Multimed Tools Appl (2019) 78:337–362
https://doi.org/10.1007/s11042-018-6239-3

Low illumination person re-identification

**Fei Ma[1] · Xiaoke Zhu[2] · Xinyu Zhang[1] · Liang Yang[1] ·
Mei Zuo[1] · Xiao-Yuan Jing[1]**

Received: 30 July 2017 / Revised: 26 April 2018 / Accepted: 4 June 2018 /
Published online: 19 June 2018

Abstract Low illumination is a common problem for recognition and tracking. Low illumination video-based person re identification (re-id) is an important application in practice. Low illumination usually results in severe loss of visual appearance and space-time information contained in pedestrian image or video, which brings large difficulty to re-identification. However, the problem of low illumination video-based person re-id (LIVPR) has not been well studied. In this paper, we propose a novel triplet-based manifold discriminative distance learning (TMD^{2}L) approach for LIVPR. By regarding each video as an image set, TMD^{2}L aims to learn a manifold-based distance metric, under which the intrinsic structure of image sets can be preserved, and the distance between truly matching sets is smaller than that between wrong matching sets. Experiment results on the new collected low illumination person sequence (LIPS) dataset, as well as two simulated datasets LI-PRID 2011 and LI-iLIDS-VID show that our proposed approach TMD^{2}L outperforms existing representative person re-id methods.

Keywords Low illumination · Person re-identification · Local linear model ·
Discriminative distance learning

1 Introduction

Person re-identification (re-id) is becoming a hot topic in computer vision and machine learning gradually. Person re-id matches pedestrians across different disjoint cameras in

✉ Xiao-Yuan Jing
jingxy_2000@126.com

Fei Ma
mafei8063@whu.edu.cn

[1] School of Computer Science, Wuhan University, Wuhan 430072, China

[2] School of Computer and Information Engineering, Henan University, Kaifeng 475001, China

different time periods [3], which plays an import role in smart city. However, the problem of person re-identification has the following challenges. First, most faces in images captured by different non-overlapping cameras are blurred in that the pedestrians are long distance away from the cameras. Therefore, only the appearance of pedestrian can be used to re-identification. Second, pedestrians' images are largely different due to illumination variations, poses, viewpoint, background clutter and occlusion. Third, the clothes of different persons may be similar or even same, e.g. the same uniform. Therefore, person re-identification has been an extremely challenging problem [23].

A great number of methods have been presented for person re-id problem, which can be divided into two categories: feature learning based methods [11, 15, 33, 36, 38, 41–43] and metric learning based methods [1, 9, 37, 45, 47]. Feature learning based methods aim to learn a distinct and robust feature representation. Literature [39] seeks to use a salient color names based color descriptor to describe colors. Color distributions in different color spaces are then obtained and fused for a feature vector representation. Work [13] is proposed to learn a feature representation local maximal occurrence, which analyzes the horizontal occurrence of local features to make a stable representation against viewpoint changes. The literature [15] is presented to learn a robust and efficient appearance descriptor for re-identification based on coarse, striped pooling of local features. Besides the methods of robust features, metric based approaches focus on learning an effective metric for person re-identification. Zheng et al. [45] presented the relative distance comparison (RDC) method to learn the optimal similarity measure between a pair of person images, which can avoid treating all features indiscriminately. Hirzer et al. [5] learned a discriminative Mahalanobis distance metric from pairs of samples belonging to different cameras. In [22], a latent metric learning method is developed for learning an effective metric, which can be solved via an iterative manner. A joint feature projection matrix and heterogeneous dictionary pair learning (PHDL) [48] approach is presented to jointly learn an intra-video projection matrix and a pair of heterogeneous image and video dictionaries. The learned projection matrix is used to reduce the influence of variations within each video. Two learned dictionaries can transform the heterogeneous image and video features into same dimensional coding coefficients. Joint dictionary and metric learning (JDML) [46] is presented to formulate robust feature representation learning and discriminative metric learning into a unified framework. The dictionary learning is utilized to obtain robust feature representation for images across different camera views. Metric learning is exploited to find optimal feature subspace that maximizes the inter-person divergence while minimizes the intra-person divergence.

The above works focus on matching pedestrians under normal illumination. In many cases, pedestrian videos are captured under low illumination conditions, e.g. pedestrian videos captured at night or under poor light conditions like underpass. Under these scenarios, the captured videos are low-illumination. We call re-identification under this kind of scenarios as low illumination video-based person re-identification (LIVPR). Figure 1 shows the problem of low illumination video-based person re-identification. Although there exists a street lamp in Fig. 1, the illumination under this scenario is too low to observe the texture of the pedestrian's clothes clearly, which is more difficult for person re-id than that under normal illumination.

1.1 Motivation

Low illumination is a common problem for video-based person re-id, e.g., monitoring at night or under poor light conditions like underpass. However, this problem is usually

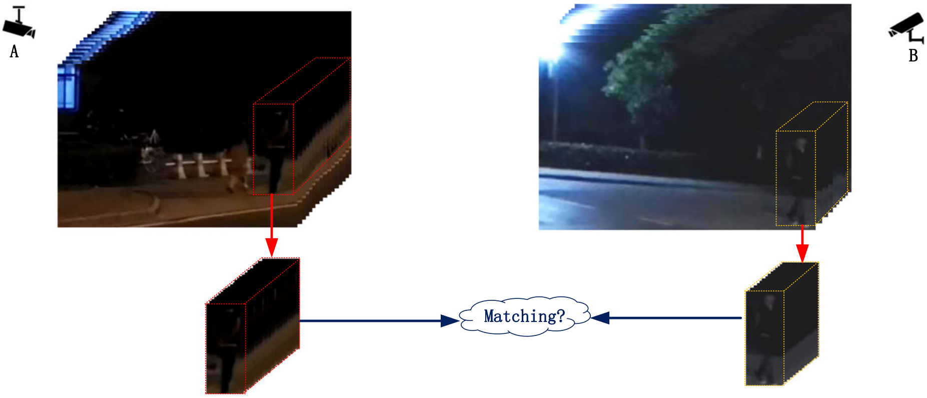


Fig. 1 A typical low illumination video-based person re-identification scenario. The pedestrian image sequences are captured under low illumination condition. One can see that the illumination is severely low, which is difficult to observe the texture of the pedestrian's clothes clearly

ignored by the existing video-based person re-identification methods, which mainly focus on solving the challenges under normal illumination condition.

Low illumination will result in severe loss of visual appearance information to the captured videos, which is harmful to the re-identification process. In particular, many appearance details, e.g., color, gradient and texture, have been lost [21] in low illumination person images. To reveal this influence, we make the following experiment. We selected randomly 10 different persons' images under normal and low illumination from two datasets respectively and computed the cosine similarity of image level visual appearance features (including LBP, LAB, RGB, HSV and HOG). Here, the low illumination scene is simulated by using the same way as [25]. The experimental result is shown in Fig. 2a. One can see that most of the cosine similarity values are smaller than 0.5, which means

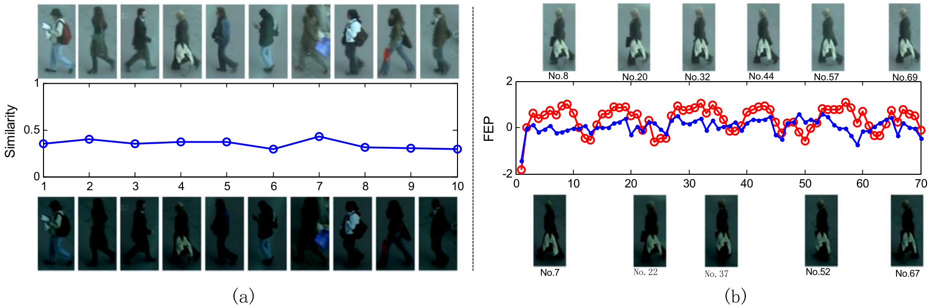


Fig. 2 Effect of low illumination. **a** Cosine similarities of the same person's appearance feature patches under normal and low illumination. The person image (64 × 128) is divided into 128 non-overlapped patches with the patch size 8 × 8. Example pairs on PRID2011 dataset under normal and simulated low illumination. For each instance, we computed the cosine similarity of image level feature patch (LBP+LAB+RGB+HSV+HOG). **b** The extraction of the walking cycles versus FEP under normal and low illumination. (*Top Row*): The first frame of each walking cycle from the third person on PRID 2011 dataset. (*Middle Row*): The normal illumination FEP (red curve) and low illumination FEP (blue curve) corresponding to 70 frames including five walking cycles. (*Bottom Row*): The first frame of each walking cycle from the third person on the LI-PRID 2011 dataset under low illumination

that the appearance information of pedestrians images under low illumination changes severely. Therefore, low illumination will increase the difficulty of distinguishing the visual appearance representations of different persons.

