SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •
Special Focus on Intelligent City and Big Data

October 2015, Vol. 58 100104:1–100104:15
doi: 10.1007/s11432-015-5404-9

Weight-based sparse coding for multi-shot person re-identification

ZHENG YanWei[1,2], SHENG Hao[1,2*], ZHANG BeiChen[1],
ZHANG Jun[3] & XIONG Zhang[1]

[1]*State Key Laboratory of Software Development Environment, School of Computer Science and Engineering, Beihang University, Beijing* 100191*, China;*
[2]*Shenzhen Key Laboratory of Data Vitalization, Research Institute in Shenzhen, Beihang University, Shenzhen* 518057*, China;*
[3]*Department of Electrical Engineering and Computer Science, University of Wisconsin-Milwaukee, Milwaukee* 53201*, USA*

Received June 30, 2015; accepted July 29, 2015; published online August 27, 2015

Abstract Person re-identification (Re-ID) is the problem of matching a person from different cameras based on appearance. It has interesting algorithm challenges and extensive practical applications. This paper presents a weight-based sparse coding approach for person re-identification. First, three hypotheses are introduced to achieve a linear combination of images based on sparse coding. Then, we convert the person re-identification problem into an optimization problem with sparse constraints. To reduce the influence of abnormal residuals caused by occlusion and body variation, a weight-based sparse coding approach is proposed to achieve the optimal weights by the ordering statistics of square residuals iteratively. Experiments on various public datasets for different multi-shot modalities have shown good performance of the proposed approach compared with other state-of-the-art ones (more than 42% and 34% at rank-1 on CAVIAR4REID and i-LIDS, respectively).

Keywords smart city, person re-identification, video surveillance, multi-shot, weight-based sparse coding

Citation Zheng Y W, Sheng H, Zhang B C, et al. Weight-based sparse coding for multi-shot person re-identification. Sci China Inf Sci, 2015, 58: 100104(15), doi: 10.1007/s11432-015-5404-9

1 Introduction

In recent years, the research and practice on smart cities have been developed rapidly. For a smart city, abundant large-scale and heterogeneous data is required to be analyzed [1]. As an efficient paradigm, data vitalization is proposed [2]. To verify this paradigm, a reference model for data service oriented architecture is developed [3]. In that model, heterogeneous data is fused in Supporting Layer for Data Vitalization Service, which includes text, image, audio, video, and geographic information data. In a real smart city system, most of the data is video data, while most of the video data is surveillance data [4]. In surveillance scenarios where long-term activities need to be modeled within a large and structured

* Corresponding author (email: shenghao@buaa.edu.cn)

environment, an important problem is to recognize a person at a different camera when the person has been previously observed at another camera [5]. This refers to person re-identification (Re-ID). Unlike passive biometrics such as face and gait, Re-ID relies on the overall appearance of individuals [6], making it a challenge task due to the changes of illumination, pose, viewpoint, background clutter, occlusion, and image resolution [7].

Currently, research on the Re-ID problem mostly focuses on two aspects: (1) Developing feature representations that are discriminative for identification and invariant to viewpoint and illumination. (2) Developing machine learning methods to classify different persons [8]. In a typical Re-ID system, the feature representation and machine learning method are devised manually. Although this is robust to low resolution, occlusions, pose, viewpoint, and illumination in some ways, it is also laborious and heuristic. Domain knowledge and many trials are required [9]. Moreover, choosing the parameters that regulate manually selected features and learning methods is hard.

Instead of using handcrafted features, some researchers attempted to generate features automatically. These methods can be classified into two types. (1) Neural network (NN) methods. Li et al. [10] tried to automatically learn optimal features. They built a filter pairing neural network (FPNN) to jointly handle misalignment, photometric and geometric transforms, occlusions, and background clutter. Ding et al. [11] produced a large number of triplet units, each of which contains an image with a matched and a mismatched reference. A convolutional neural network was proposed to generate the layered representations. However, these methods usually require a great deal of training data. (2) Sparse coding (SC) methods: Harandi et al. [12] proposed an algorithm for learning a Riemannian dictionary, which aimed to tackle the problem of sparse coding and dictionary learning in the space of symmetric positive definite matrices. However, this method is complex and sensitive to the environment.

To avoid the previously mentioned problems, a weight-based sparse coding (WSC) is proposed in this paper. Based on three hypotheses, we successfully apply the sparse coding method to Re-ID problem. Unlike the traditional sparse coding method, WSC arranges weight to every residual. The probability density function of weights is designed carefully, and is determined at every iteration when minimizing the weighted residuals. Thus, the abnormal residuals will be filtered out. WSC is simple and robust to environment changes. It also can improve the performance of traditional sparse coding method.

The contributions of this paper are as follows. First, we apply sparse coding method to the Re-ID problem based on three hypotheses, which will be introduced in Subsection 3.2. Second, a weighted method is proposed to eliminate the influence of abnormal residuals, which is vital for non-rigid matching problems, including Re-ID problem. Removing the abnormal residuals significantly improve the performance of sparse coding.

The remainder of this paper is organized as follows. Section 2 discusses the related work in person Re-ID. In Section 3, the theory of WSC is presented, including the aim of optimization, basic hypotheses, framework, and parameters estimation. The algorithm of WSC is designed in Section 4. Section 5 provides the experimental results. Section 6 concludes this paper and points out the future work.

2 Related work

There are two sets in a general Re-ID problem: a gallery set A and a probe set B. The task of Re-ID problem is to determine whether an image or a group of images of set B are the same person of set A. The matching mechanism depends on how many images are presented for each person [13]. This gives three matching modalities: (1) single-shot versus single-shot (SvsS), if each image in a set represents a different individual; (2) multiple-shot versus single-shot (MvsS), if each image in probe set B represents a different person, while in gallery set A each person has different images; (3) multiple-shot versus multiple-shot (MvsM), if both A and B contain multiple images for each individual.

There are many approaches for the SvsS modality, which include biologically inspired features based method [14,15], attributes-based method [16,17], spatiotemporal segmentation method [6], middle-level clothes attributes based method [18,19], RGB-D based method [20], visual context based method [21], feature importance evaluating method [5,22], appearance transfer learning method [23], Mahalanobis

distance learning method [24], etc.

In this paper, the proposed WSC model needs several images of one person to compose the dictionary. So, it is suitable for the MvsS and MvsM modalities. The methods for MvsS and MvsM modalities can be divided into two groups: direct and learning-based methods. The direct methods usually focus on designing features that capture the most distinguishing aspects of an individual, while the learning-based methods need a training set to learn matching parameters.

In direct methods, multiple images could be used to obtain highly discriminative features. For example, a Mean Riemannian Covariance Grid (MRCG) was proposed for combining information from multiple images, which exploited the spatial information that carried out by dense grid structure [25]. Body part is another factor that could be exploited. Bazzani et al. [26] incorporated the global and local statistical descriptions of human appearance, and proposed the Histogram Plus Epitome (HPE) to describe the human part. On this descriptor, a structured feature called Asymmetry-based HPE (AHPE) was defined to describe the whole person [27].

In learning-based methods, pairs or groups of samples that belong to the same class are always required to train the parameters. In [28], a metric using pairs of labeled samples from different cameras was used to learn the transition from one camera to the other. In [29], the metric was improved to be a Mahalanobis metric, which provides much better generalization properties.

