# Top-push Video-based Person Re-identification

Jinjie You[†,‡], Ancong Wu[†,‡], Xiang Li[†,‡], and Wei-Shi Zheng[*†,§,‡]
[†]Intelligence Science and System Lab, Sun Yat-sen University, China
[§]School of Data and Computer Science, Sun Yat-sen University, China
[‡]Guangdong Provincial Key Laboratory of Computational Science, China
youjinjie9@gmail.com, wuancong@mail2.sysu.edu.cn
lixiang651@gmail.com, wszheng@ieee.org

## Abstract

*Most existing person re-identification (re-id) models focus on matching still person images across disjoint camera views. Since only limited information can be exploited from still images, it is hard (if not impossible) to overcome the occlusion, pose and camera-view change, and lighting variation problems. In comparison, video-based re-id methods can utilize extra space-time information, which contains much more rich cues for matching to overcome the mentioned problems. However, we find that when using video-based representation, some inter-class difference can be much more obscure than the one when using still-image-based representation, because different people could not only have similar appearance but also have similar motions and actions which are hard to align. To solve this problem, we propose a top-push distance learning model (TDL), in which we integrate a top-push constrain for matching video features of persons. The top-push constraint enforces the optimization on top-rank matching in re-id, so as to make the matching model more effective towards selecting more discriminative features to distinguish different persons. Our experiments show that the proposed video-based re-id framework outperforms the state-of-the-art video-based re-id methods.*

## 1. Introduction

Person re-identification (re-id) matches persons across non-overlapping camera views at different time. Most existing works focus on matching still images represented by appearance features (*e.g.* color histograms), because of computation efficiency and limited storage space. Given a probe image, we match it against a set of gallery images, which may suffer from illumination change, viewpoint difference, complicated background and occlusions.

*Corresponding author

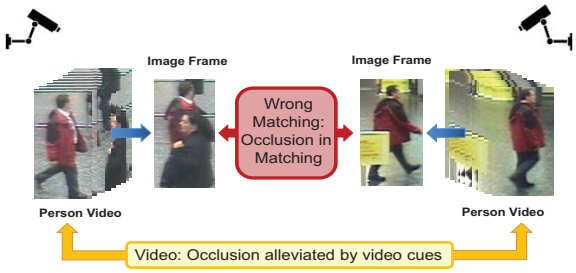
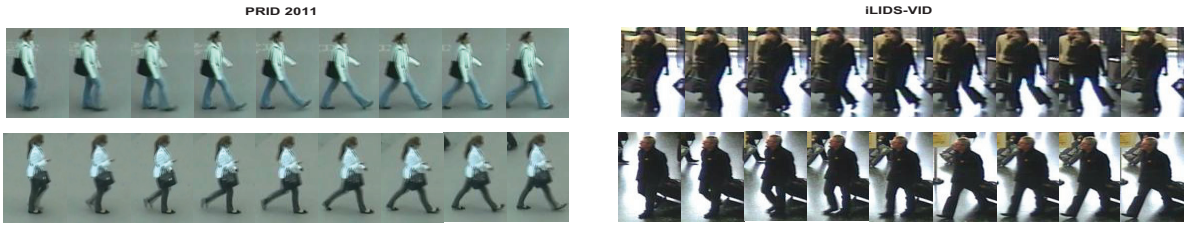



Figure 1. Video instances vs. still-image instances of the same person. It is clear that video contains much richer cues for matching.

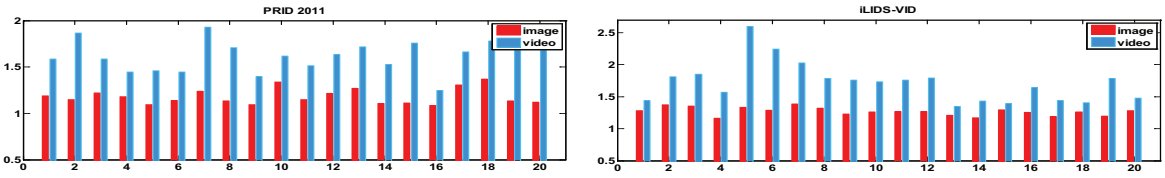
The significant visual ambiguity and appearance variation make still-image-based person re-id a challenging problem. Many methods have been developed to either extract invariant features or learn discriminative matching models [6, 32, 7, 4, 25, 43, 14, 29, 40, 27, 13, 38, 20, 28, 16, 39, 24, 26, 37, 17, 1, 21, 33, 3].

However, the still-image-based person re-id indicates that the temporal information between images of a person in each camera view is ignored. Information of a still image is sometimes not enough for recognizing a person, *e.g.* the person being occluded by objects or other persons (see Figure 1 for example). As surveillance information is recorded by videos and human operators always recognize persons in videos, it is intuitive to mine more effective information in video re-id. What more information can we obtain from videos than still images? Firstly, video is an image sequence containing space-time information, in which motion information is available. Secondly, appearance cues are more abundant in a sequence than in a still image, which can facilitate extracting more robust appearance features. Thirdly, occlusion and background influence can be eliminated to some extent. In a sequence, background variation and occlusion can be regarded as removable noises, while in still images they are troubling interferences.

