[28] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: IEEE Conf. Comput. Vision and Pattern Recognition, San Diego, CA, 2005, pp. 886–893.

[29] K. Grauman, T. Darrell, The pyramid match kernel: Discriminative classification with sets of image features, in: IEEE Int. Conf. Comput. Vision, Beijing, China, 2005, pp. 1458–1465.

[30] N. Martinel, C. Micheloni, G. L. Foresti, Kernelized saliency-based person re-identification through multiple metric learning, IEEE Trans. Image Process. 24 (12) (2015) 5645–5658.

[31] W.-S. Zheng, S. Gong, T. Xiang, Towards open-world person re-identification by one-shot group-based verification, IEEE Trans. Pattern Anal. Mach. Intell. 38 (3) (2016) 591–606.

[32] M. S. Nixon, T. Tan, R. Chellappa, Human identification based on gait, Vol. 4, Springer Science & Business Media, 2010.

[33] D. N. T. Cong, C. Achard, L. Khoudour, L. Douadi, Video sequences association for people re-identification across multiple non-overlapping cameras, in: Int. Conf. Image Anal. and Process., Vietri sul Mare, Italy, 2009, pp. 179–189.

[34] L. Bazzani, M. Cristani, A. Perina, V. Murino, Multiple-shot person re-identification by chromatic and epitomic analyses, Pattern Recognition Lett. 33 (7) (2012) 898–903.

[35] Y. Li, Z. Wu, R. J. Radke, Multi-shot re-identification with random-projection-based random forests, in: IEEE Winter Conf. Applicat. Comput. Vision, Waikoloa Beach, HI, 2015, pp. 373–380.

[36] Y. Li, Z. Wu, S. Karanam, R. J. Radke, Multi-shot human re-identification using adaptive Fisher discriminant analysis, in: Proc. Brit. Mach. Vision Conf., Swansea, UK, 2015, pp. 73.1–73.12.

[37] D. Simonnet, M. Lewandowski, S. A. Velastin, J. Orwell, E. Turkbeyler, Re-identification of pedestrians in crowds using dynamic time warping, in: *1st* Int. Workshop Re-Identification, 2012, pp. 423–432.

[38] N. Martinel, A. Das, C. Micheloni, A. K. Roy-Chowdhury, Re-identification in the function space of feature warps, IEEE Trans. Pattern Anal. Mach. Intell. 37 (8) (2015) 1656–1669.

[39] A. Chakraborty, A. Das, A. K. Roy-Chowdhury, Network consistent data association, IEEE Trans. Pattern Anal. Mach. Intell. 6 (1) (2015) 591–606.

[40] E. J. Candes, J. K. Romberg, T. Tao, Stable signal recovery from incomplete and inaccurate measurements, Communications on pure and applied mathematics 59 (8) (2006) 1207–1223.

[41] E. J. Candes, T. Tao, Near-optimal signal recovery from random projections: Universal encoding strategies?, Information Theory, IEEE Transactions on 52 (12) (2006) 5406–5425.

[42] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, Y. Ma, Toward a practical face recognition system: Robust alignment and illumination by sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 34 (2) (2012) 372–386.

[43] T. Zhang, B. Ghanem, S. Liu, N. Ahuja, Robust visual tracking via structured multi-task sparse learning, Int. J. Comput. Vision 101 (2) (2013) 367–383.

[44] J. Mairal, G. Sapiro, M. Elad, Learning multiscale sparse representations for image and video restoration, SIAM J. Multiscale Modeling and Simulation 7 (1) (2008) 214–241.

[45] Y. Hu, A. S. Mian, R. Owens, Sparse approximated nearest points for image set classification, in: IEEE Conf. Comput. Vision and Pattern Recognition, 2011, pp. 121–128.

[46] Y. Wu, J. Lim, M.-H. Yang, Object tracking benchmark, IEEE Trans. Pattern Anal. Mach. Intell. 37 (9) (2015) 1834–1848.

[47] F. Bach, R. Jenatton, J. Mairal, G. Obozinski, Optimization with sparsity-inducing penalties, Foundations and Trends® in Machine Learning 4 (1) (2012) 1–106.

[48] M. T. Harandi, C. Sanderson, R. Hartley, B. C. Lovell, Sparse coding and dictionary learning for symmetric positive definite matrices: A kernel approach, in: Eur. Conf. Comput. Vision, Florence, Italy, 2012, pp. 216–229.