An important issue is the limited number of available training samples in learning-based approaches. There are four approaches to solve this problem. (1) Higher weighting for the saliency regions. In [30], a class-aware dimensionality reduction technique called Partial Least Squares (PLS) was used to create the discriminative appearance models. When combining different feature channels, higher weights are located in regions that better distinguish a specific appearance from the remaining ones. In [31], different regions of the appearance were matched using various strategies to obtain a distinctive representation. Zhao et al. [32] proposed an unsupervised salience learning methods to find reliable and discriminative features. (2) Body part segmentation. The body was usually segmented into 3 (head, torsos and legs) or 6 (chest, head, thighs and legs) parts. In [33], Pictorial Structures (PS) was proposed for human body pose estimation, which relied on two components: one capturing the local appearance of body parts, while the other representing an articulated body structure. Based on PS, Custom Pictorial Structures was proposed to improve the localization of parts via learning the appearance of an individual [34]. (3) Patch pairs segmentation. The images are segmented into regular patches to increase the number of pairs. In [18], the patches are divided into three categories according to the discriminative and generalization ability: general, rare and effective patches. The effective patches were used to learn mid-level filters for person Re-ID. Features of these methods are all selected manually. (4) Utilize co-occurrence information. In [35], for multi-label image annotation, a semantic label embedding dictionary representation is presented that not only achieves the discriminative feature representation for each label in the image, but also mines the semantic relevance between co-occurrence labels for context information.

For sparse coding methods, various researchers have pointed out the robustness issue of the l_2-norm variance [36]. In face recognition, two methods are proposed to obtain robust sparse coding features: (1) Using the l_p-norm ($0 \leqslant p \leqslant 1$) which is less sensitive to noise and outlier to replace the l_2-norm [37]. However, the non-smoothness of l_p-norm makes it difficult to be solved by the classical optimization algorithms. (2) Downsample sparse representation by filtering out the occluded pixels, which are judged with absolute entry value bigger than a predefined threshold [38]. However, a threshold that works for all images cannot be easily found (may not exist). (3) Utilizing constraints in the video. For example, in [39], the constraints in two stages are utilized: faces from a face tack must belong to the same person and faces from a video frame can not be the same person. However, this method is not applicable on a collection of appearance images.

Our WSC method still uses the l_2-norm, which could be solved by the classical Least Absolute Shrinkage and Selection Operator (LASSO) [40]. An output feature augmented LASSO could also be used to obtain multiple output prediction [41]. Unlike [38], the occluded pixels are judged by the ordering statistics in each iteration, which will achieve generalization of threshold.

3 Weight-based sparse coding

In this section, we review the traditional sparse coding problem. Then, three hypotheses for applying sparse coding method to person Re-ID problem are proposed. To obtain robust sparse representation, the WSC model is proposed. Finally, the framework of WSC and parameter estimation is discussed.

3.1 Traditional sparse coding

The traditional sparse coding problem could be formulated as

$$\min_{\alpha} \left\|\alpha\right\|_1 \quad \text{s.t.} \quad \left\|y - D\alpha\right\|_2^2 \leqslant \epsilon, \quad \epsilon > 0, \tag{1}$$

where y is the given signal, D is the dictionary matrix, α is the coding vector of y over D (or coefficient for short), and ϵ is error bound of the sparsity.

It is equivalent to the LASSO problem, which could be formulated as

$$\min_{\alpha} \left\|y - D\alpha\right\|_2^2 \quad \text{s.t.} \quad \left\|\alpha\right\|_1 \leqslant \epsilon, \quad \epsilon > 0, \tag{2}$$

where $y = (y_1, y_2, \ldots, y_n)^{\mathrm{T}} \in \mathbb{R}^n$ is the original output signal, $D = (d_1, d_2, \ldots, d_m) \in \mathbb{R}^{n \times m}$ is the dictionary matrix with column vector d_i, and α is the coding vector.

When the residual $e = y - D\alpha$ has a Gaussian distribution, Eq. (2) is more like a Bayesian Solution of $\left\|\alpha\right\|_1 \leqslant \epsilon$ related to a prior on α. However, when the pedestrian body rotating or being partly occluded, the residual can be very large. In this case, true error distribution may be far away from the Gaussion distribution. In the following subsection, we will weight the residual to remove abnormal pixels.

3.2 Hypotheses for person Re-ID

In this paper, we apply sparse coding method to person Re-ID problem. To solve the occlusion, body orientation, and illumination changes problems, the Weight-based Sparse Coding (WSC) is proposed, which is based on the following hypotheses.

First, two images belonging to one person have more pixels with similar color than those belonging to different persons. The person Re-ID problem relies on the whole appearance of individuals. The color of the clothes plays the most important role in feature extraction. Currently the research on Re-ID focuses on short-term scene. In such conditions, the clothes of pedestrian is assumed not to change. Thus, if two images belong to one person, they would have more similar color pixels.

Second, if there are n regions with similar colors and the same shape, these regions are linearly dependent, i.e., if we denote the region color values as vectors, an arbitrary region vector could be expressed by the other $n-1$ region vectors. Thus, if there are m images that belong to one person, we can find some intersecting regions according to location and color and these regions are linearly dependent. Then, if the regions that have no similar color were removed from all these m images, one image could be expressed by the other $m-1$ images.

Third, images belonging to the same person have a smaller variance. The images of the same person would have more similar color and emerge at the same position. The difference between two images of one person will be smaller than those of two different persons.

Our WSC aims to eliminate the influence of dissimilar regions, which is reached by weighting the image pixels. The pixels in these regions are called abnormal pixels in this paper. We train some weights that were used to express an image linearly which have the smallest error compared with the real image in gallery set. The weights of abnormal pixels should be close to zero. Then, we use these weights to test whether one image in probe set could be linearly expressed by images of the same person. The person with the smallest error is the matched person.

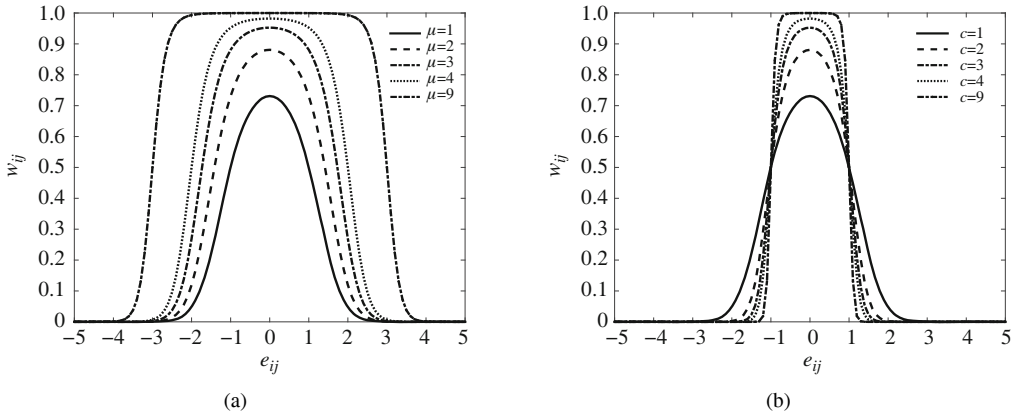

Figure 1 The curve of weight distribution function. In contrast to (6), the horizontal axis denotes e_{ij}, while the vertical axis denotes w_{ij}. (a) Curve changes with parameter μ when $c = 1$; (b) curve changes with parameter c when $\mu = 1$.

3.3 Weigt-based sparse coding

To reduce the influence of abnormal residuals, we weight the residual to control the importance of e. Thus, a weighted sparse encoder is obtained, which is shown as (3).

$$\min_{\alpha} \left\|W \circ (y - D\alpha)\right\|_2^2 \quad \text{s.t.} \quad \left\|\alpha\right\|_1 \leqslant \epsilon, \quad \epsilon > 0, \tag{3}$$

where W is cophenetic matrix (with same dimensions) with $D\alpha$, sign '$\circ$' is Hadamard or entrywise product. If $A_{m\times n} = \{a_{ij}\}$ and $B_{m\times n} = \{b_{ij}\}$ are cophenetic matrices, the Hadamard product $A \circ B$ is defined as $(A \circ B)_{ij} = a_{ij}b_{ij}$. We resize all images to the same dimensions of $p \times q \times r$. Suppose there are n images belonging to m persons that are selected as dictionary, and each person has the same number of images. An image will be reshaped into a vector with dimension of pqr, then D will be in dimension of $pqr \times mn$, α in dimension of $mn \times m$, and W in dimension of $pqr \times m$. In this paper, $W \circ (y - D\alpha)$ is also called weighted residual.

