Person Re-Identification with Discriminatively Trained Viewpoint Invariant Dictionaries

Srikrishna Karanam, Yang Li, Richard J. Radke
Department of Electrical, Computer, and Systems Engineering
Rensselaer Polytechnic Institute
Troy, NY 12180
karans3@rpi.edu, yangli625@gmail.com, rjradke@ecse.rpi.edu

Abstract

This paper introduces a new approach to address the person re-identification problem in cameras with non-overlapping fields of view. Unlike previous approaches that learn Mahalanobis-like distance metrics in some transformed feature space, we propose to learn a dictionary that is capable of discriminatively and sparsely encoding features representing different people.

Our approach directly addresses two key challenges in person re-identification: viewpoint variations and discriminability. First, to tackle viewpoint and associated appearance changes, we learn a single dictionary to represent both gallery and probe images in the training phase. We then discriminatively train the dictionary by enforcing explicit constraints on the associated sparse representations of the feature vectors. In the testing phase, we re-identify a probe image by simply determining the gallery image that has the closest sparse representation to that of the probe image in the Euclidean sense.

Extensive performance evaluations on three publicly available multi-shot re-identification datasets demonstrate the advantages of our algorithm over several state-of-the-art dictionary learning, temporal sequence matching, and spatial appearance and metric learning based techniques.

1. Introduction

Person re-identification, or re-id, in networks of cameras with non-overlapping fields of view is an important problem in surveillance applications, such as airport security [15]. As can be seen from the sample images in Figure 1, re-id is an extremely challenging problem due to illumination, pose, and viewpoint changes in images of the same person in different camera views. Due to these inter-camera variations, we cannot rely on direct matching of the appearance features.

Figure 1: Person re-identification is a challenging problem due to viewpoint changes, occlusions, illumination changes and background clutter in images of the same person in cameras with non-overlapping fields of view.

The traditional paradigm to tackle challenges posed by these inter-camera variations is to learn, in a supervised fashion, a distance metric $\mathbf{d}_M(\mathbf{y}_1, \mathbf{y}_2) = (\mathbf{y}_1 - \mathbf{y}_2)^\top \mathbf{M}(\mathbf{y}_1 - \mathbf{y}_2)$, where $\mathbf{M}$ is a positive semidefinite matrix, so that the feature vectors extracted from images of the same person are close while those extracted from different people are far apart. While this supervised learning process does take inter-camera

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.

variations into account, the Mahalanobis distance metric has limited expressive capability due to its inherent linearity. To mitigate this problem, there have been efforts to learn both inter-camera as well as feature-level transformation functions. For example, Li *et al*. [17] proposed an algorithm to learn local decision functions instead of the absolute threshold decision rule associated with the distance metrics discussed above. Xiong *et al*. [31] proposed a series of kernel-based techniques to learn non-linear feature transformation functions. However, these techniques are still limited by the performance of the learned distance metric.

Dictionaries learned from data have recently achieved impressive results in several classification and recognition problems [21, 20, 13]. This can be attributed to their strong representational power. In this paper, we learn a dictionary that is capable of discriminatively encoding the feature vectors of different people. While most previous re-id research [31, 36, 33] has focused on the single-shot problem, in surveillance applications, we typically have a track of images for each person, making realistic re-id a multi-shot problem. We exploit this fact to compute a representative feature vector from all the available images for each person. Given these feature vectors, we learn a single dictionary invariant to viewpoint changes across camera views. Additionally, we incorporate explicit constraints on the feature representations with respect to the dictionary into our problem formulation, providing the dictionary with strong discriminative ability. A key difference between our algorithm and other dictionary learning approaches is that we do not explicitly require training a separate dictionary for each class, i.e., for each person. We use three publicly available multi-shot re-id datasets to perform extensive performance evaluations against several contemporary dictionary learning, temporal sequence matching, and spatial appearance and metric learning based approaches.

1.1. Summary of our Contributions

Here, we summarize the key contributions of our work:

Viewpoint Invariance: We learn a dictionary that is invariant to viewpoint changes commonly occurring in images from surveillance cameras. We achieve this by learning a *single* dictionary that is capable of sparsely encoding images from both the gallery and probe camera views simultaneously.

Discriminability: We learn a discriminative dictionary that is capable of telling apart feature vectors from different people. While most related work in the area of discriminative dictionary learning has focused on learning incoherent, class-wise dictionaries, our unique problem formulation enables us to learn a *single* dictionary that can discriminate between data from different classes.

2. Related Work

Person re-identification: The past body of work in person re-id has revolved around two central themes - appearance modeling and metric learning. Most high-performing re-id techniques model person appearance using global color and texture histograms [6, 10, 30, 37, 24]. Local features, such as SIFT [19], extracted from small sub-regions in images, have also recently shown good matching performance [35, 36]. In metric learning, several methods have learned Mahalanobis-like distance functions [22, 12, 2]. Other approaches, such as learning decision tree ensembles [16] and salience [35], have also been explored.

