2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

Cross-view Transformation based Sparse Reconstruction for Person Re-identification

Wei-Xiong He[†], Ying-Cong Chen[†], Jian-Huang Lai[§,‡,*]
[†]School of Electronics and Information Technology, Sun Yat-sen University, China
[§]School of Data and Computer Science, Sun Yat-Sen University, China
[‡]Guandong Key Laboratory of Information Security Technology
hewx5@mail2.sysu.edu.cn, yingcong.ian.chen@gmail.com, stsljh@mail.sysu.edu.cn

***Abstract*—Based on minimum reconstruction error criterion and the intrinsic sparse property of natural data, sparse representation (SR) has shown promising performance on various image recognition tasks. However, in the field of person re-identification (re-id), the state-of-the-art is still dominated by other methods such as metric learning or CNN. It is because samples in one view may not be representative enough to represent samples from another view. As such, the reconstruction error could be excessive, and different pedestrians are indistinguishable with the coefficient produced by sparse representation. In this paper, we proposed an asymmetric sparse representation to address this problem. Samples of different camera views (gallery and probe samples) are mapped to a common latent space and the sparse coefficient is generated in this space. In this way, the representation power is enhanced and the sparse coefficient becomes more reliable. The similarities of different samples are determined by the enhanced sparse coefficient, which allows more discriminative matching across different camera views. Extensive experiments on CAVIAR4REID, iLIDS-VID and PRID 2011 datasets have demonstrated the merits of our approach.**

I. Introduction

Nowadays, most of public infrastructures such as airports, railway stations, hospitals have been equipped with camera networks for surveillance. However, these camera networks face the problem of non-overlapping field between different views, which prevents tracking pedestrians or analyzing their activities across cameras simply based on time and space cues. Therefore, it is critical to re-identify a pedestrian based on his/her appearance. Such a problem is known as the person re-identification (re-id). Because of the large variations of illumination, pose or viewpoint, the appearance of pedestrian images usually changes dramatically across different camera views (see Fig. 1). In order to eliminate the gap between different camera views, various approaches have been proposed, among which the most well investigated are pedestrian descriptors [10], [12] and metric learning [3], [4], [9], [15], [19]. However, due to the extremely complex environmental conditions, it is almost impossible to design a reliable descriptor that is both robust and discriminative. Metric learning alleviates this problem by utilizing supervision information to push relevant image pairs together while pulling irrelevant pairs apart.

Benefiting from the ability to capture sparse property of the natural images, sparse representation (SR) has been proven effective in the field of image recognition such as face recognition [17]. However, although some works tried to deal with the re-id problem with SR, they have not achieved the state-of-the-art so far. This may be because features of different camera views are very different, which limits the power to represent a probe sample from one camera view with the gallery samples from a different view. Note that most SR methods are based on the minimum reconstruction error criterion. Such poor representation power restricts the reliability of the computed reconstruction error.

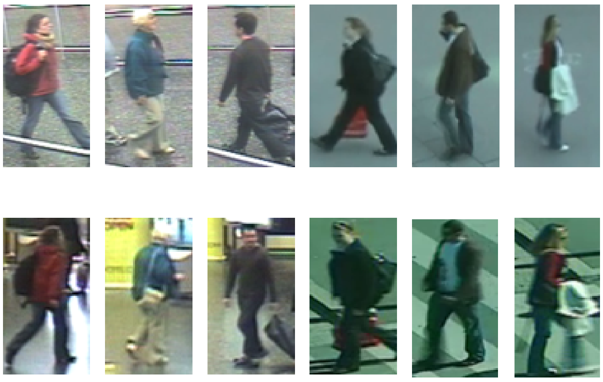


Fig. 1. Illustration of pedestrian images pairs with large appearance variation.

The main difficulty to leverage SR for the re-id problem is to reduce the gap between different views so that samples from one view could obtain better representation power for samples from other views. Recently, Chen et al. [3] proposed an asymmetric metric learning to mitigate the distribution mismatch problem in person re-id. Inspired by the idea that using different but related mappings for different camera views to reduce the discrepancy between different camera views, we propose to incorporate view-specific mappings in the SR framework. Specifically, we firstly combine features of the same pedestrian into a unitary feature vector by average pooling (for the computational issue), then jointly learn the view-specific mappings and the sparse coefficient for each pedestrian. The similarity between two pedestrians can be calculated by computing their reconstruction coefficient, which is equivalent to their matching probability after imposing explicit constraints. The key assumption is that although

