Discriminative Dictionary Learning Sparse Coding for Person Re-Identification*

Hao Sheng[1,2], Beichen Zhang[1], Yan Huang[1], Yanwei Zheng[1,2], Zhang Xiong[1]

Abstract— **Person re-identification is one of the most important issues in intelligent transportation systems. Recently, the widespread availability of cameras and a growing need for public safety have increasingly motivated interest in the problem of person re-identification in multi-camera networks. The main difficulty of person re-identification arises from the variations in human pose, different viewpoint in multi-camera, cluttered background, occlusion, and low image resolution, which lead person re-identification to a challenging problem. This paper presents a method based on sparse coding for person re-identification. To apply sparse coding method, we firstly solve the problem of aligning person images, and to enhance the discrimination of dictionary, a dictionary learning model is added into our method. Experiments on benchmark dataset (CAVIARa, ETZH, i-LIDS) demonstrate that the proposed method outperforms the state-of-the-art approaches.**

I. INTRODUCTION

Human behavior analysis is one of the most important parts in intelligent transportation systems. In the past, most of the algorithms focus on analyzing individual behavior in a simple environment monitored by a single camera. However, with growing need of public safety and incident management, group human behavior analysis is becoming more significant. Because of the limit of single-camera horizon, group human behavior analysis is usually performed in a multi-camera and complex surveillance system. In surveillance scenarios where long-term activities need be modeled within a large and structured environment, an important problem is to recognize a person at a different camera when the person has been previously observed at another camera [1]. This problem is called person re-identification (Re-ID).

Person re-identification can be converted to a retrieval problem: given one or multi images of a query person, the task is to rank the similarities (or distances) between the query person and all labeled candidates, then the classification is based on the rank. The set of query persons is traditionally called probe, and the set of labeled candidates are denoted by the gallery. There are three modalities for person re-identification: (1) single-shot versus single-shot (SvsS), if there is only one image for each person in both probe and gallery set. (2) multiple-shot versus single-shot (MvsS), if multiple images in the gallery set, while there is single image in gallery set. (3) multiple-shot versus multiple-shot (MvsM), if both gallery and probe set contain multiple images for each person. Due to the fact that the same person may be acquired at different cameras in a surveillance system or at the same camera in different times, the images belong to one person may gain great variations in pose, illumination, viewpoint, background clutter, and partial occlusion.

To solve such a challenging problem, many methods have been presented. These methods can be roughly classified into two categories: 1. Methods based on developing a descriptor of appearance-based features. These methods try to exploit robust features based on color, shape, texture, and spatial structure. The descriptors should be discriminative enough for identification as well as robust to variations described above. 2. Methods based on learning for classification. These methods cast re-identification as learning problems in which optimal metric distance is learned for classification. The above two categories of methods are devised manually. Although these methods are effective, domain knowledge and many trials are required [2]. These methods are laborious and heuristic. It is hard to find a common descriptor or a common metric model that suitable for all scenarios. To achieve better performances in classification, parameters or the model structure may change in different scenarios. In this case, choosing the parameters that regulate of manually selected features and learning methods is difficult.

To avoid the previously mentioned problems, a discriminative dictionary learning sparse coding method (DDLSC) is proposed in this paper. Different from those traditional methods, we don't concentrate on either exploiting features or developing a descriptor. Our method is based on sparse coding, which has been proved to be a powerful tool for face recognition [3]. Denote by $y = [y_1; y_2; \ldots; y_n] \in \mathbb{R}^n$ a query sample, the first phase of sparsity based classification is to represent y over a set of images which called dictionary $D = [d_1, d_2, \ldots, d_m] \in \mathbb{R}^{n \times m}$, i.e., $T \approx D\alpha$, where the representation vector α has only a few large entries. In our method, we attempt to align the person images, so that a person image can be linearly expressed by dictionary. Dictionary plays an important role in sparse coding procedures, and it's obviously that employing a more discriminating dictionary can achieve more desirable performances. This refers to dictionary learning (DL). DL aims to learn from the training samples the space where the given image could

*This study was partially supported by the National Natural Science Foundation of China, the National High Technology Research and Development Program of China and the National Science & Technology Pillar Program. Supported by the Programme of Introducing Talents of Discipline to Universities and the Open Fund of the State Key Laboratory of Software Development Environment.