In addition, due to the low illumination, it is difficult to detect a walking cycle of each person by Flow Energy (FEP) [17, 31] accurately. The extraction of the walking cycles by FEP under normal and low illumination is shown in Fig. 2b. The extracted correct walking cycles under normal illumination are shown in Fig. 2b (*Top Row*). The local maxima of FEP corresponds to the posture when the person's two legs overlap while at the local minima the two legs are the farthest away. We take the frame at the local maxima of FEP as the start frame of each walking cycle. Under normal illumination, the sequence numbers of the start frames for all walking cycles are {8,20,32,44,57,69}, respectively. As shown in Fig. 2b (*Bottom Row*), the detected start frames under low illumination are separately {7,22,37,52,67}, and there is even one walking cycle that is not detected at all. The main reason why the first frame of each walking cycle is inaccurate under low illumination is that FEP is greatly affected by the low illumination. However, existing video-based person re-id models do not provide a good solution to the LIVPR task.

Researches in [14] indicate that images of the same identity captured under low illumination usually lie on a nonlinear space, e.g., the samples caused by the pose changing and illumination changing will lie on a nonlinear space, which will result in a poor performance for recognition in linear space. Manifold learning is an effective technique to deal with the data in nonlinear distribution, which can preserve the intrinsic structure of the data in low dimension space from high dimension space. Manifold is stable to varying poses and lighting conditions [28, 35]. The linear subspace model is limiting for complex cases with variations in pose or illumination [6]. Comparing to the explicitly mapped Euclidean space [19], the discriminative learning on the original manifold space can better preserve geometry structure of samples.

Motivated by the above analysis, we intend to solve the problem of video-based person re-id under low illumination based on the manifold learning technique.

1.2 Contribution

We summarize the contributions of this paper as the following **four** points:

(1) To the best of our knowledge, this is the first attempt to solve the problem of low illumination video-based person re-id (LIVPR).

(2) We propose a triplet-based manifold discriminative distance learning (TMD^{2}L) approach for LIVPR task. Local Linear Models (LLMs) of each manifold are constructed by the furthest seed point of the maximal linear patch (MLP) [30]. The distance of different global nonlinear manifolds is presented by their corresponding LLMs. TMD^{2}L seeks to learn a discriminant metric that makes the distance between videos from the same person become smaller than that between videos from different persons.

(3) We contribute a new person sequence dataset, named Low Illumination Pedestrian Sequence (LIPS) dataset, which is collected under night scenes in campus of Wuhan University. The dataset includes 90021 images of 100 pedestrians captured by different cameras, with around 450 images per person at each camera. The LIPS dataset is challenging since it involves complicated cluttered background and occlusions under low illumination. To the best of our knowledge, this is the first pedestrian video sequence dataset under low illumination for person re-identification.

(4) We evaluate the proposed approach on three pedestrian video datasets, including the new collected LIPS dataset, as well as two simulated PRID 2011 [4] and iLIDS-VID [31]. Extensive experiments are conducted on three datasets, and experimental results demonstrate the effectiveness of the proposed approach for the LIVPR problem.

The remainder of this paper is organized as follows: Section 2 presents the most related works and gives a discussion of the relationship between our proposed approach and these works. Section 3 details the proposed TMD^{2}L approach and shows the process of performing person re-identification with the proposed approach. Section 4 elaborates the matching process of person re-identification. Section 5 details the experimental results and parameters analysis. Section 6 concludes the paper and the future work.

2 Brief review of related work

In this section, we briefly review the related person re-id methods. Existing person re-id methods can be classified into two categories: image-based and video-based person re-id methods. The former focuses on the image-to-image matching and can be further divided into two categories: feature learning methods [2, 7, 44] and distance learning methods [9, 12, 24, 37, 45]. The feature learning methods [10, 39] aim to extract a distinct and robust feature representation for matching. The distance learning methods [16, 37] focus on seeking an optimal distance metric for person re-id.

Recently, several researches [17, 31, 40, 47] started to consider solving the video-based person re-id problem. In [20], a novel recurrent neural network architecture is presented for video-based person re-id. The convolutional network, recurrent layer, and temporal pooling layer, are jointly trained to act as a feature extractor to extract temporal features by a recurrent convolutional neural network. Top-push distance learning model (TDL) [40] integrates metric learning with a top-push constraint for matching video features of persons. Under the constraint of top-push, the learned distance metric can help to look for a latent feature space to maximize the margin of inter-classes while minimize the distance between intra-class. TDL model utilizes a stochastic gradient descent projection algorithm to obtain an optimized positive semi-definite matrix. With the learned metric, TDL can reduce the ambiguities of the sample distribution.

The spatio-temporal representation Fisher vectors approach (STFV3D) [17] introduces a method of extracting space-time features from videos. It extracts the same person's walking cycles from the video by computing the Flow Energy Profile (FEP) of the lower body. FEP is the local maxima when the person's two legs overlap and the local minima when the two legs are the farthest away. It splits the same person's video sequence into small segments according to the obtained local maxima/minima of FEP. It needs more than 20 frames for each person's video sequence. The smaller segmentation of walking cycles can align the dynamic appearance of different people both spatially and temporally. Therefore, a series of body-action units can be obtained. The final space-time representation can be concatenated by learning Fisher vectors in each unit. STFV3D is the competing feature from videos for person re-identification.

Simultaneous intra-video and inter-video distance learning (SI^{2}DL) [47] learns a distance metric for intra-video and an inter-video distance metric from the training videos simultaneously. Each video can be considered as an image set, which has the existence of large intra-video and inter-video variations. SI^{2}DL learns the more discriminative distance metric by reducing the influence of these variations. The intra-video distance metric can make each video more compact. It designs a video relationship model, i.e., video triplet, which is constituted by a pair of truly matching videos and an ”impostor” video. Under the constraint of video triplet, the learned inter-video discriminant metric can make the distance between two truly matching persons smaller than that between two wrong matching persons.

Although the above methods solve some problems of person re-id effectively, they study the person re-id problems under normal illumination. Different from these methods, our approach is designed to solve the low illumination video-based person re-id problem particularly.

3 Our approach

In this section, we elaborate two key components of the proposed TMD^{2}L approach, including constructing the Local Linear Model(LLM) and learning discriminative distance metric.

3.1 Constructing local linear model

We firstly prepare each frames of each person by histogram equalization, which can eliminate the low illumination effects [28] as shown in Fig. 4b. As the analysis in Fig. 2, the low-illumination images contain less effective information than those with normal illumination. The features caused by the pose changing and illumination changing will lie on a nonlinear space [26], which will result in a poor performance for recognition in linear space. Manifold learning is an effective technique to deal with the data in nonlinear distribution, which can preserve the intrinsic structure of the data in low dimension space from high dimension space. Although the contrast of neighbor pixels under low illumination is not significant, there exist intrinsic structure in image samples. Manifold can preserve local neighborhood structures of the pedestrian data [14] in nonlinear space to some extent. In this paper, we model the image set as a manifold. For each manifold, we utilize an effective clustering method to extract a set of clusters, with each cluster being one local linear model [28]. The distance between different manifolds is computed by that between their corresponding LLMs. We utilize the furthest seed point of the maximal linear patch (MLP) [18, 30, 32] to construct the LLMs, which can avoid the problem of unbalanced clusters. Denote by $X = [x_1, x_2, ..., x_i, ...x_n]$ a video of one pedestrian and x_i is the i^{th} image sample in X. Let $\{M_1, M_2, ..., M_i, ..., M_k\}$ $(k < n)$ be the constructed LLMs, where M_i is the i^{th} LLM. The basic idea of constructing LLMs is shown in Fig. 3. Detailed steps of constructing LLMs are as follows:

(1) We firstly compute the distance between geodesic distance matrix D_G and Euclidean distance matrix D_E, which can be used to judge whether two samples are neighbors by (1). The samples x_i and x_j are neighbors if $D_G(x_i, x_j)/D_E(x_i, x_j) \leq \delta$. The threshold δ is set by referring to [28, 30]. A larger δ implies fewer local linear models vice versa. Therefore, δ affects the trade off between efficiency and accuracy [30], which can be written as follow:

$$\frac{D_G(x_i, x_j)}{D_E(x_i, x_j)} \leq \delta. \qquad (1)$$

(2) We select the furthest point x_f away from X_{mean} as a seed point each time shown in Fig. 3a, which can make the different LLMs separate away in some extent. Here,

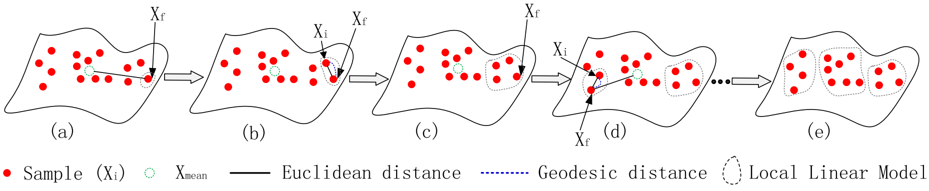


Fig. 3 Illustration of constructing LLMs. Samples are constructed to the corresponding LLMs by the ratio of geodesic distance and Euclidean distance. **a** The distribution of all original samples. X$_{mean}$ represents the center of all samples. X$_f$ denotes the furthest point away from X$_{mean}$. **b** Computing the ratio of Euclidean distance and geodesic distance between X$_i$ and X$_f$. **c** The first LLM has been constructed. **d** Constructing next LLM. **e** All LLMs have been constructed

X_{mean} is the first-order statistics of all samples in X, which can be computed with $X_{mean} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

(3) Grouping each sample of X into the corresponding LLM according to (1), as shown in Fig. 3b–c. The process of constructing the LLMs of each person is summarized as Algorithm 1.