[49] M. I. Khedher, M. A. El Yacoubi, B. Dorizzi, Multi-shot SURF-based person re-identification via sparse representation, in: IEEE Int. Conf. Advanced Video and Signal based Surveillance, Krakow, Poland, 2013, pp. 159–164.

[50] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-up robust features (SURF), Comput. Vision and Image Understanding 110 (3) (2008) 346–359.

[51] G. Lisanti, I. Masi, A. Bagdanov, A. Del Bimbo, Person re-identification by iterative re-weighted sparse ranking, IEEE Trans. Pattern Anal. Mach. Intell. 37 (8) (2015) 1629–1642.

[52] N. Martinel, C. Micheloni, Sparse matching of random patches for person re-identification, in: Proc. Int. Conf. Distributed Smart Cameras, Venice, Italy, 2014.

[53] W.-S. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, S. Gong, Partial person re-identification, in: IEEE Int. Conf. Comput. Vision, Santiago, Chile, 2015, pp. 4678–4686.

[54] B. Ma, Y. Su, F. Jurie, Local descriptors encoded by Fisher vectors for person re-identification, in: ECCV Workshops, 2012.

[55] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society. Series B (methodological) (1977) 1–38.

[56] A. Vedaldi, B. Fulkerson, VLFeat: an open and portable library of computer vision algorithms, in: Int. Conf. Multimedia, 2010.

[57] T. Jaakkola, D. Haussler, et al., Exploiting generative models in discriminative classifiers, Annu. Conf. Neural Inform. Process. Syst.

[58] J. Sánchez, F. Perronnin, T. Mensink, J. Verbeek, Image classification with the Fisher vector: Theory and practice, Int. J. Comput. Vision 105 (3) (2013) 222–245.

[59] Y. C. Eldar, M. Mishali, Robust recovery of signals from a structured union of subspaces, IEEE Trans. Inf. Theory 55 (11) (2009) 5302–5316.

[60] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, Found. and Trends Mach. Learning 3 (1) (2011) 1–122.

[61] W. Deng, W. Yin, Y. Zhang, Group sparse optimization by alternating direction method, in: SPIE Optical Engineering+ Applications, 2013, pp. 88580R–88580R.

[62] S. Karanam, Y. Li, R. J. Radke, Person re-identification with discriminatively trained viewpoint invariant dictionaries, in: IEEE Int. Conf. Comput. Vision, Santiago, Chile, 2015.

[63] K. Liu, B. Ma, W. Zhang, R. Huang, A spatio-temporal appearance representation for video-based pedestrian re-identification, in: IEEE Int. Conf. Comput. Vision, 2015.

[64] M. Hirzer, P. M. Roth, M. Köstinger, H. Bischof, Relaxed pairwise learned metric for person re-identification, in: Eur. Conf. Comput. Vision, Florence, Italy, 2012, pp. 780–793.

[65] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, S. Lin, Graph embedding and extensions: a general framework for dimensionality reduction, IEEE Trans. Pattern Anal. Mach. Intell. 29 (1) (2007) 40–51.

[66] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, H. Bischof, Large scale metric learning from equivalence constraints, in: IEEE Conf. Comput. Vision and Pattern Recognition, Providence, RI, 2012, pp. 2288–2295.

[67] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Annu. Conf. Neural Inform. Process. Syst., Lake Tahoe, NV, 2012, pp. 1097–1105.

[68] P. Zhu, L. Zhang, W. Zuo, D. Zhang, From point to set: Extend the learning of distance metrics, in: IEEE Int. Conf. Comput. Vision, Sydney, Australia, 2013, pp. 2664–2671.

[69] E. Elhamifar, G. Sapiro, R. Vidal, See all by looking at a few: Sparse modeling for finding representative objects, in: IEEE Conf. Comput. Vision and Pattern Recognition, Providence, RI, 2012, pp. 1600–1607.

[70] E. Elhamifar, G. Sapiro, R. Vidal, Finding exemplars from pairwise dissimilarities via simultaneous sparse recovery, in: Annu. Conf. Neural Inform. Process. Syst., Lake Tahoe, NV, 2012, pp. 19–27.

[71] M. Gou, X. Zhang, A. Rates-Borras, S. Asghari-Esfeden, M. Sznaier, O. Camps, Person re-identification in appearance impaired scenarios, in: Proc. Brit. Mach. Vision Conf., York, UK, 2016.