[1]Hao Sheng, Beichen Zhang, Yan Huang, Yanwei Zheng and Zhang Xiong are with State Key Laboratory of Software Development Environment, School of Computer Science and Engineering, Beihang University, Beijing 100191 , P.R.China. {shenghao, zhangbc, yanhuang, zhengyw, xiongz}@buaa.edu.cn

[2]Hao Sheng and Yanwei Zheng are with Shenzhen Key Laboratory of Data Vitalization, Research Institute in Shenzhen, Beihang University, ShenZhen 518057, P.R.China.

be well represented or coded. In our method, we proposed an objective function based on three terms of constraints. After optimizing the objective function iteratively, we obtain a trained dictionary instead of directly using the training samples as dictionary.

The paper is organized as follows. Section II discusses related works on person re-identification. Traditional sparse coding method on digit recognition is presented in Section III. The image alignment is presented in Section IV. The objective function based on three terms of constraints is presented in Section V. Section VI details the optimization of the model. The experimental setup and results are discussed in Section VII. Finally, in Section VIII we draw some conclusions and discuss new directions for research.

II. RELATED WORK

Many recent works have been presented for the problem of person re-identification. Most of them focus primarily on either new descriptors for person appearance, or on learning techniques for person re-identification.

In descriptor based methods, multiple images could be used to obtain highly discriminative features. The first work that considered the problem of appearance models for person recognition, reacquisition and tracking was that of Gray et al. [4]. In [5], [6], the authors argue that, there is a need for metrics that apply to complete systems for person re-identification problems instead of evaluating independently. They proposed the Cumulative Match Curve (CMC) as a standard protocol to compare results and introduced the VIPeR dataset for re-identification.

Descriptors of appearance usually include color, texture. A high dimension descriptor combined by multi features is always effective. For example, in [7], the author extracted a 7-dimensional local feature on each pixel in the person images. The 7-d feature is a combination involves color, texture and spatial structure. Then the features are modeled by a Gaussian model.

Some methods focus on structured patches due to the background clutter and depth information of structure. In [1], the author presents an appearance-based method called Symmetry-Driven Accumulation of Local Features (SDALF), the descriptor focus on highly-structured patches, and is modeled by finding the symmetry axes for each person image and combining different parts of human body that are represented by weighted HSV color histograms. Zhao et al. [8] used an unsupervised learned salience model for patch matching such that the reliable and discriminative matched patches can be matched for better re-identification performance. Bird et al. [9] model the appearances of individuals by stripe based rigid blobs. The image of a pedestrian is divided into ten equally spaced horizontal strips, after that, the author learn the mean feature vectors of the horizontal strips in a training step.

Distance learning is another categories. In [10], Hirzer et al. learned a metric from pairs of samples belong to inter-cameras using discriminative Mahalanobis metric learning. Another Mahalanobis metric based methods is in [11], a metric learning framework is presented to obtain a robust Mahalanobis metric for Large Margin Nearest Neighbor classification with Rejection (LMNN-R). Zheng et al. [12] introduced a Relative Distance Comparison (RDC) model. The authors argue that, a pair of true matches having a smaller distance than a wrongly match pair. This approach avoids treating all features indiscriminately and does not assume the existence of some universally distinctive and reliable features.

An earlier study based on sparse coding for person re-identification is presented in [13]. The authors 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.

III. TRADITIONAL SPARSE REPRESENTATION BASED CLASSIFICATION (SRC)

The SRC method has widely used in face recognition (FR) recently. This method is based on the hypothesis that each face image of one person can be linearly expressed by other images that belong to this person. The SRC method [3] can be simply described as follows:

$$\hat{\alpha} = \arg\min_{\alpha} \{\|y - D\alpha\|_2 + \lambda \|\alpha\|_1\} \tag{1}$$

where $\|\cdot\|_1$ is the l_1-norm, λ is a constant, $y = [y_1; y_2; \ldots; y_n] \in \mathbb{R}^n$ is a test image to be coded, $D = [d_1, d_2, \ldots, d_m] \in \mathbb{R}^{n \times m}$ is the dictionary matrix, and α is the coding coefficient of y over D. Each column of D denoted by d_i is a training sample (or features that the dimension has been reduced), and the purpose of Eq. (1) is to code y by training samples (i.e., D) and let α to be sparsely.