[*]Corresponding author

978-1-5090-4846-5/16/$31.00 ©2016 IEEE

samples from one view cannot well represent those in the other view in the original feature space, by projecting them into a common space, samples of different views become closer and thus the representation power is improved. Note that our method is inherently different from CVDCA which only learns view-specific mappings and directly use the mapped features for re-id. Our method additionally learns the sparse representations from the data, so that the structure of the gallery and probe samples can be utilized. Such structure is proven to be useful in the re-id task [8], [11], [13], and the performance is further improved after utilizing them to bring additional cues. Our method is also different from DVDL[8] which discriminatively trained a viewpoint invariant dictionary in the learned subspace from LFDA[15]. We jointly learn the cross-view transformation and the sparse coefficient, so the proposed transformation is optimal for our model. We have demonstrated that the proposed transformation obtained better performance than transformation learned from LFDA. We validate our algorithm proposed in this paper using three publicly available multi-shot re-id datasets: CAVIAR4REID[5], iLIDS-VID[16] and PRID 2011[7]. The experiment results show that our algorithm performs excellently in the multi-shot situation.

To summarize, our contributions include:

1) We addressed the cross-view reconstruction problem in person re-identification.
2) We proposed an asymmetric sparse representation to address the reconstruction problem.
3) Our method significantly outperforms the state-of-the-art on CAVIAR4REID[5], iLIDS-VID[16] and PRID 2011[7] datasets.

II. Review of Related Method

In this section, we will review the following two methods which is related to our method.

A. Sparse Reconstruction

The main idea about Sparse Reconstruction[17] is that, given sufficient samples $y_{i,1}, y_{i,2}...y_{i,n}$ from class i, a testing sample y of the same class should approximately lie in the linear span of the training samples:

$$y \approx s_{i,1}y_{i,1} + s_{i,2}y_{i,2} + \ldots + s_{i,n}y_{i,n} = Y s_i \qquad (1)$$

Where $s_i = [s_{i,1}, s_{i,1}, ..., s_{i,n}]$ represents the vector of reconstruction coefficients.

In the situation of person re-identification, a probe sample p can be approximately represented as:

$$p \approx s_1 G_1 + s_2 G_2 + \ldots + s_n G_n = Gs \qquad (2)$$

Where $G_i = [g_{i,1}, g_{i,2}, \ldots, g_{i,n_i}]$ is the sample set of the i_{th} person, $G = [G_1, G_2, \ldots, G_n]$ is the gallery sample set.After the solution vector $\tilde{s}$ is obtained, the reconstruction error is defined as:

$$e_{i,j} = \frac{\|p - g_{i,j}\tilde{s}_{i,j}\|_2}{\|p\|_2} \qquad (3)$$

Where $g_{i,j}, \tilde{s}_{i,j}, e_{i,j}$ denote the j_{th} sample of the i_{th} person, its corresponding reconstruction coefficient and reconstruction error, respectively.

Finally, the gallery images are ranked by their reconstruction error. However, there are large intra-class variations between pedestrian images across different cameras. Large reconstruction error exists between the probe images and the corresponding gallery images. Therefore, the performance of the above sparse reconstruction method tends to be unsatisfactory.

B. Cross-view Feature Mapping

The Cross-view Feature Mapping for person re-identification [3] is proposed to solve the feature discrepancy problem across non-overlapping camera views. For different cameras, the cross-view transformations is learned by the following objective function:

$$\begin{aligned} \min_{U^1,U^2,\cdots,U^N} & \\ & \sum_{p=1}^{N-1}\sum_{q=p+1}^{N}\sum_{i=1}^{n^p}\sum_{j=1}^{n^q} W_{ij}^{p,q}\|U^{pT}x_i^p - U^{qT}x_j^q\|_2^2 \\ & +\sum_{p=1}^{N}\sum_{i=1}^{n^p}\sum_{j=1}^{n^p} W_{ij}^{p,q}\|U^p x_i^{pT} - U^p x_j^{pT}\|_2^2 \\ & +\lambda\sum_{p=1}^{N-1}\sum_{q=p+1}^{N}\|U^p - U^q\|_F^2 \\ s.t. \quad & U^{kT}M^kU^k = I; k = 1,2,\cdots,N \end{aligned} \qquad (4)$$

Where $W_{ij}^{p,q}$ is the weight on each pair of samples between view p and view q, U^p is the projection of view p. The key idea of this method is to use different mappings for different views so that the discrepancy of different views is reduced.