We fit the weight distribution in each iteration when minimizing (3). When the residual is large, the weight should be close to 0, otherwise 1. Thus, it is reasonable to assume the Probability Density function (PDF) with the following character.

$$f_\theta(x) = f_\theta(-x), \tag{4}$$

$$f_\theta(x_1) > f_\theta(x_2) \ \text{ if } \ |x_1| < |x_2|, \tag{5}$$

where θ is the parameter of PDF f. Eq. (4) indicates that the PDF is symmetrical about the y-axis. Eq. (5) indicates that the PDF is a decreasing function at the right of symmetry axis, while an increasing function at left.

The sigma alike function is selected as the weight distribution function. Suppose the residual matrix $e = \{e_{ij}\}$, then the weight could be defined as

$$w_{ij} = \frac{1}{1 + e^{-c(\mu - e_{ij}^2)}}, \tag{6}$$

where parameter $c > 0$ controls the decreasing rate, and $\mu > 0$ controls the inflection point. Figure 1 shows the curves change with the parameters.

3.4 Framework of WSC

Figure 2(a) shows the optimization framework of WSC. In the optimization, a LASSO method is applied to minimize the weighted residual. Upon achieving the optimal result, coefficient α and weight matrix W are obtained, which will be used in the testing procedure.

Figure 2(b) shows the testing framework. Giving an image y_j, we compute the residual of y_j with every class y_i. The class that has minimized residual with class j is the class that class j belongs to.

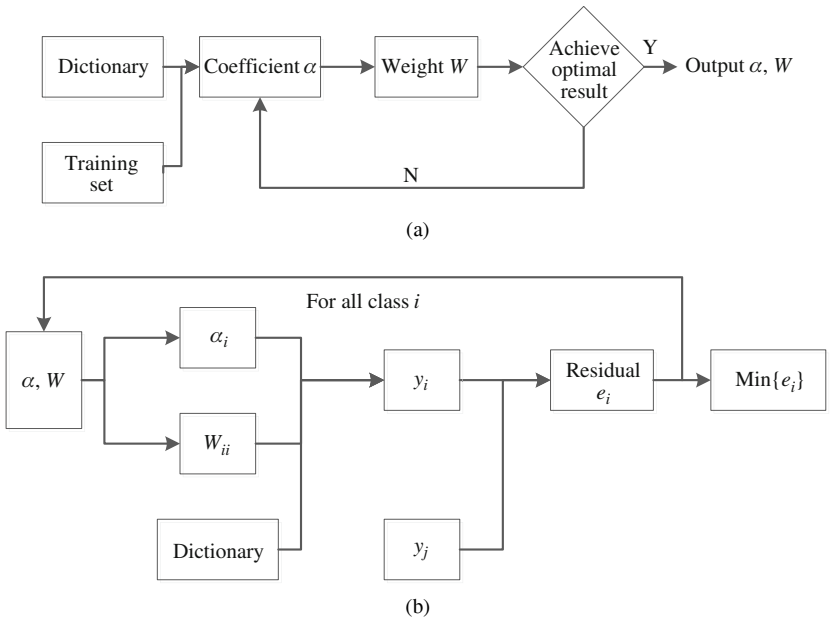

Figure 2 Framework of WSC. (a) Optimization; (b) test.

3.5 Parameter estimation

There are two parameters c and μ in (6) that need to be determined. When $\mu = e_{ij}^2$, we can get that $w_{ij} = 0.5$. We need to filter out the abnormal residuals, while reserving the normal. Hence, most of the weights of residuals should be more than 0.5. We calculate the ordering statistics $\psi = (e_{(1)}^2, e_{(2)}^2, \ldots, e_{(n)}^2)$ by sorting e_{ij}^2 in an ascending order, i.e., $e_{(1)}^2 \leqslant e_{(2)}^2 \leqslant \cdots \leqslant e_{(n)}^2$. We set μ as

$$\mu = \psi([\tau n + 0.5]), \tag{7}$$

where $\tau \in (0, 1)$ denotes the percent of reserving pixels, and $[x]$ denotes the integer part of x.

Parameter c controls the decreasing rate of weight value from 1 to 0. If c is larger, the weight value decreases quicker. Let $\lambda = c\mu$. In (6), when $e_{ij}^2 = 0$, w_{ij} reaches the maximum value.

$$w_{\max} = \frac{1}{1 + e^{-c\mu}} = \frac{1}{1 + e^{-\lambda}}. \tag{8}$$

The λ should be sufficiently large to make $w_{\max}$ close to 1. On the other hand, if λ is very large, the weight value would decrease quickly. Then, some residuals in the middle position of ψ may be assigned to an error part. In experiment, we test the interval of parameter λ that should be $[5, 8]$.

4 Algorithm of weight-based sparse coding

In this section, we will discuss the implement of WSC, including optimization and testing algorithms.

4.1 Optimization algorithm

Optimization procedure for (3) is shown in Algorithm 1. In person Re-ID problem, the dictionary is an image set randomly selected from the whole dataset. Coefficient α is also initialized randomly. Then the first residual e and weight matrix W can be calculated. After the parameters are initialized, the rest is a standard optimal procedure. After the optimization result is achieved, the coefficient α and weight matrix W are obtained, which are the parameters for testing procedure.

Algorithm 1 WSC Optimization

Require: Dictionary D, normalized training sample y, and randomly initialized coefficient α;
Ensure: $\|\alpha\|_1 \leqslant \epsilon$;
1: $t \Leftarrow 1, \alpha^{(1)} \Leftarrow \alpha$;
2: **while** $t \leqslant 100$ or achieving optimal result **do**
3: $y^{(t)} \Leftarrow D\alpha^{(t)}$;
4: $e_{ij}^{(t)} = \|y - y^{(t)}\|_2$;
5: $w_{ij}^{(t)} \Leftarrow 1/[1 + \exp\{-c(\mu - (e_{ij}^{(t)})^2)\}]$;
6: $\alpha^* \Leftarrow \arg\min_\alpha \|W \circ (y - D\alpha^{(t)})\|_2^2$ s.t. $\|\alpha\|_1 \leqslant \epsilon$;
7: $\alpha^{(t+1)} \Leftarrow \alpha^{(t)} + \sigma(\alpha^* - \alpha^{(t)})$, where $0 < \sigma < 1$ is the step size.
8: $t \Leftarrow t + 1$;
9: **end while**

Algorithm 2 WSC Testing

Require: Dictionary D, dictionary labels DL, class number CN, testing sample y, weight W from Algorithm 1, and random coefficient α from Algorithm 1;
1: $t \Leftarrow 1$;
2: **while** $t \leqslant CN$ **do**
3: $i \Leftarrow 1$;
4: **while** $i \leqslant len(DL)$ **do**
5: **if** $DL(i) == t$ **then**
6: $\alpha_i^{te} \Leftarrow \alpha_i$;
7: **else**
8: $\alpha_i^{te} \Leftarrow 0$;
9: **end if**
10: $i \Leftarrow i + 1$;
11: **end while**
12: $z \Leftarrow W \circ (y - D * \alpha^{te})$;
13: $Result(t) = z' * z$;
14: $t \Leftarrow t + 1$;
15: **end while**
16: find class: $C = \arg\min_t Result(t)$;

4.2 Testing algorithm

In testing procedure (Algorithm 2), the optimal algorithm is also needed to compute the coefficient α and weight W from probe images. After that, we calculate the square residual for the test class to each class in gallery set. The corresponding class which has the minimum square residual is the class that the test class belongs to.

