Multi-Shot Human Re-Identification Using Adaptive Fisher Discriminant Analysis

Yang Li
yangli625@gmail.com
Ziyan Wu
ziyan@alum.rpi.edu
Srikrishna Karanam
karans3@rpi.edu
Richard J. Radke
rjradke@ecse.rpi.edu

Department of Electrical, Computer, and Systems Engineering
Rensselaer Polytechnic Institute
110 8th St.
Troy, NY USA

Abstract

While much research in human re-identification has focused on the single-shot case, in real-world applications we are likely to have an image sequence from both the person to be matched and each candidate in the gallery, extracted from automated video tracking. It is desirable to take advantage of the multiple visual aspects (states) of each subject observed during training and testing. However, since each subject may spend different amounts of time in each state, equally weighting all the images in a sequence is likely to produce suboptimal performance. To address this problem, we introduce an algorithm to hierarchically cluster image sequences and use the representative data samples to learn a feature subspace maximizing the Fisher criterion. The clustering and subspace learning processes are applied iteratively to obtain diversity-preserving discriminative features. A metric learning step is then applied to bridge the appearance difference between two cameras. The proposed method is evaluated on three multi-shot re-id datasets and the results outperform state-of-the-art methods.

1 Introduction

Manually monitoring and analyzing videos from many cameras in a surveillance installation is infeasible, since it is expensive and inaccurate. Consequently, automated human re-identification (re-id) has attracted growing interest in the computer vision community. The problem is: given a person (the "probe") in one view, identify the same person from a gallery of candidates in other non-overlapping views upon his/her reappearance. This is an inherently challenging problem, because the target's appearance across cameras can be significantly different based on viewpoint changes, illumination variation, and occlusions. Currently, the major research effort for this problem is focused on the single-shot scenario: that is, each person has only one image available per camera view. For single-shot re-id, researchers have extensively studied the construction of descriptive and discriminative appearance models [1, 6, 7, 12, 20, 29, 31] as well as metric learning techniques [6, 13, 15, 21, 24, 25, 33].

(a) (b)

Figure 1: Image samples from multi-shot re-id datasets. (a) The iLIDS-VID image sequence dataset [28]; each person has an image sequence per view. (b) The iLIDS multi-shot dataset [32]; each person has several images from multiple views.

In spite of the great research progress achieved for the single-shot case, real-world re-id performance is hindered by the limited information extracted from single images. From a practical perspective, the re-id task originates from video analysis applications [16], which means there are multiple frames available for the individuals in each field of view. Thus, it is natural to use such information to improve re-id performance.

In this paper, we focus on a scenario directly connected to a real-world video analysis problem. Suppose there are two camera views; a target individual is identified in the probe view, then he or she is tracked until leaving the current view. In the gallery view, candidates are generated via pedestrian detection followed by tracking. Thus, the input of re-id is actually a set of consecutive images or image sequences, rather than simply one shot per person from each view, as shown in Figure 1(a). This is different from the broader use of the term "multi-shot", where instead of a sequence of consecutive images, there are just a few random shots from multiple views, as shown in Figure 1(b). We only consider two camera views here, but our approach can be easily extended to multiple views.

To re-identify a person, we want to learn a feature space where images belonging to the same person stay close while images belonging to different people are far apart, which can be achieved by Fisher Discriminant Analysis (FDA) [8]. However in the case of multi-shot re-identification, the samples in the training set and the testing set may not be drawn from the same distributions, which may severely affect the discriminant analysis. For example, we intuitively want to capture as many different states (i.e., visual aspects) of a person as possible. However, it may be that in the sequence of a person in the training set, a majority of frames of the person are in a single state, which may cause a biased result in the FDA. That is, features reflecting other important states will be ignored. Hence, it is necessary to select representative samples which can cover the diversity of the person by clustering before conducting FDA. On the other hand, the performance of clustering depends on the feature space. Only important and discriminative features should be used in clustering. Either conducting clustering before or after FDA, the results will be suboptimal.

In this paper, we propose an Adaptive Fisher Discriminant Analysis (AFDA) algorithm to mitigate this issue. Local Fisher Discriminant Analysis (LFDA) [27] is adapted to maximize inter-class distance and minimize intra-class difference, while preserving local structures. A Fisher Guided Hierarchical Clustering (FGHC) algorithm is integrated with LFDA to select representative samples from each class and maintain diversity based on the Fisher criterion. By iteratively updating the representative samples and the discriminative feature space, a

diversity-preserving subspace can be obtained. Since LFDA preserves local structures, we further learn a distance function that compensates for the difference caused by viewpoint variation.