Although more information can be obtained from per-




(a)




(b)

Figure 2. In (a), on each dataset, we show two video instances of different people, who are wearing similar clothes and walking similarly. In (b), we randomly selected 20 video instances of different people. For each instance, we compute its image frame level feature representation (color&LBP) and video level's (HOG3D+Color&LBP(pooling)) where for the image level's we randomly selected one image frame from a video instance. For each level's representation, we compute the largest intra-class distance $D_w$ and the smallest inter-class distance $D_b$ with respect to each sample. The x-axis is the index of the 20 random samples and the y-axis is the value of $D_w/D_b$. It can be observed that most ratio values of videos are larger than those of image frames, *i.e*. these videos have more ambiguities than images.

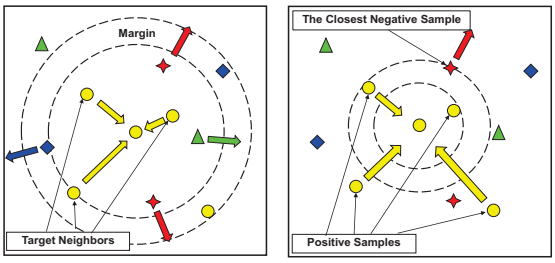
son videos, more challenges come along. Firstly, like the still-image-based approaches, video-based person representations are also similar because of similar appearance. Secondly, although the motion of a pedestrian is a kind of behavioral biometrics, that is an important discriminative cue for identifying different persons, it is unfortunate that the walking actions or other motions of different persons may be similar as well (see Figure 2 (a) for example), which means the inter-class variation may be smaller for video-based representation of a person. As shown in Figure 2 (b), we demonstrate that for some instances, it is harder to distinguish the video representations of different identities (due to large $\frac{D_w}{D_b}$ value) than the still image cases. It shows that the ratio between maximum intra-class distance and minimum inter-class distance is much larger for the video-based re-id as compared to the image-based re-id because some motion information of different people could be similar. This suggests the ambiguity of videos is more serious, and it is true that more intra-class distances are larger than the related minimum inter-class distance. The observation here would imply the discriminative information could be hidden in the minor difference of actions and motion. To mine these minor differences in the data, more stringent constraint should be exploited to look for a latent space to maximize the inter-class margin between different persons. So far, only a few video-based methods [31, 10, 11] have been developed. However, the mentioned problem for video-based person re-id still remains unsolved.

To address the above problem in video-based person re-id, we propose a top-push distance learning model (TDL) in this work. For a person video sequence, we exploit a feature representation constituted by HOG3D [12] and the average pooling of color histograms and LBP features [9]. Based on that, we propose a discriminative distance model optimized towards the realization of the top-push distance constraint combined with the minimization of intra-class variations. We employ the idea of top-push in [15] and introduce it into distance metric learning, in order to optimize the matching accuracy at the top rank for person re-id, which helps to look for a latent feature space to explicitly enlarge the inter-class margin between video sequences.

Extensive experiments have been conducted on two video datasets including PRID 2011 [8] and iLIDS-VID [31] to validate the effectiveness of the proposed TDL model. Our results demonstrate that (1) by formulating the video-based person re-id problem as a distance metric learning problem with top-push constraint modeling, significant improvement on matching accuracy can be obtained against the existing video-based person re-id techniques; and (2) our proposed TDL model outperforms not only related distance/rank learning methods but also related representative still-image-based person re-id methods applied for the video-based person re-id problem under multi-shot setting.

## 2. Related Works

The unsolved problem of person re-id caused by lighting change, viewpoint change, occlusions and intricate background has been increasingly focused on and becomes an important topic in visual surveillance in the last five years. To overcome these challenges, most of existing works can




(a) LMNN (b) TDL

Figure 3. LMNN (left) vs. our TDL (right). Compared to LMNN, since the minimum inter-class distance is considered in TDL, the imposters are more heavily penalized.

be categorized into extracting discriminant/relible features [6, 32, 7, 4, 25, 14, 36, 35] or learning robust metrics or subspaces for matching [2, 29, 40, 27, 13, 20, 28, 26, 37, 21, 22] in recent works. However, all these works use appearance features of still images to match, which may suffer from small inter-class variations caused by similar pedestrians clothing and large intra-class variations caused by occlusions. Although it is natural to extend them to handle video-based person re-id under a multi-shot setting, it is not an optimal way as shown in our experiments, moreover it takes more times for matching due to the increase of gallery size.

Recently, a few works started to consider solving the person video matching problem in re-id. Dynamic Time Warping (DTW), which is a popular sequence matching method widely used for action recognition [23], was applied for video-based person re-id [30]. Wang *et al*. [31] introduced a pictorial video segmentation approach and deployed a fragment selecting and ranking model for person matching. Srikrishna *et al*. [10] introduced a block sparse model to handle the video-based person re-id problem by the recovery problem on embedding space. However, these works assume all image sequences are synchronized, but it becomes unapplicable due to different actions taken by different people. It is also costly and difficult to obtain perfectly aligned pairwise person videos across non-overlapping camera views. All these works use either multiple images or a selected fragment of a sequence to extract feature, and thus they ignore the integrity and the richness of video features. So they are ineffective for solving the video-based person re-id problem.

We extend the use of top-push constraint from linear ranking function [15] to second-order distance metric learning in our TDL model. Both the proposed TDL model and the linear function in [15] aim to optimize the top-rank matching performance. The difference is that the TDL model is able to look for a latent subspace rather than computing only one ranking function score, so that more robust latent features can be exploited. Since, the top rank linear function learning is a RankSVM [29] like learning, which has been shown very costly on high dimensional and moderately large-scale dataset [42], the top rank linear function learning cannot be straightforward generalized to a multiple dimensional one [15]. Our experiments suggest that exploring a subspace rather than a hyperplane is more robust for person re-id.