After that, do classification via

$$class(y) = \arg\min_{i} \|y - D_i\hat{\alpha}_i\|_2 \tag{2}$$

where D_i is the sub-dictionary for class i, and $\hat{\alpha}_i$ is the corresponding coefficients in α.

IV. LINEAR EXTENSION BODIES ALIGNMENT

As we pursue sparse coding method, the bodies of persons need to be aligned among images. So far, sparse coding have been used in face recognition images, most of which are cropped and aligned by the location of eyes. However, due to the variations of pose and viewpoint, it is difficult to accurately find matching points as eyes in bodies. Employing misalignment body images directly in dictionary learning process can lead to unreliable result. Thus, a Linear Extension Bodies Alignment(LEBA) method is proposed to construct the dataset used in our sparse coding method.

Different from face recognition images which only contain small area of background, there about thirty to fifty percent is background area in person images. The first step of LEBA is trying to discard the background region. Here Deep Decompositional Network (DDN, proposed in [14])

(a) (b) (c)

Fig. 1. An example of redraw result. (a) The original image; (b) The foreground obtained by DDN; (c) The redrawed image.

is used to eliminate the impact caused by background as much as possible. DDN is robust to complex pose variations, occlusions and background clutters, and it can parse a person image into four coarse scale regions including: "head (h)", "upper-body (up-b)", "lower-body (lo-b)" and "background (bg)", represented as region set R where $R = \{h, up-b, lo-b, bg\}$. Given some person images $I_i = \{I_1, I_2, ..., I_n\}$, the background region is discarded respectively on each image by the use of bg in R:

$$I_i^{fr} = I_i \odot I_i^{bg}, \tag{3}$$

where I_i^{bg} is the background mask, the value equals to 0 when the pixel in I_i^{bg} belongs to bg and 1 other wise, $\odot$ is Hadamard product.

After discarding the background, the boundaries between upper-body and lower-body which are found between $up-b$ and $lo-b$ can be employed as bodies alignment location among persons. When images are being coded, the dimension of query image and samples in dictionary must agree. Therefore, the background region must be redrawed. Based on the hypothesis that images belong to the same person contain more pixels in similar color than images belong to different person. We fix the alignment location, and respectively resize the upper and lower bodies to the same vertical size among persons. After that, bicubic interpolation is employed to redraw the background region line by line. An example is shown in Fig. 1.

V. DISCRIMINATIVE DICTIONARY LEARNING

To improve the classification performance, a dictionary learning (DL) method is proposed in this subsection.

Suppose that we have n classes of person, the training samples can be denoted by $T_i = [T_1, T_2, \ldots, T_n]$, where T_i is the sub-set, which is comprised of the i^{th}-person images. We try to obtain a trained dictionary D and coding coefficient α, subject to $T \approx D\alpha$. In order to make expression simple, here we rewrite D as $D = [D_1, D_2, \ldots, D_n]$, and $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_n]$, where D_i is the sub-dictionary, which is comprised of dictionary atoms that belong to i^{th}-person, and α_i is the matrix that contains the coding coefficients of T_i over D. We rewrite α_i further as $\alpha_i = \left[\alpha_i^1; \alpha_i^2; \ldots; \alpha_i^n\right]$, where α_i^j is the coding coefficient of T_i over the sub-dictionary D_j. In fact, here $T_i \approx D\alpha_i = D_1\alpha_i^1 + D_2\alpha_i^2 + \cdots + D_n\alpha_i^n, i = 1, 2, \ldots, n$.

To obtain the trained dictionary, we employ three terms of constraint in our dictionary learning model. Take T_i for example, our constraints are described as follows:

i.) Sub-set T_i should be well represented by D, i.e. T_i and $D\alpha_i$ is nearly the same ($T_i \approx D\alpha_i$), which implies $\|T_i - D\alpha_i\|_F^2$ is small.

ii.) Sub-set T_i should be well represented by sub-dictionary D_i, because an acceptable sparse coding result means that a test image belongs to person i is represented almost by images belong to person i, but not by images belong to other person. Thus, $\left\|T_i - D_i\alpha_i^i\right\|_F^2$ is small.

iii.) As is discussed above, sub-set T_i shouldn't be well represented by sub-dictionary D_j, $j \neq i$, i.e. coefficients of images belong to person i (α_i^i) is relatively large, and coefficients of images belong to other person (α_i^j, $j \neq i$) is nearly zero. That implies $\sum_{j=1, j\neq i}^{n} \left\|D_j\alpha_i^j\right\|_F^2$ is small.

