Sparse Re-Id: Block Sparsity for Person Re-Identification

Srikrishna Karanam, Yang Li, Richard J. Radke
Department of Electrical, Computer, and Systems Engineering
Rensselaer Polytechnic Institute
Troy, NY 12180
{karans3,liy21}@rpi.edu, rjradke@ecse.rpi.edu

Abstract

This paper presents a novel approach to solve the problem of person re-identification in non-overlapping camera views. We hypothesize that the feature vector of a probe image approximately lies in the linear span of the corresponding gallery feature vectors in a learned embedding space. We then formulate the re-identification problem as a block sparse recovery problem and solve the associated optimization problem using the alternating directions framework. We evaluate our approach on the publicly available PRID 2011 and iLIDS-VID multi-shot re-identification datasets and demonstrate superior performance in comparison with the current state of the art.

1. Introduction

Automated person re-identification, or re-id, systems play a key role in several security and surveillance applications. Re-identifying the same person across a camera network with non-overlapping views is particularly challenging, since inter-camera illumination and appearance variations are often very pronounced. Many researchers have addressed this problem by matching appearance features in the single-shot setting [29, 18, 11], i.e., assuming only one image per person per camera view exists. However, in many real-world surveillance applications, such as an airport camera network [13], we have a set of images for each person, e.g., acquired while tracking them. In this case, re-id is actually a multi-shot problem. Unfortunately, most current re-id algorithms are designed for the single-shot setting, and are not capable of exploiting the availability of multiple images for each person.

This material is based upon work supported by the U.S. Department of Homeland Security under Award Number 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.

In this paper, we present a principled approach specifically designed to solve the multi-shot re-id problem. Our algorithm design stems from the following intuitions:

- In some learned embedding space, the feature vector of the probe image of a person approximately lies in the linear span of the corresponding images of that person in the gallery.
- If we construct a dictionary $\mathbf{D}$ whose columns are the feature vectors in the embedding space of all the images corresponding to all the persons in the gallery, the feature vector of the probe image can be expressed as a sparse linear combination of the columns of $\mathbf{D}$. Most importantly, the recovered sparse coefficient vector will have a block structure because the dictionary $\mathbf{D}$ has a block structure. This is true because we have several images for each person in the gallery, and the sets of images for each person naturally form disparate blocks.

Building upon these ideas, we present a novel formulation of the multi-shot re-id problem. We pose the problem of determining the class of a feature vector $\mathbf{p}$ of a probe image as a block sparse recovery problem. We then solve the associated block sparse minimization problem using the alternating directions framework. We evaluate our algorithm on two publicly available benchmarking datasets and demonstrate superior results when compared to the current state of the art in multi-shot re-id. The proposed method is particularly well-suited to scenarios involving background clutter and occlusions.

2. Related Work

The traditional paradigm for solving the person re-id problem is to extract appearance features of the target and each candidate and then compare the feature vectors using a distance metric. This has given rise to two different research paths: appearance modeling and metric learning. Most re-id algorithms describe the appearance using texture and color

histograms [6, 26]. To learn distance metrics, most methods focus on learning Mahalanobis-like distances [18, 11, 1]. However, these methods are designed for the single-shot setting, i.e., they rely on comparing the feature vector of one probe image with the feature vector of one gallery image. The naive way to extend such methods to the multi-shot setting is to compare every possible pair of probe and gallery images and aggregate the results.

Several methods specifically tackle the multi-shot re-id problem. For example, Cong *et al*. [3] used image sequences to build aggregated appearance descriptors. Wang *et al*. [24] proposed an algorithm that selects discriminative fragments to learn a video ranking function. Li *et al*. [14] learned discriminative random forests and aggregated classification scores for all the available images for each person to make a decision. Image sequences have also been used to perform direct sequence matching. Simonnet *et al*. [21] used dynamic time warping to perform temporal sequence matching. The multi-shot re-id problem has also been formulated as a gait recognition problem [19], where person discrimination is based on the walking style. However, these methods are likely to fail in the presence of background clutter and occlusions.

Sparse representations have become a popular framework for various computer vision tasks, most notably in face recognition [25, 23], object tracking [17, 27] and image restoration [16]. However, sparse representation based classification methods have received relatively little attention for the person re-id problem. While some methods take this approach, e.g., [8, 12, 15], they do not exploit the inherent block structure of the feature vector dictionary.

3. Algorithm Description

3.1. Feature Extraction

We describe each image using texture and color histograms, which are popular descriptors for person re-identification [20, 30]. Following the approach of Gray and Tao [7], we divide the image into six horizontal strips. In each strip, we first compute filter responses of 13 Schmid filters and 6 Gabor filters. The filter responses are then used to compute a histogram with 16 bins. To describe the color information in each strip, we compute the 16-bin histograms in the whitened RGB space, the HSV space and the YCbCr space. This results in a 432-dimensional feature vector for each strip. The feature vectors for all the 6 strips are concatenated to form a 2592-dimensional feature vector.