5 Experiment results

In this section we report experiment results performed on several person Re-ID datasets and we also compare our method with the state-of-the-art methods. All experiments are performed on publicly available datasets for MvsS and MvsM Re-ID problem, including CAVIAR4REID [42], ETHZ [43] and iLIDS [44]. There is no publicly available code for some methods which makes the side-by-side comparison difficult. Hence, we only compare our method with the public available results in this paper.

5.1 Datasets

CAVIAR4REID is a dataset for evaluating person Re-ID algorithms, which has been extracted from the well-known CAVIAR dataset mostly famous for person tracking and detection. Shopping center dataset contains 26 sequences recorded from two different points of view at resolution of 384×288 pixels. CAVIAR4REID dataset contains object (people) bounding boxes extracted from the shopping center dataset. It contains 72 unique individuals: 50 of them with both the camera views and the remaining 22 with only one camera view. The images for each camera view have variance with respect to resolution changes, light conditions, occlusions, and pose changes. So it is challenging to the Re-ID task.

ETHZ dataset is captured from moving cameras in a crowded street. It contains three sub-datasets. The images are extracted from the ground truth location of people in the video with original resolution. ETHZ1, ETHZ2 and ETHZ3 contain 83 people (4875 images), 35 people (1936 images), and 28 people (1762 images), respectively.

iLIDS is created from the pedestrians observed in two non-overlapping camera views from the i-LIDS Multiple-Camera Tracking Scenario (MCTS) dataset, which was captured at an airport arrival hall under a multi-camera CCTV network. It comprises 600 image sequences of 300 distinct individuals, with one pair of image sequences from two camera views for each person. Each image sequence has variable length ranging from 23 to 192 image frames, with an average number of 73. The iLIDS-VID dataset is very challenging due to clothing similarities among people, lighting and viewpoint variations across camera views, cluttered background and random occlusions.

5.2 Experimental setup and evaluation system

WSC focuses on MvsS and MvsM person Re-ID. We denote N as the number of images for one person in dictionary. We select $N = 3, 5, 10$ to perform on previously mentioned datasets respectively. The remainder images are divided into five sets evenly. One set is used as validation set to determine the supper-parameters, including c, μ in (6), and training step size σ in Algorithm 1. The other four sets are used as the testing set. Each image is processed in RGB color space, and resized to the same size at resolution of 30×75 pixels. Then, the image is represented as a vector with dimension of $30\times75\times3 = 6750$.

The Cumulative Matching Characteristic (CMC) curve is the standard performance measure for the person Re-ID task. CMC measures the correctly matched ratio at rank r. We initialize $rank(r)$ to zero. If there are m classes (persons) in the MvsS modality, $D\alpha$ will contain m vectors which represent the m classes, respectively. For each image y_i in probe set, we compute $a_i^{(j)} = \|W_i \circ (y_i - (D\alpha)_j)\|_2^2$ for each $j = 1, 2, \ldots, m$. We calculate $\phi = (a_{(1)}, a_{(2)}, \ldots, a_{(m)})$ by sorting $a_i^{(j)}$ in ascending order. We find $a_{(p)} \in \phi$ which has the same label with y_i, then compute $rank(p) = rank(p)+1$. Calculate all images i as previous steps, we will get the final CMC curve. For MvsM modality, we suppose there are n images for one person in the probe set. For each image k of person j, we calculate $a_i^{(j,k)} = \|W^{(j,k)} \circ (y_{j,k} - (D\alpha)_{j,k})\|_2^2$, and $a_i^{(j)} = \min\{a_i^{(j,k)}\}$. After that, the CMC curve will be computed as in MvsS modality.

5.3 Linear combination of images

Figure 3 shows the optimal procedure of 10 persons from CAVIAR4REID dataset. The top 5 rows present 5 matched persons, while the other 5 rows present 5 mismatched persons. Figure 3(a) shows the test image, (b) is the dictionary set where $N = 5$, (c) and (d) is the result of $D\alpha$ and $W \circ (D\alpha)$, respectively. When testing whether an image belongs to a person, the components for other persons of W and α are set to zero. Figure 3 (e) and (f) are the results of $D\alpha$ and $W \circ (D\alpha)$ after setting the other components to zero. The class of (f) with the minimum square residual of (a) is the class that (a) belongs to.

The dictionary contains all the 72 unique individuals in CAVIAR4REID. When $N = 5$, it totally contains $72 \times 5 = 360$ images. Results of (c) and (d) are computed by the whole dictionary, while (e) and (f) are computed by one person images of the dictionary.

Let us focus on the sixth row. The test image differs significantly from the dictionary images, but the result of Figure 3(e) is similar with the test image. In this case, the whole dictionary does represent a significantly different image. However, this may also lead to mismatched cases. For example, Figure 4 (a) and (b) show the contribution of dictionary images that represent the first and sixth person in Figure 3, respectively. The first person contributes 89.8% to represent herself, so when other components of W and α are set to zero, the result of $W \circ (D\alpha)$ is approximately invariant. This is the matched case. In the sixth person's representation, the images only contribute 18.2% to himself. The contribution of multiple persons is more than his own, i. e., the sixth person is represented by multiple persons. This leads to a mismatched case.

In Figure 5, the top 4 images of the first person is her own. Only one image of her own is not included in the top 10 images. The first image of the second row is very similar with the testing image. For the

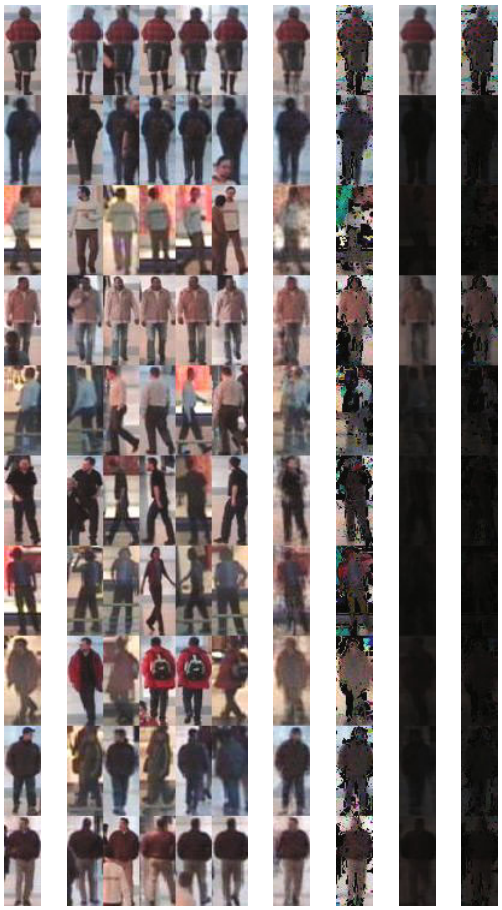

Figure 3 Optimal procedure. The first to fifth rows are matched individuals, while the sixth to tenth rows are mismatched. (a) Images in probe set, i.e., the testing images; (b) images in dictionary set, where $N = 5$; (c) result of $D\alpha$; (d) result of $W \circ (D\alpha)$; (e) result of $D\alpha$, where the components of α for the other persons are set to zero; (f) result of $W \circ (D\alpha)$, where the components of α and W for the other persons are set to zero.