Algorithm 1 Constructing local linear model

```
Input: The data set $X = [x_1, x_2, ..., x_n]$ and the center of all samples $X_{mean}$
Output: The local linear model $\{M_1, ..., M_i, ..., M_k\}, k < n$
    Initialization $M_0 = [x_1, x_2, ..., x_n], k = 1$;
    Repeat
    Step 1: Selecting the furthest point $x_f$ ($x_f \in M_0$) away from $X_{mean}$ as a seed
  point, $M_0 = M_0 - \{x_f\}$;
    Step 2: Updating $M_k = \{x_f\}$;
    Step 3: $\forall x_i \in M_0$, if $x_i$ and $x_f$ satisfy (1), then update $M_k = M_k \cup \{x_i\}$, $M_0 =$
  $M_0 - \{x_i\}$; If all $x_i$ and $x_f$ do not satisfy (1), then $k = k + 1$;
    Until $M_0 = \varnothing$
```

3.2 Triplet-based manifold discriminative distance learning (TMD^{2}L)

As described in the above subsection, the local linear models can be obtained by Algorithm 1. Denote by $M = \{M_1, ..., M_i, ..., M_n\}$ the collection of constructed LLMs of all training pedestrian videos, where n is the LLM number, M_i=$\{y_{i1}, ...y_{ij}, ...y_{in_i}\}$ is the i^{th} LLM. Here, y_{ij} is the j^{th} sample in M_i, and n_i is the sample number of M_i. The contrast of pixels in neighbor area under low illumination is not significant, which will make it difficult to distinguish the samples. To solve the problem of low distinction between persons under low illumination and improve the discriminability of LLMs, we introduce the idea of discriminant distance learning into the problem of low-illumination person re-identification, which can make the same person compact and separate the different persons away. This basic idea has been illustrated in Fig. 4. Therefore, the objective function is defined as:

$$\begin{gathered}\arg\min_{W} \; f(W, M) + \alpha g(W, M) \\ s.t. W^T W = I,\end{gathered} \tag{2}$$

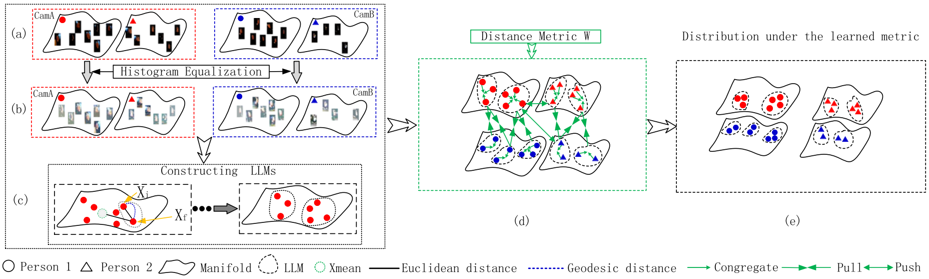


Fig. 4 Conceptual illustration of our approach. **a** Modeling image set of each person as corresponding manifold. **b** Histogram equalization is used to eliminate the illumination effects. **c** Taking one manifold as an example, to show the brief illustration of constructing the LLM. **d** Learning the discriminative metric to maximize the distance of different persons and minimize the distance of same persons after constructing the local linear models by the furthest points based on MLP. **e** Positions of different LLMs under the learned distance metric *W*

where W represents the distance metric to be learned and α is a balancing factor. $f(W, M)$ is the LLM congregating term, which can make the samples within each LLM move close to the center of this LLM, as shown in Fig. 4c. $g(W, M)$ is the triplet-based LLM discriminant term, which is used to make the distance between truly matching LLMs smaller than that between the wrong matching LLMs, as shown in Fig. 4c.

$$f(W, M) = \tfrac{1}{L} \sum_{i=1}^{n} \sum_{j=1}^{n_i} \left\| W^T (y_{ij} - m_i) \right\|_2^2 \tag{3}$$

where L is the number of all samples, n_i is the number of samples in the i^{th} LLM. m_i is the first-order statistics of the i^{th} LLM, which can be computed by

$$m_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}. \tag{4}$$

$g(W, M)$ is designed as follows:

$$g(W, M) = \sum_{<i,j,k>\in\mathcal{T}} d(W, M_i, M_j) - \beta d(W, M_i, M_k). \tag{5}$$

where β is a balancing factor of the discriminant term, $\mathcal{T}$ is the collection of LLM triplets, with each triplet consisting of two truly matching LLMs and an ”impostor” LLM. Here, we employ the similar strategy as [47] to construct LLM triplet. M_i and M_j represents the truly matching pair, and M_i and M_k is the wrong matching pair. $d(W,\mathrm{M_i},\mathrm{M_j})$ is the distance function under the learned distance metric.

$$d(W, M_i, M_j) = \frac{1}{n_i * n_j} \sum_{p=1}^{n_i} \sum_{q=1}^{n_j} \left\| W^T (y_{ip} - y_{jq}) \right\|_2^2. \tag{6}$$

3.3 The optimization of TMD^{2}L

In this subsection, we will describe an efficient solution of (2). Firstly, we simplify $f(W, M)$ as follows:

$$f(W, M) = tr\left(W^T \Sigma_F W\right), \tag{7}$$

where

$$\Sigma_F = \frac{1}{L}\sum_{i=1}^{n}\sum_{j=1}^{n_i}(y_{ij}-m_i)(y_{ij}-m_i)^T. \tag{8}$$

Secondly, we reformulate $g(W, M)$ to the following form:

$$g(W, M) = \sum_{<i,j,k>\in\mathcal{T}} tr\left(W^T(\Sigma_{G1} - \beta\ \Sigma_{G2})W\right), \tag{9}$$

where

$$\Sigma_{G1} = \frac{1}{n_i * n_j}\sum_{p=1}^{n_i}\sum_{q=1}^{n_j}(y_{ip}-y_{jq})(y_{ip}-y_{jq})^T, \tag{10}$$

$$\Sigma_{G2} = \frac{1}{n_i * n_k}\sum_{p=1}^{n_i}\sum_{q=1}^{n_k}(y_{ip}-y_{kq})(y_{ip}-y_{kq})^T. \tag{11}$$

By substituting (7) and (9) into (2), we can rewrite our objective function as:

$$\begin{aligned}&\min_W\ tr\left(W^T\Sigma_F W\right)\\ &+\ \ \alpha \sum_{<i,j,k>\in\mathcal{T}} tr\left(W^T(\Sigma_{G1} - \beta\ \Sigma_{G2})W\right)\\ &\qquad s.t.\ W^TW = I.\end{aligned} \tag{12}$$

Based on (12), we can reformulate our objective function as:

$$\begin{aligned}&\min_W\ tr\left(W^T\left(\mathcal{D}_1 + \mathcal{D}_2\right)W\right)\\ &\quad s.t.\ \ W^TW = I,\end{aligned} \tag{13}$$

where

$$\mathcal{D}_1 = \frac{1}{L}\sum_{i=1}^{n}\sum_{j=1}^{n_i}(y_{ij}-m_i)(y_{ij}-m_i)^T, \tag{14}$$

$$\begin{aligned}\mathcal{D}_2 = \ \alpha \sum_{<i,j,k>\in\mathcal{T}}&\left(\frac{1}{n_i * n_j}\sum_{p=1}^{n_i}\sum_{q=1}^{n_j}(y_{ip}-y_{jq})(y_{ip}-y_{jq})^T\right.\\ &\left.-\frac{\beta}{n_i * n_k}\sum_{p=1}^{n_i}\sum_{q=1}^{n_k}(y_{ip}-y_{kq})(y_{ip}-y_{kq})^T\right).\end{aligned} \tag{15}$$

By constructing the Lagrange function and setting the derivative of (13) w.r.t. W to zero, we can get:

$$(\mathcal{D}_1 + \mathcal{D}_2)W = \lambda W \tag{16}$$

where λ is a Lagrange multiplier. Then, W can be obtained by solving the above eigen-decomposition problem (16). It is clear that the solution of (16) yields the eigenvectors and eigenvalues of $(\mathcal{D}_1 + \mathcal{D}_2)$. The eigenvectors corresponding to the k smallest eigen-values of $(\mathcal{D}_1 + \mathcal{D}_2)$ can be selected as $W = [w_1, w_2, ..., w_k]$. The optimization of TMD^{2}L is summarized in Algorithm 2.