Several approaches have been proposed to specifically address the multi-shot re-id problem. Simonnet *et al*. [27] used dynamic time warping to perform direct image sequence matching. Wang *et al*. [29] presented a technique to automatically select the most discriminative fragments from a given set of images, and learned a video ranking function to perform re-identification. Gait recognition was used in [23], where person discrimination is based on the walking style.

Our work falls into the category of metric learning for re-identification. While dictionary learning has been shown to provide promising results in problems such as face recognition and object classification, as will be discussed next, it has received little attention for the person re-id problem. Liu *et al*. [18] jointly learned two coupled dictionaries to capture the appearance variations across the gallery and probe camera images. However, learning separate dictionaries for each camera view can pose practical difficulties as the amount and dimensionality of the training data increases.

Dictionary learning: Recently, dictionary learning has been successfully applied to various recognition problems. Shekhar *et al*. [26] employed domain adaptation to learn class-wise discriminative dictionaries. Jiang *et al*. [13] employed label consistency constraints to jointly learn a discriminative dictionary and a linear classifier. Zhang and Li [34] extended the K-SVD [1] algorithm by incorporating classification error into the problem formulation and learned class-wise dictionaries. Yang *et al*. [32] employed Fisher discrimination constraints on the associated sparse codes to learn class-wise discriminative dictionaries.

A common theme of most of these approaches is to learn separate sub-dictionaries for each data class to achieve discriminability. In contrast to these approaches, our algorithm explicitly incorporates feature-level constraints into the problem formulation, while learning a single viewpoint invariant dictionary.

3. Algorithm Description

In this section, we first briefly review the basics of dictionary learning before discussing our approach to learn discriminative viewpoint invariant dictionaries.

3.1. Dictionary Learning

In traditional reconstructive dictionary learning problems, the goal is to learn a dictionary $\mathbf{D}$ that sparsely and accurately captures the information present in all the input signals $\mathbf{y}_i, \forall i$. This is mathematically formulated as the following optimization problem:

$$\mathbf{D}^*, \mathbf{X}^* = \arg\min_{\mathbf{D}, \{\mathbf{x}_i\}} \sum_{i=1}^{n} \{\lambda \|\mathbf{x}_i\| + \|\mathbf{y}_i - \mathbf{D}\mathbf{x}_i\|_2^2\} \quad (1)$$

where $\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \end{bmatrix}$ corresponds to a sparse coding of the input signals $\mathbf{y}_i$, $\forall i$. This problem is typically solved using the alternating directions framework by alternately fixing $\mathbf{D}$ and $\mathbf{X}$ and optimizing over the other variable.

3.2. Problem Specification

Let $\mathbf{f}_{ij} \in \mathbb{R}^d$ be the representative feature vector computed from all the available images for the person with index i in the view of camera j. In all our subsequent discussion, we let $j = 1$ to denote the gallery camera and $j = 2$ to denote the probe camera. We discuss the computation of this feature vector in Section 4.2. Consider three such feature vectors $\mathbf{f}_{11}$, $\mathbf{f}_{12}$, and $\mathbf{f}_{22}$. We seek to learn a dictionary $\mathbf{D} \in \mathbb{R}^{d \times N}$ that is capable of discriminating the sparse codes corresponding to the feature vectors $\mathbf{f}_{11}$, $\mathbf{f}_{12}$, and $\mathbf{f}_{22}$. Specifically, let $\mathbf{s}_{11}$, $\mathbf{s}_{12}$, and $\mathbf{s}_{22}$ be the sparse codes of these feature vectors with respect to the dictionary $\mathbf{D}$. We compute $\mathbf{s}_{ij}$ by solving the following l_1 regularized least squares problem:

$$\mathbf{s}_{ij} = \arg\min_{\mathbf{s}} \lambda \|\mathbf{s}\|_1 + \|\mathbf{f}_{ij} - \mathbf{D}\mathbf{s}\|_2^2 \quad (2)$$

Since $\mathbf{f}_{11}$ and $\mathbf{f}_{12}$ are the feature vectors, albeit in different camera views, of the same person, our hypothesis is that $\mathbf{s}_{12}$ will have a smaller Euclidean distance to the gallery sparse code $\mathbf{s}_{11}$ than $\mathbf{s}_{22}$. The intuition here is that the images of the same person in different camera views should have similar sparse codes with respect to the learned dictionary.

3.3. Problem Formulation

Given the feature vectors $\mathbf{f}_{i1}$ and $\mathbf{f}_{i2}$, $i = 1, 2, \cdots, n$ of n persons in the gallery and the probe camera views respectively, our goal is to learn a dictionary $\mathbf{D}$ satisfying the following properties:

Property 1: $\mathbf{D}$ *should be viewpoint invariant.*

$\mathbf{D}$ should be able to accurately represent the feature vectors $\mathbf{f}_{i1}$ and $\mathbf{f}_{i2}$ even if they are computed from images with large viewpoint changes. This is a desirable property for the person re-id problem because, in practice, viewpoint changes among probe and gallery cameras are often quite pronounced. The intuition here is that we learn a common dictionary to represent both the gallery and the probe images. To satisfy this property, we formulate the dictionary learning process as the following minimization problem:

$$\begin{aligned} \mathbf{D}^*, \mathbf{S}_1^*, \mathbf{S}_2^* = \arg\min_{\mathbf{D}, \{\mathbf{s}_{i1}\}, \{\mathbf{s}_{i2}\}} \sum_{i=1}^{n} \{\lambda_1 \|\mathbf{s}_{i1}\|_1 + \\ + \|\mathbf{f}_{i1} - \mathbf{D}\mathbf{s}_{i1}\|_2^2 + \lambda_2 \|\mathbf{s}_{i2}\|_1 \\ + \|\mathbf{f}_{i2} - \mathbf{D}\mathbf{s}_{i2}\|_2^2\} \end{aligned} \quad (3)$$