Different from existing distance metric learning methods, our proposed TDL model is specially motivated from the observation that the inter-class variation is much smaller on video level than that on still image level, so the top-push constraint, a more effective relative comparison, is employed to explicitly avert this problem in a latent feature space. In particular, our approach is related to Weinberger *et al*.'s LMNN method [34]. LMNN aims at optimizing KNN classification by using the local structure of the data. For each instance, a local neighborhood is established, including the $k$ nearest neighbors sharing the same label (target neighbors). Samples that invade this perimeter with a different label (impostors) are penalized (see Figure 3 (a)). Our method seems similar to LMNN; however, an important difference is that a more stringent top-push constraint is used to guide the distance learning, which notably benefits the top-rank matching results in person re-id (see Figure 3 (b)). Ours is also related to the relative distance comparison (RDC) [41]. While RDC is limited by the scale of relative comparison, the proposed TDL can largely reduce the number of relative comparisons in the context of top-push modeling. In addition, compared to LDA [5], our model replaces the maximum of inter-class distance by the minimization of hinge loss of top-push comparison, so that our model has imposed much more powerful constraint on the inter-class modeling. The significant improvement against LMNN, RDC and LDA will be shown in the experiment part.

## 3. Approach

The feature representation of a person video in our model has two main components: space-time features and appearance features. For extracting the space-time features, we employ the HOG3D descriptor [12] to represent the person video. The HOG3D feature contains spatial gradient and temporal dynamic information. For extracting the appearance features, we first use color histograms and LBP features [9] to describe a person appearance in each image frame. To obtain stable appearance cues and suppress noises caused by occlusions, we express the appearance features of a person video by average pooling of features of all frames from that video. The average pooling of color histograms and LBP features can represent rich appearance information of a person in video. The space-time features and the appearance features describe different information of a person in video, and those two types of features are complementary. Therefore in our model, the two features

are combined to address the challenging video-based person re-id problem caused by background change, occlusions and motions.

We denote the training set by $X = \{(\vec{x}_i, y_i)\}_{i=1}^{s}$, where $\vec{x}_i \in \mathbb{R}^d$ is the feature vector extracted from a video of person labeled $y_i$. We denote the distance between any two feature vectors $\vec{x}_i$ and $\vec{x}_j$ by $\mathcal{D}(\vec{x}_i, \vec{x}_j)$.

### 3.1. Enhancing Top-rank Matching by Top-push Distance Learning

For person re-id, it is always expected that, for a query image, the top-rank matching of gallery images is correct. This means the distance between any matched gallery sample and the query should be smaller than the one between any unmatched one and the query. Therefore, in our distance metric learning modeling, we are concerning the relative comparison between the distance of a positive pair and the minimum distance of all related negative pairs, rather than comparing the positive pair with each of the related negative pair. In formulation, that is, for each example $\vec{x}_i$, we wish to realize the following comparison:

$$\mathcal{D}(\vec{x}_i, \vec{x}_j) + \rho < \min_{y_k \neq y_i} \mathcal{D}(\vec{x}_i, \vec{x}_k),\ y_i = y_j, \tag{1}$$

where $\rho$ is a slack parameter. In this work, we set $\rho = 1$. To quantify the above comparison, we aim to minimize a hinge loss function incurred by the positive pairs whose distances are not smaller than the smallest distance of negative pairs with respect to input $\vec{x}_i$:

$$\min \sum_{\vec{x}_i, \vec{x}_j, y_i = y_j} \max\big\{\mathcal{D}(\vec{x}_i, \vec{x}_j) - \min_{y_k \neq y_i} \mathcal{D}(\vec{x}_i, \vec{x}_k) + \rho, 0\big\}. \tag{2}$$

The minimization of the loss of the above comparison refers to inter-class separation, which however does not address the intra-class variation. Therefore, we also wish to strengthen the correlation of samples of any positive pair by minimizing the distance between samples of the same class in the meanwhile, i.e.,

$$\min \sum_{\vec{x}_i, \vec{x}_j, y_i = y_j} \mathcal{D}(\vec{x}_i, \vec{x}_j). \tag{3}$$

Therefore, the objective function of top-push distance learning is formulated below:

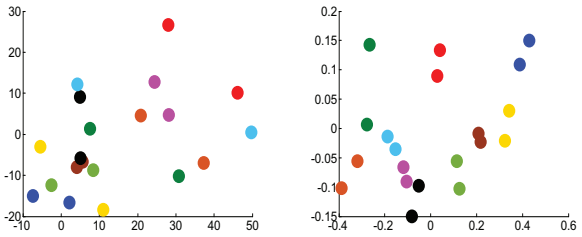
$$\begin{aligned} f(D) =& (1-\alpha) \sum_{\vec{x}_i, \vec{x}_j, y_i = y_j} \mathcal{D}(\vec{x}_i, \vec{x}_j) \\ &+ \alpha \sum_{\vec{x}_i, \vec{x}_j, y_i = y_j} \max\big\{\mathcal{D}(\vec{x}_i, \vec{x}_j) - \min_{y_k \neq y_i} \mathcal{D}(\vec{x}_i, \vec{x}_k) + \rho, 0\big\}, \end{aligned} \tag{4}$$

where $\alpha \in [0, 1]$ refers to a weighting parameter that balances the two terms. We call the second term the *top-push constraint*. Through the optimization, the first term penalizes large distances between positive pairs, and meanwhile the second term penalizes small distances between each sample and the closest sample that is differently labeled. The learning induced by this cost function are illustrated in Figure 3 for an input. We call our approach the top-push distance learning (TDL). In TDL, we specially consider the optimization of Mahalanobis distance under Criterion 4, i.e., considering

$$\mathcal{D}(\vec{x}_i, \vec{x}_j) = (\vec{x}_i - \vec{x}_j)^\top \mathbf{M} (\vec{x}_i - \vec{x}_j), \tag{5}$$

where $\mathbf{M} \succeq 0$ is a positive semi-definite matrix.

