Neurocomputing 290 (2018) 121–129

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Spatial-temporal representatives selection and weighted patch descriptor for person re-identification

Aihua Zheng[a], Foqin Wang[a], Amir Hussain[a,b], Jin Tang[a], Bo Jiang[a,*]

[a] Anhui University, No. 111 Jiulong Road, Hefei, China
[b] University of Stirling, Stirling, FK9 4LA, Scotland, UK

ARTICLE INFO

Article history:
Received 5 April 2017
Revised 25 December 2017
Accepted 7 February 2018
Available online 15 February 2018

Communicated by Dr. K. Li

Keywords:
Multi-shot person re-identification
Informative representatives
Spatial-temporal
Weighted patch descriptor

ABSTRACT

How to represent the sequential person images is a crucial issue in multi-shot person re-identification. In this paper, we propose to select the spatial-temporal informative representatives to describe the image sequence. Specifically, we address representatives selection as a row-sparsity regularized minimization problem which can be effectively solved via convex programming. The sparsity of the representatives is controlled by a regularization parameter based on both spatial and temporal dissimilarities. Furthermore, we design a weighted patch descriptor by employing the random walk with restart model to propagate the patch weights on the person image. Finally, we utilize the cross-view quadratic discriminant analysis as the metric learning to mitigate the cross-view gaps among different cameras. Extensive experiments on three benchmark datasets iLIDS-VID, PRID 2011 and SAIVT-SoftBio demonstrate the promising performance of the proposed method.

1. Introduction

Person Re-identification (Re-ID) aims to recognize the same individual crossing non-overlapping camera networks, which is a crucial step of surveillance systems in modern society. Despite of years of effort, it still faces big challenges due to the occlusions and the cross-view gaps (visual differences while crossing different cameras) caused by the changes of illumination, viewpoint, person pose and so on.

In spite of great achievement on single-shot Re-ID where only a single image is recorded for each person per camera view, the limited information of a single image impedes its performance. Multi-shot Re-ID, where normally the sequential frames are recorded for each person per camera view, is more natural in real-life surveillance systems, and expected to boost the performance of Re-ID. Therefore, we focus on multi-shot Re-ID in this paper.

Despite of richer information in multiple images, there are additional challenges in multi-shot Re-ID. On the one hand, the majority of the sequential frames contain redundant information with similar visual appearance. Therefore, it is crucial to summarize and interpret the person images by informative representatives. Some existing works selected the representatives via clustering. Hassen et al. [1] proposed to select key frames based on Mean-shift clustering [2]. Li et al. [3] designed a Fisher discriminant analysis guided hierarchical clustering method for multi-shot Re-ID. However, they clustered the person images only based on the spatial dissimilarity. To our knowledge, none of the existing methods take into account the temporal relationships between the sequential person images.

On the other hand, the distractors or junks are ubiquitous due to false detection or tracking which generates the bounding boxes as person images for Re-ID. As a result, background clutters and occlusions may corrupt the appearance descriptors of the person and the further learning model for identification. Therefore, it is essential to highlight the person body against the background or occlusions on the person images.

Based on above discussion, we propose a novel spatial-temporal representatives selection (STRS) method based on the weighted patch descriptor for multi-shot Re-ID. The main contribution in this paper can be summarized as follows:

- We propose to select the spatial-temporal representatives for multi-shot Re-ID. Specifically, we employ the row-sparsity regularized minimization program to select the informative representatives based on both spatial and temporal priors.
- In order to suppress undesired background clutters in a bounding box, we design a weighted patch descriptor to enhance the discrimination between the person and background based on the random walk with restart model.

* Corresponding author at: Anhui University, No. 111 Jiulong Road, Hefei, China.
E-mail address: zeyiabc@163.com (B. Jiang).

https://doi.org/10.1016/j.neucom.2018.02.039
0925-2312/

Table 1
Notations.

$\mathbf{X} \in \mathbb{R}^{d\times N}$	A person video
$\mathbf{x}_i \in \mathbb{R}^d$	Feature vector of the i-th frame
N	Number of frames in $\mathbf{X}$
d	Dimensionality of $\mathbf{x}_i$
$\mathbf{X}^S \in \mathbb{R}^{d\times S}$	Representatives selected from $\mathbf{X}$
$\mathbf{D} \in \mathbb{R}^{N\times N}$	Spatial distance matrix
$\mathbf{T} \in \mathbb{R}^{N\times N}$	Temporal distance matrix
$\mathbf{Z} \in \mathbb{R}^{N\times N}$	Indicator matrix of representatives
$\mathbf{\Lambda} \in \mathbb{R}^{N\times N}$	Lagrange multipliers
$\mathbf{f}_i \in \mathbb{R}^d$	Feature vector of the i-th patch
$\mathbf{W} \in \mathbb{R}^{mn\times mn}$	Edge weights on graph of patches
mn	Number of patches
$\mathbf{A} \in \mathbb{R}^{mn\times mn}$	Transition matrix
$\mathbf{r}^{row} \in \mathbb{R}^{mn}$	Restart distributions in rows
$\mathbf{r}^{col} \in \mathbb{R}^{mn}$	Restart distributions in columns
$\mathbf{r} \in \mathbb{R}^{mn}$	Restart distributions
$\pi \in \mathbb{R}^{mn}$	Weights of the patches
$\mathbf{X}^a \in \mathbb{R}^{d\times N_a}$	Selected representatives of person video from camera a
$\mathbf{X}^b \in \mathbb{R}^{d\times N_b}$	Selected representatives of person video under camera b
n_a	Number of persons under camera a
n_b	Number of persons under camera b
N_a	Number of all selected representatives under camera a
N_b	Number of all selected representatives under camera b
$\mathbf{H} \in \mathbb{R}^{d\times r}$	Subspace projection matrix
y_i	Gallery label of the i-th person
l_i	Probe label of the i-th person
ψ_i	Number of selected representatives of the i-th person under camera a
τ_i	Number of selected representatives of the i-th person under camera b

The rest of this paper is organized as follows. Section 2 reviews the literatures on multi-shot Re-ID. Section 3 elaborates the proposed approach, followed by the evaluation of our approach comparing with the state-of-the-arts in Section 4. Finally, Section 5 concludes our paper.

First of all, the notation definition in the following sections can be referred as Table 1.

2. Related work

The traditional paradigm of Re-ID falls into two research paths: (1) appearance modeling to leverage the various changes and occlusions between cameras and, (2) learning models to mitigate the appearance gaps between the low-level features and high-level semantics.

2.1. Appearance modeling

Appearance modeling on single-shot Re-ID has been well explored in the past decade. Liu et al. [4] and Zhao et al. [5] selected discriminative features which adaptively exploited features based on the person appearance. Liao et al. [6] designed a local maximal occurrence descriptor for person Re-ID. Shi et al. [7] encoded the person via horizontal stripes in multi-level to capture both visual cues and spatial structure. A straightforward strategy to employ the single-shot appearance models on the image sequences in the multi-shot Re-ID task is the averaging pooling. However, appearance modeling for multi-shot Re-ID concerns more on temporal aspect. Gheissari et al. [8] developed a spatiotemporal segmentation algorithm to provide a structural information with stable appearance invariance for person images. Farenzena et al. [9] proposed symmetry-driven accumulation of local features based on the distribution rules of a human body. Cheng et al. [10] designed a visual descriptor, named Custom Pictorial Structure (CPS), to learn the appearance of an person by improving the localization of its parts from multiple images. Bedagkar-Gala and Shah [11] designed an adaptive part-based spatiotemporal model based on the color and facial features to characterize person appearance. Bazzani et al. [12] statistically explored the global chromatic and local patch appearance on the informative image set of a person. Bak et al. [13] and Wu et al. [14] used the pose priors to solve pose variation based on two strict assumptions. However, they modeled the appearance of the person images on the whole bounding boxes, while the background clutters and occlusions may corrupt the appearance descriptors and the forthcoming learning models.