We note that the term $\|\mathbf{f}_{i1} - \mathbf{D}\mathbf{s}_{i1}\|_2^2 + \|\mathbf{f}_{i2} - \mathbf{D}\mathbf{s}_{i2}\|_2^2$ in the above formulation ensures that the dictionary $\mathbf{D}$ represents both $\mathbf{f}_{i1}$ and $\mathbf{f}_{i2}$, thereby satisfying our requirement that it be viewpoint invariant.

Property 2: $\mathbf{D}$ *should be discriminative.*

$\mathbf{D}$ should be able to discriminate between feature vectors of the same person and the feature vectors of different people. In our work, we enforce this discriminability by imposing explicit constraints on the sparse codes that represent these feature vectors with respect to the dictionary. Specifically, if $d_i = \|\mathbf{s}_{i1} - \mathbf{s}_{i2}\|$ represents the Euclidean distance between the sparse codes corresponding to the gallery and the probe feature vectors of person i, and $d_{ij} = \|\mathbf{s}_{i1} - \mathbf{s}_{j2}\|$ represents the Euclidean distance between the sparse codes corresponding to the gallery feature vector of person i and the probe feature vector of person j, we explicitly require the following conditions to hold:

$$d_i < d_{ij}, \forall j \neq i, \forall i \quad (4)$$

Requiring that the learned dictionary satisfies both the properties, we have the following overall minimization problem:

$$\begin{aligned} \min_{\mathbf{D}, \{\mathbf{s}_{i1}\}, \{\mathbf{s}_{i2}\}} \quad & \textstyle\sum_{i=1}^{n} \{\lambda_1 \|\mathbf{s}_{i1}\|_1 + \|\mathbf{f}_{i1} - \mathbf{D}\mathbf{s}_{i1}\|_2^2 + \\ & \lambda_2 \|\mathbf{s}_{i2}\|_1 + \|\mathbf{f}_{i2} - \mathbf{D}\mathbf{s}_{i2}\|_2^2\} \\ \text{s.t.} \quad & \|\mathbf{s}_{i1} - \mathbf{s}_{i2}\|_2 < \|\mathbf{s}_{i1} - \mathbf{s}_{j2}\|_2, \\ & \forall j \neq i, \forall i \end{aligned} \quad (5)$$

We note that while the objective in the optimization problem above is convex, the constraints are defined in a manner that is not consistent with disciplined convex programming [9]. Therefore, we introduce two constants c_1 and c_2 and reformulate the problem as:

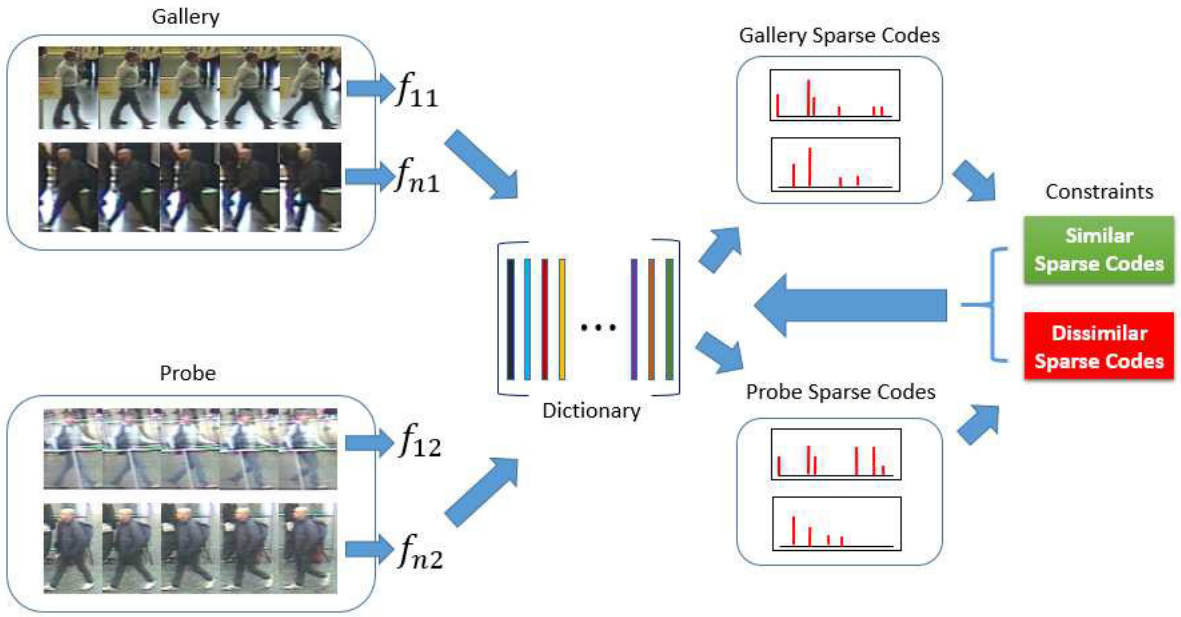

Figure 2: A visual summary of our training process. Given image sequences in both gallery and probe camera views for n persons, we first compute their representative feature vectors. We then iteratively train a discriminative viewpoint invariant dictionary by imposing explicit constraints on the corresponding gallery and probe sparse codes.