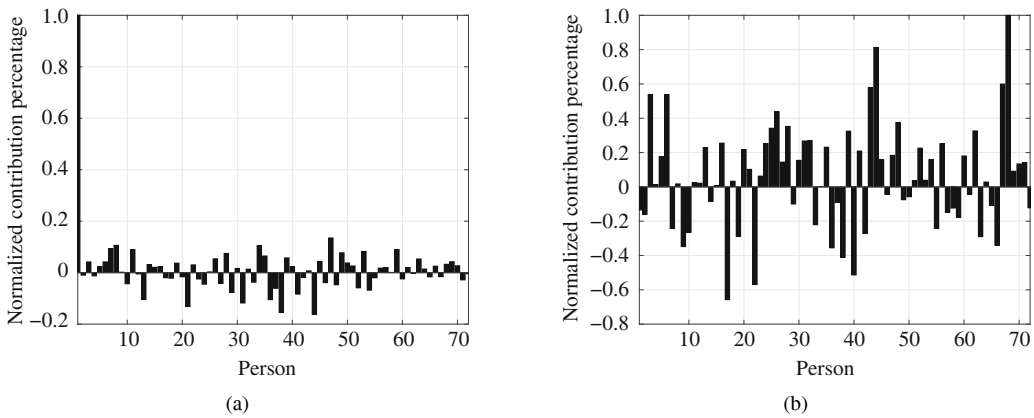

Figure 4 The normalized contribution of dictionary images that represents a testing image. (a) Representation contribution to the first image of Figure 3; (b) representation contribution to the sixth image of Figure 3.

sixth person, none of his own image is included in the top 10 dictionary images. From left to right of the fourth row, the image is more and more similar with the testing image.

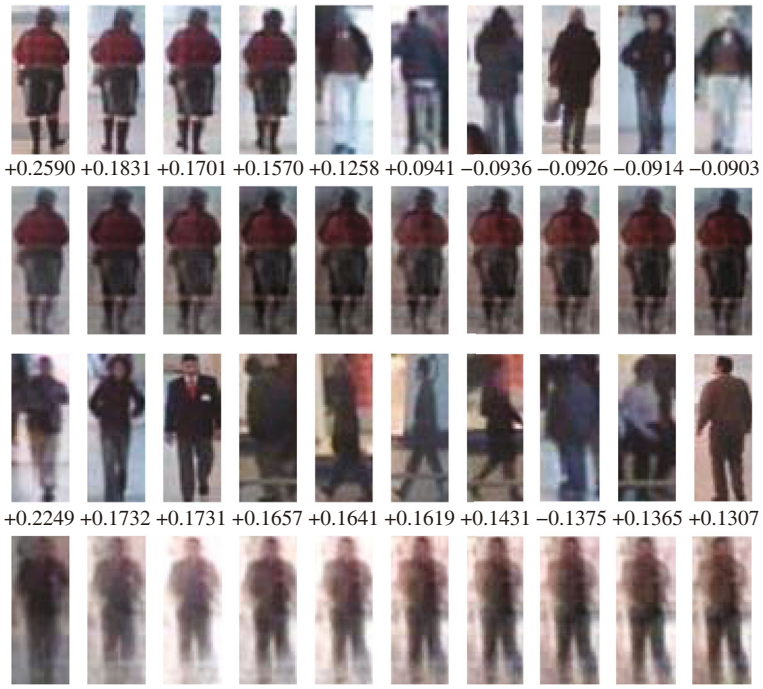

Figure 5 Dictionary images rank sort by contribution in ascending order. The first and third rows are top ten images of the first and sixth person in Figure 3, respectively, where the images are sorted by the absolute value of corresponding α in an ascending order. The labeled number of the image is the corresponding coefficient of α. The images of the second row are results of $D\alpha$ that are calculated by top $5, 10, 15, \ldots, 50$ images and coefficients of the first person, while the fourth row is the sixth person.

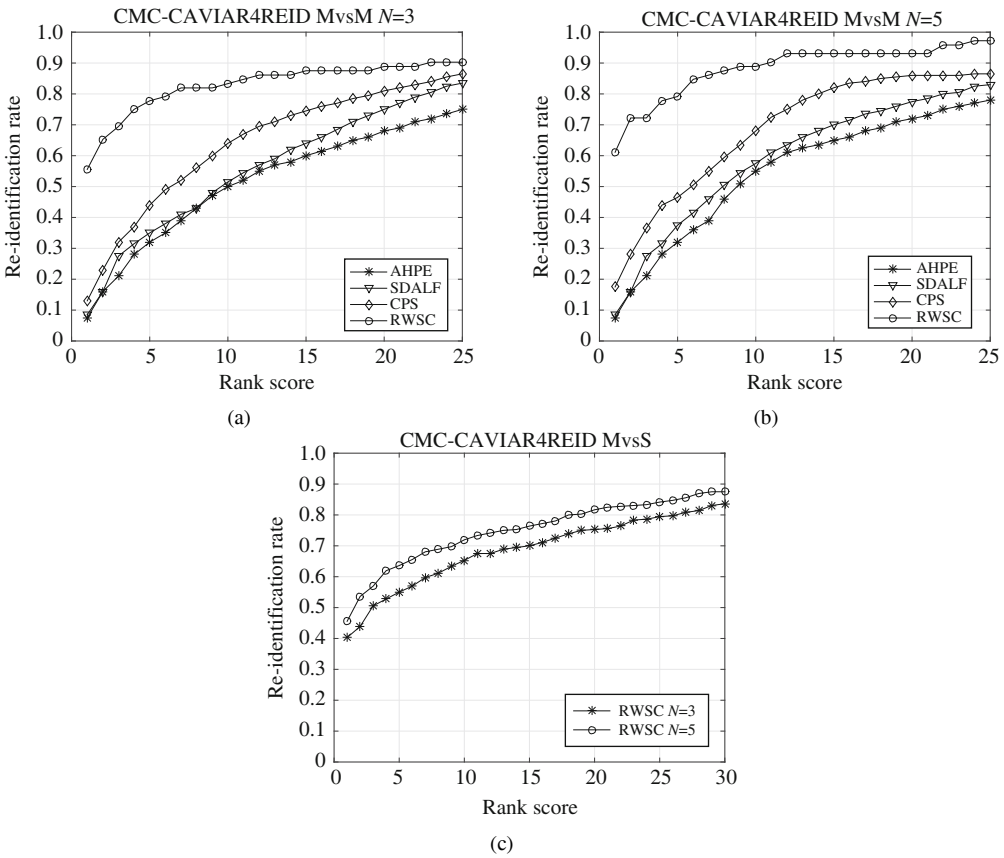

Figure 6 Results on CAVIAR4REID dataset. (a) MvsM when $N = 3$; (b) MvsM when $N = 5$; (c) MvsS when $N = 3, 5$.