2 Related Work

Several researchers have proposed to leverage multi-shot information for appearance modeling in re-id, though not in the sense we describe here. Gheissari *et al*. [11] introduced a spatio-temporal segmentation algorithm to generate stable edgel structures defining appearance invariant regions. Bedagkar-Gala and Shah proposed another spatio-temporal method [3] to build color cluster correspondence across consecutive frames; the matched clusters are further grouped temporally into representative meta-colors. These clusters are extracted from each body part to form a spatio-temporal model. Farenzena *et al*. [7] formulated a part-based appearance model that consists of overall chromatic information, regional color arrangement and recurrent structures. Multi-shot is used in the sense of accumulating features or selecting the most similar frame as a representative. Bak *et al*. [2] proposed a patch-based model where a feature covariance matrix is extracted from each image patch and averaged over consecutive frames. Cheng *et al*. [6] extracted body parts based on pictorial structures. Using multiple images, each localized pixel is fit to a Gaussian distribution and iteratively improves the body part model. Wang *et al*. [28] selected discriminative fragments from each image sequence and then extracted space-time based features, which are used to learn a multi-instance ranking model to perform re-id. Multi-shot information is also applied in gait analysis [18, 19]. While it is able to generalize a walking model for pedestrians, this technique might not be suitable for surveillance videos, since there usually exist occlusions and the resolution and frame rate is low.

Recently, several metric learning techniques have also been presented that directly consider multi-shot information. Pedagadi *et al*. [24] applied Local Fisher Discriminant Analysis (LFDA) [27] to project the data points onto a lower dimensional embedding space based on the Fisher criterion while preserving data locality. It handles single-shot and multi-shot cases in the same way; the latter has slightly more sample points available during the learning process. Zhang *et al*. [30] considered multiple image sample points of a person to follow a multimodal distribution, and formulated a loss function to describe the distance between feature point clusters, defined via k-nearest neighbors. García *et al*. [10] learned pairwise feature dissimilarity spaces (PFDS) based on the viewpoint similarity between two people. Li *et al*. [17] proposed an ensemble of random forest classifiers for re-id, where the multi-shot information is used to customize the trained random forests for a specific target.

In the rest of this paper, we first present algorithm details in Section 3. Then we present extensive experimental results in Section 4 to analyze the proposed algorithm and compare it with state-of-the-art techniques.

3 Proposed Algorithm

3.1 Feature Extraction

We adopt the ensemble of localized features proposed by Gray and Tao [12] as the image descriptor. Specifically, the bounding box around a human in a given frame is evenly divided into six horizontal strips, inside of which color and texture histograms are extracted.

We employed RGB, HSV and YCbCr channels for color histograms. Because of the varying illumination conditions, color may render differently across camera views. We perform a simple color normalization within the bounding box in the RGB channel: $(R',G',B') = \left(\frac{R-mean(R)}{std(R)}, \frac{G-mean(G)}{std(G)}, \frac{B-mean(B)}{std(B)}\right)$.

For texture features, histograms are extracted from the response of two filter families, Schmid [26] and Gabor [9] filters, which describe rotational and horizontal/vertical textures respectively. We applied 13 Schmid filters and 8 Gabor fiters. The histograms are computed and normalized within each image strip for all the color and texture channels, then concatenated into one feature vector. In all the experiments, we employ 16 bins for each channel's histogram. The total number of features is 2592.

3.2 Local Fisher Discriminant Analysis

The feature vectors employed in the re-id problem are usually high-dimensional, leading to sparse sample points in the feature space. Thus it is necessary to find an intrinsic low-dimensional space, such that samples from the same person stay close to each other, while remaining far apart from those belonging to different people. This coincides with the Fisher criterion [8]:

$$J = \mathrm{Tr}\left((\boldsymbol{T}^\top \boldsymbol{S}^w \boldsymbol{T})^{-1} \boldsymbol{T}^\top \boldsymbol{S}^b \boldsymbol{T}\right) \tag{1}$$

where $\boldsymbol{T}$ is the linear tranformation matrix that projects data samples onto a low-dimensional subspace. $\boldsymbol{S}^w$ and $\boldsymbol{S}^b$ are the within-class scatter matrix and between-class scatter matrix respectively.