III. Proposed Method

A. Problem Specification

Let $G = [G_1, G_2, ..., G_{n_g}]$ denote the feature matrix extracted from the gallery images, where n_g denotes the number of pedestrian, $G_i = [g_{i,1}, g_{i,2}, ..., g_{i,n_i}]$ is the feature set of the i_{th} pedestrian, and n_i is the number of images belonging to the i_{th} pedestrian. The definition of feature matrix $P = [P_1, P_2, ..., P_{n_p}]$, extracted from the probe images, is similar to G.

We firstly apply average pooling to the gallery feature matrix G as [8]. Then the mean gallery feature matrix $\overline{G} = [\overline{G}_1, \overline{G}_2, \cdots, \overline{G}_{n_g}]$ is obtained, where $\overline{G}_i$ is the mean feature of images with label i in the gallery. The mean probe feature matrix $\overline{P}$ is calculated too.

B. Problem Formulation

Our idea comes from the intuition that relative images will have smaller reconstruction error while bigger reconstruction error exists among images with different labels in the transformed space.

We firstly consider the relation among a probe sample $\overline{P}_i$, its relative sample $\overline{G}_i$ and irrelative sample $\overline{G}_j$ in the original space:

$$\begin{aligned} \overline{P}_i - \overline{G}_i &= \epsilon_{intra} \\ \overline{P}_i - \overline{G}_j &= \epsilon_{inter} \end{aligned} \qquad (5)$$

Where ϵ_{intra} and ϵ_{inter} are the difference vectors between $\overline{P}_i$ and $\overline{G}_i$, $\overline{P}_i$ and $\overline{G}_j$, respectively.

The reconstruction error between image pairs is defined as:

$$\begin{aligned} e_{intra} &= \frac{\|\overline{P}_i - \overline{G}_i s_i\|_2}{\|\overline{P}_i\|_2} \leq |1 - s_i| + \frac{\|\epsilon_{intra}\|_2}{\|\overline{P}_i\|_2} \\ e_{inter} &= \frac{\|\overline{P}_i - \overline{G}_j s_j\|_2}{\|\overline{P}_i\|_2} \leq |1 - s_j| + \frac{\|\epsilon_{inter}\|_2}{\|\overline{P}_i\|_2} \end{aligned} \tag{6}$$

Where s_i is the intra-class reconstruction coefficient and s_j is the inter-class reconstruction coefficient.

Since the variations between pedestrian images across different cameras are large, even images from the same pedestrian have large gaps, the assumption $\epsilon_{intra} < \epsilon_{inter}$ is not always be satisfied. Therefore, the upper limit of e_{inter} maybe larger than e_{intra}.

Inspired by [3], we apply the cross-view transforation T_A and T_B to the probe samples and the gallery samples respectively, aiming to reduce the upper limit of e_{intra} and increase the upper limit of e_{inter}. The reconstruction error in the transformed space is defined as Eq.(7).

$$\begin{aligned} e_{intra} &= \frac{\|T_A\overline{P}_i - T_B\overline{G}_i s_i\|_2}{\|T_A\overline{P}_i\|_2} \leq |1 - s_i| + \frac{\|\epsilon'_{intra}\|_2}{\|T_A\overline{P}_i\|_2} \\ e_{inter} &= \frac{\|T_A\overline{P}_i - T_B\overline{G}_j s_j\|_2}{\|T_A\overline{P}_i\|_2} \leq |1 - s_j| + \frac{\|\epsilon'_{inter}\|_2}{\|T_A\overline{P}_i\|_2} \end{aligned} \tag{7}$$

Where ϵ'_{intra} and ϵ'_{inter} are the difference vectors between $\overline{P}_i$ and $\overline{G}_i$, $\overline{P}_i$ and $\overline{G}_j$ in the transformed space, respectively.

Normally, it is considered that the samples with same label become closer in the transformed space and samples from different class will be far away. Therefore, it is easier to achieve the goal $e_{intra} < e_{inter}$.