Given the feature vectors $\{\mathbf{g}_{ij}\}$ for the gallery images and $\{\mathbf{p}_{ij}\}$ for the probe images, where j denotes the j^{th} image of the i^{th} unique person, computed as described above, we then learn a transformed embedding space using local Fisher discriminant analysis (LFDA) [22]. For the sake of notational convenience, let us define the matrix $\mathbf{F} \in \mathbb{R}^{f \times N}$ of all the feature vectors $\{\mathbf{g}_{ij}\}$ and $\{\mathbf{p}_{ij}\}$ as $\mathbf{F} = \begin{bmatrix}\{\mathbf{g}_{ij}\} & \{\mathbf{p}_{ij}\}\end{bmatrix}$. Here, $f = 2592$, and N is the total number of images available. The traditional Fisher discriminant analysis (FDA), which minimizes the within-class and maximizes the between-class scatter, fails to give satisfactory results if the input data is multi-modal. Indeed, in the multi-shot re-id problem, the data is multimodal since each person in the gallery view and the probe view has multiple images. To this end, we employ LFDA, wherein locality preserving projections [9] are used to ensure the feature vectors of each person are close in the embedding space, thereby preserving the local structure of the data. Specifically, we first define an affinity matrix $\mathbf{A}$ that captures the closeness of the feature vectors $\mathbf{F}_{*a}$ and $\mathbf{F}_{*b}$, where $\mathbf{F}_{*a}$ is the a^{th} column of $\mathbf{F}$. The value $\mathbf{A}_{ab} = 1$ if $\mathbf{F}_{*a}$ and $\mathbf{F}_{*b}$ are close to each other; otherwise it is set to 0. Here, we use the k-nearest neighbors rule with $k = 7$ to determine this closeness.

We then define the local within-class and between-class scatter matrices $\mathbf{S}_w$ and $\mathbf{S}_b$ as

$$\mathbf{S}_w = \frac{1}{2}\sum_{a,b=1}^{N} \mathbf{A}^w_{ab}(\mathbf{F}_{*a} - \mathbf{F}_{*b})(\mathbf{F}_{*a} - \mathbf{F}_{*b})^\top \quad (1)$$

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

where $\mathbf{A}^w_{ab}$ and $\mathbf{A}^b_{ab}$ are defined as

$$\mathbf{A}^w_{ab} = \begin{cases} \frac{\mathbf{A}_{ab}}{n_c} & \text{if class}(\mathbf{F}_{*a}) = \text{class}(\mathbf{F}_{*b}) = c \\ 0 & \text{if class}(\mathbf{F}_{*a}) \neq \text{class}(\mathbf{F}_{*b}) \end{cases} \quad (3)$$

$$\mathbf{A}^b_{ab} = \begin{cases} \mathbf{A}_{ab}(\frac{1}{N} - \frac{1}{n_c}) & \text{if class}(\mathbf{F}_{*a}) = \text{class}(\mathbf{F}_{*b}) = c \\ \frac{1}{N} & \text{if class}(\mathbf{F}_{*a}) \neq \text{class}(\mathbf{F}_{*b}) \end{cases} \quad (4)$$

where n_c is the number of images available for the person with index c. The transformation $\mathbf{T} \in \mathbb{R}^{f \times d}$ to the d-dimensional embedding space is then learned as

$$\mathbf{T} = \underset{\mathbf{T}}{argmax}\ \text{trace}\{(\mathbf{T}^\top \mathbf{S}_w \mathbf{T})^{-1}\mathbf{T}^\top \mathbf{S}_b \mathbf{T}\} \quad (5)$$

Our training process is illustrated in Figure 1.

3.2. Block Sparsity for Re-Identification

3.2.1 Problem Formulation

Let K denote the number of unique people in the gallery, and n_i be the number of available images for the person with index i, denoted as P_i in the following. Let $\hat{\mathbf{g}}_{ij} \in \mathbb{R}^d$, $j = 1, \ldots, n_i$ be the d-dimensional feature vector in

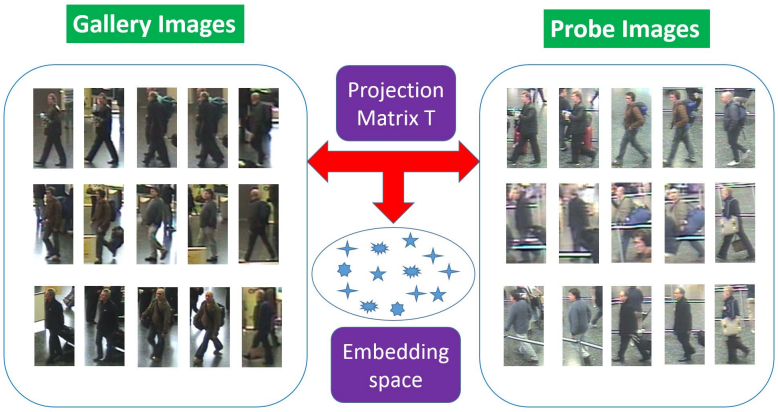

Figure 1: In the training stage, we project the gallery and probe images to a common embedding space.