Maximizing the above equation to find the matrix $\boldsymbol{T}$ is known as Fisher Discriminant Analysis (FDA) [8], and it works well for unimodal data. However, in the re-id problem, the goal is to match people based on images captured from different cameras. Because of viewpoint and environmental variations, the appearance of the same person from these two cameras will be inherently different. Thus, the data samples in the feature space naturally split into two clusters, which means the data from each person is multi-modal. In such case, if the FDA technique is applied, it will try to merge the clusters for each person. This is undesirable since it may make separating different classes more difficult. We applied Local Fisher Discriminant Analysis (LFDA) [27] to mitigate this issue.

LFDA combines the idea of Fisher Discriminant Analysis (FDA) [8] and Locality Preserving Projections (LPP) [22] to satisfy the Fisher criterion while preserving local structures. The within-class scatter matrix and the between-class scatter matrix in Equation 1 can be expressed as

$$\boldsymbol{S}^w = \frac{1}{2}\sum_{i,j=1}^{N} \boldsymbol{W}^w_{i,j}(x_i - x_j)(x_i - x_j)^\top, \tag{2}$$

$$\boldsymbol{S}^b = \frac{1}{2}\sum_{i,j=1}^{N} \boldsymbol{W}^b_{i,j}(x_i - x_j)(x_i - x_j)^\top, \tag{3}$$

where

$$\boldsymbol{W}^w_{i,j} = \begin{cases} \boldsymbol{A}_{i,j}/n_c & \text{if } y_i = y_j = c, \\ 0 & \text{if } y_i \neq y_j, \end{cases} \tag{4}$$

$$\boldsymbol{W}^b_{i,j} = \begin{cases} \boldsymbol{A}_{i,j}(1/N - 1/n_c) & \text{if } y_i = y_j = c, \\ 1/N & \text{if } y_i \neq y_j. \end{cases} \tag{5}$$

For FDA, the weighting $\boldsymbol{A}_{i,j} \in \{0,1\}$, i.e., if two sample points $\{x_i, x_j\}$ are from the same class c, their contribution to the scatter matrices takes a constant weight. To preserve the local structure of the data, LFDA imposes the affinity matrix $\boldsymbol{A}$ defined in [22] to the matrices $\boldsymbol{W}^w$ and $\boldsymbol{W}^b$, where $\boldsymbol{A}_{ij} \in [0,1]$ is a continuous-valued affinity between x_i and x_j. A larger value indicates a higher similarity.

The LFDA transformation matrix $\boldsymbol{T}$ can be obtained by maximizing Equation 1. The solution can be obtained by solving the generalized eigenvalue problem:

$$\boldsymbol{S}^b \varphi = \lambda \boldsymbol{S}^w \varphi, \tag{6}$$

where $\{\varphi_i\}_{i=1}^d$ are the eigenvectors and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ are the associated eigenvalues. Then the transformation matrix $\boldsymbol{T} = [\varphi_1 | \varphi_2 | \ldots | \varphi_m]$.

3.3 Fisher Guided Hierarchical Clustering

We now take a closer look at using image sequences. At typical video frame rates, adjacent images exhibit few differences, but across the whole sequence there are several distinctive states. Each state can be represented by a cluster of images. An example is shown in Figure 2. A subject is tracked in one of the iLIDS [23] surveillance video datasets and the extracted images are clustered into five groups. In such cases, if all the data samples of a person are treated equally, our representation will be highly biased, because there is redundant information unevenly distributed across different states of the person. For some important states (e.g., specific poses or lighting conditions), it is possible that only a few samples are available. In these cases, with the objective function in Equation 1, important features may be ignored.

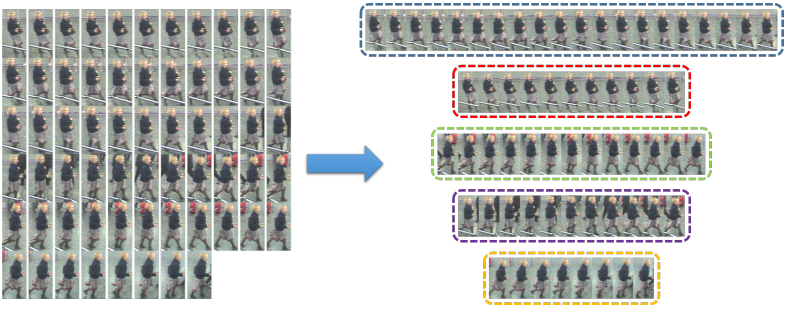

Figure 2: A tracking image sequence is clustered into 5 groups.