Table 1 Performance at rank-1 with respect to the state-of-the-art on CAVIAR4REID

Modality	MvsM ($N = 3$)	MvsM ($N = 5$)	MvsS ($N = 3$)	MvsS ($N = 5$)
AHPE [27]	0.075	0.075	–	–
SDALF [13]	0.085	0.083	–	–
CPS [34]	0.13	0.175	–	–
WSC	**0.556**	**0.611**	0.403	0.456

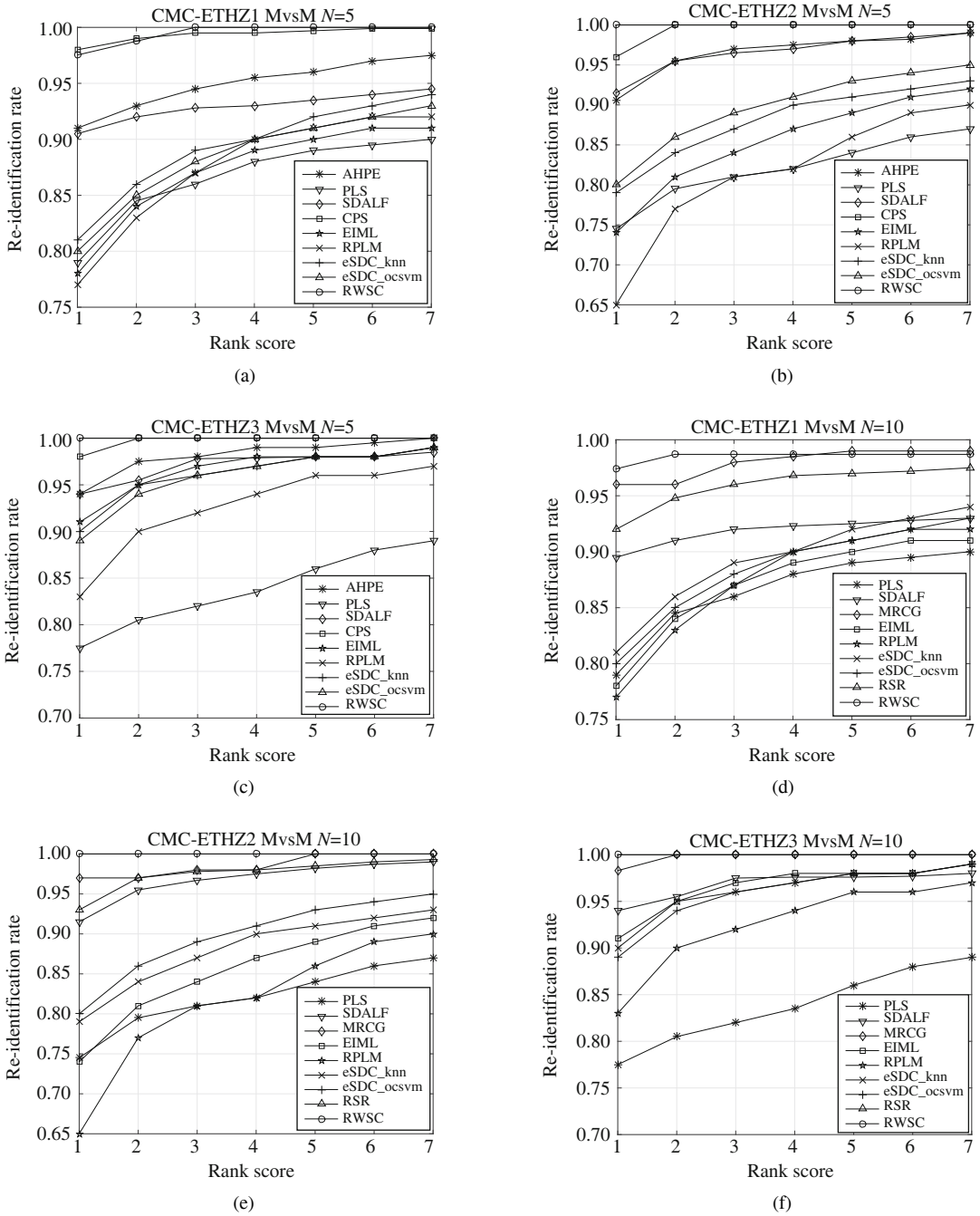

Figure 7 MvsM on ETHZ 1, 2, 3 datasets. (a) MvsM on ETHZ1 when $N = 5$; (b) MvsM on ETHZ2 when $N = 5$; (c) MvsM on ETHZ3 when $N = 5$; (d) MvsM on ETHZ1 when $N = 10$; (e) MvsM on ETHZ2 when $N = 10$; (f) MvsM on ETHZ3 when $N = 10$.

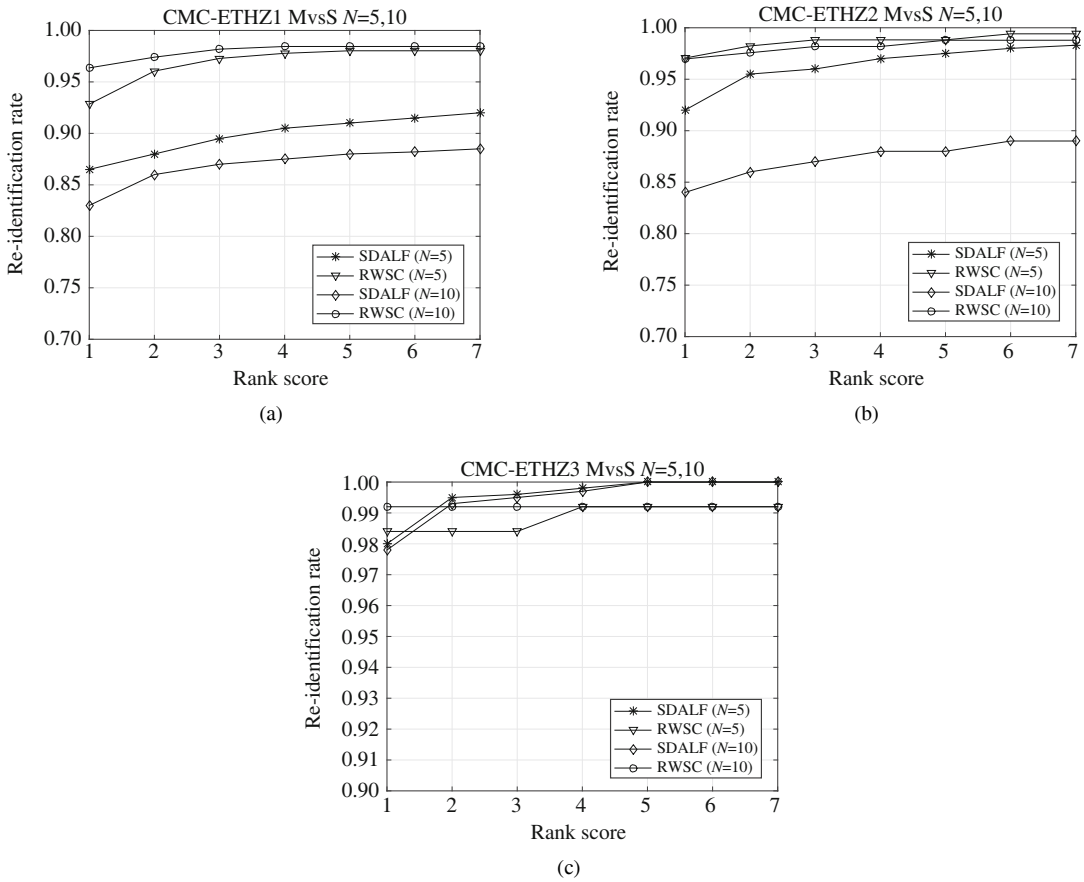

Figure 8 MvsS on ETHZ datasets. (a) MvsS on ETHZ1; (b) MvsS on ETHZ2; (c) MvsS on ETHZ3.