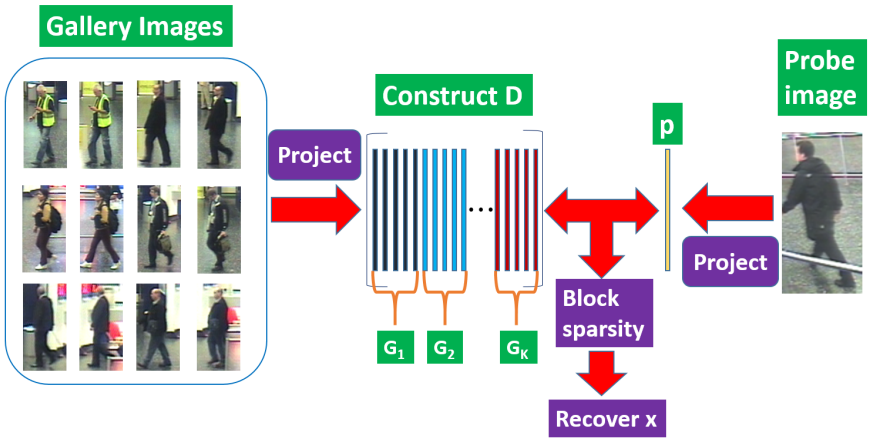

Figure 2: Given the probe feature vector, we project it to the learned embedding space and solve a block sparse recovery problem to determine its class.

the learned embedding space of the j^{th} image of P_i in the gallery.

We define the person-specific dictionary $\mathbf{G}_i \in \mathbb{R}^{d \times n_i}$ as

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

We then construct the gallery dictionary $\mathbf{D} \in \mathbb{R}^{d \times N}$ as:

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

where $N = \sum_{i=1}^{K} n_i$ is the total number of images of all people present in the gallery.

Now consider $\hat{\mathbf{p}} \in \mathbb{R}^d$, the feature vector in the learned embedding space of a particular image of P_i in the probe camera view. We hypothesize that $\hat{\mathbf{p}}$ approximately lies in the subspace spanned by the feature vectors $\hat{\mathbf{g}}_{ij}$, i.e.,

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

where $x_{ij} \in \mathbb{R}$, $j = 1, 2, \ldots, n_i$.

Put a different way, we model

$$\hat{\mathbf{p}} \approx \mathbf{G_1}\mathbf{x}[1] + \mathbf{G_2}\mathbf{x}[2] + \cdots + \mathbf{G_K}\mathbf{x}[K] \tag{9}$$

where $\mathbf{x}[i] = \begin{bmatrix} x_{i1} & x_{i2} & \cdots & x_{in_i} \end{bmatrix} \in \mathbb{R}^{n_i}$ represents the block of coefficients corresponding to the person with index i, and our hypothesis is that in the most desirable solution, the contribution from the coefficient block $\mathbf{x}[i]$ dominates the contributions from the coefficient block $\mathbf{x}[j], j \neq i$.

We note that this hypothesis is stronger than the model

$$\hat{\mathbf{p}} \approx \mathbf{D}\mathbf{x} \tag{10}$$

where $\mathbf{x} \in \mathbb{R}^N$ and the hypothesis is that $\mathbf{x}$ is sparse. That is, instead of expecting the solution vector to have as few non-zero coefficients as possible, our approach requires these coefficients to be concentrated in one of the person-specific blocks of the feature dictionary.

Following [5], we pose our problem as the following

l_1/l_2 optimization :

$$\begin{aligned} \min_{\mathbf{x}} \quad & \textstyle\sum_{i=1}^{K} \|\mathbf{x}[i]\|_2 \\ \text{s.t.} \quad & \hat{\mathbf{p}} = \mathbf{D}\mathbf{x} \end{aligned} \tag{11}$$

Intuitively, this problem formulation attempts to minimize the l_2 norm, or the energy, of the blocks in the coefficient vector $\mathbf{x} = \begin{bmatrix}\mathbf{x}[1] & \mathbf{x}[2] & \cdots & \mathbf{x}[K]\end{bmatrix}$. Subsequently, given the recovered block sparse coefficient vector, $\mathbf{x}_s$, we determine the identity of the person represented by the feature vector $\hat{\mathbf{p}}$ by simply determining the block that results in the least residual error. Specifically, we compute the residual $r_i = \|\hat{\mathbf{p}} - \mathbf{G}_i\mathbf{x}_s[i]\|$, $i = 1, 2, \ldots, K$, and assign the index of the least residual as the identity of the person. Figure 2 illustrates the overall approach.

3.2.2 Occlusions and Data Corruptions

The images of people captured from surveillance cameras are often occluded by other people and/or objects, as can be seen from the sample images shown in Figure 3. Unlike other related multi-shot re-id techniques, our formulation allows us to explicitly model occlusions. Specifically, we introduce an error term $\mathbf{e} \in \mathbf{R}^d$ into the problem formulation of Equation 10. Our linear approximation model now becomes:

$$\hat{\mathbf{p}} = \mathbf{D}\mathbf{x} + \mathbf{e} \tag{12}$$

Figure 3: Occluded people in the iLIDS-VID dataset.