Thus, using all the data samples to find a transformation matrix as discussed in Section 3.2 will lead to suboptimal results. Instead, a sample selection scheme is necessary to avoid biased and redundant information. We propose a novel algorithm to select representative samples via hierarchical clustering based on the Fisher criterion function, and then use the selected data to train the Fisher discriminant transformation matrix. By iteratively conducting clustering and LFDA, better performance can be achieved. The details are presented in Algorithm 1.

First, all the data samples are used in LFDA to initialize the transformation matrix $\boldsymbol{T}$. Then, we project all the data onto a low-dimensional subspace via the current $\boldsymbol{T}$, where a hierarchical clustering is performed for all the data points of each person per camera. The basic idea is that data points extracted from one image sequence initially form one cluster,

Algorithm 1: Adaptive Fisher Discriminative Analysis (AFDA)

Input: $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_P]$ where $\boldsymbol{X}_i$ contains data samples of person i
$\boldsymbol{Y} = [\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_P]$ where $\boldsymbol{Y}_i$ contains class labels corresponding to data samples of person i
Output: $\boldsymbol{T}$: transformation matrix
$\boldsymbol{T}_{new}$ = LFDA($\boldsymbol{X}, \boldsymbol{Y}$);
Initialize $J_{new} = 1$, $J = 0$, $iter = 0$;
while *($J_{new} > J$) and ($iter < MAX\_ITERATION$)* **do**
  $J = J_{new}$, $\boldsymbol{T} = \boldsymbol{T}_{new}$;
  foreach *person i* **do**
    Initialize cluster labels $\boldsymbol{C}_i = \{\boldsymbol{C}_i^1, \boldsymbol{C}_i^2, \cdots \boldsymbol{C}_i^n\}$ of data samples based on n camera views;
    foreach *cluster j in $\boldsymbol{C}_i$* **do**
      $\boldsymbol{C}_i'^j$ = FGHC($\boldsymbol{X}_i^{c_i^j}$, $\boldsymbol{C}_i^j$, c_i^j, $\boldsymbol{T}$), where $\boldsymbol{X}_i^{c_i^j}$ contains samples from $\boldsymbol{X}_i$ with label c_i^j;
      Update $\boldsymbol{C}_i^j$ in $\boldsymbol{C}_i$ with $\boldsymbol{C}_i'^j$;
    end
    foreach *label m in unique($\boldsymbol{C}_i$)* **do**
      $\bar{\boldsymbol{x}}_i^m = \frac{1}{n_m} \sum_{\boldsymbol{x}_i^k \in \boldsymbol{X}_i^m} \boldsymbol{x}_i^k$, where $n_m = \|\boldsymbol{X}_i^m\|$;
      $\bar{y}_i^m = i$;
    end
    $\boldsymbol{X}_i = \{\bar{\boldsymbol{x}}_i^1, \bar{\boldsymbol{x}}_i^2, \cdots, \bar{\boldsymbol{x}}_i^{N_i}\}$, $\boldsymbol{Y}_i = \{\bar{y}_i^1, \bar{y}_i^2, \cdots, \bar{y}_i^{N_i}\}$, where $N_i = \|unique(\boldsymbol{C}_i)\|$;
  end
  $\boldsymbol{T}_{new}$ = LFDA($\boldsymbol{X}, \boldsymbol{Y}$);
  Calculate J_{new} with Equation 1 using $\boldsymbol{T}_{new}$;
  $iter = iter + 1$;
end
Return $\boldsymbol{T}$;

which is hierarchically subclustered based on the Fisher criterion. These data points will then be replaced by the sample mean of the generated clusters, as a new representation of the current image sequence. The detailed clustering algorithm is presented in Algorithm 2 and will be discussed shortly. After processing all the training subjects, a new transformation matrix $\boldsymbol{T}_{new}$ can be learned using LFDA again with the updated data points. We compare the Fisher criterion computed within the new feature subspace by $\boldsymbol{T}_{new}$ with that of $\boldsymbol{T}$. If it is increased, we update $\boldsymbol{T}$ to $\boldsymbol{T}_{new}$ and repeat the clustering process in the new subspace; otherwise we return the matrix $\boldsymbol{T}$ as the final transformation matrix.