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

where $c_1 \ll c_2$.

We further note that this problem is not convex in $\mathbf{D}$, $\{\mathbf{s}_{i1}\}$, and $\{\mathbf{s}_{i2}\}$ simultaneously, but is convex in one of the variables while holding the other two fixed. Therefore, as discussed in the next section, we use the method of alternating directions to solve this problem.

3.4. Solving the optimization problem

We employ the alternating directions framework to solve the problem of Equation 6. Specifically, we alternatively optimize over $\mathbf{s}_{i1}$, $\mathbf{s}_{i2}$, and $\mathbf{D}$ one at a time, while fixing the other two. This process is described next.

3.4.1 Update steps for $\mathbf{s}_{ij}$

We first fix $\mathbf{D}$ and $\mathbf{s}_{i2}$, $\forall i$ and optimize over $\mathbf{s}_{i1}$, $\forall i$. In this case, the optimization problem reduces to:

$$\begin{aligned} \min_{\{\mathbf{s}_{i1}\}} \quad & \textstyle\sum_{i=1}^{n}\{\lambda_1\|\mathbf{s}_{i1}\|_1 + \|\mathbf{f}_{i1} - \mathbf{D}\mathbf{s}_{i1}\|_2^2\} \\ \text{s.t.} \quad & \|\mathbf{s}_{i1} - \mathbf{s}_{i2}\|_2 < c_1, \forall i \\ & \|\mathbf{s}_{i1} - \mathbf{s}_{j2}\|_2 < c_2, \forall j \neq i, \forall i \end{aligned} \tag{7}$$

To solve this problem, we optimize over a particular $\mathbf{s}_{p1}$ at a time, while fixing all other sparse codes $\mathbf{s}_{i1}$, $\forall i \neq p$. Therefore, the above optimization problem now reduces to:

$$\begin{aligned} \min_{\mathbf{s}_{p1}} \quad & \lambda_1\|\mathbf{s}_{p1}\|_1 + \|\mathbf{f}_{i1} - \mathbf{D}\mathbf{s}_{p1}\|_2^2 \\ \text{s.t.} \quad & \|\mathbf{s}_{p1} - \mathbf{s}_{p2}\|_2 < c_1 \\ & \|\mathbf{s}_{p1} - \mathbf{s}_{j2}\|_2 < c_2, \forall j \neq p \end{aligned} \tag{8}$$

We note that this problem conforms to disciplined convex programming, and can be solved using CVX, a package for specifying and solving convex programs [8, 7].

We next fix $\mathbf{D}$ and $\mathbf{s}_{i1}^*$, $\forall i$ and optimize over $\mathbf{s}_{i2}$, $\forall i$. $\mathbf{s}_{i1}^*$, $\forall i$ correspond to the optimal gallery camera sparse codes obtained in the first step above. In this case, the optimization problem reduces to:

$$\begin{aligned} \min_{\{\mathbf{s}_{i2}\}} \quad & \textstyle\sum_{i=1}^{n}\{\lambda_1\|\mathbf{s}_{i2}\|_1 + \|\mathbf{f}_{i2} - \mathbf{D}\mathbf{s}_{i2}\|_2^2\} \\ \text{s.t.} \quad & \|\mathbf{s}_{i2} - \mathbf{s}_{i1}^*\|_2 < c_1, \forall i \\ & \|\mathbf{s}_{i2} - \mathbf{s}_{j1}^*\|_2 < c_2, \forall j \neq i, \forall i \end{aligned} \tag{9}$$

We note that this problem has a similar structure as the update problem for $\mathbf{s}_{i1}$, $\forall i$, and solve it in a similar manner as before to obtain the optimal probe camera sparse codes $\mathbf{s}_{i2}^*$, $\forall i$.