To visualize the effectiveness of TDL, a comparison between the data distributions of the original feature space and the latent feature space learned by TDL is shown in Figure 4. The change of distribution indicates that the input data samples of the same person are ambiguous, while TDL does reduce the ambiguities and the data distribution is much favorable for classification.



(a) Before TDL (b) After TDL

Figure 4. Illustration of the effectiveness of TDL, where 10 different persons in PRID 2011 dataset were selected for demonstration. Points of different colors indicate different persons. The left are the person data points in the original 2-D space and the right is the projected person data points in the 2-D space learned by TDL. The projection matrix $\mathbf{L}$ is obtained by decomposing matrix $\mathbf{M}$ into $\mathbf{M} = \mathbf{L}^\top \mathbf{L}$.

### 3.2. Optimization

To simplify our notation, we denote the outer product of pairwise differences by

$$\mathbf{X}_{i,j} = (\vec{x}_i - \vec{x}_j)(\vec{x}_i - \vec{x}_j)^\top. \tag{6}$$

Based on Eq. (6), we can reformaulate $\mathcal{D}(\vec{x}_i, \vec{x}_j)$ as follows:

$$\mathcal{D}(\vec{x}_i, \vec{x}_j) = tr(\mathbf{M}\mathbf{X}_{i,j}) \tag{7}$$

and therefore we can reformulate the objective function Eq.(4) as:

$$\begin{aligned} f(\mathbf{M}) =& (1-\alpha) \sum_{\vec{x}_i, \vec{x}_j, y_i = y_j} tr(\mathbf{M}\mathbf{X}_{i,j}) \\ &+ \alpha \sum_{\vec{x}_i, \vec{x}_j, y_i = y_j} \max\big\{tr(\mathbf{M}\mathbf{X}_{i,j}) - \min_{y_k \neq y_i} tr(\mathbf{M}\mathbf{X}_{i,k}) + \rho, 0\big\}. \end{aligned} \tag{8}$$

**Algorithm 1** The Optimisation Algorithm for TDL.

**Initialize:**

Initialize metric with the identity matrix $\mathbf{M}_0 := \mathbf{I}$;
The triggered set $\mathcal{N}(\mathbf{M}_0) := \{\}$ ;
The gradient $\mathbf{G}_t := (1-\alpha)\sum_{i,j}\mathbf{X}_{i,j}$ ;
The counter $t := 0$.

1: **while** (not converged) **do**
2: Search the smallest between-class distance by Eq.(3).
3: Construct the triggered set $\mathcal{N}(\mathbf{M}_t)$ by indices $(i,j,k)$ determined by the second term of Eq.(4).
4: Compute $\mathbf{G}_t$ by Eq.(9).
5: Compute $\mathbf{M}_{t+1} := \mathbf{M}_t - \lambda\mathbf{G}_t$.
6: Project $\mathbf{M}_{t+1}$ onto the cone of all positive semi-definite matrices $\mathcal{P}_+(\mathbf{M}_{t+1})$.
7: $t := t+1$.
8: **end while**
9: **return** $\mathbf{M}_t$.

Our model applies a stochastic gradient descent projection method to compute an optimized positive semi-definite matrix $\mathbf{M}$ in Eq.(8). In particular, at the $t$-th iteration, Eq.(8) is piecewise linear with respect to $\mathbf{M}$. At step $t$, given $\mathbf{M} = \mathbf{M}_t$, we define a set of indices $(i,j,k) \in \mathcal{N}(\mathbf{M}_t)$, if and only if the indices $(i,j,k)$ trigger the second term of Eq.(8). The stochastic gradient $\mathbf{G}_t$ of $f(\mathbf{M})$ at step $t$ is computed by:

$$\mathbf{G}_t = \frac{\partial f}{\mathbf{M}}|_{\mathbf{M}=\mathbf{M}_t} = (1-\alpha)\sum_{i,j}\mathbf{X}_{i,j} + \alpha\sum_{(i,j,k)\in\mathcal{N}(\mathbf{M}_t)}(\mathbf{X}_{i,j}-\mathbf{X}_{i,k}). \tag{9}$$

The optimization of Eq.(8) must satisfy the constraint that the matrix $\mathbf{M}_{t+1}$ remains positive semi-definite. For this purpose, we project $\mathbf{M}_{t+1}$ onto the cone of all positive semi-definite matrices $\mathcal{P}_+$ after each gradient descent step. To be specific, we first perform the eigen-decomposition on $\mathbf{M}_{t+1}$:

$$\mathbf{M}_{t+1} = \mathbf{V}_{t+1}\mathbf{D}_{t+1}\mathbf{V}_{t+1}^\top. \tag{10}$$

In order to apply the projection, we will update the diagonal matrix $\mathbf{D}_{t+1}$ by removing all the negative eigenvalues, and then reconstruct $\mathbf{M}_{t+1}$ by Eq.(10).