2.2. Learning models

The pioneer learning models for single-shot Re-ID include KISSME [15], LMNNR [16] and ITML [17]. Recently, Liao et al. [6] designed a cross-view quadratic discriminant analysis to learn the metric on the derived subspace. You et al. [18] proposed a top-push distance learning model which enforced the optimization to select more discriminative features to distinguish persons. The naive way to extend such single-shot learning methods to the multi-shot case is to evaluate every possible image pair as the training or testing set and aggregate the results. Recently, some methods specifically handle the multi-shot learning problem. Cong et al. [19] introduced a graph-based approach to learn the manifold structure while preserving the properties of the video sequences lower dimensional subspace for Re-ID. Simonnet et at. [20] explored the multi-shot Re-ID as the temporal sequence matching via dynamic time warping (DTW). Zhang et al. [21] measured the similarity between the image subsets by introducing an energy-based loss function. Pedagadi et al. [22] embedded the image features into a lower dimension space via Local Fisher Discriminant Analysis (LFDA) [23] for multi-shot Re-ID with more training samples. Li et al. [24] proposed to train a random forest within pairwise constraints to address the multi-shot Re-ID in the reduced random projection subspace. Li et al. [25] proposed to learn the local metric field by exploring the discriminative potentiality of a new set-to-set distance. Wang et al. [26,27] automatically selected discriminative video fragments and simultaneously learnt a video ranking Re-ID. However, most of existing methods explored the sequential person images on the entire sequence, where the redundant information among the adjacent frames may bias the learning models.

3. Our approach

In this paper, we propose a novel spatial-temporal representatives selection based on the weighted patch descriptor for multi-shot Re-ID. Our approach consists of three steps. First, we select the informative spatial-temporal representatives from the sequential person images via the row-sparsity regularized minimization optimization problem. Second, we construct a weighted patch descriptor based on the random walk with restart model [28,29] to suppress the undesirable background clutters and occlusions. Finally, we peform multi-shot Re-ID using Cross-view Quadratic Discriminant Analysis (XQDA) method [6] to mitigate the cross-view gaps among the non-overlapping cameras.

3.1. Spatial-temporal representatives selection

We shall elaborate our spatial-temporal representatives selection (STRS) method in this section.

3.1.1. Model formulation

Given $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N] \in \mathbb{R}^{d\times N}$ as the sequential person images containing N frames in d-dimensional feature space, and $\mathbf{D}_{ij}$ as the nonnegative spatial distance (dissimilarity) between frame $\mathbf{x}_i$ and $\mathbf{x}_j$, we consider the problem of selecting a few representatives $\mathbf{X}^S \subseteq \mathbf{X}$ that can efficiently describe the whole image sequence $\mathbf{X}$ based on the dissimilarities $\mathbf{D}_{ij}$ between images. Let $\mathbf{Z}_{ij}$ indicate the probability of representing the i-th frame by the j-th frame. From the probability perspective, we expect $\mathbf{Z}_{ij}$ to be nonnegative together with the nature of $\sum_{i=1}^{N}\mathbf{Z}_{ij} = 1$ [30,31]. Therefore, we construct the formulation as,

$$\begin{aligned}\min_{\mathbf{Z}} \quad & \sum_{i=1}^{N}\sum_{j=1}^{N}\mathbf{D}_{ij}\mathbf{Z}_{ij} + \lambda\|\mathbf{Z}\|_{\mathbf{0,2}} \\ s.t. \quad & \mathbf{1}^\top\mathbf{Z} = \mathbf{1}^\top, \mathbf{Z} \geq \mathbf{0}\end{aligned} \tag{1}$$

where $\|\mathbf{Z}\|_{0,2} = \sum_{i=1}^{N}\|\sqrt{\sum_{j=1}^{N}\mathbf{Z}_{ij}^2}\|_0$. The first term reflects the total cost of encoding the person images by selected representatives while the second term, which denotes the number of non-zero columns in matrix $\mathbf{Z}$, enforces the sparsity of the representatives. λ is the trade-off parameter controls the sparsity of the representatives. The equality constraint $\mathbf{1}^\top\mathbf{Z} = \mathbf{1}^\top$ together with nonnegative constraint $\mathbf{Z} \geq \mathbf{0}$ guarantee the probability nature of $\mathbf{Z}_{ij}$.

Since $\|\mathbf{Z}\|_{0,2}$ is hard to enforce, one popular way is to approximate $l_{0,2} - norm$ by $l_{1,2} - norm$ [30,31]. Therefore, Eq. (1) can be rewritten as,

$$\begin{aligned}\min_{\mathbf{Z}} \quad & \sum_{i=1}^{N}\sum_{j=1}^{N}\mathbf{D}_{ij}\mathbf{Z}_{ij} + \lambda\|\mathbf{Z}\|_{\mathbf{1,2}} \\ s.t. \quad & \mathbf{1}^\top\mathbf{Z} = \mathbf{1}^\top, \mathbf{Z} \geq \mathbf{0}\end{aligned} \tag{2}$$

where $\|\mathbf{Z}\|_{1,2} = \sum_{i=1}^{N}\|\sqrt{\sum_{j=1}^{N}\mathbf{Z}_{ij}^2}\|_1$.

The above model has been successfully explored to select the representatives based on the content dissimilarities between video frames [30,31]. However, despite of the spatial dissimilarity, sequential frames are highly correlated along temporal shifting, which plays an important role in data mining and learning areas [32,33]. As we observed, adjacent person images are generally with high similarity. Therefore, in addition to the spatial dissimilarity, we further expect to select the temporally sparse representatives. Let $\mathbf{T}_{ij} = |j - i|$ be the temporal dissimilarity between frame $\mathbf{x}_i$ and $\mathbf{x}_j$, we propose to incorporate temporal relationship in representatives selection. Based on the above discussion, we formulate our Spatial-temporal Representatives Selection (STRS) problem as,

$$\begin{aligned}\min_{\mathbf{Z}} \quad & \sum_{i=1}^{N}\sum_{j=1}^{N}\mathbf{D}_{ij}\mathbf{Z}_{ij} + \alpha\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbf{T}_{ij}\mathbf{Z}_{ij} + \lambda\|\mathbf{Z}\|_{\mathbf{1,2}} \\ s.t. \quad & \mathbf{1}^\top\mathbf{Z} = \mathbf{1}^\top, \mathbf{Z} \geq \mathbf{0}\end{aligned} \tag{3}$$

where α is the balance parameter to leverage the contribution of the temporal aspect.

Using matrix operation, Eq. (3) can be formulated more compactly as,

$$\begin{aligned}\min_{\mathbf{Z}} \quad & \mathbf{Tr}(\mathbf{D}^\top\mathbf{Z}) + \alpha\mathbf{Tr}(\mathbf{T}^\top\mathbf{Z}) + \lambda\|\mathbf{Z}\|_{\mathbf{1,2}} \\ s.t. \quad & \mathbf{1}^\top\mathbf{Z} = \mathbf{1}^\top, \mathbf{Z} \geq \mathbf{0}\end{aligned} \tag{4}$$

where $\mathbf{Tr}(\cdot)$ denotes the trace operator.

3.1.2. Optimization

Our STRS in Eq. (4) is a convex problem. The global optimal solution can be efficiently computed using the following Alternating Direct Method of Multipliers (ADMM) algorithm [34,35].