Table 2 Performance at rank-1 with respect to the state-of-the-art on ETHZ

Dataset	ETHZ1				ETHZ2				ETHZ3			
Modality	MvsM		MvsS		MvsM		MvsS		MvsM		MvsS	
	$N=5$	$N=10$	$N=5$	$N=10$	$N=5$	$N=10$	$N=5$	$N=10$	$N=5$	$N=10$	$N=5$	$N=10$
AHPE [27]	0.91	–	–	–	0.906	–	–	–	0.94	–	–	–
PLS [30]	0.79	0.79	–	–	0.745	0.745	–	–	0.775	0.775	–	–
SDALF [13]	0.902	0.896	0.865	0.83	0.916	0.915	0.92	0.84	0.937	0.941	0.98	0.978
CPS [34]	**0.977**	–	–	–	0.973	–	–	–	0.98	–	–	–
MRCG [25]	–	0.96	–	–	–	0.97	–	–	–	0.983	–	–
EIML [29]	0.78	0.78	–	–	0.74	0.74	–	–	0.91	0.91	–	–
RPLM [28]	0.77	0.77	–	–	0.65	0.65	–	–	0.83	0.83	–	–
eSDC_knn [32]	0.81	0.81	–	–	0.79	0.79	–	–	0.90	0.90	–	–
eSDC_ocsvm [32]	0.80	0.80	–	–	0.80	0.80	–	–	0.89	0.89	–	–
RSR [12]	–	0.92	–	–	–	0.93	–	–	–	–	–	–
WSC	0.975	**0.974**	**0.928**	**0.964**	**1**	**1**	**0.971**	**0.970**	**1**	**1**	**0.984**	**0.992**

5.4 Results on multiple fixed cameras

The images of CAVIAR4REID dataset are captured from two fixed cameras. In the fixed camera scenario, the backgrounds of different persons from the same camera view are similar. This may lead to a mismatch between different persons.

On this dataset we compare WSC with the AHPE [27], SDALF [13], and CPS [34] method. In Figure 6 (a) and (b), we report the CMC curves for WSC and the state-of-the-art for MvsM ($N \in \{3, 5\}$ respectively). In MvsS modality, there is no publicly available result on CAVIAR4REID. We only report

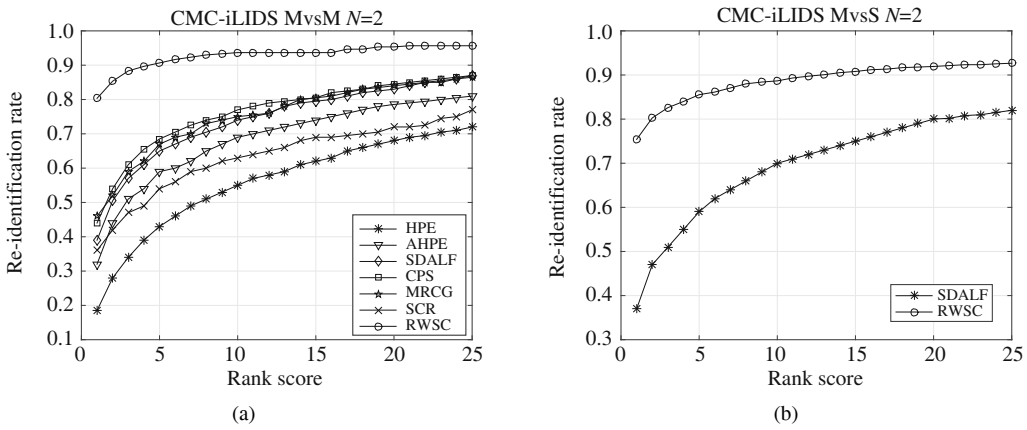

Figure 9 Results on iLIDS dataset. (a) MvsM when $N = 2$; (b) MvsS when $N = 2$.

Table 3 Performance at rank-1 with respect to the state-of-the-art on iLIDS

Modality	MvsM ($N = 2$)	MvsS ($N = 2$)
HPE [26]	0.185	–
AHPE [27]	0.32	–
SDALF [13]	0.39	0.37
CPS [34]	0.44	–
MRCG [25]	0.46	–
SCR [45]	0.36	–
WSC	**0.803**	**0.754**

our results in Figure 6(c).

In Table 1 we report rank-1 results for the MvsS and MvsM ($N \in \{3, 5\}$) modalities. WSC outperforms current methods up to 42.6% and 43.6% for MvsM $N = 3$ and $N = 5$, respectively. For MvsS modality, WSC method achieves 40.3% and 45.6% at rank-1 when $N = 3$ and $N = 5$, respectively.

5.5 Results on moving cameras

The images of ETHZ datasets are captured from moving cameras. In the moving camera scenario, change of viewpoint is large. This may lead to a mismatch between images for the same person.

On ETHZ $1, 2, 3$ datasets we compare WSC with AHPE [27], PLS [30], SDALF [13], CPS [34], MRCG [25], EIML [29], RPLM [28], eSDC_knn [32], eSDC_ocsvm [32], and RSR [12] method. In Figures 7 and 8, we report the CMC curves for MvsM and MvsS ($N \in \{5, 10\}$) modalities, respectively. In Table 2 we report rank-1 results for the MvsS and MvsM ($N \in \{5, 10\}$) modalities on ETHZ $1, 2, 3$ datasets. WSC reaches 97.5% at rank-1 on ETHZ1. This is comparative with respect of the state-of-the-art methods. On ETHZ2 and ETHZ3, WSC reaches 100% at rank-1.

5.6 Results on image sequences of video

The images of ETHZ datasets are frame sequences of videos from two camera views. In this scenario, the backgrounds of different images for the same person, and the backgrounds of different persons from the same camera view are all similar. Unlike CAVIAR4REID dataset, there are multiple images with similar backgrounds for one person. This may improve the performance of matching. There is also challenge of viewpoint changes, but it is not so serious because the change between two consecutive frames is small.

On iLIDS we compare WSC with HPE [26], AHPE [27], SDALF [13], CPS [34], MRCG [25], and SCR [45] method. In Figure 9 (a) and (b), we report the CMC curves for WSC and the state-of-the-art for MvsM ($N = 2$) and MvsS ($N = 2$) modality, respectively. In Table 3, we report rank-1 results for the MvsM and MvsS ($N = 2$) modalities. WSC outperforms current methods up to 34.3% and 38.4% for MvsM and MvsS, respectively.

6 Conclusion and future work

In this paper we proposed an approach to person Re-ID that is based on robust weighted sparse coding. We showed how to apply sparse coding method to the Re-ID problem, and filter out the abnormal residuals through weights. Our WSC method has the following advantages. (1) WSC obtains the state-of-the-art performance on multi-shot vs multi-shot and multi-shot vs single-shot Re-ID modalities. (2) WSC requires no labeled data for training, which is competitive with respect to learning-based methods. (3) WSC requires no foreground/background or body part segmentation.

Future work includes: (1) We used the naive images as dictionary, which may not be the optimal or even suboptimal basis. A trained optimal basis may lead to better performance. (2) In Subsection 5.3, we found that a person may be represented by other persons. Adding a constraint to the optimization goal that mainly uses one person's images to represent an image may produce better solutions.

Acknowledgements

This work was partially supported by National Natural Science Foundation of China (Grant No. 61472019), National High-tech R&D Program of China (863 Program) (Grant No. 2013AA01A603) and National Aerospace Science Foundation of China (Grant No. 2013ZC51). Supported by the Programme of Introducing Talents of Discipline to Universities and the Open Fund of the State Key Laboratory of Software Development Environment (Grant No. #SKLSDE-2015ZX-21).