The algorithm is summarized in **Algorithm 1**. We denote the gradient step size by $\lambda > 0$. In practice, it worked starting with $\lambda = 1e-03$. Then, at each iteration, we increased $\lambda$ by a factor of 1.01 if the loss function decreased and decreased $\lambda$ by a factor of 0.5 if the loss function increased.

**Matching.** The learned metric can be exploited to perform person re-id by matching a probe person video sequence $\vec{x}_p$ against a gallery set $\{\vec{x}_g\}$ in another camera view. The distance between a probe video sequence $\vec{x}_p$ and a gallery video sequence $\vec{x}_g$ is computed by

$$\mathcal{D}(\vec{x}_p, \vec{x}_g) = (\vec{x}_p - \vec{x}_g)^\top \mathbf{M}(\vec{x}_p - \vec{x}_g). \tag{11}$$

# 4. Experiments

## 4.1. Datasets and settings

**Datasets.** Our experiments were conducted on two publicly available video datasets for video-based person re-id: the PRID 2011 dataset [8] and the iLIDS-VID dataset [31]. The PRID 2011 dataset consists of video pairs recorded from two different but static surveillance cameras. 385 persons were recorded in camera view A, and 749 persons in camera view B. Among all persons, 200 persons were recorded in both camera views. Each video is comprised of 5 to 675 image frames, with an average of 100 for each. To guarantee the effective length of the video, we selected 178 persons with more than 27 frames in our experiments. This dataset was captured in uncrowded outdoor scenes with relatively simple and clean background and rare occlusions, and several different poses of person are available in each camera view (Figure 5(a) ). The iLIDS-VID dataset contains 600 video of 300 randomly sampled people. Each person has one pair of video from two camera views. Each video is comprised of 23 to 192 image frames, with an average of 73 for each. Compared with the PRID 2011 dataset, it was captured in an airport arrival hall under a multi-camera CCTV network. The challenges of this dataset largely lie in clothing similarities, lighting and viewpoint changes across camera views, complicated background and occlusions (Figure 5(b) ).

**Settings.** In our experiments, we adopted a single-shot experiment setting. All datasets were randomly divided into training set and testing set by half so that there were $p = 89$ and $p = 150$ individuals in the testing sets of PRID 2011 and iLIDS-VID respectively. In the testing stage, the videos from one camera were used as the gallery set while the ones from another camera as the probe set. The cumulative matching characteristic (CMC) curve is used to measure the performance of each method on each dataset. A rank $k$ matching rate indicates the accuracy of the matching between the probe video $\vec{x}_p$ and the gallery videos $\{\vec{x}_g\}_{g=1}^k$ in the top $k$ rank list. To obtain statistically reliable results, we repeated the procedure 10 times and reported the average results.

## 4.2. Feature Extraction

To obtain more abundant and robust features for representing a person video, we explored a combined person video feature representation. We expressed each sample with appearance feature on image frame level and space-time feature on video level. Specifically, at the image frame level, each frame of the person video was resized to $128\times48$

(a) PRID 2011 (b) iLIDS-VID

Figure 5. Example pairs of image sequences of the same person appearing in different camera views.

| Methods | PRID 2011 | | | | iLIDS-VID | | | |
|---|---|---|---|---|---|---|---|---|
| | Rank-1 | Rank-5 | Rank-10 | Rank-20 | Rank-1 | Rank-5 | Rank-10 | Rank-20 |
| TDL | **56.74** | **80.00** | **87.64** | **93.59** | **56.33** | **87.60** | **95.60** | **98.27** |
| SDALF [4] | 5.2 | 20.7 | 32.0 | 47.9 | 6.3 | 18.8 | 27.1 | 37.3 |
| Salience [38] | 25.8 | 43.6 | 52.6 | 62.0 | 10.2 | 24.8 | 35.5 | 52.9 |
| RPRF [19] | 19.3 | 38.4 | 51.6 | 68.1 | 14.5 | 29.8 | 40.7 | 58.1 |
| SRID [10] | 35.1 | 59.4 | 69.8 | 79.7 | 24.9 | 44.5 | 55.6 | 66.2 |
| DVDL [11] | 40.6 | 69.7 | 77.8 | 85.6 | 25.9 | 48.2 | 57.3 | 68.9 |
| Color&LBP+DVR [31] | 37.6 | 63.9 | 75.3 | 88.3 | 34.5 | 56.7 | 67.5 | 77.5 |

Table 1. Comparison with the state-of-the-art methods on PRID 2011 and iLIDS-VID datasets. Results are shown as matching rates (%) at Rank = 1, 5, 10, 20. Best results are in boldface font.

pixels and divided into patches with size $8 \times 16$ with $50\%$ overlap both in the horizontal and vertical directions. That is to say, there were $155$ patches for extracting color histograms and LBP features [9]. For each patch, histograms of color channels in HSV and LAB color spaces and LBP descriptor were computed. All the appearance feature descriptors within the image frame were concatenated together to form a 1705-dimensional feature vector. At the video level, we extracted a 1200-dimensional HOG3D feature vector for each person video [12]. In the end, we described the whole person video using a 2905-dimensional vector by connecting this HOG3D feature with average pooling of color histograms and LBP features over all image frames of the video.