We can also observe from Eq.(7) that when s_k change from 0 to 1, the term $|1 - s_k|$ decreases from 1 to 0. Therefore, in order to minimize the upper limit of e_{intra} and maximize the upper limit of e_{inter}, the cross-view transforation should be equipped with the property that let s_i be close to 1 and s_j become 0, simultaneously.As a result, $\overline{P}_i$ is hoped to be represent as Eq.(8).

$$\begin{aligned} T_A\overline{P}_i \approx T_B\overline{G}_1 \cdot 0 + T_B\overline{G}_2 \cdot 0 + ... + T_B\overline{G}_i \cdot 1 \\ + ... + T_B\overline{G}_{n_g} \cdot 0 = T_B\overline{G}\hat{s} \end{aligned} \tag{8}$$

Where $\hat{s} = [0, 0, ..., 1, 0, ..., 0]^T$.

Furthermore, for any probe image $\overline{P}_i$, we convert its reconstruction coefficient vector to the matching probability by imposing explicit constraints:

$$\begin{aligned} \sum_{j=1}^{n_g} s_{i,j} = 1 \\ s_{i,j} \geq 0, \forall j = 1, 2, ..., n_g \end{aligned} \tag{9}$$

Where $s_{i,j}$ is the reconstruction coefficient between probe image $\overline{P}_i$ and the gallery image $\overline{G}_j$. After imposing explicit constraints in Eq.(9), $s_{i,j}$ approximately represents the matching probability between $\overline{P}_i$ and $\overline{G}_j$.

Following the inference above, we define the overall minimization problem as:

$$\begin{aligned} \min_{T_A, T_B, S} \quad & \sum_{i=1}^{n_p} (\|T_A\overline{P}_i - T_B\overline{G}s_i\|_2^2 + \lambda_1\|s_i - \hat{s}_i\|_2^2) \\ & + \lambda_2\|T_A - T_B\|_F^2 \\ s.t. \quad & \sum_{j=1}^{n_g} s_{i,j} = 1, \forall i = 1, 2, \cdots, n_p \\ & s_{i,j} \geq 0, \forall i, j \end{aligned} \tag{10}$$

Where $S = [s_1, s_2, \cdots, s_{n_p}]$, $\hat{s}_i$ is the reconstruction coefficient vector obtained by Eq.(8) with $\overline{P}_i$. The first item of Eq.(10) ensures that the gallery images represent the probe $\overline{P}_i$ and the second item tends to cause s_i close to $\hat{s}_i$. The sparse property is reflected in the term $\lambda_1\|s_i - \hat{s}_i\|_2^2$ because only one coefficient in $\hat{s}_i$ is nonzero. We also add the term $\lambda_2\|T_A - T_B\|_F^2$ as [3] to control the difference between the cross-view transformations, since there could be relation between the contents captured by any two camera views though discrepancy exists across disjoint camera views. Those relate contents contain the existence of the same person and probably similar environments.

Note that our model is different from [3], which combines the models of metric learning and nearest neighbor but independently optimize these model and lead to a suboptimal solution. Our model jointly learns the sparse model and the cross-view transformation T_A and T_B, thus the solution of our methods is optimal.

C. Solving the optimization problem

We employ the alternating directions framework to solve the problem of Eq.(10). Specifically, we alternatively optimize over T_A, T_B and S one at a time, while fixing the other two. The optimization problem can be solved by conducting the following steps iteratively until convergency.

1.Fix T_A, T_B and optimize over S, the optimization problem becomes

$$\begin{aligned} \min_{S} \quad & \sum_{i=1}^{n_p} (\|T_A\overline{P}_i - T_B\overline{G}s_i\|_2^2 + \lambda_1\|s_i - \hat{s}_i\|_2^2) \\ s.t. \quad & \sum_{j=1}^{n_g} s_{i,j} = 1, \forall i = 1, 2, \cdots, n_p \\ & s_{i,j} \geq 0, \forall i, j \end{aligned} \tag{11}$$

While T_A, T_B are fixed, s_p and $s_q (p \neq q)$ are independent. Hence we optimize a column s_p of S at a time, s_p will be updated by solving Eq.(12).

$$\begin{aligned} \min_{s_p} \quad & \|T_A\overline{P}_i - T_B\overline{G}s_p\|_2^2 + \lambda_1\|s_p - \hat{s}_p\|_2^2 \\ s.t. \quad & \sum_{j=1}^{n_g} s_{p,j} = 1 \\ & s_{p,j} \geq 0, \forall j = 1, 2, \cdots, n_g \end{aligned} \tag{12}$$

We use CVX[6] to solve this problem which conforms to disciplined convex programming.