Algorithm 2 Optimization process of TMD^{2}L

```
Input: The video set $X = [x_1, x_2, ..., x_N]$, $N$ is the total number of videos.
Output: The learned eigenvectors $W$.
    Initialize: Initialize metric with the matrix $W = I$;
    Step 1: Constructing the LLMs of the training samples by Algorithm 1;
    Step 2: Calculating the distance of each LLM as $\mathcal{D}_1$ by (14); Computing the
  triplet-based LLMs discriminant distance as $\mathcal{D}_2$ by (15);
    Step 3: Performing eigen-decomposition on $(\mathcal{D}_1 + \mathcal{D}_2)$.
```

3.4 Computational cost

The time complexity of our model TMD^{2}L mainly comes from three phases, especially constructing LLMs, dividing the negative samples and learning the discriminant metric phase. In the phase of constructing LLMs phase, the main time cost is to group the samples of each person into the corresponding LLM, which is $O(N^2)$. In the phase of dividing the negative samples, the time cost focuses on the calculation for Σ_{G1} and Σ_{G2}, which is $O(p \times N_{G1} + p \times N_{G2})$. N_{G1} and N_{G2} are the number of Σ_{G1} and Σ_{G2}, respectively. p is the dimension of features. In the phase of learning the discriminant metric, the time cost is to compute the eigen-decomposition, which is $O(p^3)$. The space complexity focuses on constructing LLMs and computing the eigen-decomposition. The space complexity of constructing LLMs is $O(N * p)$. The space complexity of eigen-decomposition in (16) is $O(N * k) + O(k^2)$. k is the number of eigenvalues.

4 Re-identification

In this section, we elaborate the person re-identification with the LLMs and the learned discriminant metric. The steps of detailed re-identification are as follows:

1) *Constructing local linear models:* For the probe and gallery videos, we construct LLMs using Algorithm 1. Let $P = \{M_p^1, ..., M_p^i, ..., M_p^{n_p}\}$ be a probe video with n_p LLMs, and M_p^i is the i^{th} LLM of P. $G = \{G_1, ..., G_j, ..., G_m\}$ be a set of the m gallery videos, where $G_j = \{M_g^{j1}, ..., M_g^{jk}, ..., M_g^{jn_k}\}$ with n_k LLMs and M_g^{jk} is the k^{th} LLM of G_j.
2) *Computing the distance:* With the learned discriminant metric W, the distance $d(M_p^i, M_g^{jk})$ between the probe LLMs and the LLMs of each gallery is computed by (6) and the smallest distance represents the distance between the probe video and each gallery video.
3) *Re-Identifying the Probe video in Gallery videos:* Sorting the obtained distances, and the gallery video with the smallest distance is the true matching of P. The procedure of our approach TMD^{2}L for matching is summarized in Algorithm 3.

Algorithm 3 TMD^{2}L for Person Re-identification

```
1. Constructing the LLMs
    Constructing the LLMs of each person from the training samples and testing samples
    by Algorithm 1, respectively; The testing samples are divided into the probe sets and
    the gallery sets.
2. Learning discriminant distance metric
    The discriminant distance metric $W$ can be learned by Algorithm 2;
3. Re-identification
    Computing the distance between the probe and gallery sets, and then sorting the dis-
    tance in ascending order, finally the gallery video with the smallest distance is the
    truly matching for the probe video.
```

5 Experiments

To evaluate the effectiveness of our approach, we conduct extensive experiments on three pedestrian video datasets, including the new collected LIPS dataset, and two simulated publicly available datasets(PRID 2011 and iLIDS-VID), as shown in Fig. 6 and Table 1.

5.1 Datasets and settings

In this section, we briefly introduce the collected person re-id dataset under night scene in Wuhan University campus and two publicly available datasets. Table 1 provides a statistical summary of three datasets. We also annotate the attributes of each dataset by the following list: "cameras" denotes the number of different disjoint cameras. "label" means the way of segmenting pedestrians from the original frames by manually or automatically. "total frames" is the total number of persons appearing in both cameras of the corresponding dataset.

(1) **Newly collected low illumination pedestrian sequence dataset:** There exists no publicly available low illumination person re-id dataset. To fill this gap, we contribute a new low illumination pedestrian sequence (LIPS) dataset, which is collected under night scenes in campus of Wuhan University. The flowchart of collecting new pedestrian dataset is shown in Fig. 5. Firstly, we use two cameras to record the videos of pedestrians walking through the road on campus. Secondly, the videos including pedestrians are converted into image sequences. We extract the persons with bounding box from each frame by a self-developed software manually. Finally, we normalize the extracted image into the same size and encode the image sequences corresponding persons.

Table 1 Information of Three Datasets

Dataset	People	Total Frames	Average length	Occlusion degree	Year	Cameras	Label
LIPS	100	90,021	450	Partial	2017	2	hand
LI-iLIDS-VID	300	42,459	100	Partial	2014	2	hand
LI-PRID 2011	200	40,033	73	Few	2011	2	hand

One real night-scene person dataset LIPS, and two simulated datasets (LI-iLIDS-VID and LI-PRID 2011)

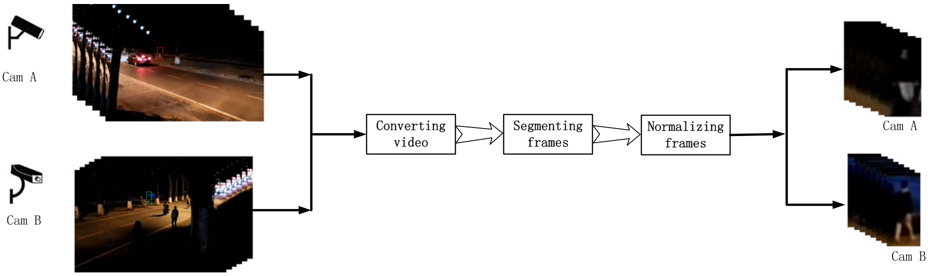


Fig. 5 The flowchart of collecting and normalizing new person dataset at night scene. The original videos of pedestrian are captured by two disjoint cameras. Firstly, we transform the video to image sequences. Secondly, persons with bounding boxes are segmented by manually. Finally, the samples are normalized by the same size, which is 64×128. In the left subgraph, the pedestrian is marked by red(green) rectangle. In the right subgraph, the image sequences belong to the corresponding pedestrian

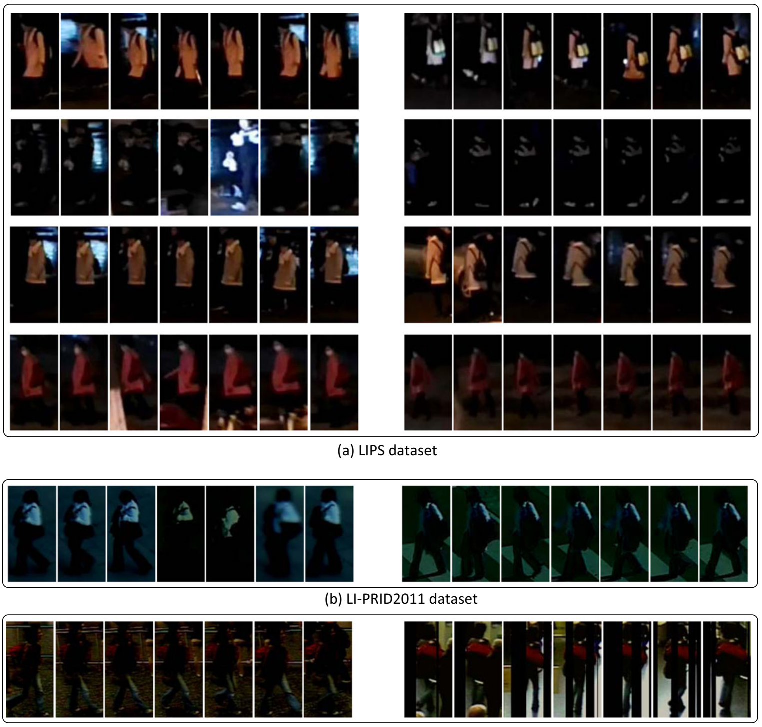
The dataset includes 90021 images of 100 pedestrians captured by two non-overlapping cameras. The length of each image sequence ranges from 193 to 833 frames, with an average number of 450 for each person at each camera. Most images are captured under very low illumination (see Fig. 6a). Many persons are occluded by other objects, e.g., other pedestrians, cars or trees. The illumination of some frames in persons video is severe low or huge changed by the cars light. All image sequences are normalized to 64 × 128 pixels after segmenting the person from each video. The key frames of each person in newly collected dataset are available at.[1]

(2) **Two simulated datasets:** There exist two publicly available video person datasets, whose attributes are shown in Table 1, however they are under normal illumination. For PRID 2011 dataset, we utilize image sequence pairs that have more than 20 frames for evaluation. This dataset was created in co-operation with the Austrian Institute of Technology. In experiments, half of all sequence pairs sets are randomly selected for training, and the remaining sequence pairs are used for testing. The PRID 2011 dataset includes video pairs captured by two different outdoor cameras. 385 persons were recorded in one camera, and 749 persons in the other camera. 245 persons appear in both cameras simultaneously. The number of images in each image sequence ranges 5 to 675 frames, with an average of 100 for each person.

The iLIDS-VID dataset consists of 600 videos. This dataset was captured in an airport arrival hall under a multi-camera CCTV network. Each person has one pair of video from two different cameras. The number of frames in each video ranges from 23 to 192, with an average of 73 for each person. We utilize sequences pairs that have more than 20 frames for evaluation. Then, all sequence pairs are randomly divided into two parts of equal size, with one for training, and the other for testing. The iLIDS-VID dataset is challenging since it involves complicated cluttered background and occlusions.