We first rewrite the problem of Eq. (4) as,

$$\begin{aligned}\min_{\mathbf{Z},\mathbf{Y}} \quad & \mathbf{Tr}(\mathbf{D}^\top\mathbf{Y}) + \alpha\mathbf{Tr}(\mathbf{T}^\top\mathbf{Y}) + \lambda\|\mathbf{Z}\|_{\mathbf{1,2}} \\ s.t. \quad & \mathbf{1}^\top\mathbf{Y} = \mathbf{1}^\top, \mathbf{Y} \geq \mathbf{0}, \mathbf{Z} = \mathbf{Y}\end{aligned} \tag{5}$$

Then, ADMM [34,35] solves a sequence of sub-problems as,

$$\begin{aligned}\min_{\mathbf{Z},\mathbf{Y}} \quad & \mathbf{Tr}(\mathbf{D}^\top\mathbf{Y}) + \alpha\mathbf{Tr}(\mathbf{T}^\top\mathbf{Y}) + \lambda\|\mathbf{Z}\|_{\mathbf{1,2}} + \frac{\mu}{\mathbf{2}}\|\mathbf{Z} - \mathbf{Y}\|_{\mathbf{F}}^{\mathbf{2}} + \langle\mathbf{\Lambda}, \mathbf{Z} - \mathbf{Y}\rangle \\ s.t. \quad & \mathbf{1}^\top\mathbf{Y} = \mathbf{1}^\top, \mathbf{Y} \geq \mathbf{0}\end{aligned} \tag{6}$$

where $\langle\mathbf{P}, \mathbf{B}\rangle = \mathbf{Tr}(\mathbf{P}^\top\mathbf{B})$, $\mathbf{\Lambda}$ is Lagrange multipliers, μ is a penalty parameter.

There are two main parts of the whole ADMM algorithm, i.e., solving the sub-problems (Step 1 and Step 2) and updating parameters (Step 3).

Step 1. Solving $\mathbf{Y}$ while fixing $\mathbf{Z}$. The problem becomes,

$$\begin{aligned}\min_{\mathbf{Z},\mathbf{Y}} \quad & \mathbf{Tr}(\mathbf{D}^\top\mathbf{Y}) + \alpha\mathbf{Tr}(\mathbf{T}^\top\mathbf{Y}) + \frac{\mu}{2}\|\mathbf{Z} - \mathbf{Y}\|_F^2 + \langle\mathbf{\Lambda}, \mathbf{Z} - \mathbf{Y}\rangle \\ s.t. \quad & \mathbf{1}^\top\mathbf{Y} = \mathbf{1}^\top, \mathbf{Y} \geq \mathbf{0}\end{aligned} \tag{7}$$

This problem can effective solved as discussed in [31].

Step 2. Solving $\mathbf{Z}$ while fixing $\mathbf{Y}$. The problem becomes,

$$\min_{\mathbf{Z},\mathbf{Y}} \quad \lambda\|\mathbf{Z}\|_{1,2} + \frac{\mu}{2}\|\mathbf{Z} - \mathbf{Y}\|_F^2 + \langle\mathbf{\Lambda}, \mathbf{Z} - \mathbf{Y}\rangle \tag{8}$$

which is equivalent to,

$$\min_{\mathbf{Z},\mathbf{Y}} \quad \lambda\|\mathbf{Z}\|_{1,2} + \frac{\mu}{2}\|\mathbf{Z} - \mathbf{Y} + \frac{1}{\mu}\mathbf{\Lambda}\|_{\mathbf{F}}^{\mathbf{2}} \tag{9}$$

The optimal $\mathbf{Z}^*$ is given by,

$$\mathbf{Z}^* = \Theta_{\frac{\mu}{\lambda}}\left(\mathbf{Y} - \frac{1}{\mu}\mathbf{\Lambda}\right) \tag{10}$$

where Θ is the $\ell_{2,1}$ minimization operator [36].

Step 3. Updating parameters $\mathbf{\Lambda}$ and μ as,

$$\begin{aligned}\mathbf{\Lambda} &\Leftarrow \mathbf{\Lambda} + \mu(\mathbf{Z} - \mathbf{Y}) \\ \mu &\Leftarrow \rho\mu\end{aligned} \tag{11}$$

where $\rho > 1$.

The algorithm iteratively conducts Step 1–Step 3 until convergence. The complete algorithm is summarized in Algorithm 1. Given the sequential image sequence of a person, we can construct the selected representatives $\mathbf{X}^S \subseteq \mathbf{X}$ according to the indices of the nonzero rows of $\mathbf{Z}$.

Algorithm 1 Optimization procedure of STRS in Eq. (4).

Input: **D**, **T**

Initialize $\mathbf{Y}=\mathbf{Z}=\mathbf{I}$, $\mathbf{\Lambda}=\mathbf{0}$, set $\lambda=0.1$, $\alpha=0.9$, $\mu=10^{-1}$.

Output: **Z**

1: **while** not converges **do**

2: Solve **Y** while fixing **Z** as,
$\min_{\mathbf{Z},\mathbf{Y}} \ \mathbf{Tr}(\mathbf{D}^\top\mathbf{Y})+\alpha\mathbf{Tr}(\mathbf{T}^\top\mathbf{Y})+\frac{\mu}{2}\|\mathbf{Z}-\mathbf{Y}\|_F^2+\langle\mathbf{\Lambda},\mathbf{Z}-\mathbf{Y}\rangle$
$s.t. \quad \mathbf{1}^\top\mathbf{Y}=\mathbf{1}^\top, \mathbf{Y}\geq\mathbf{0}$

3: Solve **Z** while fixing **Y** as,
$\min_{\mathbf{Z},\mathbf{Y}} \ \lambda\|\mathbf{Z}\|_{1,2}+\frac{\mu}{2}\|\mathbf{Z}-\mathbf{Y}\|_F^2+\langle\mathbf{\Lambda},\mathbf{Z}-\mathbf{Y}\rangle$

4: Update parameters $\mathbf{\Lambda}$, μ as,
$\mathbf{\Lambda}\Leftarrow\mathbf{\Lambda}+\mu(\mathbf{Z}-\mathbf{Y})$, $\mu\Leftarrow\rho\mu$

5: **end while**

3.2. Weighted patch descriptor

Although above model can select the spatial-temporal representatives from person images, a single person image itself, which is commonly represented by a rectangular bounding box, contains both person body and background clutters. These background clusters usually lead to inaccurate dissimilarity computation and thus degrade our representatives selection results. In order to highlight the person area while suppress the background or occlusion area on the person image, we design a weighted patch descriptor for representatives selection and the forthcoming metric learning. First, we employ the random walk with restart model [28,29] to propagate the patch weights across different patches of image. Then, we incorporate the obtained patch weights into a patch-based feature descriptor, local maximal occurrence [6] to obtain a kind of robust weighted feature descriptor.

3.2.1. Patch weights propagation

Given the person image with $m\times n$ patches, we construct a graph $\mathbf{G}(\mathbf{V},\mathbf{E})$ whose nodes represent $m\times n$ patches and edges denote the relationship among patches. If nodes/patches $v_i, v_j\in\mathbf{V}, i,j=1,\ldots,mn$, are 8-neighbors, the corresponding edge weight $w_{ij}\in\mathbf{W}$, which reflects the similarity between patches v_i and v_j is defined as,

$$w_{ij}=\exp(-\gamma\|f_i-f_j\|^2) \tag{12}$$

where γ indicates a scaling parameter, f_i and f_j are the feature vectors of patch v_i and v_j. According to the random walk with restart model [28,29], the probability that a walker moves from node v_i to node v_j is normalized as,

$$a_{ij}=\frac{w_{ij}}{\sum_i w_{ij}} \tag{13}$$

where $a_{ij}\in\mathbf{A}$ is the transition matrix.