3.4.2 Update step for $\mathbf{D}$

Finally, we fix $\mathbf{s}_{i1}$ and $\mathbf{s}_{i2}$, $\forall i$ and optimize over $\mathbf{D}$. In this case, it is straightforward to see that the optimization prob-

lem reduces to:

$$\mathbf{D}^* = \arg\min_{\mathbf{D}} \sum_{i=1}^{n} \{\|\mathbf{f}_{i1} - \mathbf{D}\mathbf{s}_{i1}^*\|_2^2 + \|\mathbf{f}_{i2} - \mathbf{D}\mathbf{s}_{i2}^*\|_2^2\} \quad (10)$$

where the $\mathbf{s}_{ij}^*$ are obtained from the update steps above. This problem can be converted to the following Frobenius norm minimization problem:

$$\mathbf{D}^* = \arg\min_{\mathbf{D}} \|\mathbf{F}_1 - \mathbf{D}\mathbf{S}_1^*\|_F^2 + \|\mathbf{F}_2 - \mathbf{D}\mathbf{S}_2^*\|_F^2 \quad (11)$$

where $\mathbf{F}_i = \begin{bmatrix}\mathbf{f}_{1i} & \mathbf{f}_{2i} & \cdots & \mathbf{f}_{ni}\end{bmatrix}$ and $\mathbf{S}_i = \begin{bmatrix}\mathbf{s}_{1i} & \mathbf{s}_{2i} & \cdots & \mathbf{s}_{ni}\end{bmatrix}$. We note that this problem is convex in $\mathbf{D}$ and has a unique global minimum, given by:

$$\mathbf{D}^* = (\mathbf{F}_1\mathbf{S}_1^\top + \mathbf{F}_2\mathbf{S}_2^\top)(\mathbf{S}_1\mathbf{S}_1^\top + \mathbf{S}_2\mathbf{S}_2^\top)^{-1} \quad (12)$$

Our training procedure is visually summarized in Figure 2.

3.5. Re-Identification

Given the gallery feature vectors $\mathbf{f}_{i1}, i = 1, 2, \cdots, p$, we propose the following steps to re-identify a person represented by the probe feature vector $\mathbf{f}_{u2}$.

1. For each gallery feature vector $\mathbf{f}_{i1}$, compute the corresponding sparse codes with respect to the learned dictionary $\mathbf{D}$ as:

   $$\mathbf{s}_{i1} = \arg\min_{\mathbf{s}} \lambda\|\mathbf{s}\|_1 + \|\mathbf{f}_{i1} - \mathbf{D}\mathbf{s}\|_2^2, \forall i \quad (13)$$

   In our work, we use the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) [3] to solve this l_1-regularized least squares problem.

2. Similarly, compute the sparse code $\mathbf{s}_{u2}$ of the unknown probe feature vector $\mathbf{f}_{u2}$ with respect to $\mathbf{D}$.

3. Now, compute the Euclidean distance between $\mathbf{s}_{u2}$ and each of $\mathbf{s}_{i1}$ to form the distance vector d:

   $$d(i) = \|\mathbf{s}_{u2} - \mathbf{s}_{i1}\|, \forall i \quad (14)$$

4. Finally, the class u of the probe person is obtained as the index of the minimum value in d.

4. Experiments and Results

4.1. Datasets

We empirically validate the algorithm proposed in this paper using three publicly available multi-shot re-id datasets: PRID 2011 [11], iLIDS-VID [29], and CAVIAR4REID [5].

PRID 2011: The PRID 2011 dataset consists of image sequences for 200 people in two non-overlapping camera views. The images were captured in an uncrowded outdoor environment with significant viewpoint and illumination variations.

iLIDS-VID: The iLIDS-VID dataset consists of image sequences for 300 people in two non-overlapping camera views. The images were captured at a crowded airport arrival hall with significant background clutter, occlusions, and viewpoint and illumination variations.

CAVIAR4REID: This dataset consists of multiple images for 72 people in two non-overlapping camera views, of which 50 people appear in both views. The images were captured at a shopping center with significant occlusions and viewpoint and illumination variations.

4.2. Feature extraction

Let $\mathbf{I}_i, i = 1, 2, \cdots, n$ represent the n available images for the person with index i. We divide each $\mathbf{I}_i$ into 6 horizontal stripes following Gray and Tao [10]. In each stripe, we compute texture and color histograms. Specifically, we compute responses of 13 Schmid and 6 Gabor filters to form the 16 bin texture histogram. We then compute 16-bin histograms in the YCbCr, HSV, and whitened RGB spaces to form the color descriptor. We concatenate all the histograms to form the 2592-dimensional descriptor $\mathbf{x}_i$ for each image $\mathbf{I}_i$.

Given the descriptors $\mathbf{x}_i \in \mathbb{R}^d$, $i = 1, 2, \cdots, n$ for the n available images for each person, we transform these features into a new space with a transformation matrix $\mathbf{T}$ learned using Local Fisher Discriminant Analysis (LFDA) [28]. Typically, the image sequence for each person displays multi-modality due to occlusions and background and illumination variations across all the images. Therefore, LFDA is particularly suitable in this case as it attempts to preserve the local structure of the data during the embedding process. If $\hat{\mathbf{x}}_i \in \mathbb{R}^{d'}$, $i = 1, 2, \cdots, n$, represents the feature set for a particular person in the transformed space, we compute the mean of these feature vectors to form the single representative feature vector $\mathbf{f} \in \mathbb{R}^{d'}$ for that person.

4.3. Evaluation protocol and implementation details

We randomly split the image sequences in each test dataset into equal-sized training and testing sets. For a fair comparative evaluation, following [29], for the PRID 2011 dataset, we use image sequences from 178 people containing more than 21 frames. We use the sequences in the training set to learn the feature transformation matrix $\mathbf{T}$ and the dictionary $\mathbf{D}$. Using $\mathbf{T}$, we project the test set to the embedding space, and compute the re-identification performance. We repeat this process for 10 such train-test splits and report the overall average performance.