We require that all sub-set T_i, $i = 1, 2, \ldots, n$ subject to the constraints above. Finally, we define our dictionary learning model as:

$$G_{(D,\alpha)} = \arg\min_{(D,\alpha)} \sum_{i=1}^{n} \begin{pmatrix} \|T_i - D\alpha_i\|_F^2 \\ +\lambda_1 \left\|T_i - D_i\alpha_i^i\right\|_F^2 \\ +\lambda_2 \sum_{j=1, j\neq i}^{n} \left\|D_j\alpha_i^j\right\|_F^2 \end{pmatrix} \tag{4}$$

VI. OPTIMIZATION OF DL MODEL

In this section, we describe how to optimize the dictionary learning model.

Eq. 4 is a multi-variable optimization problem, which can be solved by optimizing D and α alternatively as some other multi-variable optimization problem. The procedures of optimization can be divided into four steps:

Step 1. Initialization D and α.

We require training set T, and training set labels tl as input. We randomly initialize all atoms in D as unit vectors. Then initialize coding coefficients α as zero.

Step 2. Fix D and optimize α.

Eq. 4 will reduce to

$$G_\alpha = \arg\min_{\alpha} \sum_{i=1}^{n} \begin{pmatrix} \|T_i - D\alpha_i\|_F^2 \\ +\lambda_1 \left\|T_i - D_i\alpha_i^i\right\|_F^2 \\ +\lambda_2 \sum_{j=1, j\neq i}^{n} \left\|D_j\alpha_i^j\right\|_F^2 \end{pmatrix} \tag{5}$$

Many standard convex optimization techniques are available to solve the optimization of Eq. 5. In our model, we use the algorithm in [15].

Step 3. Fix α and optimize D.

We try to update D atom by atom. For example, suppose that $d_l \in D$ is an atom, and β_l is the corresponding coding coefficient. When optimizing d_l, all other atoms are fixed.

We try to extract $d_l\beta_l$. Eq. 4 can be converted to

$$G_{d_l} = \arg\min_{d_l} \begin{pmatrix} \sum_{i=1}^{n} \|T_i - \sum_{d_p \in D}^{d_p \neq d_l} d_p\beta_p - d_l\beta_l\|_F^2 + \\ \lambda_1 \sum_{i=1}^{n} \|T_i - \sum_{d_q \in D_i}^{d_q \neq d_l} d_q\beta_q - d_l\beta_l\|_F^2 + \\ \lambda_2 \sum_{i=1, i\neq m}^{n} \| \sum_{d_r \in D, d_r \notin D_i}^{d_r \neq d_l} d_r\beta_r + d_l\beta_l\|_F^2 \end{pmatrix} \tag{6}$$

Notice that when all other atoms are fixed, $T_i - \sum_{d_p \in D}^{d_p \neq d_l} d_p\beta_p$, $T_i - \sum_{d_q \in D_i}^{d_q \neq d_l} d_q\beta_q$, and $\sum_{d_r \in D, d_r \notin D_i}^{d_r \neq d_l} d_r\beta_r$ are fixed. Thus, we rewrite them as Y_1, Y_2, and Y_3, respectively. We require that each atom in D is a unit vector, i.e. $d_l^T d_l = 1$. Eq. 6 can be rewrited as

$$G_{d_l} = \arg\min_{d_l} \begin{pmatrix} \sum_{i=1}^{n} \|Y_1 - d_l\beta_l\|_F^2 + \\ \lambda_1 \sum_{i=1}^{n} \|Y_2 - d_l\beta_l\|_F^2 + \\ \lambda_2 \sum_{i=1, i\neq m}^{n} \|Y_3 + d_l\beta_l\|_F^2) \end{pmatrix}, d_l^T d_l = 1 \tag{7}$$

Eq. 7 is a single-variable optimization problem. This optimization problem can be solved by using Langrage multiplier.