It is noted that the patches around the center of the image have higher probabilities to belong to the person body [29]. Based on this observation, we define the restart distribution as,

$$\mathbf{r}=((1-\beta)\mathbf{r}^{row}+\beta\mathbf{r}^{col})/2 \tag{14}$$

where β is the hyper-parameter, $\mathbf{r}^{row}$ and $\mathbf{r}^{col}$ indicate the restart distributions in rows and columns respectively. As illustrated in Fig. 1(a), given $\mathbf{\Omega}^{bnd}$ and $\mathbf{\Omega}^{in}$ as the boundary patch set and inner patch set of the image respectively, $\mathbf{r}^{row}$ and $\mathbf{r}^{col}$ are defined as,

$$\mathbf{r}_i^{row}=\begin{cases} Dis(\mathbf{f}_i,\mathbf{f}_i^{row}), & v_i\in\mathbf{\Omega}^{in},\\ 0, & v_i\in\mathbf{\Omega}^{bnd}.\end{cases} \tag{15}$$

$$\mathbf{r}_i^{col}=\begin{cases} Dis(\mathbf{f}_i,\mathbf{f}_i^{col}), & v_i\in\mathbf{\Omega}^{in},\\ 0 \quad , & v_i\in\mathbf{\Omega}^{bnd}.\end{cases} \tag{16}$$

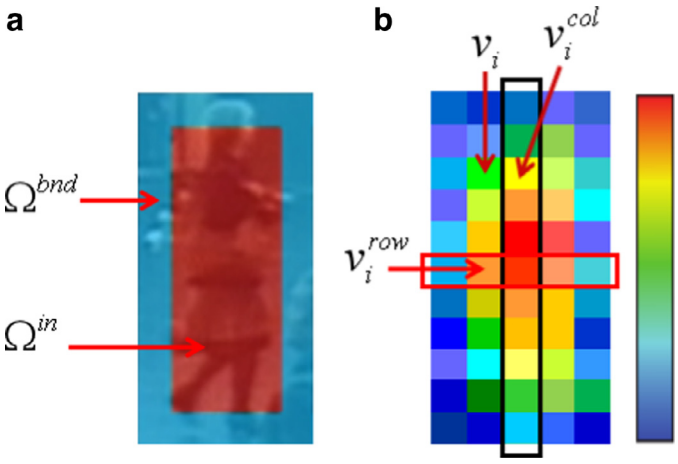


Fig. 1. Illustration of restart distributions in rows and columns. (a) The boundary patch set and inner patch set and (b) illustration of patches v_i^{row} and v_i^{col}.

where $Dis(\mathbf{f}_i,\mathbf{f}_i^{row})$ (or $Dis(\mathbf{f}_i,\mathbf{f}_i^{col})$) denotes the Euclidian distance between the feature vectors of patch v_i and the patch v_i^{row} of the same column (or v_i^{col} of the same row) as patch v_i along the center patch row (or column), as illustrated in Fig. 1(b).

The key idea of random walk with restart (RWR) model [28,29] is that the walker is forced to return to specified nodes based on a restart distribution **r**. Formally, given **r**, RWR iteratively updates the current probability distribution $\boldsymbol{\pi}^t$ as follows,

$$\boldsymbol{\pi}^{t+1}\Leftarrow\epsilon\mathbf{A}\boldsymbol{\pi}^t+(1-\epsilon)\mathbf{r} \tag{17}$$

where $(1-\epsilon)$ is the restart probability. The converged distribution $\boldsymbol{\pi}^*$ satisfies,

$$\boldsymbol{\pi}^*=\epsilon\mathbf{A}\boldsymbol{\pi}^*+(1-\epsilon)\mathbf{r} \tag{18}$$

Eq. (18) can be effectively solved as,

$$\boldsymbol{\pi}^*=(1-\epsilon)(\mathbf{I}-\epsilon\mathbf{A})^{-1}\mathbf{r} \tag{19}$$

We render $\boldsymbol{\pi}^*$ as the weights of corresponding patches. Fig. 2 visualizes some example results of the calculated patch weights. From which we can see, it can help us enhance the person body (higher weights) and suppress the influence of the background or occlusions (lower weights), which is expected to guide a more effective matching.

3.2.2. Weighted feature extraction

Once obtained the patch weights $\boldsymbol{\pi}^*=[\pi_1,\ldots,\pi_{mn}]$, we construct the weighted feature descriptor by integrating the weights into a patch-based feature descriptor, local maximal occurrence [6]. Specifically, given $\mathbf{f}_i$ as the feature vector of the i-th patch on a person image which is extracted in the same manner as [6], we construct the weighted patch descriptor as $[\pi_1\mathbf{f}_1,\ldots,\pi_{mn}\mathbf{f}_{mn}]$. Furthermore, a three-scale pyramid scheme is considered via down-sampling and local average pooling operation (more details refer to [6]). The final weighted descriptor $\mathbf{x}_i$ of each image has 26,960 dimensions. Noted that, the representatives selection is based on the weighted descriptor $\mathbf{x}_i, i=1,\ldots,N$.

3.3. Metric learning for representatives based multi-shot Re-ID

After selecting the informative representatives on the weighted patch descriptors, a metric learning step is essential to mitigate the cross-view gaps caused by the changes such as illumination, viewpoint, person pose and so on while crossing different cameras. Different from the verification problem which judges

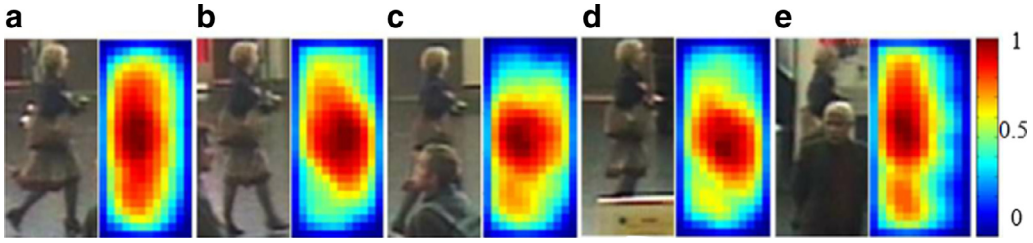


Fig. 2. The examples of weights visualization. (a) Clean background, (b) less occlusion, (c) medium occlusion, (d) background occlusion and (e) Larger occlusion. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

whether the image pair from two non-overlapping cameras derives from the same person or not via binary classification methods [37–39], the common strategy for Re-ID is to learn a metric distance function to rank the gallery images in one camera view against each probe image from another camera view. In this paper, we employ the Cross-view Quadratic Discriminant Analysis (XQDA) method [6] as the metric learning for Re-ID. Let $\{\mathbf{X}^a, \mathbf{X}^b\}$, where $\mathbf{X}^a = [\mathbf{X}_1^{S_a}, \mathbf{X}_2^{S_a}, \ldots, \mathbf{X}_{n_a}^{S_a}] \in \mathbb{R}^{d\times N_a}$ and $\mathbf{X}^b = [\mathbf{X}_1^{S_b}, \mathbf{X}_2^{S_b}, \ldots, \mathbf{X}_{n_b}^{S_b}] \in \mathbb{R}^{d\times N_b}$ denote the cross-view training set of N_a and N_b selected representatives of n_a and n_b persons from camera a and camera b respectively. They can be further expanded as: $\mathbf{X}^a = [\mathbf{x}_1^{S_a}, \mathbf{x}_2^{S_a}, \ldots, \mathbf{x}_{N_a}^{S_a}]$ and $\mathbf{X}^b = [\mathbf{x}_1^{S_b}, \mathbf{x}_2^{S_b}, \ldots, \mathbf{x}_{N_b}^{S_b}]$, while $\mathbf{x}_i^{S_a}, \mathbf{x}_i^{S_b} \in \mathbb{R}^d$ denoting the weighted feature vector of the i-th selected representative in camera a and camera b respectively. Due to the unreliable result and inefficient procedure caused by the large dimensional features [40], XQDA [6] learns a subspace $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_r] \in \mathbb{R}^{d\times r}$, and the distance function in the r-dimensional subspace for the cross-view dissimilarity measurement.