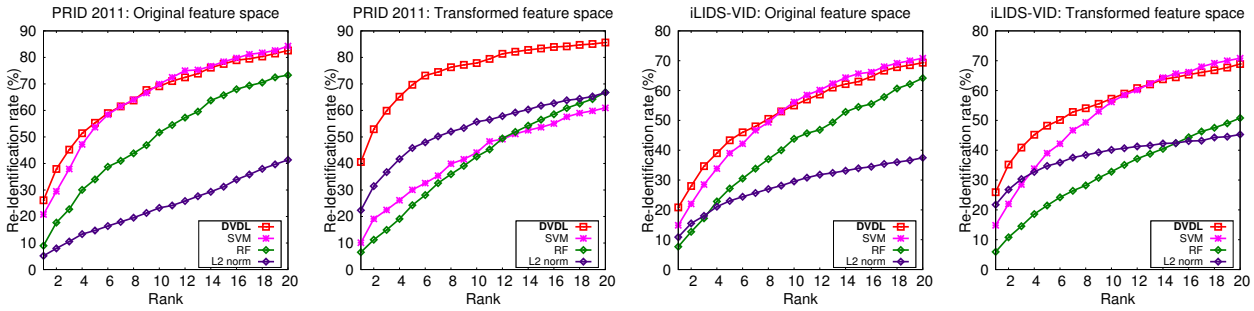

Figure 3: Evaluating the impact of the learned dictionary in comparison with baseline distance metrics on the PRID 2011 and iLIDS-VID datasets.

We set the dimension of the transformed space $d' = 150$ for the PRID 2011 dataset and $d' = 250$ for the iLIDS-VID dataset using cross-validation on the training set. We set the regularization parameters $\lambda_1 = 0.001$ and $\lambda_2 = 0.001$. The number of iterations in the procedure described in Section 3.4 was set to 5. The constants c_1 and c_2 were set to 0.1 and 100 respectively.

We first evaluate the learned dictionary by comparing its performance with distance metrics learned using standard classifiers. We then compare the results of our algorithm with several techniques in three key areas that are relevant to our work in the person re-id problem: discriminative dictionary learning, temporal sequence matching, and spatial appearance feature representation and distance metric learning. We chose techniques that had open-source code and gave state-of-the-art performance. We abbreviate our algorithm as **DVDL**.

4.4. Evaluating the learned dictionary

To evaluate the impact of the learned dictionary as a distance metric, we also learned baseline distance metric functions using the SVM [24] and Random Forests (RF) [16] classifiers. We performed two sets of experiments, first with features in the original texture and color space, and then with features in the transformed LFDA space. We also compare our results with the baseline L2 norm in both the original feature space as well as the transformed feature space. The results are illustrated in the cumulative match characteristic (CMC) curves in Figure 3 for both experiments.

We note from the plots that while using L2 norm as the distance function in the LFDA space gives better results when compared to that in the original feature space, our learned dictionary outperforms LFDA even in the original feature space. We further note that our approach results in a rank-1 performance improvement of 5.3% and 18.2% in the original and transformed feature spaces respectively for the PRID 2011 dataset. The corresponding improvements for the iLIDS-VID dataset are 6% and 4.2%. These results clearly validate the impact of our dictionary as a distance function regardless of the features used.

4.5. Comparison with discriminative dictionary learning techniques

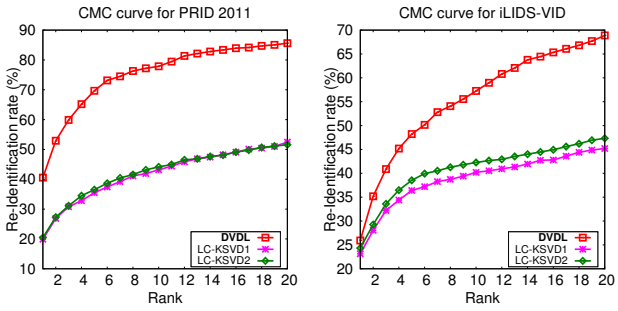

Figure 4: The cumulative match characteristic curves for PRID 2011 and iLIDS-VID in comparison with the state of the art in discriminative dictionary learning techniques.

We evaluate the performance of the label consistent K-SVD based dictionary learning algorithm proposed in [13]. We consider both LC-KSVD1 and LC-KSVD2. The results are shown in the CMC curves in Figure 4 and summarized in Table 1. We can see from the results that our dictionary learning algorithm results in significantly better performance on both the PRID 2011 and the iLIDS-VID datasets.