Take the first term for example, after adding Langrage term, the first term can be converted to

$$\begin{aligned} J_{d_l,\gamma} &= \arg\min_{d_l} \|Y_1 - d_l\beta_l\|_F^2 - \gamma(d_l^T d_l - 1) \\ &= \arg\min_{d_l} tr(Y_1 - d_l\beta_l)(Y_1 - d_l\beta_l)^T - \gamma d_l^T d_l + \gamma \\ &= \arg\min_{d_l} tr(Y_1 Y_1^T - Y_1\beta_l^T d_l^T - d_l\beta_l Y_1^T \\ &\quad + d_l\beta_l\beta_l^T d_l^T - \gamma d_l^T d_l + \gamma) \end{aligned} \tag{8}$$

Differentiating $J_{d_l,\gamma}$ with respect to d_l, d_l^T and γ, then let them be 0, we have

$$-Y_1\beta_l^T + d_l\beta_l\beta_l^T - \gamma d_l = 0 \tag{9}$$

then,

$$d_l = Y_1\beta_l^T / \left\|Y_1\beta_l^T\right\|_2 \tag{10}$$

where $\|\cdot\|_2$ is the l_2-norm.

The remaining terms are similar to Eq.10, and finally, the solution of Eq. 4 under constrain $\boldsymbol{d}_l^{\mathrm{T}}\boldsymbol{d}_l = 1$ is

$$\boldsymbol{d}_l = \frac{\sum_{i=1}^{n}(Y_1\boldsymbol{\beta}_l^{\mathrm{T}} + \lambda_1 Y_2\boldsymbol{\beta}_l^{\mathrm{T}}) - \lambda_2 \sum_{i=1, i\neq m}^{n} Y_3\boldsymbol{\beta}_l^{\mathrm{T}}}{\left\|\sum_{i=1}^{n}(Y_1\boldsymbol{\beta}_l^{\mathrm{T}} + \lambda_1 Y_2\boldsymbol{\beta}_l^{\mathrm{T}}) - \lambda_2 \sum_{i=1, i\neq m}^{n} Y_3\boldsymbol{\beta}_l^{\mathrm{T}}\right\|_2} \tag{11}$$

Step 4. Iteration

If the value of $G_{(D,\alpha)}$ are close enough to last iteration or the number of iterations reachs to maximum, output D; else, return to ***Step 2***.

The whole procedure of optimization is summarized in Table I.

TABLE I
ALGORITHM OF DISCRIMINATIVE DICTIONARY LEARNING

Discriminative Dictionary Learning
Require : Training set T, Training set labels tl
1: $D^{(0)} \leftarrow$ randomly initialize, $\alpha^{(0)} \leftarrow 0$, $t \leftarrow 1$;
2: **while** $(G^{(t)} - G^{(t-1)} \geqslant \epsilon$ && $t \leqslant$ max_iter_num) **do**
3: $t \leftarrow t + 1$;
4: $\alpha^{(t)} \leftarrow$ new α; (via algorithm in [15].)
5: **for** (all d_l in $D^{(t-1)}$)
6: $d_l = \frac{\sum_{i=1}^{n}(Y_1\boldsymbol{\beta}_l^{\mathrm{T}} + \lambda_1 Y_2\boldsymbol{\beta}_l^{\mathrm{T}}) - \lambda_2 \sum_{i=1, i\neq m}^{n} Y_3\boldsymbol{\beta}_l^{\mathrm{T}}}{\left\|\sum_{i=1}^{n}(Y_1\boldsymbol{\beta}_l^{\mathrm{T}} + \lambda_1 Y_2\boldsymbol{\beta}_l^{\mathrm{T}}) - \lambda_2 \sum_{i=1, i\neq m}^{n} Y_3\boldsymbol{\beta}_l^{\mathrm{T}}\right\|_2}$;
7: **end for**
8: $D^{(t)} \leftarrow D^{(t-1)}$;
9: $G^{(t)} \leftarrow G_{(D^{(t)},\alpha^{(t)})}$;
10: **end while**
Output : The trained dictionary $D^{(t)}$.