Simulating night-scene datasets To further verify the evaluation of our approach, inspired by [12, 34], we generate two simulated low-illumination datasets (LI-PRID 2011 and LI-iLIDS-VID) for evaluation, which are based on PRID 2011 [4] and iLIDS-VID [31]. Figure 6b and c show the simulated images from LI-PRID 2011 and LI-iLIDS-VID datasets.

[1] https://sites.google.com/site/whulips



(a) LIPS dataset

(b) LI-PRID2011 dataset

(c) LI-iLIDS-VID dataset

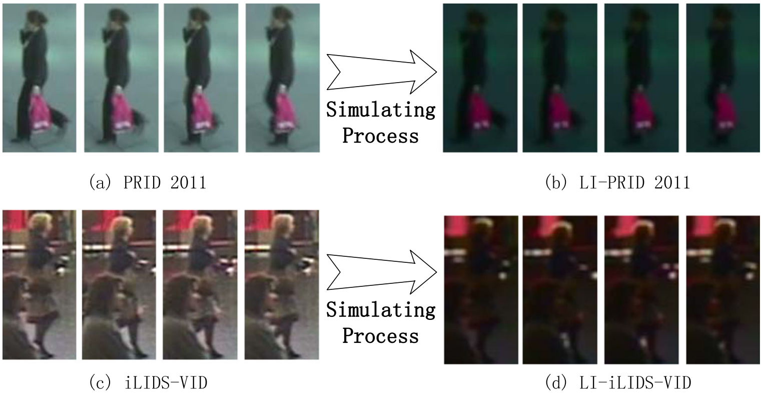
Fig. 6 Example pairs of several key frames on three datasets. The same person walks through two different cameras in each row. **a** LIPS dataset is a newly collected person dataset under low illumination. Two simulated datasets: **b** LI-PRID2011 dataset, and **c** LI-iLIDS-VID dataset

In experiments, the low illumination scene is simulated by using [25]. The procedure of simulating night scene is as follows. I denotes the original normal-illumination image from two existing datasets. T denotes the simulated night-scene image. Firstly, R, G and B denote three channels of color image respectively. X, Y and Z are the temporary variables, which are computed by (17).

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.5149 & 0.3244 & 0.1607 \\ 0.2654 & 0.6704 & 0.0642 \\ 0.0248 & 0.1248 & 0.8504 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}. \tag{17}$$

The scotopic luminance $V = Y\left[1.33(1 + \frac{Y+Z}{X}) - 1.68\right]$. Secondly, we compute the each pixel by $c_T = c_I - kVc_{blue}$. $c_I(c_T)$ presents the pixel of the original image I(the target simulated image T). $k = 0.93$ and $c_{blue} = 0.9$ are set empirically. Finally, we filter the obtained image by Gaussian, where the default value for filter template is 5×5 and the standard deviation is 1.6. The illustration of simulating low-illumination datasets are shown in Fig. 7. The simulated night-scene image samples corresponding to two existing datasets are shown in Fig. 6b–c.

(3) **Parameter Settings**. The important parameters in our model include α, β and δ. α controls the balance of two items. β controls the balance of the triplet including negative sample pair and positive sample pair. δ denotes the ratio of geodesic distance



(a) PRID 2011

(b) LI-PRID 2011

(c) iLIDS-VID

(d) LI-iLIDS-VID

Fig. 7 The original and simulated illumination datasets. The process of simulating night scene is detailed in Simulating night-scene datasets. **a** PRID2011 and **c** iLIDS-VID are the original datasets under normal illumination; **b** LI-PRID2011 and **d** LI-iLIDS-VID are the simulated low-illumination datasets

and Euclidean distance, which controls the number of neighbor points. In particular, we set the balancing factor $\alpha = 0.3$ on LIPS, LI-PRID 2011 and LI-iLIDS-VID. The balancing factor of the discriminant term β is set as 0.6 on LIPS, LI-PRID 2011 and LI-iLIDS-VID. The threshold δ is set as 1.4 for all datasets.

(4) **Compared Methods**. To evaluate the performance of TMD^{2}L, we select seven state-of-the-art related methods as the compared methods, including STFV3D [17], RDC [45], KISSME [9], TDL [40], SI^{2}DL [47], JDML [46] and PHDL [48] . In experiments, we perform competing methods with the codes provided by the authors. In experiment results, we use the average cumulative match characteristic (CMC) curves [29] to show the top ranked matching rates.

(5) **Feature Extraction.** In preparation phase, we conduct histogram equalization [28] on image sequences of each person. The person image (128× 64 pixels) is divided into 128 non-overlapped patches with the patch size 8 × 8. We extract the HSV, LAB and LBP features from each patch and is represented by a 25600-dimensional feature vector. We perform PCA to keep above 90% data energy, and obtain the 600-dimensional feature vector.

5.2 Evaluation on the LIPS dataset

Following the evaluation protocol in [9], we randomly select half of the dataset, i.e., 50 video pairs, for training, and use the remaining 50 pairs for testing. Table 2 and Fig. 8a report the top ranked matching rates of TMD^{2}L and the competing methods on the new LIPS dataset. We can see that our TMD^{2}L achieves better performance than the other methods. In particular, our TMD^{2}L improves the rank 1 matching rate by at least 3.81%(57.81%-54.00%). The main advantages of our proposed method have two folds: 1) TMD^{2}L can preserve local neighborhood structures of the pedestrian data to some extent. Although the entire pixel values of the images are declining severely, there exists a certain intrinsic structure of images in the data space corresponding to different persons. 2) TMD^{2}L learns a discriminative distance metric. The distance metric is an effective technique for person re-identification. The discriminative distance metric contains a better discriminative capability, which can reduce the between-video and within-video variations simultaneously.

Table 2 Top r ranked matching rate (%) on LIPS dataset

Method	$r = 1$	$r = 5$	$r = 10$	$r = 20$
RDC	19.53	46.83	56.34	61.98
STFV3D	23.22	46.23	59.56	75.48
KISSME	26.53	52.96	73.99	84.96
PHDL	44.51	75.62	83.75	92.18
TDL	49.31	86.96	94.91	98.83
SI^{2}DL	53.63	87.84	95.56	97.91
JMDL	54.00	87.50	94.50	98.50
TMD^{2}L	**57.81**	**89.65**	**96.76**	**98.98**

Best results are in boldface font

5.3 Evaluation on the LI-PRID 2011 dataset

We also evaluated the proposed approach on the simulated LI-PRID 2011 dataset. The results in Table 3 and Fig. 8b show that our approach outperforms the other methods. More specifically, our TMD^{2}L improves the average rank 10 matching rate by at least 6.04% (79.98%-73.94%) on LI-PRID 2011. The results are worse than the original. This is because the illumination of most pedestrians is much low. The disadvantages are two folds: 1) The effective information under low illumination is lost severely. Pixel values of pedestrians are too low under low illumination so that the difference of some pixels in each frame are lower than that under normal illumination. The feature information, e.g., texture and gradient, is lost severely. 2) The extreme points of FEP are not obvious. The space-time feature cannot be aligned by the walking cycles correctly. In this case, even some pedestrians' walking cycles would be partitioned wrongly, leading to that the methods based on walking cycles (e.g. STFV3D) perform poorly .

5.4 Evaluation on the LI-iLIDS-VID dataset

In Table 4 and Fig. 8c, we reported the comparison of our TMD^{2}L with five state-of-the-art person re-identification methods on LI-iLIDS-VID dataset. The results show that our TMD^{2}L outperforms the other competing methods. For instance, our proposed TMD^{2}L model improved the average rank 10 matching rate by at least 5.33%(74.06%-68.73%) on

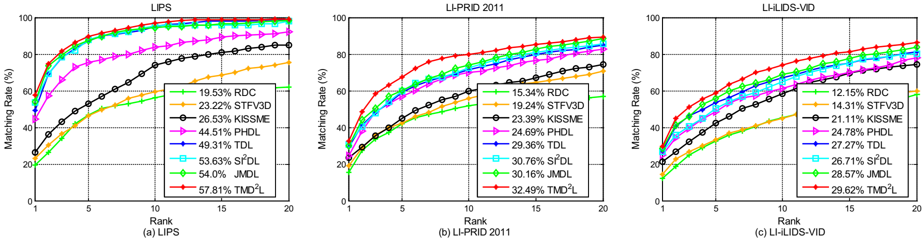


Fig. 8 Performance comparison using CMC curves on three datasets. **a**: LIPS dataset. **b**: LI-PRID2011 dataset. **c**:LI-iLIDS-VID dataset

Table 3 Top *r* ranked matching rate (%) on simulated LI-PRID 2011 dataset

Method	$r = 1$	$r = 5$	$r = 10$	$r = 20$
RDC	15.34	41.98	51.98	56.93
STFV3D	19.24	42.45	55.66	70.56
KISSME	23.39	44.91	59.99	74.43
PHDL	24.69	57.07	69.84	82.79
TDL	29.36	59.11	71.44	84.93
SI^{2}DL	30.76	59.98	72.39	85.63
JMDL	30.16	60.35	73.94	88.39
TMD^{2}L	**32.49**	**67.39**	**79.98**	**89.43**

Best results are in boldface font

LI-iLIDS-VID dataset. The results are worse than the original. This database is very challenging under normal illumination. This is because the many pedestrian in videos were captured with significant background clutter and occluded by other people or objects. The effective feature information corresponding to same person is lost.