In Algorithm 1, the clustering results are updated within every iteration hierarchically. Here we again apply the Fisher criterion to guide the clustering process, shown in Algorithm 2. Given the data samples of an image sequence, we first calculate the Fisher criterion J for the current clustering scheme. Then, the k-means ($k = 2$) clustering algorithm is performed on each existing cluster. Every split is examined by comparing the new Fisher criterion value J_{new} with J, and will be accepted only if it gives better clustering results of the data points. The process is repeated for each cluster until there is no further splits. To avoid over-clustering the data, we set a upper bound for the number of clusters. This clustering scheme has several advantages. First, it is relatively simple and efficient. Second, it can automatically find a suitable number of clusters. Third, this method is also based on the Fisher criterion, which can lead to more representative data points for the LFDA algorithm.

In the proposed algorithm, the feature space used in the intra-class clustering and the sample distribution of each class used in subspace learning are updated iteratively and collaboratively, ensuring local structure and diversity are preserved while selecting discriminative features between different people. We also note that there are two applications of the Fisher criterion in the proposed method: one is applied to the data samples within each class locally, in order to obtain the most representative data points for each class; the other is applied to all the data globally, for determining a discriminant feature subspace. By alternating these two

Algorithm 2: Fisher Guided Hierarchical Clustering (FGHC)

Input: $\boldsymbol{X}$: samples
$\boldsymbol{C}$: cluster label of the samples
c: the label whose corresponding samples will be clustered
$\mathbf{T}$: projection matrix
Output: $\boldsymbol{C'}$: updated cluster labels
Initialize $\boldsymbol{C'} = \boldsymbol{C}$;
Calculate Fisher criterion J with Equation 1 using $\boldsymbol{T}$;
K-means clustering on $\boldsymbol{X_c}$ with $K = 2$, where $\boldsymbol{X_c}$ are samples with the label c, generating new labels c_1 and c_2;
Update labels in $\boldsymbol{C'}$ with c_1 and c_2;
Calculate Fisher criterion J_{new} with Equation 1 using $\boldsymbol{T}$ and $\boldsymbol{C'}$;
if $(J_{new} > J)$ *&* *(# of clusters in* $\boldsymbol{C'} \leq$*MAX_CLUSTER_NUM)* *&* $(\boldsymbol{C'} \neq \boldsymbol{C})$ **then**
  $\boldsymbol{C'_1} = FGHC(\boldsymbol{X}, \boldsymbol{C'}, c_1, \boldsymbol{T})$;
  $\boldsymbol{C'_2} = FGHC(\boldsymbol{X}, \boldsymbol{C'}, c_2, \boldsymbol{T})$;
  Update cluster labels in $\boldsymbol{C'}$ with $\boldsymbol{C'_1}$ and $\boldsymbol{C'_2}$;
else
  Return $\boldsymbol{C}$;
end
Return $\boldsymbol{C'}$;

processes, a better feature subspace can be learned via LFDA fitting the representative data points and the results can be iteratively improved.

3.4 Metric Learning

As discussed in Section 3.2, the feature vectors extracted from the same person will display multi-modalaliy because of the different camera views, and LFDA will preserve such structure. Thus, a metric learning step is necessary to further compensate for the difference between the two cameras.

Suppose the person p captured in camera a has a feature vector set $\boldsymbol{X}_p^a = \{x_1^a, x_2^a, \ldots, x_{n_p}^a\}$ where $x_i^a \in \mathbb{R}^d$. These data samples are clustered using Algorithm 2. We then take the set of sample means of each cluster as the representative points for person p in camera a; that is $\bar{\boldsymbol{X}}_p^a = \{\bar{x}_1^a, \bar{x}_2^a, \ldots, \bar{x}_{n'_p}^a\}$, where n'_p is the number of clusters. Now we can use any metric learning technique to learn the difference between the two cameras. In our experiment, we employed the RankSVM algorithm [14, 25]. The idea is to minimize the norm of a vector w that satisfies the following ranking relationship