$$d_{\mathbf{H}}(\mathbf{X}^a, \mathbf{X}^b) = (\mathbf{X}^a - \mathbf{X}^b)^T\mathbf{H}(\mathbf{\Sigma}_I^{'-1} - \mathbf{\Sigma}_E^{'-1})\mathbf{H}^T(\mathbf{X}^a - \mathbf{X}^b), \tag{20}$$

where $\mathbf{\Sigma}_I^{'-1} = \mathbf{H}^T\mathbf{\Sigma}_I\mathbf{H}$ and $\mathbf{\Sigma}_E^{'-1} = \mathbf{H}^T\mathbf{\Sigma}_E\mathbf{H}$, $\mathbf{\Sigma}_I$ and $\mathbf{\Sigma}_E$ are the covariance matrices of the intrapersonal variations Ω_I and the extrapersonal variations Ω_E. The subspace $\mathbf{H}$ projection matrix $\mathbf{h}$ can be given to solve Eq. (20) due to the difficulty to directly optimize $d_{\mathbf{H}}$. We can optimize Eq. (20) via,

$$J(\mathbf{h}) = \frac{\mathbf{h}^T\mathbf{\Sigma}_E\mathbf{h}}{\mathbf{h}^T\mathbf{\Sigma}_I\mathbf{h}} \tag{21}$$

Then, the solution of $J(\mathbf{h})$ is given as,

$$\max_{\mathbf{h}} \mathbf{h}^T\mathbf{\Sigma}_E\mathbf{h} \quad s.t.\ \mathbf{h}^T\mathbf{\Sigma}_I\mathbf{h} = 1 \tag{22}$$

The computation of $\mathbf{\Sigma}_I$ and $\mathbf{\Sigma}_E$ is provided as follows,

$$n_I\Sigma_I = \bar{\mathbf{X}}^a\bar{\mathbf{X}}^{a\top} + \bar{\mathbf{X}}^b\bar{\mathbf{X}}^{b\top} - \mathbf{Q}\mathbf{G}^\top - \mathbf{G}\mathbf{Q}^\top,$$
$$n_E\Sigma_E = N_b\mathbf{X}^a\mathbf{X}^{a\top} + N_a\mathbf{X}^b\mathbf{X}^{b\top} - \mathbf{S}\mathbf{R}^\top - \mathbf{R}\mathbf{S}^\top - n_I\Sigma_I, \tag{23}$$

where $\bar{\mathbf{X}}^a = (\sqrt{\psi_1}\mathbf{x}_1^{S_a}, \sqrt{\psi_1}\mathbf{x}_2^{S_a}, \ldots, \sqrt{\psi_1}\mathbf{x}_{\tau_1}^{S_a}, \ldots, \sqrt{\psi_{n_a}}\mathbf{x}_{N_a}^{S_a})$, $\bar{\mathbf{X}}^b = (\sqrt{\tau_1}\mathbf{x}_1^{S_b}, \sqrt{\tau_1}\mathbf{x}_2^{S_b}, \ldots, \sqrt{\tau_1}\mathbf{x}_{\psi_1}^{S_b}, \ldots, \sqrt{\tau_{n_b}}\mathbf{x}_{N_b}^{S_b})$, $\mathbf{Q} = (\sum_{y_1}\mathbf{x}_i^{S_a}, \sum_{y_2}\mathbf{x}_i^{S_a}, \ldots, \sum_{y_i}\mathbf{x}_i^{S_a}, \ldots, \sum_{y_{n_a}}\mathbf{x}_i^{S_a})$, $\mathbf{G} = (\sum_{l_1}\mathbf{x}_j^{S_b}, \sum_{l_2}\mathbf{x}_j^{S_b}, \ldots, \sum_{l_j}\mathbf{x}_j^{S_b}, \ldots, \sum_{l_{n_b}}\mathbf{x}_j^{S_b})$, $\mathbf{S} = \sum_{i=1}^{N_a}\mathbf{x}_i^{S_a}$, $\mathbf{T} = \sum_{j=1}^{N_b}\mathbf{x}_j^{S_b}$. y_i and l_j are the gallery and probe labels, respectively. $\psi_i = |\mathbf{X}_i^{S_a}|$ and $\tau_i = |\mathbf{X}_i^{S_b}|$ denote the number of the selected representatives of the i-th person from $\mathbf{X}^a$ and $\mathbf{X}^b$, respectively.

4. Experimental results

We evaluate our method on three benchmark datasets including iLIDS-VID [26], PRID 2011 [41] and SAIVT-SoftBio [42] comparing to the state-of-the-art algorithms for multi-shot Re-ID. We use the standard measurement named Cumulative Match Characteristic (CMC) curve to figure out the matching results, where the matching rate at rank-n indicate the percentage of correct matchings in top n candidates according to the learnt distance function Eq. (20).

4.1. Experiment setup

4.1.1. Datasets

iLIDS-VID [26] is created from the pedestrians observed in two non-overlapping camera views from the iLIDS Multiple-Camera Tracking Scenario (MCTS), which was captured at an airport arrival hall under a multi-camera CCTV network. It consists of 600 image sequences for 300 randomly sampled people. The length of each image sequence varies from from 23 to 192 frames, with an average number of 73. It is a very challenging dataset due to large viewpoint and illumination variations, occlusions and similar clothing among person across cameras.

PRID 2011 [41] consists of 400 image sequences of 200 outdoor persons in two adjacent cameras. The length of each image sequence varies from 5 to 675 image frames, with an average number of 100. Following the protocol in [3,42,43], we only evaluate 178 persons with length >21 frames. The images in this dataset involve viewpoint, illumination, and background variations. Compared to iLIDS-VID dataset, it was captured with clean background and rare occlusions.

SAIVT-SoftBio [42] consists of 152 persons captured from eight surveilance cameras in a building environment. Since not every person appears in each camera view, following the literatures [42,43], we select cameras 3/8 including 99 person pairs with similar viewpoints and cameras 5/8 including 103 person pairs with large viewpoint changes. The length of the image sequence in selected camera pairs varies from 10 to 992, with average number of 200 frames. Images captured from camera pair 5/8 are more challenging than those from cameras 3/8 due to the larger viewpoint changes.

4.1.2. Parameters

For iLIDS-VID and PRID 2011, We randomly select half of the persons as training and the other half as testing. For SAIVT-SoftBio, following the principle in [43], we randomly select one third samples as training and the remnant as testing. There are five important parameters in our method. During weight calculation, the scaling parameter γ controls the similarity between patches, ϵ controls the proportion of initial weight referring to the restart distributions, while β is the hyper-parameter determining the proportion of the restart distributions in rows and columns. During spatial-temporal representatives selection, λ is the trade-off parameter controlling the sparsity of the representatives. Smaller λ, more representatives will generate. α is the trade-off between dissimilarity matrix $\mathbf{D}$ and temporal matrix $\mathbf{T}$. We adjust one parameter while fixing other parameters and then obtain the best performance for our approach. The parameters are empirically set as: $\{\gamma, \epsilon, \beta, \lambda, \alpha\} = \{0.004, 0.1, 0.85, 0.1, 0.9\}$.