2.Fix T_B, S, we define Γ as

$$\Gamma = \sum_{i=1}^{n_p}(\|T_A\overline{P}_i - T_B\overline{G}s_i\|_2^2 + \lambda_1\|s_i - \hat{s}_i\|_2^2) \tag{13}$$

let $\frac{\partial\Gamma}{\partial T_A} = 0$, we get

$$T_A = T_B(\overline{G}S\overline{P}^T + \lambda_2 I)(\overline{P}\,\overline{P}^T + \lambda_2 I)^{-1} \tag{14}$$

3.Fix T_A, S, let $\frac{\partial\Gamma}{\partial T_B} = 0$, we obtain

$$T_B = T_A(\overline{P}S^T\overline{G}^T + \lambda_2 I)(\overline{G}SS^T\overline{G}^T + \lambda_2 I)^{-1} \tag{15}$$

Algorithm 1 shows the procedure for solving Eq.(10).

Algorithm 1 The Optimization of Eq.(10)

```
Input: mean matrix $\overline{P}$ and $\overline{G}$, parameter $\lambda_1, \lambda_2$
Output: the cross-view transformation $T_A$ and $T_B$
Initialize: $T_A = I, T_B = I, S$
while not converge do
Step 1: Update each column of $S^1$ by solving Eq.(12)
Step 2: Update $T_A$ by Eq.(14)
Step 3: Update $T_B$ by Eq.(15)
end while
```

D. Re-Identification

Given a mean probe sample $\overline{P}_i^t$ form testing mean probe set $\overline{P}^t$ and the testing mean gallery set $\overline{G}^t$. We obtain the matching probability vector by solving Eq.(16) and rank the gallery sample according the matching probability vector.

$$\begin{aligned} \min_{s} \quad & \|T_A\overline{P}_i^t - T_B\overline{G}^t s_i\|_2^2 \\ s.t. \quad & \sum_{j=1}^{n_g^t} s_{i,j} = 1 \\ & s_{i,j} \geq 0, \forall j = 1, 2, \cdots, n_g^t \end{aligned} \tag{16}$$

Where n_g^t is number of ID in the testing gallery set. Fig. 2 shows the process of calculating the matching probabilities between $\overline{P}_i^t$ and each sample in $\overline{G}^t$ and ranking the samples according to the matching probabilities.

IV. Experiment

In this section, we report the performance of the proposed method on CAVIAR4REID dataset, iLIDS-VID dataset and PRID 2011 dataset. All the experiments are repeated 10 times to get an average result. The methods and their results for comparing are in strict accordance with their related papers.

A. Features and Parameter Settings

Features: We extract the LOMO[10] descriptor for each image. This descriptor is robust to the illumination variations and the viewpoint changes. Considering the efficiency of our algorithm, we apply PCA to reduce the dimension of features.

Parameter Settings: In the following experiments, we set $\lambda_1 = 1, \lambda_2 = 0.3$.