5.5 Further analysis of TMD^{2}L

5.5.1 Influence of parameters

There are two important parameters, i.e., the balancing factors α and β, in objective function, we implemented TMD^{2}L model by selecting the parameters on three datasets. In particular, we take the experiment on the LIPS dataset as an example. We conduct experiments by changing the value of α from 0.1 to 1 step by 0.1 and β from 0.01 to 1 step by 0.05. The top ranked matching rates are shown in Fig. 9a and b. As illustrated, the performance of our approach is not very sensitive to the choice of α in the range (0.2, 0.6) and β in the range (0.3, 0.8). When $\alpha = 0.3$ and $\beta = 0.6$, our TMD^{2}L achieved the best result.

On the simulated datasets, we take the simulated LI-PRID 2011 dataset as an example. α is a tuning parameter, which controls the balance of two items in objective function. β controls the balance of the triplet including negative sample pair and positive sample pair.

Table 4 Top *r* ranked matching rate (%) on simulated LI-iLIDS-VID dataset

Method	$r = 1$	$r = 5$	$r = 10$	$r = 20$
RDC	12.15	32.59	45.14	57.91
STFV3D	14.31	33.16	44.87	59.84
KISSME	21.11	42.22	58.22	74.44
PHDL	24.78	48.67	61.20	77.83
TDL	27.27	53.76	67.01	81.29
SI^{2}DL	26.71	50.38	64.93	80.25
JMDL	28.57	55.77	68.73	83.66
TMD^{2}L	**29.62**	**58.75**	**74.06**	**86.25**

Best results are in boldface font

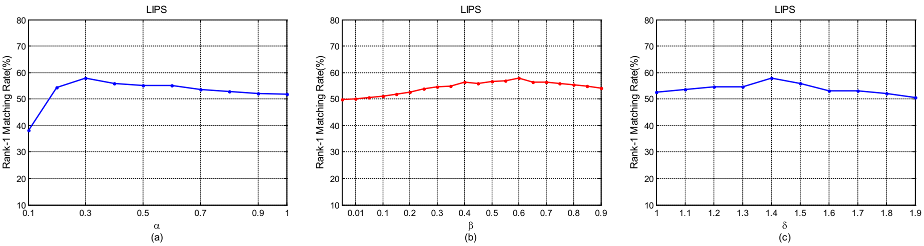


Fig. 9 Rank 1 matching rates versus different values of α, β and δ on LIPS dataset. **a** Performance:α. **b** Performance:β. **c** Performance:δ

We conduct experiments by changing the value of α from 0.1 to 1 step by 0.1 and β from 0.01 to 1 step by 0.05. The results of matching rate are shown in Fig. 10a–b. As illustrated, the performance of our approach is stable to the choice of α in the range (0.2, 0.7) and β in the range (0.4, 0.9). When $\alpha = 0.3$ and $\beta = 0.6$, our TMD^{2}L achieved the best result. Similar conclusions can be observed on the LI-iLIDS-VID dataset.

5.5.2 *Influence of threshold parameter*

In the phase of constructing the LLMs, the geodesic distance is always no smaller than Euclidean distance, therefore the threshold δ is no less than 1 in (1). Here, we take the experiment on the LIPS dataset as an example. We conduct experiments by changing the value of δ from 1 to 2 step by 0.1. When $\delta = 1.4$, our TMD^{2}L achieved the best result as shown in Fig. 9c. As illustrated, the performance of our approach is not very sensitive to the choice of δ in the range (1.2, 1.7).

On the simulated dataset, we take the simulated LI-PRID 2011 dataset as an example. δ denotes the ratio of geodesic distance and Euclidean distance, which controls the number of neighbor points. We conduct experiments by changing the value of δ from 1 to 2 step by 0.1. The results of matching rate are shown in Fig. 10c. As illustrated, the performance of our approach is stable to the choice of δ in the range (1.2, 1.7). When $\delta = 1.4$, our approach TMD^{2}L achieved the highest. Similar conclusions can be observed on the LI-iLIDS-VID dataset.

5.5.3 *Effect of manifold technique*

In this section, we discuss the effect of manifold in our model. The manifold technique can preserve the intrinsic structure of the data and be stable to varying poses and lighting

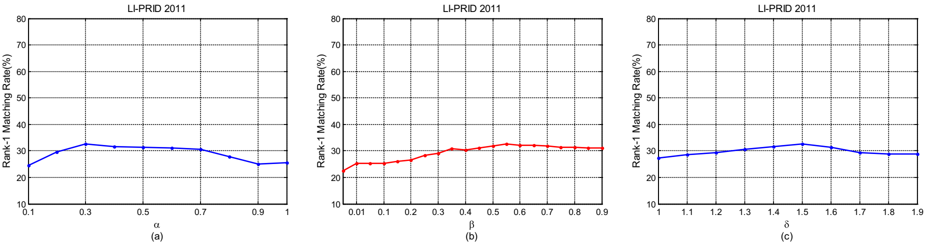


Fig. 10 Rank 1 matching rates versus different values of α, β and δ on LI-PRID 2011 dataset. **a** Performance:α. **b** Performance:β. **c** Performance:δ

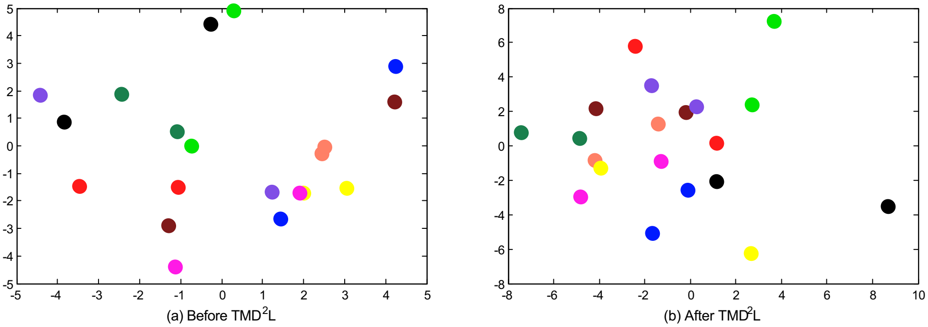


Fig. 11 Illustration of the effectiveness of triplet-based manifold, where ten different persons in LIPS dataset were selected for demonstration. Points with same color denote the same person pair from two cameras. **a** The sample data points in the original 2-D space, and **b** the projected sample data points in the new subspace learned by TMD^{2}L

conditions [28, 35]. Although the low-illumination images contain less effective information than those with normal illumination and lie in non-linear subspace, there exist the intrinsic structures in low-illumination images. Manifold learning technique can preserve local neighborhood structures [26] of the pedestrian data in nonlinear space to some extent. To visualize the effectiveness of triplet-based manifold, a demonstration between the feature distributions of the original feature subspace and the latent feature subspace learned by our approach is shown in Fig. 11. We perform PCA [27] on the features to obtain two major principal components that are used to show the sample distribution. The distribution of samples belonging to the same class is scattered before performing manifold, while the samples belonging to the same class are clustered together with performing manifold for our model. We can see that the original samples of the same person are ambiguous, while our TMD^{2}L can improve the discriminability, and thus the distribution of samples is more favorable for matching.

To demonstrate the effect of manifold in our TMD^{2}L on the matching rates, we remove the local linear model construction out of our model, and reformulate our model as a linear distance learning model. We call the modified method of TMD^{2}L as TMD^{2}L-M. We conduct the experiments on LIPS, LI-PRID 2011 and LI-iLIDS-VID. The results of TMD^{2}L and TMD^{2}L-M are reported in Table 5. One can see that the matching rate decreases significantly without the manifold. Therefore the manifold in our model is effective.