Table 2
Comparison results on iLIDS-VID and PRID 2011 (in %).

Dataset	iLIDS-VID				PRID 2011				
Rank Method	1	5	10	20	1	5	10	20	Reference
SDALF	6.3	18.8	27.1	37.3	5.2	20.7	32.0	47.9	2010 CVPR [9]
Salience	10.2	24.8	35.5	52.9	25.8	43.6	52.6	62.0	2013 CVPR [44]
RankSVM	18.6	43.3	57.1	71.2	22.4	51.9	66.8	80.7	2002 SIGKDD [45]
RPFR	14.5	29.8	40.7	58.1	19.3	38.4	51.6	68.1	2015 WACV [24]
LFDA	21.1	34.8	41.3	48.7	22.3	41.7	51.6	62.0	2006 ICML [23]
SRID	24.9	44.5	55.6	66.2	35.1	59.4	69.8	79.7	2015 CVPRW [46]
DVDL	25.9	48.2	57.3	68.9	40.6	69.7	77.8	85.6	2015 ICCV [47]
AFDA	37.5	62.7	73.0	81.8	43.0	72.7	84.6	91.9	2015 BMVC [3]
DVR	39.5	61.1	71.7	81.0	40.0	71.7	84.5	92.2	2016 TPAMI [27]
OURS	**64.5**	**86.8**	**93.4**	**97.3**	**84.2**	**96.3**	**98.3**	**99.7**	Proposed

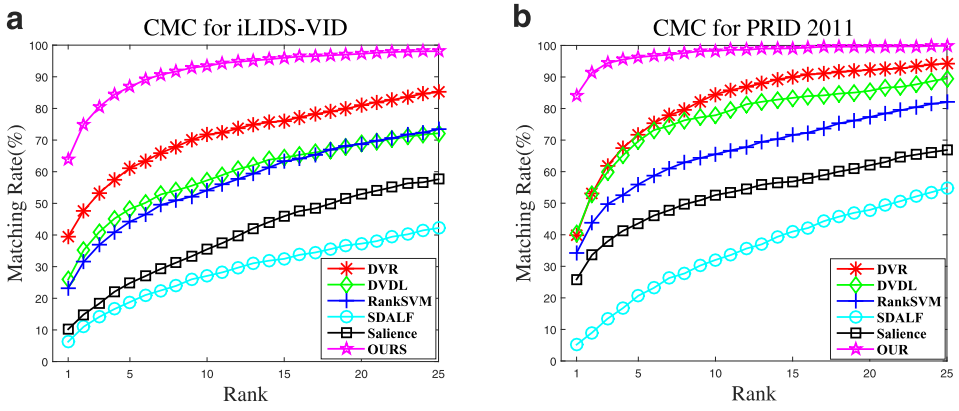


Fig. 3. The cumulative match characteristic curves on iLIDS-VID and PRID 2011 in comparison with the state-of-the-arts.

Table 3
Comparison results on SAIVT-SoftBio (in %).

Dataset	SAIVT-SoftBio(Cameras 3/8)				SAIVT-SoftBio(Cameras 5/8)				
Rank Method	1	5	10	20	1	5	10	20	Reference
LFDA	12.2	36.8	54.6	74.9	9.3	27.1	41.2	60.6	2006 ICML [23]
RankSVM	32.4	68.4	82.0	92.9	14.9	40.5	57.9	75.0	2002 SIGKDD [45]
PFDS	33.2	60.5	74.0	87.2	18.6	32.9	53.0	85.3	2014 ICPR [43]
Fused	36.4	60.3	76.0	87.6	20.0	33.0	50.4	67.8	2012 DICTA [42]
AFDA	43.0	72.7	84.6	91.9	30.9	61.6	77.3	91.1	2015 BMVC [3]
OURS	**83.4**	**98.8**	**99.6**	**99.9**	**75.7**	**90.8**	**95.8**	**97.1**	Proposed

4.2. Evaluations on benchmarks

The performance of the proposed approach on the three benchmark datasets comparing with the state-of-the-art algorithms is reported in this section.

iLIDS-VID. The comparison results on iLIDS-VID dataset is reported in Table 2 and Fig. 3(a). As we can see, our approach achieves the best performance. Specifically, the Rank 1 and Rank 5 performances of ours are 64.5% and 86.80%, respectively, where as the second best results are 39.5% and 61.1%, respectively.

PRID 2011. The results on PRID 2011 dataset are shown in Table 2 and Fig. 3(b). Compared to the iLIDS-VID dataset, this dataset is easier to achieve better performance due to the relatively more clean background and fewer occlusions. Our approach can achieve 84.2% by Rank 1 which is almost twice of the second best method AFDA [3].

SAIVT-SoftBio. The results on SAIVT-SoftBio dataset are shown in Table 3. We adopt the same experimental protocols as Fused [42] and PFDS [43]. Clearly, our approach significantly outperforms the state-of-the-art algorithms. Specifically, in Cameras 3/8 case, the Rank 1 and Rank 5 have reached 83.4% and 98.8%. While for the more challenging case, Cameras 5/8, the results of Rank 1 and Rank 5 are 75.7% and 90.8%, respectively.

Table 4
Parameter evaluation on iLIDS-VID (in %).

Param	Setting	Rank1	Param	Setting	Rank1	Param	Setting	Rank1
γ	0.003	64.1	β	0.8	63.9	λ	0.05	61.0
	0.004	64.5		0.85	64.5		0.1	64.5
	0.005	64.4		0.9	64.3		0.2	64.0
ϵ	0.05	63.7				α	1.0	64.0
	0.1	64.5					0.9	64.5
	0.15	63.0					0.8	64.0

4.3. Component analysis

In order to evaluate the component contribution of our method, we evaluate the component of the weighted patch descriptor and the spatial-temporal representative selection. Fig. 4 reports the component analysis. Generally speaking: (1) Spatial-temporal representatives selection outperforms either spatial or temporal representatives selection on both original feature and the weighted

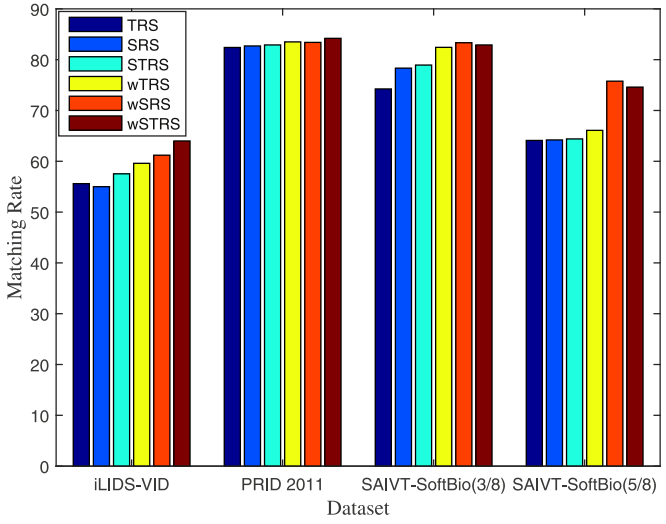


Fig. 4. Component analysis of proposed method. *SRS*, *TRS* and *STRS* denote Spatial Representatives Selection (by setting $\alpha = 0$), Temporal Representatives Selection (by setting $\mathbf{D}_{ij} = 0$) and Spatial-temporal Representatives Selection, respectively, on original features (by setting all $\pi_i = 1$). ***wSRS***, ***wTRS*** and ***wSTRS*** denote ***SRS***, ***TRS*** and ***STRS*** on the weighted patch features, respectively.

feature, which indicates the contribution of taking into account of both spatial and temporal dissimilarities. (2) wSTRS, wSRS and wTRS outperforms STRS, SRS and TRS respectively, which implies the benefit of the weighted patch descriptor. (3) It seems that neither patch weights nor temporal information achieves distinct improvement on PRID 2011, the reason might be the images in PRID 2011 were captured with relatively clean background and rare occlusions. In a sense, our approach is more competitive for complex environments.