$$w^\top (|\bar{z}_i^a - \bar{z}_i^b| - |\bar{z}_i^a - \bar{z}_j^b|) > 0, \quad i, j = 1, 2, \ldots, P \text{ and } i \neq j \tag{7}$$

where $\bar{z}_i^a$ is the sample mean of the representative feature vectors of person i in camera a, projected by the learned transformation matrix $\boldsymbol{T}$,

$$\bar{z}_p^a = \sum_i^{n'_p} \boldsymbol{T}^\top \bar{x}_i, \tag{8}$$

and P is the total number of training subjects. It should be noted here that $|\cdot|$ is an element-wise absolute difference operator. The RankSVM method finds w by solving the problem

$$\begin{aligned} &\arg\min_{w,\xi} \left(\tfrac{1}{2}||w||^2 + C \sum_{i=1}^{P} \xi_i\right) \\ &s.t.\ w^\top (|\bar{z}_i^a - \bar{z}_i^b| - |\bar{z}_i^a - \bar{z}_j^b|) \geq 1 - \xi_i, \quad \xi_i \geq 0 \end{aligned} \tag{9}$$

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

4 Experiments

In this section, we present experiments conducted on three image-sequence-based public benchmark datasets, PRID 2011 [9], iLIDS-VID [28] and SAIVT-SoftBio [9], and compare the performance of the proposed algorithm with state-of-the-art re-id algorithms [7, 13, 28, 31].

4.1 Evaluation Datasets

PRID 2011 [9] This dataset is extracted from videos of two surveillance cameras installed outside a building. It contains 200 people appearing in both camera views, and each person has one image sequence from each camera. This dataset suffers from viewpoint changes, but has relatively clear background and few occlusions.

iLIDS-VID [28] This dataset is extracted from an airport multi-camera surveillance network [23]. It includes 300 pedestrians appearing in two non-overlapping camera views. The challenges of this dataset include occlusions and viewpoint and illumination variations across different views.

SAIVT-SoftBio [9] This dataset consists of 150 people captured from eight surveillance cameras in a building environment. Each person may only appear in a subset of cameras. Objects in this dataset rarely get occluded, but the viewpoint and illumination vary substantially across cameras.

4.2 Experimental Settings

In all the experiments, we randomly split the dataset into training and testing sets. For iLIDS-VID and PRID 2011, we adopted the protocol in [28] and use the same number of people for training and testing. For SAIVT-SoftBio, we followed the protocol in [10], using a third of all people for training and the remainder for testing. During the training stage, the projection dimension size is set as 100, the maximum cluster number for each person is 10 and the maximum iteration number is 8.

In the testing stage, image sequences from one camera form the probe set while those from the other camera are used as the gallery set. As described in Section 3.4, each image sequence is clustered and the sample means of each cluster are used as the representative data points. The final single feature vector for an image sequence is obtained by taking the average of its representative data points. Each feature vector in the probe set is then compared with all the candidate vectors in the gallery set, and a ranking is produced based on the re-id result. Given this ranking for all probes, cumulative match characteristic (CMC) curves are generated to report results, where the matching rate at rank n means the percentage of probe images whose correct match appears within the top n candidates. All the experimental results are averaged over 10 random splits.

4.3 Evaluation of Algorithm Components

As discussed in Section 3, LFDA will perform poorly when applied to image sequences directly. We first conducted experiments comparing the performance of the proposed algorithm and LFDA. We employed a similar protocol to [24], where all the feature vectors in

the training set are used to find the transformation matrix and the mean of the feature vectors of each image sequence is used during testing. The iLIDS-VID dataset is used in this experiment, and results are shown in Figure 3.

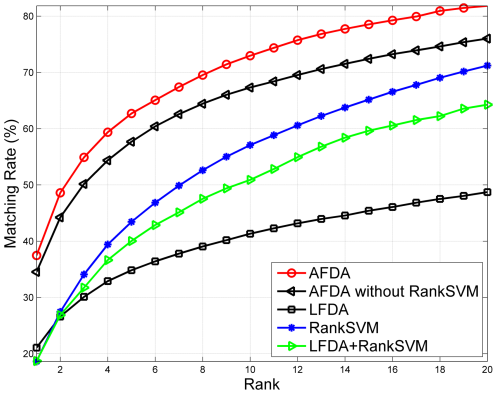

Figure 3: Evaluation on each component of the proposed algorithm on iLIDS-VID.