The minimization problem of Equation 11 can then be expressed as:

$$\begin{aligned} \min_{\mathbf{x},\mathbf{e}} \quad & \sum_{i=1}^{K} \|\mathbf{x}[i]\|_2 + \|\mathbf{e}\|_1 \\ \text{s.t.} \quad & \hat{\mathbf{p}} = \mathbf{D}\mathbf{x} + \mathbf{e} \end{aligned} \tag{13}$$

Given the recovered block sparse coefficient vector $\mathbf{x}_s$ and the error vector $\mathbf{e}_s$, we compute the residual $r = \|\hat{\mathbf{p}} - \mathbf{G}_i\mathbf{x}_s[i] - \mathbf{e}\|$, $i = 1, 2, \ldots, K$ and determine the identity of the person as before.

3.2.3 Block Sparse Recovery using Alternating Directions

Given $\hat{\mathbf{p}}$ and $\mathbf{D}$, we use the alternating directions framework to obtain the solution to the problem of Equation 13. First, by introducing a slack variable $\mathbf{s} \in \mathbb{R}^N$, we re-formulate the problem of Equation 13 as:

$$\begin{aligned} \min_{\mathbf{s},\mathbf{x},\mathbf{e}} \quad & \sum_{i=1}^{K} \|\mathbf{s}[i]\|_2 + \|\mathbf{e}\|_1 \\ \text{s.t.} \quad & \mathbf{s} = \mathbf{x} \\ & \hat{\mathbf{p}} = \mathbf{D}\mathbf{x} + \mathbf{e} \end{aligned} \tag{14}$$

We now introduce Lagrange multipliers $\mathbf{m}_1 \in \mathbf{R}^N$ and $\mathbf{m}_2 \in \mathbf{R}^d$ to convert the constrained minimization problem of Equation 14 into the following unconstrained minimization problem:

$$\begin{aligned} \min_{\mathbf{s},\mathbf{x},\mathbf{e}} & \sum_{i=1}^{K} \|\mathbf{s}[i]\|_2 + \|\mathbf{e}\|_1 \\ & - \mathbf{m}_1^\top(\mathbf{s} - \mathbf{x}) - \mathbf{m}_2^\top(\mathbf{D}\mathbf{x} + \mathbf{e} - \hat{\mathbf{p}}) \\ & + \frac{\eta_1}{2}\|\mathbf{s} - \mathbf{x}\|^2 + \frac{\eta_2}{2}\|\mathbf{D}\mathbf{x} + \mathbf{e} - \hat{\mathbf{p}}\|^2 \end{aligned} \tag{15}$$

We add the quadratic penalty terms $\frac{\eta_1}{2}\|\mathbf{s} - \mathbf{x}\|^2$ and $\frac{\eta_2}{2}\|\mathbf{D}\mathbf{x} + \mathbf{e} - \hat{\mathbf{p}}\|^2$ to the cost function due to their smoothness property. We seek to minimize the cost function in Equation 15 over the three variables $\mathbf{s}$, $\mathbf{x}$, and $\mathbf{e}$. To this end, we employ the alternating directions framework by iteratively minimizing the cost function with respect to one variable at a time, while keeping the other two variables 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:

$$\begin{aligned} \min_{\mathbf{x}} \quad & -\mathbf{m}_1^\top(\mathbf{s} - \mathbf{x}) - \mathbf{m}_2^\top(\mathbf{D}\mathbf{x} + \mathbf{e} - \hat{\mathbf{p}}) \\ & + \frac{\eta_1}{2}\|\mathbf{s} - \mathbf{x}\|^2 + \frac{\eta_2}{2}\|\mathbf{D}\mathbf{x} + \mathbf{e} - \hat{\mathbf{p}}\|^2 \end{aligned} \tag{16}$$

It is easy to see that this $\mathbf{x}$ sub-problem has a closed-form solution, given by:

$$\mathbf{x}^* = (\eta_1 I + \eta_2 \mathbf{D}^\top\mathbf{D})^{-1}(\eta_2\mathbf{D}^\top(\hat{\mathbf{p}} - \mathbf{e}) + \eta_1\mathbf{s} + \mathbf{m}_2^\top\mathbf{D} - \mathbf{m}_1) \tag{17}$$

Next, we fix $\mathbf{s}$ and $\mathbf{x}$, and minimize the cost function with respect to $\mathbf{e}$. In this case, the cost function reduces to:

$$\min_{\mathbf{e}} \|\mathbf{e}\|_1 - \mathbf{m}_2^\top(\mathbf{D}\mathbf{x}^* + \mathbf{e} - \hat{\mathbf{p}}) + \frac{\eta_2}{2}\|\mathbf{D}\mathbf{x}^* + \mathbf{e} - \hat{\mathbf{p}}\|^2 \tag{18}$$

where $\mathbf{x}^*$ is the solution to the $\mathbf{x}$ sub-problem above. This $\mathbf{e}$ sub-problem also has a closed-form solution, given by:

$$\mathbf{e}^* = \text{shrink}\left(\frac{\mathbf{m}_2}{\eta_2} - \mathbf{D}\mathbf{x}^* - \hat{\mathbf{p}}, \frac{1}{\eta_2}\right) \tag{19}$$

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

Finally, we fix $\mathbf{x}$ and $\mathbf{e}$, and minimize the cost function with respect to $\mathbf{s}$. In this case, the cost function reduces to:

$$\min_{\mathbf{s}} \sum_{i=1}^{K} \|\mathbf{s}[i]\|_2 - \mathbf{m}_1^\top(\mathbf{s} - \mathbf{x}^*) + \frac{\eta_1}{2}\|\mathbf{s} - \mathbf{x}^*\|^2 \quad (20)$$