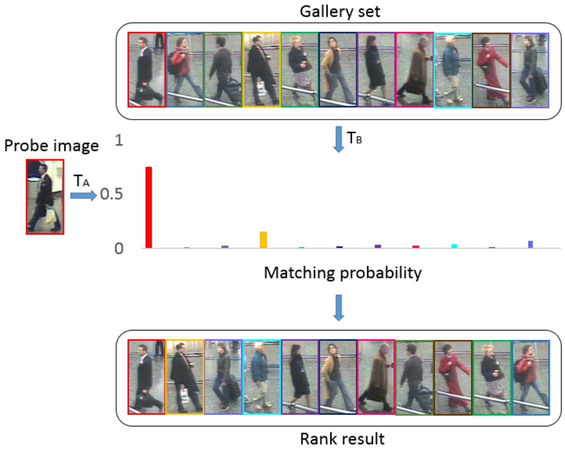


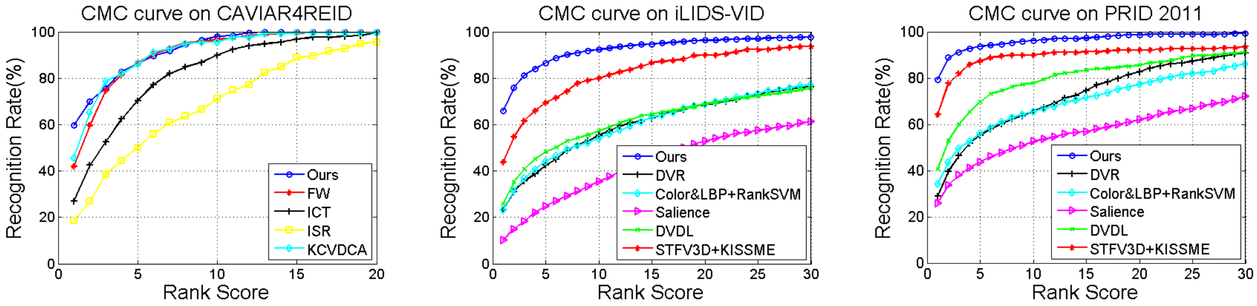
Fig. 2. The process of calculating the matching probability. The color of image border donates the label of image, a bar in the figure donates the matching probability between the probe image and the gallery image which has the same color with the bar.

B. Experiment on CAVIAR4REID

CAVIAR4REID contains 72 pedestrians of which 50 are viewed in disjoint camera views while the other 22 are not. Each image in the CAVIAR4REID dataset has variable scales from 17×39 to 72×144. The experiment is carried out with 10 images for each person. Since there are 22 pedestrians whose images were only captured in a single view in the dataset, we did not select them for experiments and used the rest 50 pedestrians for evaluation. We compare our method with the related methods KCVDCA[3] and ISR[11], the performance of further method such as FW[14] and ICT[1] is also carried out for comparing. The results of the comparison with the State-of-the-Art are shown in Table. I and Fig. 3(a). Note that in [11] the gallery contains images from both views, which may not be realistic in many cases such as cross-camera tracking. We re-evaluate its performance under our setting that gallery and probe images are strictly from different views. We can see that under this setting, ISR does not perform very well. This is consistent with our analysis that images of one view are not representative to those of the other view. Our method significantly outperforms ISR since the sparse coefficient is computed in the enhanced space where the representation power is improved. Our method also outperforms KCVDCA which ranks the gallery samples by using the nearest neighbor module and ignores the structure of the gallery and probe samples.

C. Experiment on iLIDS-VID

The iLIDS-VID dataset consists of 600 image sequences from two camera views. Each image sequence has variable length ranging from 23 to 192 frames. This dataset is very challenging because of clothing similarities among people, lighting and viewpoint variations across camera views, cluttered background and occlusions. Following the evaluation protocol of [16], we randomly choose 150 pedestrian image sequences for training and use the other sequences to form the



(a) CAVIAR4REID (b) iLIDS-VID (c) PRID 2011

Fig. 3. The cumulative match characteristic curves for CAVIAR4REID, iLIDS-VID and PRID 2011 datasets.

TABLE I
TOP RANKED MATCHING RATE(%) ON CAVIAR4REID COMPARED TO THE STATE-OF-THE-ART.

dataset	CAVIAR4REID			
rank	1	5	10	20
Ours	**59.6**	86.4	**98.0**	**100.0**
KCVDCA[3]	45.6	86.0	95.6	99.6
FW[14]	41.9	**86.5**	96.7	100.0
ISR[11]	18.4	50.0	71.2	95.6
ICT[1]	26.8	70.4	90.0	99.6

TABLE II
TOP RANKED MATCHING RATE(%) ON ILIDS-VID COMPARED TO THE STATE-OF-THE-ART.

dataset	iLIDS-VID			
rank	1	5	10	20
Ours	**65.7**	**86.5**	**92.3**	**96.3**
DVDL[8]	25.9	48.2	57.3	68.9
Color&LBP+RankSVM[2]	23.2	44.2	54.1	68.8
DVR[16]	23.3	42.4	55.3	68.6
STFV3D[12]+KISSME[9]	43.8	69.3	80.0	90.0
Salience[18]	10.2	24.8	35.4	52.9

TABLE III
TOP RANKED MATCHING RATE(%) ON PRID 2011 COMPARED TO THE STATE-OF-THE-ART.

dataset	PRID 2011			
rank	1	5	10	20
Ours	**79.3**	**93.7**	**96.2**	**98.7**
DVDL[8]	40.6	69.5	77.8	85.6
Color&LBP+RankSVM[2]	34.3	56.0	65.5	77.3
DVR[16]	28.9	55.3	65.5	82.8
STFV3D[12]+KISSME[9]	64.1	87.3	89.9	92.0
Salience[18]	25.8	43.6	52.6	62.0

testing set. We compare our algorithm with temporal sequence matching techniques including DVR[16], STFV3D[12] and the dictionary learning method DVDL[8]. Following [16], the results of salience[18] and the method which averages the Color and LBP features of each frame in a sequence and uses rankSVM[2] as the distance metric is calculated for comparing. the previous methods, the comparison is shown in table II. It can be seen that we improve the rank-1 recognition rate from 43.8% to 65.7%, showing the significant superiority in solving the person re-identification problem. We also plot the CMC curve in Fig. 3(b) to show the overall evaluation in this dataset.