It can be seen that by iteratively clustering the image sequences and using the representative data samples, we were able to learn a much better subspace for re-id. The results of using RankSVM alone and LFDA followed by RankSVM are also provided in Figure 3. While RankSVM improves the baseline LFDA results, our proposed clustering approach offers a significant increase in performance. We also note that if the metric learning stage (*e.g.*, RankSVM) is ignored, the performance will be worse, but is still much better than simply applying LFDA.

4.4 Evaluation on Benchmark Datasets

We compared the proposed algorithm with state-of-the-art algorithms across three benchmark datasets. The results for the PRID and iLIDS-VID datasets are shown in Table 1. The proposed algorithm achieves the best performance. Specifically, Rank 1 for PRID is 43.0% and Rank 5 increases to 72.7%, while the best performance across the other methods are 41.7% and 64.5% respectively. Rank 1 for iLIDS-VID is 37.5% and rank 5 is 62.7%, while the second best results are 34.5% and 56.7% respectively. As mentioned earlier, the PRID dataset is comparatively easy because of relatively clear background and fewer occlusions, so that re-id algorithms generally achieve better results on this dataset.

Table 1: Experiment results comparison on 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 [4]	5.2	20.7	32.0	47.9	6.3	18.8	27.1	37.3
Salience [31]	25.8	43.6	52.6	62.0	10.2	24.8	35.5	52.9
DVR [23]	28.9	55.3	65.5	82.8	23.3	42.4	55.3	68.4
Color&LBP [31] + DVR [23]	37.6	63.9	75.3	89.4	34.5	56.7	67.5	77.5
SDALF + DVR [23]	31.6	58.0	70.3	85.3	26.7	49.3	61.0	71.6
Salience + DVR [23]	41.7	64.5	77.5	88.8	30.9	54.4	65.1	77.1
RankSVM [31]	22.4	51.9	66.8	80.7	18.6	43.3	57.1	71.2
LFDA [23]	22.3	41.7	51.6	62.0	21.1	34.8	41.3	48.7
AFDA	**43.0**	**72.7**	**84.6**	**91.9**	**37.5**	**62.7**	**73.0**	**81.8**

The results for SAIVT-SoftBio are shown in Table 2. While this dataset is not widely evaluated, we compared our algorithm with the reported results of the Fused [4] and PFDS

[11] algorithms, adopting the same experimental protocol. Two camera pairs are evaluated: cameras 3/8 which include 99 people from similar viewpoints, and cameras 5/8 which have 103 people with large view angle changes. The data for the comparison algorithms were carefully extracted from the plots in the original papers as the raw numbers were not provided. The proposed algorithm displays significant improvement in both cases. Specifically, for cameras 5/8, it achieves Rank 1 performance of 30.9% and Rank 5 of 61.6%, while for the easier case, cameras 3/8, the results of Rank 1 and 5 are 44.4% and 77.4% respectively.

Table 2: Experiment results comparison on the SAIVT-SoftBio dataset

Dataset	Cameras 3/8				Cameras 5/8			
Rank	Rank 1	Rank 5	Rank 10	Rank 20	Rank 1	Rank 5	Rank 10	Rank 20
Fused [8]	36.4	60.3	76.0	87.6	20.0	33.0	50.4	67.8
PFDS [11]	33.2	60.5	74.0	87.2	18.6	32.9	53.0	85.3
RankSVM [14]	32.4	68.4	82.0	92.9	14.9	40.5	57.9	75.0
LFDA [21]	12.2	36.8	54.6	74.9	9.3	27.1	41.2	60.6
AFDA	**44.4**	**77.4**	**89.4**	**95.9**	**30.9**	**61.6**	**77.3**	**91.1**

5 Conclusion

We presented a novel algorithm, Adaptive Fisher Discriminant Analysis, to effectively make use of image sequences in human re-id problems. Discriminant analysis and clustering are directly integrated, so that local structure and sample diversity are effectively preserved when finding discriminative feature subspaces between samples from different individuals. By iteratively updating the feature space and representative samples for each person, the algorithm can determine a better feature subspace before feeding into a metric learning method to establish the relationship between cameras. Since the proposed algorithm effectively mitigates the problem of biased sample distributions, more discriminative features can be selected and more robust classifiers can be trained. The proposed algorithm is tailored for the multi-shot scenario and shows significant improvement compared with state-of-the-art techniques. In future work we plan to further explore space-time based descriptors and approaches to describe scenes and the relationships between different views.