## 4.3. Evaluation of Comparison

### 4.3.1 Comparison with the State-of-the-art Methods

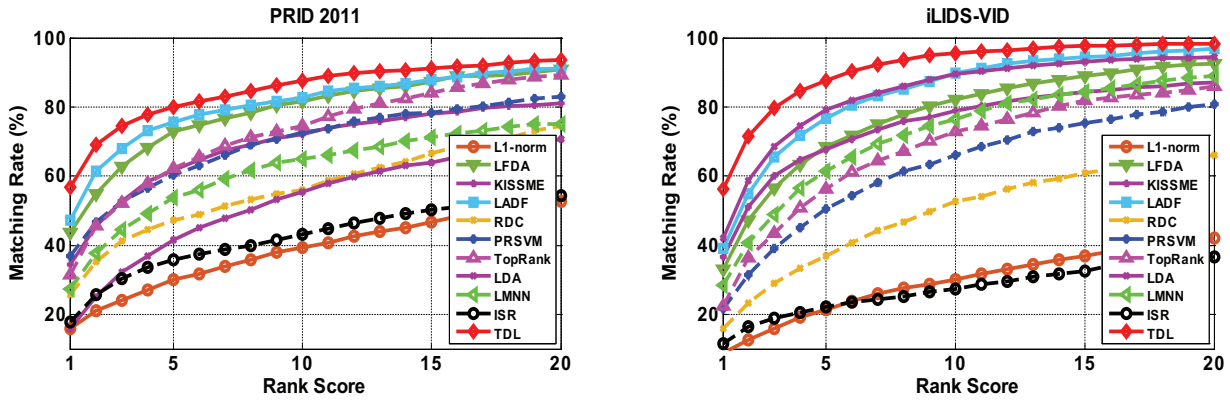
In Table 1, we reported the comparison of our proposed TDL model with the existing six state-of-the-art video-based person re-id methods on PRID 2011 and iLIDS-VID datasets, including SDALF [4], Salience [38], RPRF [19], SRID [10], DVDL [11] and Color&LBP+DVR [31]. DVDL is a dictionary learning method based on multi-shot re-id datasets. DVR is a method based on ranking model, which also selects discriminative video fragment from a candidates pool in the training process. The results show clearly that with the proposed TDL model, the matching performance on both datasets is improved significantly. For instance, on iLIDS-VID dataset, our TDL improved the Rank-1 matching rate by $21.8\%$ compared to the second best method Color&LBP+DVR.

Another interesting but indeed fact can be observed is that TDL outperformed others much better on iLIDS-VID. We examined that this is probably because more intra-class distances could be much larger than inter-class ones under more occlusions on iLIDS-VID. While the compared distance models do not explicitly and directly quantify the relation between each intra-class distance and the related minimum inter-class distance, the proposed TDL employs the top-push strategy and makes the distance model quantify more effective features and thus performs more stably.

### 4.3.2 Comparison with Related Methods

There are several existing distance/subspace learning models usually applied for person re-id. For fair comparison, all compared methods used the same feature representation of person videos described in Sec. 4.2. We first compared our TDL with representative rank/distance/subspace learning methods for video-based matching, *e.g*. TopRank [15], linear discriminant analysis (LDA) [5] and LMNN [34]. Our results (Figure 6 and Table 2) show clearly that the proposed TDL model obtains better matching rates than these methods. More specifically, on PRID 2011 dataset, the Rank-1 matching rate is 56.74% for TDL, whilst 31.69% for TopRank, 15.84% for LDA, and 27.19% for LMNN. These results show that these related methods performed poorly for video-based person re-id. As seen from the comparison video-based matching results in Figure 6 and Table 2, the improvement was particularly significant on iLIDS-VID dataset, which is more challenging due to more ambiguities caused by occlusions and illumination. With the top-push constraint, the ambiguities can be better removed.

The video-based matching results of several representa-




(a) PRID 2011 (b) iLIDS-VID

Figure 6. Video-based matching rates (%) of different methods on PRID 2011 and iLIDS-VID.