5.5.4 Effect of local model

In this section, we analyze the effect of constructing local models by different clustering methods, e.g., K-means [8] and the furthest point based MLP [30]. K-means algorithm often utilizes Euclidean distance as metric. Our method, i.e.,the furthest point based MLP, gets

Table 5 Top r ranked matching rate (%) of TMD^{2}L and TMD^{2}L-M on three datasets

Methods	LIPS				LI-PRID 2011				LI-iLIDS-VID			
	$r = 1$	$r = 5$	$r = 10$	$r = 20$	$r = 1$	$r = 5$	$r = 10$	$r = 20$	$r = 1$	$r = 5$	$r = 10$	$r = 20$
TMD^{2}L-M	54.73	86.57	94.69	96.86	28.51	64.43	76.86	87.31	25.54	54.69	71.13	84.31
TMD^{2}L	57.81	89.65	96.76	98.98	32.49	67.39	79.98	89.43	29.62	58.75	74.06	86.25

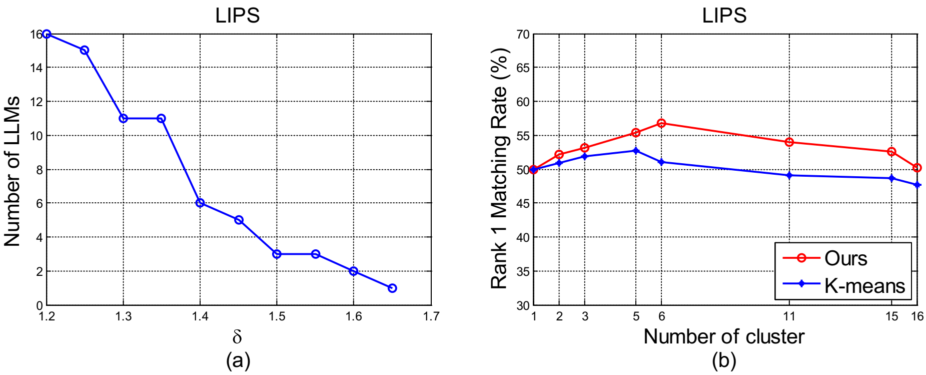


Fig. 12 The effect of local models on the LIPS dataset. **a** The relationship of the number of LLMs and the ratio δ on the LIPS dataset. **b** The number of cluster affects the matching rates with different clustering algorithms

the LLMs by (1), which is the ratio of Euclidean distance and geodesic distance. We take the image set of the first person on LIPS dataset as an example. δ ranges from 1.2 to 1.7. Similar results can be obtained from the other persons. As shown in Fig. 12a, one can see that the larger δ is, the fewer the number of LLMs is. In experiments, the matching rates achieve the highest when $\delta = 1.4$ with the average number of all persons.

The results in Fig. 12b demonstrate that our clustering method outperforms K-means algorithm. We can observe that the matching rates are higher than those of K-means when the number of cluster is bigger than 1. When the number of cluster is 1, we learn the intra projection from the whole image set of each person. Similar conclusions can be obtained from other datasets.

5.5.5 *Effect of low illumination*

In this section, we discuss the effect of low illumination. The effective space-time feature approaches, e.g., DVR and STFV3D, are based on the correct walking cycles, which are extracted by FEP [17, 31]. The energy of FEP is affected by the low illumination. The illumination at night scene is low that causes inaccurate walking cycle extraction of the pedestrians. The error segments of walking cycles cannot align the space-time feature, and

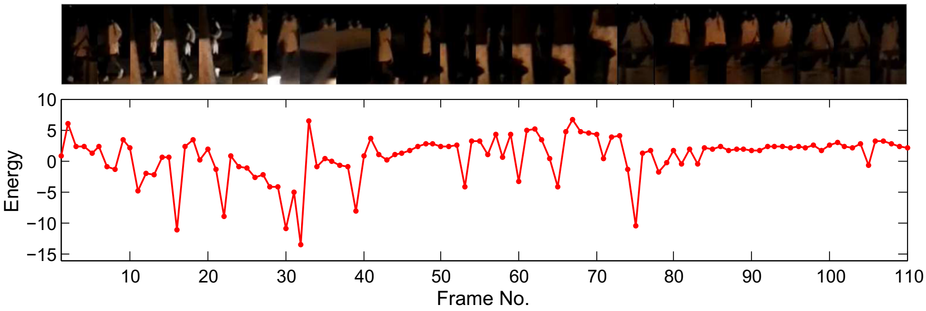


Fig. 13 The key frames versus FEP on LIPS dataset. (*Top Row*): The key frames from the first person on LIPS dataset. (*Bottom Row*): The energy of FEP corresponding to 110 image sequences

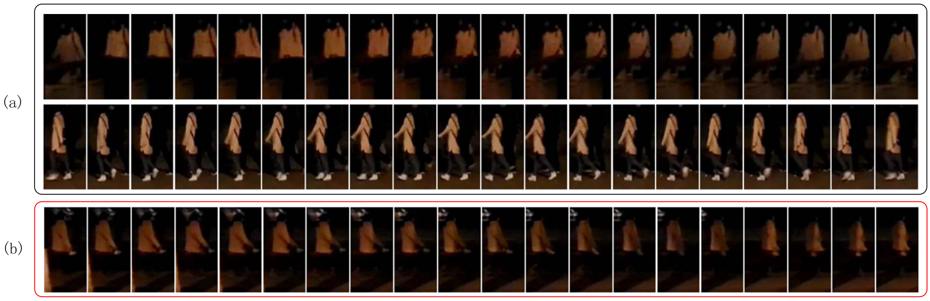


Fig. 14 Failure examples on LIPS dataset. The color of clothes is changing under lamp light at night. **a** One person appears in two cameras. (*First Row*): the image sequences belong to camera A in the probe set for testing. (*Second Row*): the image sequences of the same person belong to camera B in the gallery set. **b** The image sequences of different person belonging to camera B are the wrong matching samples

then it will result in the incorrect matching between the probe video and the gallery videos. Here, we take the experiments on the real scene (LIPS) dataset as an example and analyze the failure examples for incorrect extraction of walking cycles. We selected the first person from LIPS as an example. The result of walking cycles extraction is shown in Fig. 13, which includes five walking cycles. We can see that the low illumination affects the FEP significantly and the energy varies largely, and thus results in inaccurate walking cycles extraction.

5.5.6 Failure examples

In this section, we discuss about the failure examples in experiments. We take the new collected dataset LIPS as an example. As shown in Fig. 14, the color of clothes is changing since one person walks through different lamps. The changing color causes trouble for re-id. As shown in Fig. 14a, one person walks through different cameras in different time periods. As can be seen from Fig. 14b, the color of another person is more similar to the first row of Fig. 14a than that of the second row in Fig. 14a. The reason may be that the color of clothes belonging to different pedestrians is becoming much similar under different lamp light, while there exist much differences in the visual appearance of the same person under the different light source.

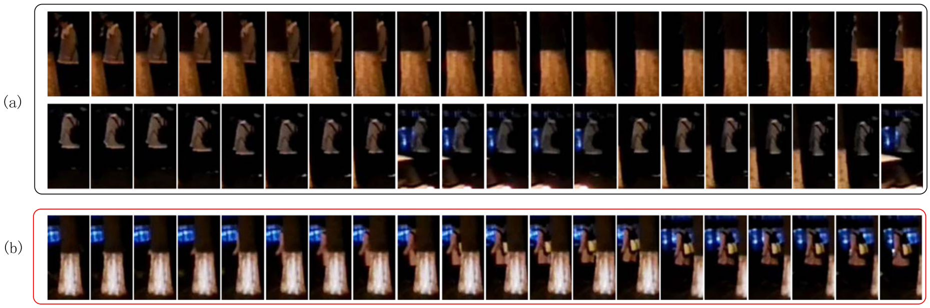


Fig. 15 Failure examples on LIPS dataset. The cluttered background causes inaccurate walking cycle extraction and trouble for matching. **a** One person appears in two cameras. (*First Row*): the image sequences belong to camera A in the probe set for testing. (*Second Row*): the image sequences of other person belong to camera B in the gallery set. **b** The image sequences belonging to camera B are the wrong matching samples

In the second case, the scenario with occlusion and cluttered background is a common problem for person re-id and tracking since the pedestrians are moving under the complex environment. The occlusions (e.g., occluded by trees and cars) result in the severe loss of effective information as shown in Fig. 15, which make it difficult match. And the occlusion will cause inaccurate walking cycle extraction and affect the space-time information for re-id methods based on space-time features. As can be seen from Fig. 15b, the occlusion by tree is severe, which is trouble for re-id.

6 Conclusion

There exists no low illumination pedestrian dataset, to fill the gap, we contribute a new low illumination pedestrian dataset (LIPS) . To address the challenges associated with low illumination video-based person re-id problem, we propose a novel triplet-based manifold discriminative metric learning model. The effectiveness of TMD^{2}L has been demonstrated by extensive experiments on three datasets including a newly collected dataset (LIPS) and two simulated datasets.

Future works include two aspects as follows: 1) There exist some occlusions for pedestrians, e.g., occluded by other persons, trees or cars. Occlusions will lead to the loss of effective information, which can be difficult for person re-id. We will consider extending our model to solve the problem of occlusion for pedestrians. 2) The distance between pedestrian and cameras is different while the pedestrian is walking in the street. The different distance will lead to the different scales of pedestrians' image sequences, which will be a new problem for person re-id. Different scales of images will be difficult for alignment of images belonging to the same person.