VII. EXPERIMENTAL RESULT

In this section, we evaluate the performances of our method on three public availabe datasets, including *CAVIAR4REID*, *ETHZ*, and *i-LIDS*. We compare DDLSC with several state-of-the-art person re-identification methods.

A. Experimental Setup

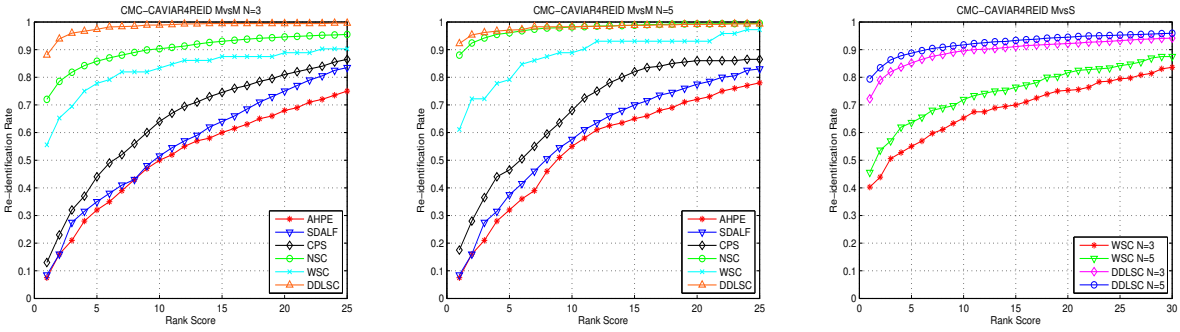
The success of sparse representation based classification owes to the fact that an image can be linearly represented or coded by some other representative samples from the same class. The number of samples for each class makes a great influence on the classification result. Dictionary with only single sample for each person can severely impact the represent result. In consideration of this case, our DDLSC method is suitable for MvsS and MvsM modalities. Therefore, we focus on MvsS and MvsM modalities, and compare DDLSC with several state-of-the-art methods.

Note that not all techniques report results on all three datasets or on all two modalities (MvsS and MvsM). For example, most of state-of-the-art methods report results only for MvsM modality on *CAVIAR4REID*. To make the most comprehensive comparison, we test our method on all modalities and include all reported results from the above methods, when available.

In our experiment, each image is resized to the same size (30×75 pixels), and processed in RGB space. After redrawed, each image is represented as a vector with dimension of $3 \times 30 \times 75 = 6750$.

We denote N as the number of samples for each person in training set. We randomly choose N images from each individuals as the training set (the total number of images in training set is $N\times$number of individuals), and the remaining images as the testing set. We use training set to obtain the dictionary via methods in Algorithm 1.

During the experiment, we find that $\lambda_1 = \lambda_2 = 1$ makes the best trade-off performance on classification between



(a) (b) (c)

Fig. 2. Comparative performance evaluation on *CAVIAR4REID* dataset. (a) MvsM when N = 3; (b) MvsM when N = 5; (c) MvsS when N = 3 and N = 5.

TABLE II

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 [16]	0.91	–	–	–	0.906	–	–	–	0.94	–	–	–
PLS [17]	0.79	0.79	–	–	0.745	0.745	–	–	0.775	0.775	–	–
SDALF [1]	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 [18]	0.977	–	–	–	0.973	–	–	–	0.98	–	–	–
MRCG [19]	–	0.96	–	–	–	0.97	–	–	–	0.983	–	–
EIML [20]	0.78	0.78	–	–	0.74	0.74	–	–	0.91	0.91	–	–
RPLM [10]	0.77	0.77	–	–	0.65	0.65	–	–	0.83	0.83	–	–
RSR [21]	–	0.92	–	–	–	0.93	–	–	–	–	–	–
WSC [13]	0.975	0.974	0.928	0.964	**1**	1	**0.971**	0.970	**1**	1	**0.984**	**0.992**
DDLSC	**0.998**	**1**	**0.952**	**0.966**	0.990	**1**	0.945	**0.972**	0.992	**1**	0.974	0.986

MvsM modality and MvsS modality. Accordingly, we fixed $\lambda_1 = \lambda_2 = 1$ for all experiments reported in this paper. As for maximum number of iterations I, experiments show that most of the obejective function value G will converge when the number of iterations reach to 35, thus, we set $I = 35$.