4.6. Comparison with temporal sequence matching techniques

Following the evaluation in [29], we use dynamic time warping to compute the similarity between two sequences. To describe the appearance in every frame of the sequence, we use both Color and Local Binary Pattern (LBP) [12] features, and the HoGHoF [14] features. The Color and LBP features encode color and texture information, whereas the HoGHoF features encode motion and texture information. We also compared our approach with the DVR model [29],

Table 1: Comparison with the state of the art in discriminative dictionary learning: 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
LC-KSVD1 [13]	19.9	35.5	43.1	52.5	23.1	36.4	40.2	45.2
LC-KSVD2 [13]	20.5	36.5	44.2	51.6	24.3	38.5	42.3	47.3
DVDL	**40.6**	**69.7**	**77.8**	**85.6**	**25.9**	**48.2**	**57.3**	**68.9**

Table 2: Comparison with the state of the art in temporal sequence matching: 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
Color & LBP [12] + DTW [25]	14.6	33	42.6	47.8	9.3	21.7	29.5	43
HoGHoF [14] + DTW [25]	17.2	37.2	47.4	60	5.3	16.1	29.7	44.7
DVR [29]	28.9	55.3	65.5	82.8	23.3	42.4	55.3	68.4
DVDL	**40.6**	**69.7**	**77.8**	**85.6**	**25.9**	**48.2**	**57.3**	**68.9**

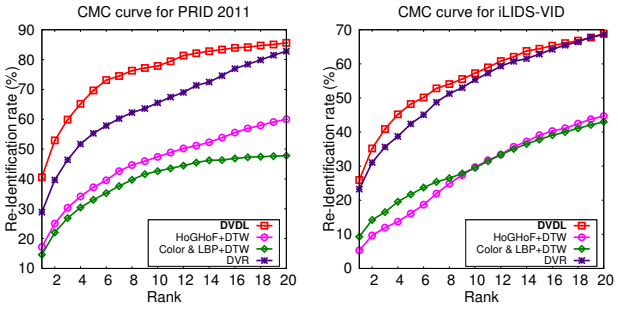

Figure 5: The cumulative match characteristic curves for PRID 2011 and iLIDS-VID in comparison with the state of the art in temporal sequence matching techniques.

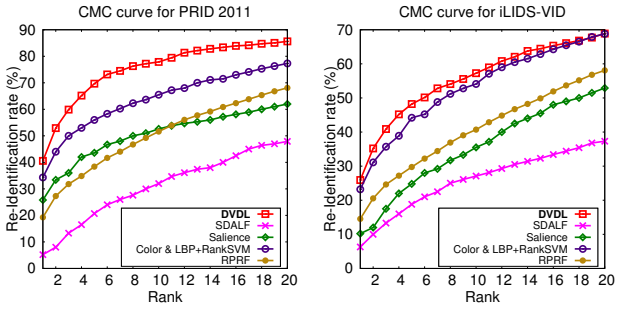

Figure 7: The cumulative match characteristic curves for PRID 2011 and iLIDS-VID in comparison with the state of the art in spatial feature representation techniques.

which selects and ranks fragments from image sequences. The results are shown in the CMC curves in Figure 5 and summarized in Table 2. Clearly, as can be seen from the graphs and the table, our approach results in consistently better performance when compared to the other three approaches. Specifically, the rank-1 performance of our algorithm is 40.6% and 25.9% on PRID 2011 and iLIDS-VID respectively, whereas the corresponding numbers for the next best performing approach are 28.9% and 23.3%.

4.7. Comparison with spatial feature representation and metric learning techniques

We also compared our performance against several spatial feature representation and metric learning methods. SDALF [6] computes HSV histograms, maximally stable color regions (MSCR) and recurrent highly structured patches (RHSP) in several local patches. Salience [35] uses color histograms and SIFT computed in dense local patches and learns salience to match persons. RPRF [16] uses global color and texture histograms and learns a distance metric function using random forests to match feature vectors. We also evaluate the performance of Color and LBP features. Following [29], we average the Color and LBP features of each frame in a sequence and use rankSVM [4] as the distance metric to compute the re-id performance. The results obtained for these methods are shown in the CMC curves in Figure 7 and summarized in Table 3. Clearly, as can be seen from the graphs and the table, our approach results in consistently better performance when compared to the other approaches. Specifically, the

Table 3: Comparison with the state of the art in spatial feature representation: 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 [35]	25.8	43.6	52.6	62	10.2	24.8	35.5	52.9
RPRF [16]	19.3	38.4	51.6	68.1	14.5	29.8	40.7	58.1
Color & LBP [12] + RankSVM [4]	34.3	56	65.5	77.3	23.2	44.2	54.1	68.8
DVDL	**40.6**	**69.7**	**77.8**	**85.6**	**25.9**	**48.2**	**57.3**	**68.9**

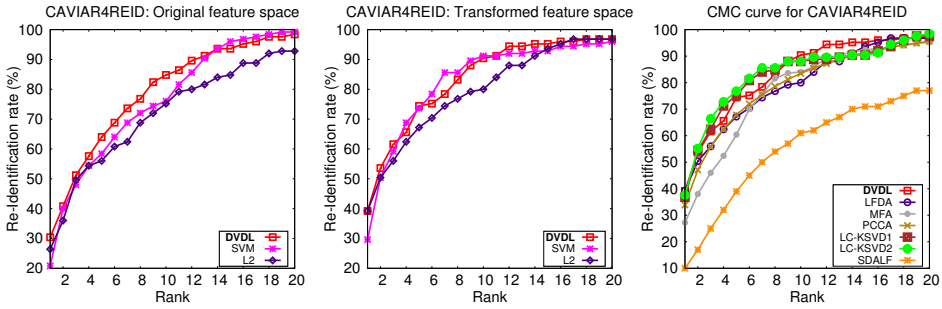

Figure 6: The cumulative match characteristic curves for the CAVIAR4REID dataset.

rank-1 performance of our algorithm is 40.6% and 25.9% on PRID 2011 and iLIDS-VID respectively, whereas the corresponding numbers for the next best performing approach are 34.3% and 23.2%.