We note that this $\mathbf{s}$ sub-problem also has a closed-form solution, where the coefficients for each block $i = 1, 2, \ldots, K$ are given by the so-called block shrink [4] operation:

$$\mathbf{s}^*[i] = \max\left(\|\mathbf{x}^*[i] + \frac{\mathbf{m}_1[i]}{\eta_1}\| - \frac{1}{\eta_1}, 0\right)\frac{\mathbf{x}^*[i] + \frac{\mathbf{m}_1[i]}{\eta_1}}{\|\mathbf{x}^*[i] + \frac{\mathbf{m}_1[i]}{\eta_1}\|^2} \quad (21)$$

Finally, we update the Lagrange multipliers as $\mathbf{m}_1 = \mathbf{m}_1 - \eta_1(\mathbf{s}^* - \mathbf{x}^*)$ and $\mathbf{m}_2 = \mathbf{m}_2 - \eta_1(\mathbf{D}\mathbf{x}^* + \mathbf{e}^* - \hat{\mathbf{p}})$. This iterative procedure is summarized in Algorithm 1.

Algorithm 1: An alternating directions algorithm to solve the minimization problem of Equation 13

Input : $\hat{\mathbf{p}}, \mathbf{D} \in \mathbb{R}^{m \times n}$
Output: $\mathbf{x}^*, \mathbf{e}^*$
Initialize $\mathbf{s} = \mathbf{0}, \mathbf{e} = \mathbf{0}, \mathbf{m}_1 = \mathbf{0}, \mathbf{m}_2 = \mathbf{0}$;
$\eta_1 = \frac{2m}{\|\hat{\mathbf{p}}\|_1}, \eta_2 = \eta_1$;
for $t \leftarrow 1, 2, \ldots$ **do**
  $\mathbf{x}_t = (\eta_1 I + \eta_2 \mathbf{D}^\top \mathbf{D})^{-1}(\eta_2 \mathbf{D}^\top(\hat{\mathbf{p}} - \mathbf{e}_{t-1}) + \eta_1 \mathbf{s}_{t-1} + \mathbf{m}_2^\top \mathbf{D} - \mathbf{m}_1)$;
  $\mathbf{e}_t = \text{shrink}(\frac{\mathbf{m}_2}{\eta_2} - \mathbf{D}\mathbf{x}_t - \hat{\mathbf{p}}, \frac{1}{\eta_2})$;
  $\mathbf{s}_t[i] = \max\left(\|\mathbf{x}_t[i] + \frac{\mathbf{m}_1[i]}{\eta_1}\| - \frac{1}{\eta_1}, 0\right)\frac{\mathbf{x}_t[i] + \frac{\mathbf{m}_1[i]}{\eta_1}}{\|\mathbf{x}_t[i] + \frac{\mathbf{m}_1[i]}{\eta_1}\|^2}$, $i = 1, 2, \ldots, K$;
  $\mathbf{m}_1 = \mathbf{m}_1 - \eta_1(\mathbf{s}_t - \mathbf{x}_t)$;
  $\mathbf{m}_2 = \mathbf{m}_2 - \eta_1(\mathbf{D}\mathbf{x}_t + \mathbf{e}_t - \hat{\mathbf{p}})$
end
$\mathbf{x}^* = \mathbf{x}_t$;
$\mathbf{e}^* = \mathbf{e}_t$;

3.2.4 Re-identification

Let $\hat{\mathbf{p}}_{ij} \in \mathbb{R}^d$, be the feature vector in the embedding space of the j^{th} image of P_i in the probe camera view. To re-identify this person, we solve the problem of Equation 13 for each $\hat{\mathbf{p}}_{ij}$, $j = 1, 2, \ldots, m$, where m is the number of images available in the probe camera view for P_i. In each case, we compute the residual vector $r_j(i) = \|\hat{\mathbf{p}}_{ij} - \mathbf{G}_i\mathbf{x}_s[i] - \mathbf{e}_s\|$, $i = 1, 2, \ldots, K$. We then sum all the residual vectors $\mathbf{r}_j \in \mathbb{R}^K$ for each image of the person to form the net residual vector $\mathbf{R} = \sum_{j=1}^{K} \mathbf{r}_j$. We then determine the identity of the person as the index of the minimum value in this net residual vector $\mathbf{R}$. This process is summarized in Algorithm 2.