References

1. Chen J, Wang Y, Tang YY (2016) Person re-identification by exploiting spatio-temporal cues and multi-view metric learning. IEEE Signal Process Lett 23(7):998–1002
2. Farenzena M, Bazzani L, Perina A, Murino V, Cristani M (2010) Person re-identification by symmetry-driven accumulation of local features. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 2360–2367
3. Gong S, Cristani M, Yan S, Loy CC (2014) Person re-identification. Springer
4. Hirzer M, Beleznai C, Roth PM, Bischof H (2011) Person re-identification by descriptive and discriminative classification. In: Scandinavian conference on image analysis, pp 91–102
5. Hirzer M, Roth P, Koestinger M, Bischof H (2012) Relaxed pairwise learned metric for person re-identification. In: European Conference on computer vision (ECCV), pp 780–793
6. Hu H (2015) Sparse discriminative multimanifold Grassmannian analysis for face recognition with image sets. IEEE Trans Circ Syst Vid Technol 25(10):1599–1611
7. Jing X-Y, Zhu X, Wu F, Hu R, You X, Wang Y, Feng H, Yang J-Y (2017) Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning. IEEE Trans Image Process 26(3):1363–1378
8. Kim T-K, Arandjelovic O, Cipolla R (2007) Boosted manifold principal angles for image set-based recognition. Pattern Recogn 40(9):2475–2484
9. Koestinger M, Koestinger M, Wohlhart P, Roth PM, Bischof H (2012) Large scale metric learning from equivalence constraints. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 2288–2295
10. Kviatkovsky I, Adam A, Rivlin E (2013) Color invariants for person reidentification. IEEE Trans Pattern Anal Mach Intell 35(7):1622–1634
11. Li W, Zhao R, Xiao T, Wang X (2014) Deepreid: deep filter pairing neural network for person re-identification. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 152–159

12. Li X, Zheng W-S, Wang X, Xiang T, Gong S (2015) Multi-scale learning for low-resolution person re-identification. In: IEEE Conference on computer vision (ICCV), pp 3765–3773
13. Liao S, Hu Y, Zhu X, Li SZ (2015) Person re-identification by local maximal occurrence representation and metric learning. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 2197–2206
14. Lin T, Zha H (2008) Riemannian manifold learning. IEEE Trans Pattern Anal Mach Intell 30(5):796–809
15. Lisanti G, Masi I, Bagdanov AD, Del Bimbo A (2015) Person re-identification by iterative re-weighted sparse ranking. IEEE Trans Pattern Anal Mach Intell 37(8):1629–1642
16. Liu H, Qi M, Jiang J (2015) Kernelized relaxed margin components analysis for person re-identification. IEEE Signal Process Lett 22(7):910–914
17. Liu K, Ma B, Zhang W, Huang R (2015) A spatio-temporal appearance representation for viceo-based pedestrian re-identification. In: IEEE Conference on computer vision (ICCV), pp 3810–3818
18. Liu M, Shan S, Wang R, Chen X (2016) Learning expressionlets via universal manifold model for dynamic facial expression recognition. IEEE Trans Image Process 25(12):5920–5932
19. Lu Y, Wang R, Shan S, Chen X (2016) Multiple-shot person re-identification via Riemannian discriminative learning. In: Asian Conference on computer vision. Springer (ACCV), pp 408–425
20. McLaughlin N, Martinez del Rincon J, Miller P (2016) Recurrent convolutional network for video-based person re-identification. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 1325–1334
21. Rao Y, Hou L, Wang Z, Chen L (2014) Illumination-based nighttime video contrast enhancement using genetic algorithm. Multimed Tools Appl 70(3):2235–2254
22. Sun C, Wang D, Lu H (2017) Person re-identification via distance metric learning with latent variables. IEEE Trans Image Process 26(1):23–34
23. Tao D, Li X, Wu X, Maybank SJ (2007) General tensor discriminant analysis and Gabor features for gait recognition. IEEE Trans Pattern Anal Mach Intell 29(10):1700–1715
24. Tao D, Guo Y, Song M, Li Y, Yu Z, Tang YY (2016) Person re-identification by dual-regularized kiss metric learning. IEEE Trans Image Process 25(6):2726–2738
25. Thompson WB, Shirley P, Ferwerda JA (2002) A spatial post-processing algorithm for images of night scenes. J Graph Tools 7(1):1–12
26. Tunç B, Gökmen M (2011) Manifold learning for face recognition under changing illumination. Telecommun Syst 47(3–4):185–195
27. Turk M, Pentland A (1991) Eigenfaces for recognition. J Cogn Neurosci 3(1):71–86
28. Wang R, Chen X (2009) Manifold discriminant analysis. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 429–436
29. Wang X, Doretto G, Sebastian T, Rittscher J, Tu P (2007) Shape and appearance context modeling. In: IEEE Conference on computer vision (ICCV), pp 1–8
30. Wang R, Shan S, Chen X, Gao W (2008) Manifold-manifold distance with application to face recognition based on image set. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 1–8
31. Wang T, Gong S, Zhu X, Wang S (2014) Person re-identification by video ranking. In: European Conference on computer vision (ECCV), pp 688–703
32. Wang W, Wang R, Huang Z, Shan S, Chen X (2015) Discriminant analysis on Riemannian manifold of Gaussian distributions for face recognition with image sets. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 2048–2057
33. Wang F, Zuo W, Lin L, Zhang D, Zhang L (2016) Joint learning of single-image and cross-image representations for person re-identification. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 1288–1296
34. Wang Z, Hu R, Yu Y, Jiang J, Liang C, Wang J (2016) Scale-adaptive low-resolution person re-identification via learning a discriminating surface. IJCAI, pp 2669–2675
35. Wen J, Fowler JE, He M, Zhao Y-Q, Deng C, Menon V (2016) Orthogonal nonnegative matrix factorization combining multiple features for spectral–spatial dimensionality reduction of hyperspectral imagery. IEEE Trans Geosci Remote Sens 54(7):4272–4286
36. Xiao T, Li H, Ouyang W, Wang X (2016) Learning deep feature representations with domain guided dropout for person re-identification. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 1249–1258
37. Xiong F, Gou M, Camps O, Sznaier M (2014) Person re-identification using kernel-based metric learning methods. In: European Conference on computer vision (ECCV), pp 1–16
38. Yan Y, Ni B, Song Z, Ma C, Yan Y, Yang X (2016) Person re-identification via recurrent feature aggregation. In: European Conference on computer vision (ECCV), pp 701–716
39. Yang Y, Yang J, Yan J, Liao S, Yi D, Li SZ (2014) Salient color names for person re-identification. In: European Conference on computer vision (ECCV), pp 536–551

40. You J, Wu A, Li X, Zheng W-S (2016) Top-push video-based person re-identification. In: IEEE Conference on computer vision and pattern recognition (ICCV), pp 1345–1353
41. Zhang R, Lin L, Zhang R, Zuo W, Zhang L (2015) Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification. IEEE Trans Image Process 24(12):4766–4779
42. Zhao R, Ouyang W, Wang X (2013) Person re-identification by salience matching. In: IEEE Conference on computer vision (ICCV), pp 2528–2535
43. Zhao R, Ouyang W, Wang X (2013) Unsupervised salience learning for person re-identification. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 3586–3593
44. Zhao R, Ouyang W, Wang X (2014) Learning mid-level filters for person re-identification. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 144–151
45. Zheng W-S, Gong S, Xiang T (2013) Reidentification by relative distance comparison. IEEE Trans Pattern Anal Mach Intell 35(3):653–668
46. Zhou Q, Zheng S, Ling H, Su H, Wu S (2017) Joint dictionary and metric learning for person re-identification. Pattern Recogn, 196–206
47. Zhu X, Jing X-Y, Wu F, Feng H (2016) Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics. In: International joint conference on artificial intelligence, pp 3552–3559
48. Zhu X, Jing X-Y, Wu F, Wang Y, Zuo W, Zheng W-S (2017) Learning heterogeneous dictionary pair with feature projection matrix for pedestrian video retrieval via single query image. In: Association for the advancement of artificial intelligence (AAAI), pp 4341–4348

Fei Ma received the B.S. degree in computer science and technology from Qufu Normal University, Qufu, China, in 2003, the M.S. degree in computer software and theory from Yunnan Normal University in 2006. He is currently pursuing the Ph.D. degree in pattern recognition and intelligence system from the Wuhan University, Wuhan, China. His research interests include pattern recognition, computer vision, and machine learning.

Xiaoke Zhu received the B.S. degree in computer science and technology from Henan University, Kaifeng, China, in 2004, the M.S. degree in computer application technology from Henan University in 2010. He is currently pursuing the Ph.D. degree in pattern recognition and intelligence system from the Wuhan University, Wuhan, China. His research interests include pattern recognition, computer vision, and machine learning.

Xinyu Zhang is currently pursuing the Ph.D. degree in pattern recognition and intelligence system from the Wuhan University, Wuhan, China. His research interests include pattern recognition, computer vision, and machine learning.

Liang Yang is currently pursuing the M.S. degree in pattern recognition and intelligence system from the Wuhan University, Wuhan, China. His research interests include pattern recognition, computer vision, and machine learning.

Mei Zuo is currently pursuing the M.S. degree in pattern recognition and intelligence system from the Wuhan University, Wuhan, China. His research interests include pattern recognition, computer vision, and machine learning.

Xiao-Yuan Jing received the Doctoral degree of Pattern Recognition and Intelligent System in the Nanjing University of Science and Technology, 1998. He was a Professor with the Department of Computer, Shenzhen Research Student School, Harbin Institute of Technology, 2005. Now he is a Professor with the State Key Laboratory of Software Engineering, School of Computer, Wuhan University, and with the College of Automation, Nanjing University of Posts and Telecommunications, China.