We employ The Cumulative Matching Characteristic (CMC) curve to measure re-identification performance. A CMC curve represents the expectation of finding the correct match in the top r matches, where r is the rank considered in the final ranking result.

We repeat all experiments for 10 times, and we use the average results to compare with state-of-the-art methods.

B. Results on CAVIAR4REID

CAVIAR4REID is extracted from the well-known *CAVIAR* dataset mostly famous for person tracking and detection evaluations. This dataset contains 72 unique individuals recorded from two different points of view.

For MvsM, we compared the recognition rate of DDLSC with several state-of-the-art methods including AHPE [16], SDALF [1], CPS [18], NSC [22] and WSC [13]. The CMC curves are reported in Fig.2. Our method outperform competing methods at rank-1 in all modalities on this dataset. We improve on the state-of-the-art by 16 percentage points for MvsM (N = 3) and by 4 percentage points for MvsM (N = 5).

For MvsS modality, we compared DDLSC with WSC [13]. DDLSC achieves 72.2% and 79.4% at rank-1 when N = 3 and N = 5, respectively.

C. Results on ETHZ

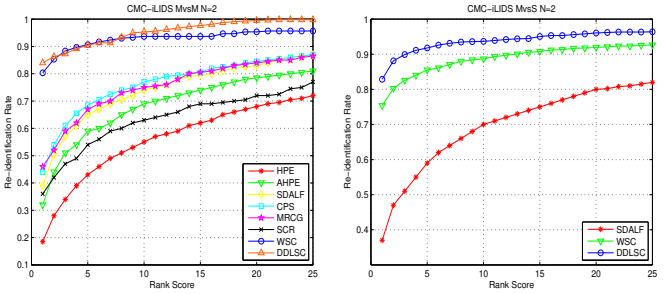
ETHZ contains three sub-datasets including *ETHZ1*, *ETHZ2* and *ETHZ3*. This dataset is captured from moving cameras in a crowed street. *ETHZ1*, *ETHZ2* and *ETHZ3* contains 83 people (4,875 images), 35 people (1,936 images), and 28 people (1,762 images), respectively.

The performances at rank-1 for the MvsS and MvsM (N = 5 and N = 10) modalities on *ETHZ* 1, 2, 3 datasets are summarized in Table II.

Our method outperforms other methods on *ETHZ1*. DDLSC reachs to a high recognition rate, and it is comparative with WSC. A noticeable result is that DDLSC reach to 100% recognition rate on *ETHZ1* (MvsM, N = 10) and *ETHZ3*(MvsM, N = 10) at rank-1.

D. Results on iLIDS

This dataset was created from the pedestrians observed in two non-overlapping camera views from the *i-LIDS* Multiple-Camera Tracking Scenario (MCTS) dataset. 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.



(a) (b)

Fig. 3. Results on *i-LIDS* dataset. (a) MvsM when N = 2; (b) MvsS when N = 2.

In Figure. 3 (a) and (b), we report the CMC curves for DDLSC and the state-of-the-art for MvsM (N = 2) and MvsS (N = 2) modality, respectively.

As is shown in Fig. 3, DDLSC outperforms current methods up to 4% for MvsM, and 6% for MvsS, respectively. DDLSC reachs to 100% at rank-25 for MvsM, while in MvsS modality, DDLSC approachs to 96% at around rank-25.

VIII. CONCLUSIONS

In this paper we propose an approach for person re-identification that is based on sparse coding. We showed how to apply sparse coding method to re-identification problem. After redrawed, we can obtain a more reliable linearly represent result for a person image. In addition, a discrimitive dictionary learning model is proposed. Experimental results demonstrate the effectiveness of the proposed approach for person re-identification problem.

Future work includes: (1) in this paper, an image is represented as a vector with dimension of 6750, which causes an expensive computation cost. We may abstract a subspace that the dimension has been reduced. We can combine feature descriptor with sparse coding, where an image can be represented as a vector with low dimension with much less computational cost. (2) the dictionary learning model can be replaced to other constraints, which may leads to a better performance.