Algorithm 2: The proposed multi-shot re-id framework

Input : Feature vectors $\hat{\mathbf{p}}_{ij} \in \mathbb{R}^d$, $j = 1, 2, \ldots, m$, of the person P_i in the probe camera view, Gallery person dictionaries $\mathbf{G}_i$, $i = 1, 2, \ldots, K$
Output: Class c of person P_i
$\mathbf{R} = \mathbf{0}$;
for $j \leftarrow 1, 2, \ldots, m$ **do**
  Solve problem of Equation 13 for $\hat{\mathbf{p}}_{ij}$ to get $\mathbf{x}^*$ and $\mathbf{e}^*$;
  Compute residuals vector $r_j(i) = \|\hat{\mathbf{p}}_{ij} - \mathbf{G}_i\mathbf{x}^*[i] - \mathbf{e}^*\|$, $i = 1, 2, \ldots, K$;
  $\mathbf{R} = \mathbf{R} + \mathbf{r}_j$;
end
$c =$ index of the minimum value in $\mathbf{R}$;

4. Experiments and Results

We experimentally validate the efficacy of the proposed multi-shot person re-identification formulation on the following publicly available multi-shot datasets: iLIDS-VID [24] and PRID 2011 [10].

iLIDS-VID [24]: This dataset was created from two non-overlapping camera views at an airport. In each camera view, the dataset consists of image sequences of variable length for 300 people. The images in this dataset suffer from extreme lighting and viewpoint variations, occlusions and cluttered background.

PRID 2011 [10]: This dataset was created from two adjacent camera views capturing outdoor scenes. In each camera view, the dataset consists of image sequences of variable length for 200 people. The images in this dataset involve viewpoint, illumination, and background variations.

4.1. Evaluation protocol

For each dataset, we randomly split the data into equal-sized training and testing sets. For each split, we randomly select 10 images for each person in both the gallery view and the probe view. The training set is used to learn the projection matrix in Equation 5. Using this projection matrix, the test set is then projected onto the d-dimensional embedding space. All the images from one camera view form the gallery dictionary $\mathbf{D}$, and the images from the other view are used as probe images. We repeat this process 10 times

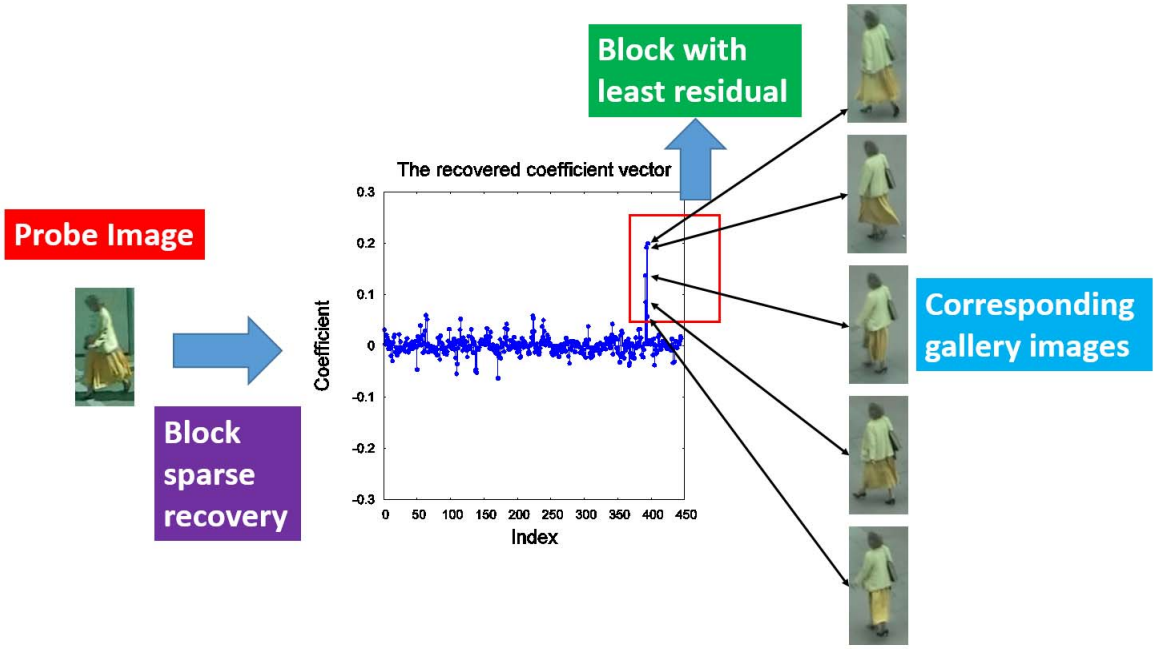

Figure 4: An illustrative example demonstrating the efficacy of our re-id framework.

Table 1: Evaluating the impact of the embedding space.

Dataset	PRID 2011				iLIDS-VID			
Rank	Rank 1	Rank 5	Rank 10	Rank 20	Rank 1	Rank 5	Rank 10	Rank 20
SRID in original feature space	9.0	27.7	40.4	56.4	12.6	28.4	37.9	50.2
SRID in embedding space	**35.1**	**59.4**	**69.8**	**79.7**	**24.9**	**44.5**	**55.6**	**66.2**

to compute the average performance for this particular train-test split. We repeat this experimental procedure for 10 such train-test splits and report the overall average performance. We compare our results with several recently proposed approaches that report state-of-the-art performance: SDALF [6], Salience [28], RPRF [14], DVR [24] and a multi-shot extension, as described in [24], of a combination of RankSVM [2] and Color and LBP features [11]. In all these competing approaches, we use the same experimental settings as described in [24]. We abbreviate our proposed approach as **SRID**.

We choose the dimension of the embedding space using cross-validation. Figure 5 plots the Rank 1 performance on the validation set for both PRID 2011 and iLIDS-VID datasets as a function of d. Since $d = 750$ results in the best rank 1 performance on both datasets, we fixed $d = 750$ for the remaining experiments.