Acknowledgement

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] Sławomir Bak, Etienne Corvée, Francois Brémond, and Monique Thonnat. Person re-identification using Haar-based and DCD-based signature. In *AVSS*, 2010.

[2] Sławomir Bak, Etienne Corvée, Francois Brémond, and Monique Thonnat. Boosted human re-identification using Riemannian manifolds. *Image and Vision Computing*, 30(6):443–452, 2012.

[3] Apurva Bedagkar-Gala and Shishir K Shah. Part-based spatio-temporal model for multi-person re-identification. *Pattern Recognition Letters*, 33(14):1908–1915, 2012.

[4] Alina Bialkowski, Simon Denman, Patrick Lucey, Sridha Sridharan, and Clinton B Fookes. A database for person re-identification in multi-camera surveillance networks. In *DICTA*, 2012.

[5] John Blitzer, Kilian Q Weinberger, and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. In *NIPS*, 2005.

[6] Dong Seon Cheng, Marco Cristani, Michele Stoppa, Loris Bazzani, and Vittorio Murino. Custom pictorial structures for re-identification. In *BMVC*, 2011.

[7] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010.

[8] Ronald A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936.

[9] Itzhak Fogel and Dov Sagi. Gabor filters as texture discriminator. *Biological Cybernetics*, 61(2):103–113, 1989.

[10] Jorge García, Niki Martinel, Gian Luca Foresti, Alfredo Gardel, and Christian Micheloni. Person orientation and feature distances boost re-identification. In *ICPR*, 2014.

[11] Niloofar Gheissari, Thomas B Sebastian, and Richard Hartley. Person reidentification using spatiotemporal appearance. In *CVPR*, 2006.

[12] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008.

[13] Martin Hirzer, Peter M Roth, Martin Köstinger, and Horst Bischof. Relaxed pairwise learned metric for person re-identification. In *ECCV*, 2012.

[14] Thorsten Joachims. Optimizing search engines using clickthrough data. In *SIGKDD*, 2002.

[15] Srikrishna Karanam, Yang Li, and Richard 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.

[16] Yang Li, Ziyan Wu, Srikrishna Karanam, and Richard J. Radke. Real-world re-identification in an airport camera network. In *Proceedings of the International Conference on Distributed Smart Cameras*, 2014.

[17] Yang Li, Ziyan Wu, and Richard J. Radke. Multi-shot re-identification with random-projection-based random forests. In *WACV*, 2015.

[18] Zongyi Liu and Sudeep Sarkar. Improved gait recognition by gait dynamics normalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(6):863–876, 2006.

[19] Jiwen Lu, Gang Wang, and Thomas S. Huang. Gait-based gender classification in unconstrained environments. In *ICPR*, pages 3284–3287, 2012.

[20] Bingpeng Ma, Yu Su, and Frédéric Jurie. Bicov: a novel image representation for person re-identification and face verification. In *BMVC*, 2012.

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

[22] Partha Niyogi and Xiaofei He. Locality preserving projections. In *NIPS*, 2004.

[23] UK Home Office. i-LIDS multiple camera tracking scenario definition. Technical report, 2008.

[24] Sateesh Pedagadi, James Orwell, Sergio Velastin, and Boghos Boghossian. Local Fisher discriminant analysis for pedestrian re-identification. In *CVPR*, 2013.

[25] Bryan Prosser, Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Person re-identification by support vector ranking. In *BMVC*, 2010.

[26] Cordelia Schmid. Constructing models for content-based image retrieval. In *CVPR*, 2001.

[27] Masashi Sugiyama. Local Fisher discriminant analysis for supervised dimensionality reduction. In *ICML*, 2006.

[28] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *ECCV*. Springer, 2014.

[29] Ziyan Wu, Yang Li, and Richard J. Radke. Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(5):1095–1108, 2015.

[30] Guanwen Zhang, Yu Wang, Jien Kato, Takafumi Marutani, and Kenji Mase. Local distance comparison for multiple-shot people re-identification. In *ACCV*. Springer, 2013.

[31] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013.

[32] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Associating groups of people. In *BMVC*, 2009.

[33] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Person re-identification by probabilistic relative distance comparison. In *CVPR*, 2011.