References

1 Theodoridis E, Mylonas G, Chatzigiannakis I. Developing an iot smart city framework. In: Proceedings of the 4th International Conference on Information, Intelligence, Systems and Applications, Piraeus-Athens, 2013. 1–6
2 Xiong Z, Luo W, Chen L, et al. Data vitalization: a new paradigm for large-scale dataset analysis. In: Proceedings of the 16th IEEE International Conference on Parallel and Distributed Systems, Shanghai, 2010. 251–258
3 Xiong Z, Zheng Y W, Li C. Data vitalization's perspective towards smart city: a reference model for data service oriented architecture. In: Proceedings of the 14th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, Chicago, 2014, 865–874
4 Filipponi L, Vitaletti A, Landi G, et al. Smart city: an event driven architecture for monitoring public spaces with heterogeneous sensors. In: Proceedings of the 4th International Conference on Sensor Technologies and Applications, Venice, 2010. 281–286
5 Farenzena M, Bazzani L, Perina A, et al. Person re-identification by symmetry-driven accumulation of local features. In: Proceedings of the 23rd IEEE International Conference on Computer Vision and Pattern Recognition, San Francisco, 2010. 2360–2367
6 Gheissari N, Sebastian T B, Hartley R. Person reidentification using spatiotemporal appearance. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, New York, 2006. 1528–1535
7 Gong S, Cristani M, Loy C C, et al. The re-identification challenge. In: Gong S, Cristani M, Yan S, et al., eds. Person Re-Identification. London: Springer, 2014. 1–20
8 Wang X, Zhao R. Person re-identification: system design and evaluation overview. In: Gong S, Cristani M, Yan S, et al., eds. Person Re-Identification. London: Springer, 2014. 351–370
9 Tang H, Huang T S. 3D facial expression recognition based on automatically selected features. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Anchorage, 2008. 1–8
10 Li W, Zhao R, Xiao T, et al. Deepreid: deep filter pairing neural network for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, Columbus, 2014. 152–159
11 Ding S, Lin L, Wang G, et al. Deep feature learning with relative distance comparison for person re-identification. Patt Recog, 2015, 48: 2993–3003
12 Harandi M T, Sanderson C, Hartley R. Sparse coding and dictionary learning for symmetric positive definite matrices: a kernel approach. In: Proceedings of the 12th European Conference on Computer Vision, Florence, 2012. 216–229
13 Bazzani L, Cristani M, Murino V. SDALF: modeling human appearance with symmetry-driven accumulation of local features. In: Gong S, Cristani M, Yan S, et al., eds. Person Re-Identification. London: Springer, 2014. 43–69
14 Ma B, Su Y, Jurie F. Bicov: a novel image representation for person re-identification and face verification. In: Proceedings of the 23rd British Machine Vision Conference, Surrey, 2012. 1–11
15 Ma B, Su Y, Jurie F. Covariance descriptor based on bio-inspired features for person re-identification and face verification. Image Vis Comput, 2014, 32: 379–390

16 Layne R, Hospedales T M, Gong S, et al. Person re-identification by attributes. In: Proceedings of the 23rd British Machine Vision Conference, Surrey, 2012. 1–11
17 Layne R, Hospedales T M, Gong S. Towards person identification and re-identification with attributes. In: Proceedings of the 23rd European Conference on Computer Vision, Surrey, 2012. 402–412
18 Zhao R, Ouyang W, Wang X. Learning mid-level filters for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, Columbus, 2014. 144–151
19 Li A, Liu L, Wang K, et al. Clothing attributes assisted person re-identification. IEEE Trans Circ Syst Video Technol, 2014, 25: 869–878
20 Satta R, Pala F, Fumera G, et al. Real-time appearance-based person re-identification over multiple KinectTM cameras. In: Proceedings of the 8th International Conference on Computer Vision Theory and Applications, Barcelona, 2013. 407–410
21 Zheng W S, Gong S, Xiang T. Associating groups of people. In: Proceedings of the 20th British Machine Vision Conference, London, 2009. 1–11
22 Liu C, Gong S, Loy C C. On-the-fly feature importance mining for person re-identification. Patt Recog, 2014, 47: 1602–1615
23 Avraham T, Gurvich I, Lindenbaum M, et al. Learning implicit transfer for person re-identification. In: Proceedings of the 23rd European Conference on Computer Vision, Surrey, 2012. 381–390
24 Roth P M, Hirzer M, Kostinger M, et al. Mahalanobis distance learning for person re-identification. In: Gong S, Cristani M, Yan S, et al., eds. Person Re-Identification. London: Springer, 2014. 247–267
25 Bak S, Corvee E, Bremond F, et al. Multiple-shot human re-identification by mean riemannian covariance grid. In: Proceedings of the 8th IEEE International Conference on Advanced Video and Signal-Based Surveillance, Klagenfurt, 2011. 179–184
26 Bazzani L, Cristani M, Perina A, et al. Multiple-shot person re-identification by HPE signature. In: Proceedings of the 20th International Conference on Pattern Recognition, Istanbul, 2010. 1413–1416
27 Bazzani L, Cristani M, Perina A, et al. Multiple-shot person re-identification by chromatic and epitomic analyses. Patt Recog Lett, 2012, 33: 898–903
28 Hirzer M, Roth P M, Kostinger M, et al. Relaxed pairwise learned metric for person re-identification. In: Proceedings of the 23rd European Conference on Computer Vision, Surrey, 2012. 780–793
29 Hirzer M, Roth P M, Bischof H. Person re-identification by efficient impostor-based metric learning. In: Proceedings of the 9th International Conference on Advanced Video and Signal-Based Surveillance, Beijing, 2012. 203–208
30 Schwartz W R, Davis L S. Learning discriminative appearance-based models using partial least squares. In: Proceeding of the XXII Brazilian Symposium on Computer Graphics and Image Processing, Brazil, 2009. 322–329
31 Bak S, Charpiat G, Corvee E, et al. Learning to match appearances by correlations in a covariance metric space. In: Proceedings of the 23rd European Conference on Computer Vision, Surrey, 2012. 806–820
32 Zhao R, Ouyang W, Wang X. Unsupervised salience learning for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, Portland, 2013. 3586–3593
33 Felzenszwalb P F, Huttenlocher D P. Pictorial structures for object recognition. Int J Comput Vis, 2005, 61: 55–79
34 Cheng D S, Cristani M, Stoppa M, et al. Custom pictorial structures for re-identification. In: Proceedings of the 22nd British Machine Vision Conference, Dundee, 2011. 1–11
35 Cao X, Zhang H, Guo X, et al. SLED: semantic label embedding dictionary representation for multi-label image annotation. IEEE Trans Image Process, 2015, 24: 2746–2759
36 Meng D Y, Zhao Q, Xu Z B. Improve robustness of sparse PCA based on L1-norm maximization. Patt Recog, 2012, 45: 487–497
37 Zhao Q, Meng D Y, Xu Z B. Robust sparse principal component analysis. Sci China Inf Sci, 2014, 57: 092115
38 Li Y L, Meng L, Feng J F, et al. Downsampling sparse representation and discriminant information aided occluded face recognition. Sci China Inf Sci, 2014, 57: 032112
39 Zhou C, Zhang C, Li X, et al. Video face clustering via constrained sparse representation. In: Proceedings of the IEEE International Conference on Multimedia and Expo, Chengdu, 2014. 1–6
40 Tibshirani R. Regression shrinkage and selection via the lasso. J Roy Stat Soc Ser B-Stat Method, 1996, 58: 267–288
41 Zhang C, Han Y, Guo X, et al. Output feature augmented lasso. In: Proceedings of the IEEE International Conference on Data Mining, Shenzhen, 2014. 680–686
42 Cheng D S, Cristani M, Stoppa M, et al. Custom pictorial structures for re-identification. In: Proceedings of the 22nd British Machine Vision Conference, Dundee, 2011. 1–11
43 Ess A, Leibe B, Van Gool L. Depth and appearance for mobile scene analysis. In: Proceedings of the 11th IEEE International Conference on Computer Vision, Brazil, 2007. 1–8
44 Wang T, Gong S, Zhu X, et al. Person re-identification by video ranking. In: Proceedings of the 13th European Conference on Computer Vision, Zurich, 2014. 688–703
45 Bak S, Corvee E, Brmond F, et al. Person re-identification using spatial covariance regions of human body parts. In: Proceedings of the 7th IEEE International Conference on Advanced Video and Signal Based Surveillance, Boston, 2010. 435–440