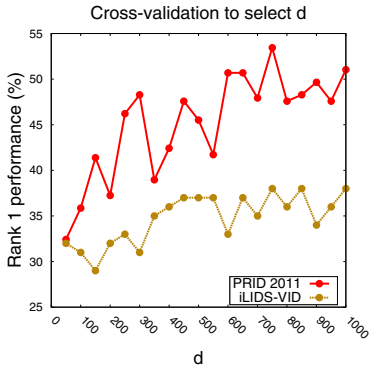

Figure 5: Rank 1 performance on the validation set.

4.2. Results and Discussion

We begin the discussion of our results with an illustrative example. Consider the probe image shown in Figure 4. This test case is taken from the PRID 2011 [10] dataset. For the purposes of this particular example, we considered 5 images for each person in both the gallery view and the probe view. Following the testing procedure described in the previous sections, we first compute the feature vector for this image and project it to the learned embedding space.

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

Dataset	PRID 2011				iLIDS-VID			
Rank	Rank 1	Rank 5	Rank 10	Rank 20	Rank 1	Rank 5	Rank 10	Rank 20
SDALF [6]	5.2	20.7	32	47.9	6.3	18.8	27.1	37.3
Salience [28]	25.8	43.6	52.6	62	10.2	24.8	35.5	52.9
RPRF [14]	19.3	38.4	51.6	68.1	14.5	29.8	40.7	58.1
DVR [24]	28.9	55.3	65.5	**82.8**	23.3	42.4	55.3	68.4
Color & LBP [11] + RankSVM [2]	34.3	56	65.5	77.3	23.2	44.2	54.1	**68.8**
SRID	**35.1**	**59.4**	**69.8**	79.7	**24.9**	**44.5**	**55.6**	66.2

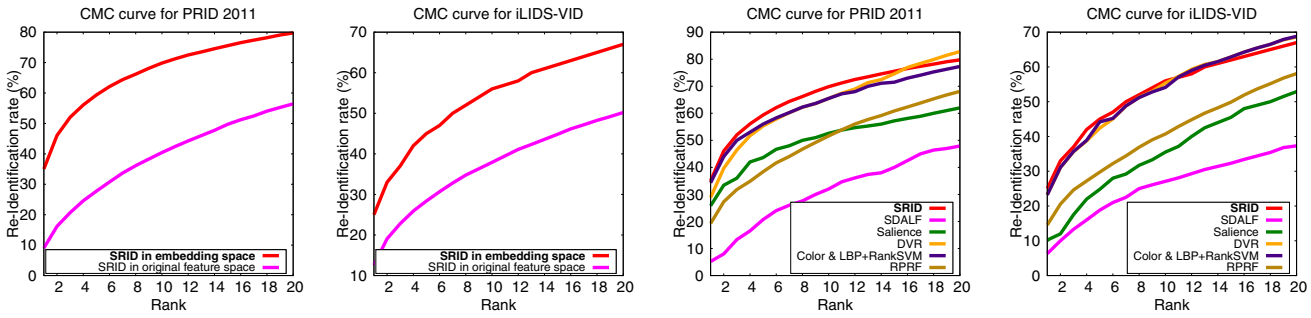

Figure 6: CMC curves for the PRID 2011 and iLIDS-VID datasets

Now, given this projected vector $\hat{\mathbf{p}}$ and the matrix $\mathbf{D}$, we compute the block sparse coefficient vector $\mathbf{x}^*$ using the alternating directions framework described in Section 3.2.3. The elements of this recovered vector are also shown in the graph in Figure 4. As can be seen from the graph, there exists one block of elements that contains significant coefficients, while most of the other terms are relatively insignificant, in line with our hypothesis described in Section 3.2. In fact, this is the block that corresponds to the correct set of gallery images, thereby forming evidence for correct re-identification in this case.

4.3. Evaluating the embedding space

We first evaluate the impact of projecting the feature vectors to the learned embedding space. To this end, we perform experiments following the protocol discussed in Section 4.1 on the iLIDS-VID and PRID 2011 datasets twice: first, applying SRID in the embedding space, and second, applying SRID in the original feature space. The corresponding cumulative match characteristic (CMC) curves for both the test datasets are shown in Figures 6a and 6b and the results are summarized in Table 1. From the table, we see that with SRID in the embedding space, Rank 1 performance has been boosted by about 26% and 12% on the PRID 2011 and iLIDS-VID datasets respectively. We can also note from the plot that SRID in the embedding space gives consistently better results at all ranks when compared with SRID in the original feature space. This experiment clearly validates our original hypothesis of formulating a linear approximation model in some learned embedding space rather than the original feature space.

4.4. Comparison with the state of the art

We next compare the rank-ordered re-identification results of our approach with that of the current state of the art in multi-shot re-id. The CMC curves for both the evaluation datasets are shown in Figures 6c and 6d, and the results are summarized in Table 2. As can be seen from the results, our proposed approach offers superior performance at ranks 1, 5 and 10 when compared with the other competing approaches on both PRID 2011 and iLIDS-VID. Specifically, the rank 1 performance of SRID is 35.1% and 24.9% on the PRID 2011 and iLIDS-VID datasets respectively, whereas the corresponding best numbers among all the competing methods are 34.3% and 23.3% respectively.