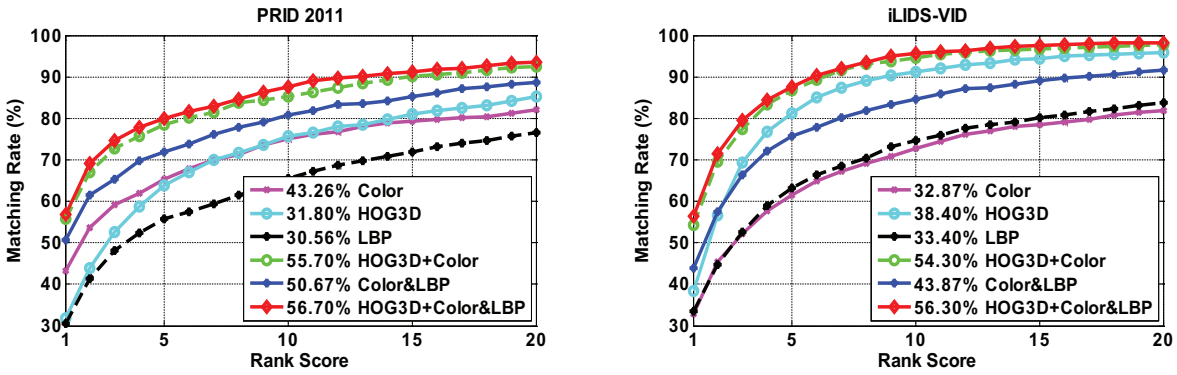
| Settings | Methods / Rank | PRID 2011 | | | | iLIDS-VID | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Rank-1 | Rank-5 | Rank-10 | Rank-20 | Rank-1 | Rank-5 | Rank-10 | Rank-20 |
| **Video-based matching** | TDL | **56.74** | **80.00** | **87.64** | **93.59** | **56.33** | **87.60** | **95.60** | **98.27** |
| | L1-norm | 15.84 | 30.00 | 39.33 | 52.70 | 8.90 | 21.40 | 30.07 | 42.07 |
| | LFDA [28] | 43.70 | 72.80 | 81.69 | 90.89 | 32.93 | 68.47 | 82.20 | 92.60 |
| | KISSME [13] | 34.38 | 61.68 | 72.13 | 81.01 | 36.53 | 67.80 | 78.80 | 87.07 |
| | LADF [20] | 47.30 | 75.50 | 82.69 | 91.12 | 39.00 | 76.80 | 89.00 | 96.80 |
| | RDC [41] | 25.62 | 47.30 | 56.07 | 74.38 | 15.80 | 36.93 | 52.60 | 66.00 |
| | PRSVM [29] | 36.97 | 60.45 | 72.47 | 83.03 | 21.53 | 50.60 | 66.00 | 80.80 |
| | ISR [24] | 17.64 | 35.84 | 43.03 | 54.38 | 11.60 | 22.13 | 27.40 | 36.67 |
| | TopRank [15] | 31.69 | 62.24 | 75.28 | 89.44 | 22.53 | 56.13 | 72.73 | 85.93 |
| | LDA [5] | 15.84 | 41.46 | 55.51 | 70.67 | 42.06 | 79.13 | 89.40 | 94.47 |
| | LMNN [34] | 27.19 | 53.71 | 64.94 | 75.17 | 28.33 | 61.40 | 76.47 | 88.93 |
| **Multiple image frames matching** | TDL | **30.22** | 59.10 | **74.04** | **88.43** | 9.81 | 27.52 | **46.10** | **62.19** |
| | L1-norm | 12.36 | 29.44 | 40.56 | 56.40 | 3.67 | 10.33 | 16.03 | 26.93 |
| | LFDA [28] | 26.40 | 56.07 | 69.89 | 81.12 | 7.80 | 23.93 | 36.47 | 50.80 |
| | KISSME [13] | 28.54 | **59.78** | 72.13 | 83.26 | **10.67** | **28.33** | 39.80 | 57.00 |
| | LADF [20] | 8.20 | 20.45 | 29.89 | 42.25 | 4.33 | 14.00 | 21.20 | 32.13 |
| | ISR [24] | 10.50 | 20.83 | 31.83 | 44.17 | 8.04 | 20.50 | 31.33 | 43.50 |
| | LDA [5] | 27.64 | 58.09 | 69.66 | 82.47 | 10.27 | 27.40 | 39.80 | 55.27 |
| | LMNN [34] | 14.38 | 38.09 | 50.22 | 67.19 | 4.47 | 13.20 | 21.60 | 35.47 |

Table 2. Comparison with related methods on PRID 2011 and iLIDS-VID datasets. The matching rate (%) at Rank $i$ means the accuracy of the matching within the top $i$ gallery classes.

tive still-image-based person re-id methods are also shown in Figure 6 and Table 2, including L1-norm, LFDA [28], KISSME [13], LADF [20], PRSVM [29], RDC [41] and ISR [24]. One can observe that our TDL model always outperformed all the compared re-id methods on both datasets. The improvement is particularly significant on iLIDS-VID dataset, and TDL is 17.33% higher than the best compared the method at Rank-1. In addition, among these six re-id methods, RDC is closely related to our model, but RDC is limited by the scale of relative comparison. In our experiments, the computational cost of our model was only 3% of the one of RDC. These results highlight the effectiveness of the proposed model.

One may wonder when using multiple image frames, whether existing still-image-based methods can achieve better performance. To answer the question, in this section, we adopted a multi-frame setting to conduct the experiments, in which 5 images of each person were randomly selected from all frames as gallery. We used the combined appearance features (Color&LBP&HOG) [18] as representation of still image frames. We performed experiments on the two datasets and the results are also reported in Table 2.

Since RDC, PRSVM and TopRank suffered from the huge computational cost with increasing size of training set under multi-frame or multi-shot setting, these methods cannot be run on a server with 64GB RAM. To be more specifical, when conducting iLIDS-VID (consists of 300 persons) under multi-shot setting, not just more persons were involved but also more images were used (10 frames for each person in the training set), so that the number of triplets for




(a) PRID 2011 (b) iLIDS-VID

Figure 7. Evaluation of different feature components in TDL on PRID 2011 and iLIDS-VID. The rank 1 matching rate of each method is provided in the legend.

relative comparison increases dramatically (more than $10^8$). RDC and PRSVM are designed to utilize all the triplets for training, and it is clear that RDC and PRSVM are costly and not computational trackable.

Compared to the video-based matching results, it is evident that all the still-image-based methods performed poorly, worse than their video-feature-based versions. This suggests that space-time video information is an important cue to augment the feature representation for person re-id; that is, video-based matching is more effective than multiple image frames matching.

### 4.4. Further Evaluation of TDL

#### 4.4.1 Effects of Different Feature Components

The feature representation used in our proposed model consists of two components: space-time features (HOG3D) and appearance features (Color&LBP (pooling)). In Figure 7 , we evaluated the effects of each component respectively. The results show that all of them are effective on their own, and when they are combined, the best performance is achieved. This validates that these feature components are complementary and should be fused.