ACKNOWLEDGMENT

This study was partially supported by the National Natural Science Foundation of China (No. 61472019), the National High Technology Research and Development Program of China (No. 2013AA01A603) and the National Science & Technology Pillar Program (No.2015BAF14B01). Supported by the Programme of Introducing Talents of Discipline to Universities and the Open Fund of the State Key Laboratory of Software Development Environment under grant #SKLSDE-2015KF-01.

REFERENCES

[1] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 2010, pp. 2360–2367.

[2] H. Tang and T. S. Huang, "3d facial expression recognition based on automatically selected features," in *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*. IEEE, 2008, pp. 1–8.

[3] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 2, pp. 210–227, 2009.

[4] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, vol. 3, no. 5. Citeseer, 2007.

[5] T. Gandhi and M. M. Trivedi, "Panoramic appearance map (pam) for multi-camera based person re-identification," in *Video and Signal Based Surveillance, 2006. AVSS'06. IEEE International Conference on*. IEEE, 2006, pp. 78–78.

[6] N. Gheissari, T. B. Sebastian, and R. Hartley, "Person reidentification using spatiotemporal appearance," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2. IEEE, 2006, pp. 1528–1535.

[7] B. Ma, Q. Li, and H. Chang, "Gaussian descriptor based on local features for person re-identification," in *Computer Vision-ACCV 2014 Workshops*. Springer, 2014, pp. 505–518.

[8] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 3586–3593.

[9] N. D. Bird, O. Masoud, N. P. Papanikolopoulos, and A. Isaacs, "Detection of loitering individuals in public transportation areas," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 6, no. 2, pp. 167–177, 2005.

[10] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof, "Relaxed pairwise learned metric for person re-identification," in *Computer Vision–ECCV 2012*. Springer, 2012, pp. 780–793.

[11] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja, "Pedestrian recognition with a learned metric," in *Computer Vision–ACCV 2010*. Springer, 2011, pp. 501–512.

[12] W.-S. Zheng, S. Gong, and T. Xiang, "Reidentification by relative distance comparison," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 3, pp. 653–668, 2013.

[13] Y. W. Zheng, S. Hao, B. C. Zhang, J. Zhang, and X. Zhang, "Weight-based sparse coding for multi-shot person re-identification," *Sciece China Information Sciences*, vol. 58, no. 10, pp. 1–15, 2015.

[14] P. Luo, X. Wang, and X. Tang, "Pedestrian parsing via deep decompositional network," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 2648–2655.

[15] S. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "A method for large-scale l1-regularized least squares problems with applications in signal processing and statistics," *IEEE J. Select. Topics Signal Process*, vol. 1, no. 4, pp. 606–617, 2007.

[16] L. Bazzani, M. Cristani, A. Perina, and V. Murino, "Multiple-shot person re-identification by chromatic and epitomic analyses," *Pattern Recognition Letters*, vol. 33, no. 7, pp. 898–903, 2012.

[17] W. R. Schwartz and L. S. Davis, "Learning discriminative appearance-based models using partial least squares," in *Computer Graphics and Image Processing (SIBGRAPI), 2009 XXII Brazilian Symposium on*. IEEE, 2009, pp. 322–329.

[18] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification." in *BMVC*, vol. 1, no. 2. Citeseer, 2011, p. 6.

[19] S. Bak, E. Corvee, F. Bremond, and M. Thonnat, "Multiple-shot human re-identification by mean riemannian covariance grid," in *Advanced Video and Signal-Based Surveillance (AVSS), 2011 8th IEEE International Conference on*. IEEE, 2011, pp. 179–184.

[20] M. Hirzer, P. M. Roth, and H. Bischof, "Person re-identification by efficient impostor-based metric learning," in *Advanced Video and Signal-Based Surveillance (AVSS), 2012 IEEE Ninth International Conference on*. IEEE, 2012, pp. 203–208.

[21] M. T. Harandi, C. Sanderson, R. Hartley, and B. C. Lovell, "Sparse coding and dictionary learning for symmetric positive definite matrices: A kernel approach," in *Computer Vision–ECCV 2012*. Springer, 2012, pp. 216–229.

[22] Y. Liu, S. S. Ge, C. Li, and Z. You, "-ns: A classifier by the distance to the nearest subspace," *Neural Networks, IEEE Transactions on*, vol. 22, no. 8, pp. 1256–1268, 2011.