5. Conclusions and Future Work

In this work, we presented a novel formulation for the multi-shot person re-identification problem. We conjectured that the feature vector of a probe image in a learned embedding space approximately lies in the linear span of

the corresponding gallery feature vectors. We constructed a dictionary $\mathbf{D}$ of the gallery feature vectors, and exploited its inherent block structure by posing re-id as a block sparse minimization problem, which we solved using the alternating directions framework. We evaluated our approach on two publicly available benchmarking multi-shot re-id datasets and demonstrated superior results when compared to the current state of the art.

As discussed in Section 4.1, for the purposes of evaluating our approach, we randomly selected 10 images to form the probe and the gallery. We did this primarily to deal with the high computational complexity associated with the iterative solution to the problem of Equation 13 discussed in Section 3.2.3. In future work, we plan on exploring two lines of work. First, we will investigate faster procedures to solve the problem of Equation 13. Next, we will focus on developing algorithms that make a more informed choice as to which of the available images are most discriminative for re-identification.

References

[1] S. Bak, G. Charpiat, E. Corvee, F. Bremond, and M. Thonnat. Learning to match appearances by correlations in a covariance metric space. In *ECCV*, pages 806–820. 2012.

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

[3] D. N. T. Cong, C. Achard, L. Khoudour, and L. Douadi. Video sequences association for people re-identification across multiple non-overlapping cameras. In *ICIAP*, pages 179–189. 2009.

[4] W. Deng, W. Yin, and Y. Zhang. Group sparse optimization by alternating direction method. In *SPIE Optical Engineering+ Applications*, pages 88580R–88580R, 2013.

[5] Y. C. Eldar and M. Mishali. Robust recovery of signals from a structured union of subspaces. *IEEE T-IT*, 55(11):5302–5316, 2009.

[6] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, pages 2360–2367, 2010.

[7] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, pages 262–275. 2008.

[8] M. T. Harandi, C. Sanderson, R. Hartley, and B. C. Lovell. Sparse coding and dictionary learning for symmetric positive definite matrices: A kernel approach. In *ECCV*, pages 216–229. 2012.

[9] X. He and P. Niyogi. Locality preserving projections. In *NIPS*, volume 16, pages 153–160, 2004.

[10] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *Image Analysis*, pages 91–102. 2011.

[11] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *ECCV*, pages 780–793. 2012.

[12] M. I. Khedher, M. A. El Yacoubi, and B. Dorizzi. Multi-shot surf-based person re-identification via sparse representation. In *AVSS*, pages 159–164, 2013.

[13] Y. Li, Z. Wu, S. Karanam, and R. J. Radke. Real-world re-identification in an airport camera network. In *ICDSC*, pages 35:1–35:6, 2014.

[14] Y. Li, Z. Wu, and R. J. Radke. Multi-shot re-identification with random-projection-based random forests. In *WACV*, pages 373–380, 2015.

[15] G. Lisanti, I. Masi, A. Bagdanov, and A. Del Bimbo. Person re-identification by iterative re-weighted sparse ranking. *IEEE T-PAMI*, PP(99), 2014.

[16] J. Mairal, G. Sapiro, and M. Elad. Learning multiscale sparse representations for image and video restoration. *SIAM MMS*, 7(1):214–241, 2008.

[17] X. Mei and H. Ling. Robust visual tracking using l_1 minimization. In *ICCV*, pages 1436–1443, 2009.

[18] A. Mignon and F. Jurie. PCCA: A new approach for distance learning from sparse pairwise constraints. In *CVPR*, pages 2666–2672, 2012.

[19] M. S. Nixon, T. Tan, and R. Chellappa. *Human identification based on gait*, volume 4. Springer Science & Business Media, 2010.

[20] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by support vector ranking. In *BMVC*, volume 2, page 6, 2010.

[21] D. Simonnet, M. Lewandowski, S. A. Velastin, J. Orwell, and E. Turkbeyler. Re-identification of pedestrians in crowds using dynamic time warping. In *ECCV Workshops*, pages 423–432, 2012.

[22] M. Sugiyama. Local fisher discriminant analysis for supervised dimensionality reduction. In *ICML*, pages 905–912, 2006.

[23] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma. Toward a practical face recognition system: Robust alignment and illumination by sparse representation. *IEEE T-PAMI*, 34(2):372–386, 2012.

[24] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by video ranking. In *ECCV*, pages 688–703. 2014.

[25] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE T-PAMI*, 31(2):210–227, 2009.

[26] Z. Wu, Y. Li, and R. Radke. Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features. *IEEE T-PAMI*, PP(99), 2014.

[27] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja. Robust visual tracking via structured multi-task sparse learning. *IJCV*, 101(2):367–383, 2013.

[28] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *CVPR*, pages 3586–3593, 2013.

[29] W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. In *CVPR*, pages 649–656, 2011.

[30] W.-S. Zheng, S. Gong, and T. Xiang. Reidentification by relative distance comparison. *IEEE T-PAMI*, 35(3):653–668, 2013.