4.4. Parameter evaluation

Table 4 reports the parameter evaluation. Generally speaking, our method is not sensitive to the parameters. The most significant parameter is λ which controls the sparsity of the representatives and significantly effects the performance of our method. As demonstrated in Fig. 5(b), we can achieve more sparse representatives and better performance by introducing the temporal aspect to Fig. 5(a). One may suggest to reduce the number of representatives by increasing λ instead. Although we can achieve more sparse representatives by increasing λ from 0.1 to 0.2 as shown in Fig. 5(c), the representatives are however not informative enough and therefore the matching rate is not competitive enough as the case in Fig. 5(b) by introducing the temporal aspect.

5. Conclusion

We have proposed a novel spatial-temporal representatives selection model for multi-shot person re-identification. The informative representatives are selected for each person based on their spatial and temporal dissimilarities. A convex objective function is formulated to find the optimal solution. Furthermore, we have designed a weighted patch descriptor by employing the random walk with restart weight propagation on the local maximal occurrence descriptor. Experimental results on the benchmark datasets demonstrate the superior performance of the proposed model. Our future work will focus on exploring the temporal consistency between the person images and the neural network based pattern recognition methods [48,49] for Re-ID.

Acknowledgment

This study was funded by the National Natural Science Foundation of China (61502006, 61602001, 61572030), Shenzhen Innovation Program (JCYJ20150401145529008) and the Natural Science Foundation of Anhui Province (1508085QF127, 1708085QF139).

References

[1] Y.H. Hassen, W. Ayedi, T. Ouni, Multi-shot person re-identification approach based key frame selection, in: Proceedings of International Conference on Machine Vision (ICMV), 2015, p. 98751H.
[2] Y. Cheng, Mean shift, mode seeking, and clustering, IEEE Trans. Pattern Anal. Mach. Intell. 17 (8) (1995) 790–799.
[3] Y. Li, Z. Wu, S. Karanam, R.J. Radke, Multi-shot human re-identification using adaptive fisher discriminant analysis, in: Proceedings of British Machine Vision Conference (BMVC), 2015, pp. 73.1–73.12.
[4] C. Liu, S. Gong, C.C. Loy, X. Lin, Person re-identification: what features are important? in: Proceedings of European Conference on Computer Vision (ECCV), 2012, pp. 391–401.
[5] R. Zhao, W. Ouyang, X. Wang, Learning mid-level filters for person re-identification, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 144–151.
[6] S. Liao, Y. Hu, X. Zhu, S.Z. Li, Person re-identification by local maximal occurrence representation and metric learning, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2197–2206.
[7] S.C. Shi, C.C. Guo, J.H. Lai, S.Z. Chen, X.J. Hu, Person re-identification with multi-level adaptive correspondence models, Neurocomputing 168 (C) (2015) 550–559.

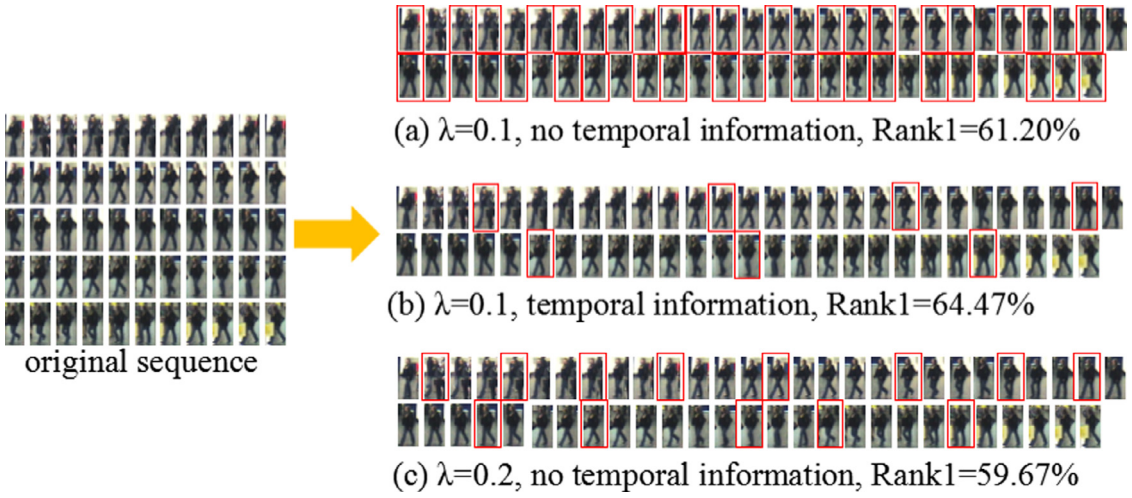
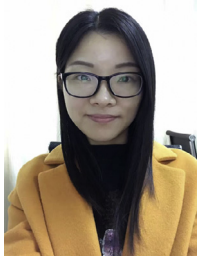
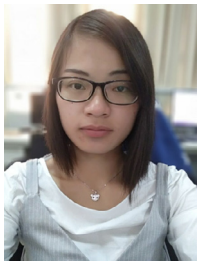


Fig. 5. The examples of selected representatives against λ. The frames with the red bounding boxes indicate the selected representatives. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