4.8. CAVIAR4REID

In this section, we report results on the CAVIAR4REID dataset. Here, we set $d' = 500$ and split the available data into equal-sized training and testing sets. We first evaluate the impact of the learned dictionary by comparing its performance with baseline distance functions. The results are shown in Figures 6a and 6b, from which it is evident that our dictionary generally performs better than the distance function based on Euclidean norm and that learned using the SVM classifier. Finally, we compare the performance of our approach with existing metric learning techniques [31]: MFA, LFDA, and PCCA. Here, we also consider SDALF, LC-KSVD1, and LC-KSVD2. The results are shown in Figure 6c, from which we note that our approach provides similar performance at lower ranks (1-5) and better performance at higher ranks (10 and later).

5. Conclusions and Future Work

We presented an effective approach to solve the person re-identification problem in non-overlapping cameras with multiple shots. We posed the problem of re-identifying a particular person in a probe camera as finding the person in the gallery camera that has the closest sparse code with respect to a learned dictionary $\mathbf{D}$ in the Euclidean sense. We trained the dictionary to simultaneously encode images from both the gallery and the probe cameras. Furthermore, we presented a method to discriminatively train the dictionary by imposing explicit constraints on the gallery and probe sparse codes. We evaluated our algorithm on three publicly available multi-shot re-id datasets, performed extensive comparisons against several contemporary methods, and demonstrated its advantages.

We note that there is still much work to be done in making both our training and testing process computationally efficient. In our dictionary training process, we use the CVX package for MATLAB to solve the optimization problems. While CVX is easy to use and is fairly efficient for small-scale problems, it does not scale well as the amount of training data becomes large. To this end, our next line of work will involve developing specialized and efficient algorithms to solve our associated optimization problems.

References

[1] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE T-SP*, 54(11):4311–4322, 2006.

[2] S. Bak, G. Charpiat, E. Corvee, F. Bremond, and M. Thon-

nat. Learning to match appearances by correlations in a covariance metric space. In *ECCV*. 2012.

[3] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

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

[5] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In *BMVC*, 2011.

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

[7] M. Grant and S. Boyd. Graph implementations for non-smooth convex programs. In V. Blondel, S. Boyd, and H. Kimura, editors, *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag Limited, 2008.

[8] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 2.1, Mar. 2014.

[9] M. Grant, S. Boyd, and Y. Ye. Disciplined convex programming. In L. Liberti and N. Maculan, editors, *Global Optimization: From Theory to Implementation*. Springer, 2006.

[10] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*. 2008.

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

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

[13] Z. Jiang, Z. Lin, and L. S. Davis. Label consistent K-SVD: Learning a discriminative dictionary for recognition. *IEEE T-PAMI*, 35(11):2651–2664, 2013.

[14] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*. IEEE, 2008.

[15] Y. Li, Z. Wu, S. Karanam, and R. J. Radke. Real-world re-identification in an airport camera network. In *ICDSC*, 2014.

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

[17] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith. Learning locally-adaptive decision functions for person verification. In *CVPR*. IEEE, 2013.

[18] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu. Semi-supervised coupled dictionary learning for person re-identification. In *CVPR*, 2014.

[19] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[20] J. Mairal, F. Bach, and J. Ponce. Task-driven dictionary learning. *IEEE T-PAMI*, 34(4):791–804, 2012.

[21] J. Mairal, J. Ponce, G. Sapiro, A. Zisserman, and F. R. Bach. Supervised dictionary learning. In *NIPS*, 2009.

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

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

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

[25] L. R. Rabiner and B.-H. Juang. *Fundamentals of speech recognition*, volume 14. PTR Prentice Hall Englewood Cliffs, 1993.

[26] S. Shekhar, V. M. Patel, H. V. Nguyen, and R. Chellappa. Generalized domain-adaptive dictionaries. In *CVPR*, 2013.

[27] 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*, 2012.

[28] M. Sugiyama. Local fisher discriminant analysis for supervised dimensionality reduction. In *ICML*, 2006.

[29] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by video ranking. In *ECCV*. 2014.

[30] 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*, 37(5):1095 – 1108, 2015.

[31] F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In *ECCV*. Springer, 2014.

[32] M. Yang, D. Zhang, and X. Feng. Fisher discrimination dictionary learning for sparse representation. In *ICCV*, 2011.

[33] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li. Salient color names for person re-identification. In *ECCV*. Springer, 2014.

[34] Q. Zhang and B. Li. Discriminative K-SVD for dictionary learning in face recognition. In *CVPR*, 2010.

[35] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013.

[36] R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identification. In *CVPR*, 2014.

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