#### 4.4.2 Influence of Parameters

We implemented our TDL model by selecting the parameter $\alpha$ on PRID 2011 and iLIDS-VID datasets. The results of area under CMC curve (AUC) were plotted in Figure 8 (a) and (b). As illustrated, when $\alpha$ was around 0.1, the model achieved the best result. The figures suggest the performance of using and not using top-push constraint in TDL is distinct. When it is not integrated, the optimization problem Eq. (8) becomes trivial since $\mathbf{M} = \mathbf{O}$ where $\mathbf{O}$ is a zero matrix is the optimal solution which cannot be effective for classification. We also observe that when $\alpha = 1$, i.e., discarding the intra-class variation minimization, it will also lead to overfitting in top-push. Thus a proper $\alpha$, for instance 0.1 here is a balance.

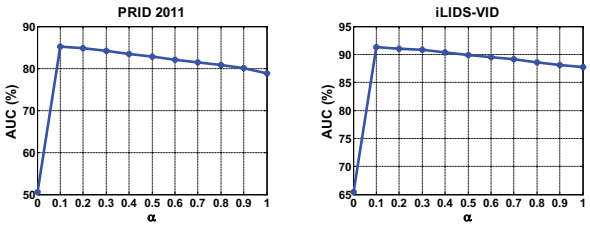



Figure 8. Parameter sensitivity analysis on PRID 2011 and iLIDS-VID

## 5. Conclusion

In this work, we have proposed a top-push distance learning (TDL) model to address the video-based person re-identification problem. While video-based representation contains more abundant space-time information than still-image based representation, there are more ambiguities in video-based features than still-image-based features. So we introduce a top-push constraint to quantify ambiguous video representation. Due to the employment of top-push constraint, the formed distance model can be more effective on top-rank performance of video-based person re-id. This is validated on through extensive experiments conducted on two video datasets including PRID 2011 and iLIDS-VID.

## Acknowledgments

This work was supported in part by the Computational Science Innovative Research Team Program, Guangdong Provincial Government of China, in part by the Natural Science Foundation of China under Grant 61472456, Grant 61522115, and Grant 6151101169, in part by the Guangzhou Pearl River Science and Technology Rising Star Project under Grant 2013J2200068, in part by the Guangdong Natural Science Funds for Distinguished Young Scholar under Grant S2013050014265, and in part by Guangdong Program for Support of Top-notch Young Professionals (No. 2014TQ01X779).

## References


[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *CVPR*, 2015.

[2] Y. Chen, W.-S. Zheng, and J. Lai. Mirror representation for modeling view-specific transform in person re-identification. In *IJCAI*, 2015.

[3] Y.-C. Chen, W.-S. Zheng, J. Lai, and P. Yuen. An asymmetric distance model for cross-view feature mapping in person re-identification. *IEEE TCSVT*, 2015.

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

[5] K. Fukunaga. *Introduction to statistical pattern recognition*. Academic press, 2013.

[6] N. Gheissari, T. Sebastian, and R. Hartley. Person reidentification using spatiotemporal appearance. In *CVPR*, 2006.

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

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

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

[10] S. Karanam, Y. Li, and R. Radke. Sparse re-id: Block sparsity for person re-identification. In *CVPR Workshop*, 2015.

[11] S. Karanam, Y. Li, and R. J. Radke. Person re-identification with discriminatively trained viewpoint invariant dictionaries. In *ICCV*, 2015.

[12] A. Klaser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, 2008.

[13] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012.

[14] I. Kviatkovsky, A. Adam, and E. Rivlin. Color invariants for person reidentification. *IEEE TPAMI*, 35(7), 2013.

[15] N. Li, R. Jin, and Z.-H. Zhou. Top rank optimization in linear time. In *NIPS*, 2014.

[16] W. Li and X. Wang. Locally aligned feature transforms across views. In *CVPR*, 2013.

[17] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014.

[18] X. Li, W.-S. Zheng, X. Wang, T. Xiang, and S. Gong. Multi-scale learning for low-resolution person re-identification. In *ICCV*, 2015.

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

[20] 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*, 2013.

[21] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015.

[22] S. Liao and S. Z. Li. Efficient psd constrained asymmetric metric learning for person re-identification. In *ICCV*, 2015.

[23] Z. Lin, Z. Jiang, and L. S. Davis. Recognizing actions by shape-motion prototype trees. In *ICCV*, 2009.

[24] G. Lisanti, I. Masi, A. Bagdanov, and A. Del Bimbo. Person re-identification by iterative re-weighted sparse ranking. 37(8):1629–1642, 2014.

[25] B. Ma, Y. Su, and F. Jurie. Local descriptors encoded by fisher vectors for person re-identification. In *ECCV*, 2012.

[26] L. Ma, X. Yang, and D. Tao. Person re-identification over camera networks using multi-task distance metric learning. *IEEE TIP*, 23(8):3656–3670, 2014.

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

[28] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *CVPR*, 2013.

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

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

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

[32] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu. Shape and appearance context modeling. In *ICCV*, 2007.

[33] X. Wang, W.-S. Zheng, X. Li, and J. Zhang. Cross-scenario transfer person re-identification. *IEEE TCSVT*, 2015.

[34] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *JMLR*, 10:207–244, 2009.

[35] S. Wu, Y.-C. Chen, X. Li, A. Wu, J. You, and W.-S. Zheng. An enhanced deep feature representation for person re-identification. In *WACV*, 2016.

[36] Z. Wu, Y. Li, and R. Radke. Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features. *IEEE TPAMI*, 2014.

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

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

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

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

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

[42] W.-S. Zheng, S. Gong, and T. Xiang. Towards open-world person re-identification by one-shot group-based verification. *IEEE TPAMI*, 2015.

[43] W.-S. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, and S. Gong. Partial person re-identification. In *ICCV*, 2015.