[8] N. Gheissari, T.B. Sebastian, R. Hartley, Person reidentification using spatiotemporal appearance, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2006, pp. 1528–1535.
[9] M. Farenzena, L. Bazzani, A. Perina, Person re-identification by symmetry–driven accumulation of local features, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010, pp. 2360–2367.
[10] D.S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, V. Murino, Custom pictorial structures for re-identification, in: Proceedings of British Machine Vision Conference (BMVC), 2011, pp. 68.1–68.11.
[11] A. Bedagkar-Gala, S.K. Shah, Part-based spatio-temporal model for multi-person re-identification, Pattern Recognit. Lett. 33 (14) (2012) 1908–1915.
[12] L. Bazzani, M. Cristani, A. Perina, V. Murino, Multiple-shot person re-identification by chromatic and epitomic analyses, Pattern Recognit. Lett. 33 (7) (2012) 898–903.
[13] S. Bąk, F. Martins, F. Brémond, Person re-identification by pose priors, Proceedings of IS&T/SPIE Electronic Imaging (2015) 93990H.
[14] Z. Wu, Y. Li, R.J. Radke, Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features, IEEE Trans. Pattern Anal. Mach. Intell. 37 (5) (2015) 1095–1108.
[15] M. Köstinger, M. Hirzer, P. Wohlhart, P.M. Roth, H. Bischof, Large scale metric learning from equivalence constraints, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 2288–2295.
[16] M. Dikmen, E. Akbas, T.S. Huang, N. Ahuja, Pedestrian recognition with a learned metric, in: Proceedings of Asian Conference on Computer Vision (ACCV), 2010, pp. 501–512.
[17] J.V. Davis, B. Kulis, P. Jain, S. Sra, I.S. Dhillon, Information-theoretic metric learning, in: Proceedings of International Conference on Machine Learning (ICML), 2007, pp. 209–216.
[18] J. You, A. Wu, X. Li, W.S. Zheng, Top-push video-based person re-identification, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1345–1353.
[19] D.N.T. Cong, C. Achard, L. Khoudour, L. Douadi, Video sequences association for people re-identification across multiple non-overlapping cameras, in: Proceedings of International Conference on Image Analysis and Processing (ICIAP), 2009, pp. 179–189.
[20] D. Simonnet, M. Lewandowski, S.A. Velastin, J. Orwell, E. Turkbeyler, Re-identification of pedestrians in crowds using dynamic time warping, in: Proceedings of IEEE International Conference on Computer Vision (ICCV), 2012, pp. 423–432.
[21] G. Zhang, Y. Wang, J. Kato, T. Marutani, K. Mase, Local distance comparison for multiple-shot people re-identification, in: Proceedings of Asian Conference on Computer Vision (ACCV), 2012, pp. 677–690.
[22] S. Pedagadi, J. Orwell, S. Velastin, B. Boghossian, Local fisher discriminant analysis for pedestrian re-identification, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3318–3325.
[23] M. Sugiyama, T. Idé, S. Nakajima, J. Sese, Local fisher discriminant analysis for supervised dimensionality reduction, in: Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), 2006, pp. 333–344.
[24] Y. Li, Z. Wu, R.J. Radke, Multi-shot re-identification with random-projection-based random forests, in: Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV), 2015, pp. 373–380.
[25] W. Li, Y. Wu, M. Mukunoki, M. Minoh, Locality based discriminative measure for multiple-shot person re-identification, in: Proceedings of IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2015, pp. 312–317.
[26] T. Wang, S. Gong, X. Zhu, S. Wang, Person re-identification by video ranking, in: Proceedings of European Conference on Computer Vision (ECCV), 2014, pp. 688–703.
[27] T. Wang, S. Gong, X. Zhu, S. Wang, Person re-identification by discriminative selection in video ranking, IEEE Trans. Pattern Anal. Mach. Intell. 38 (12) (2016) 2501–2514.
[28] C. Oh, B. Ham, K. Sohn, Probabilistic correspondence matching using random walk with restart, in: Proceedings of British Machine Vision Conference (BMVC), 2012, pp. 1–10.
[29] H.-U. Kim, D.-Y. Lee, J.-Y. Sim, C.-S. Kim, Sowp: spatially ordered and weighted patch descriptor for visual tracking, in: Proceedings of IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3011–3019.
[30] E. Elhamifar, G. Sapiro, R. Vidal, Finding exemplars from pairwise dissimilarities via simultaneous sparse recovery, in: Proceedings of Advances in Neural Information Processing Systems (NIPS), 2012, pp. 19–27.
[31] E. Elhamifar, G. Sapiro, S. Sastry, Dissimilarity-based sparse subset selection, IEEE Trans. Pattern Anal. Mach. Intell. 11 (38) (2014) 2182–2197.
[32] L. Shao, R. Gao, Y. Liu, H. Zhang, Transform based spatio-temporal descriptors for human action recognition, Neurocomputing 74 (6) (2011) 962–973.
[33] E. Fuchs, T. Gruber, H. Pree, B. Sick, Temporal data mining using shape space representations of time series, Neurocomputing 74 (1–3) (2010) 379–393.
[34] D. Gabay, B. Mercier, A dual algorithm for the solution of nonlinear variational problems via finite element approximation, Comput. Math. Appl. 2 (1) (1976) 17–40.
[35] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, Found. Trends Mach. Learn. 3 (1) (2010) 1–122.
[36] G. Liu, Z. Lin, Y. Yu, Robust subspace segmentation by low-rank representation, in: Proceedings of International Conference on Machine Learning (ICML), 2010, pp. 663–670.
[37] H. Faris, I. Aljarah, S. Mirjalili, Training feedforward neural networks using multi-verse optimizer for binary classification problems, Appl. Intell. 45 (2) (2016) 1–11.
[38] R.S. Nickerson, Binary-classification reaction time: a review of some studies of human information-processingcapabilities, Psychon. Monogr. Suppl. 4 (17) (1972) 275–318.
[39] R.W. Proctor, Y.S. Cho, Polarity correspondence: a general principle for performance of speeded binary classification tasks, Psychol. Bull. 132 (3) (2006) 416–442.
[40] D.-S. Huang, W. Jiang, A general cpl-ads methodology for fixing dynamic parameters in dual environments, IEEE Trans. Syst. Man Cybern. Part B (Cybern.) 42 (5) (2012) 1489–1500.
[41] M. Hirzer, C. Beleznai, P.M. Roth, H. Bischof, Person re-identification by descriptive and discriminative classification, in: Proceedings of Scandinavian Conference on Image Analysis (SCIA), 2011, pp. 91–102.
[42] A. Bialkowski, S. Denman, S. Sridharan, C. Fookes, P. Lucey, A database for person re-identification in multi-camera surveillance networks, in: Proceedings of Digital Image Computing Techniques and Applications (DICTA), 2012, pp. 1–8.
[43] J. Garcia, N. Martinel, G.L. Foresti, A. Gardel, C. Micheloni, Person orientation and feature distances boost re-identification, in: Proceedings of International Conference on Pattern Recognition (ICPR), 2014, pp. 4618–4623.
[44] R. Zhao, W. Ouyang, X. Wang, Unsupervised salience learning for person re-identification, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3586–3593.
[45] T. Joachims, Optimizing search engines using clickthrough data, in: Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2002, pp. 133–142.
[46] S. Karanam, Y. Li, R. Radke, Sparse re-id: block sparsity for person re-identification, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 33–40.
[47] S. Karanam, Y. Li, R.J. Radke, Person re-identification with discriminatively trained viewpoint invariant dictionaries, in: Proceedings of IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4516–4524.
[48] D.-S. Huang, Systematic Theory of Neural Networks for Pattern Recognition, Publishing House of Electronic Industry, China, Beijing, 1996, pp. 70–78.
[49] H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei, W. Zheng, S.Z. Li, Embedding deep metric for person re-identification: a study against large variations, in: Proceedings of European Conference on Computer Vision (ECCV), 2016, pp. 1–17.



Aihua Zheng received her B. Eng. degrees and finished her Master-Doctor combined program in computer science and technology from Anhui University of China in 2006 and 2008, respectively. And received her Ph.D. degree in computer science from the University of Greenwich of UK in 2012. She is currently a Lecturer in Anhui University. Her main research areas are visual based signal processing and pattern recognition.



Foqin Wang received her B. Eng. degree in computer science and technology in 2016 from Anhui University, Hefei, China. She is currently pursuing the M.S. degree in computer science and technology at Anhui University. Her current research is person re-identification.

Amir Hussain received the B. Eng. degree and the Ph.D. degree in Electronic & Electrical Engineering from University of Strathclyde, Scotland, UK, in 1992 and 1996, respectively. He is a Professor in Computing Science, School of Natural Sciences, University of Stirling, Scotland, UK. His research interests include cognitive computation, machine learning and computer vision.

Jin Tang received the B. Eng. degree in automation and the Ph.D. degree in computer science from Anhui University, Hefei, China, in 1999 and 2007, respectively. He is a Professor with the School of Computer Science and Technology, Anhui University. His research interests include computer vision, pattern recognition, and machine learning.

Bo Jiang received the B.S. degrees in mathematics and applied mathematics and the M. Eng. and Ph.D. degrees in computer science from Anhui University of China in 2009, 2012 and 2015, respectively. He is currently an associated professor in computer science at Anhui University. His current research interests include image feature extraction and matching, data representation and learning.