D. Experiment on PRID 2011

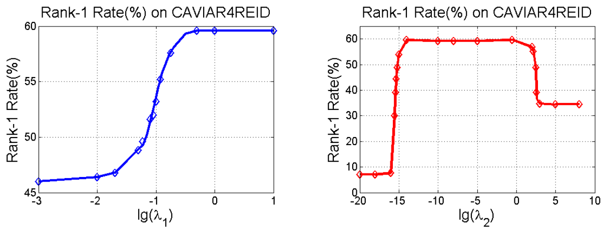
The PRID 2011 dataset consists of image sequences for 200 people in two non-overlapping camera views, each sequences has variable length ranging from 5 to 675 image frames. The images were captured in an uncrowded outdoor environment with significant viewpoint and illumination variations, which lead to the difficulty of person re-identification. Similarly, we follow the evaluation protocol of [7] and select the image sequences whose length are not less than 26, so only 188 image sequences are used. 89 pedestrian image sequences are randomly chosen for training while the others are used for testing. We also compare our algorithm with the methods evaluated on iLIDS-VID. Table III shows the matching score of our method achieves the best rate at rank 1-20 and Fig. 3(c) shows that our method achieves the best performance on this dataset in general.

E. Analysis of our model

In this section, we further evaluate the proposed method in the following two aspects.

1) Parameter analysis: λ_1 and λ_2 are the parameters in our model. λ_1 let the coefficient vector be close to the theoretic vector in the training process. λ_2 decides the similarity between T_A and T_B. We vary λ_1 and λ_2 to show the robustness of our model. The rank-1 recognition rate with different λ_1 and λ_2 on CAVIAR4REID dataset are plotted in Fig.4.

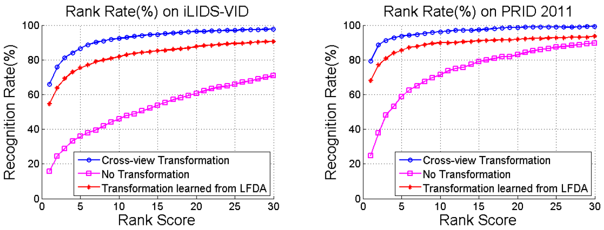
Fig.4(a) shows that when λ_1 reachs a threshold, the performance of our method will become stable and excellent, which shows the validity of the term $\lambda_1\|s_i - \hat{s}_i\|_2^2$ in Eq.(10). Differently, the curve of Fig. 4(b) shows that there are three states of our algorithm if λ_2 is changed. If λ_2 is too small, the transformation is likely to be overfitting. When λ_2 is on a suitable scale, our algorithm has the best performance. Finally, if λ_2 is too large, significant punishment arises when T_A and T_B are different, the transformation is difficult to update. Therefore, our algorithm degenerates into the situation that no transformation is used because we initialize T_A and T_B to the



(a) Rank-1 with various λ_1.

(b) Rank-1 with various λ_2.

Fig. 4. Rank-1 identification rate(%) on CAVIAR4REID.



(a) iLIDS-VID dataset.

(b) PRID 2011 dataset.

Fig. 5. Performance comparison using different transformation.

unit matrix. Actually, the parameters are easy to set because our method has the best performance even the parameters are varied in a large range.

2) Evaluation of Cross-view Transformation: We evaluate the validity of the proposed cross-view transformation in two ways. Firstly, the result of using the proposed cross-view transformation is compared with the situation that no transformation is used. The unitary mapping from LFDA[15] is also evaluated. Fig. 5 shows that the proposed cross-view transformation outperforms the other two situations. Furthermore, the reconstruction error is considered as an evaluation criterion. Fig. 6 indicates that the proposed cross-view transformation is able to reduce the reconstruction error between relative image pairs by comparing with the result of our method with no transformation.

V. Conclusion

In this paper, we propose an effective approach to solve the person re-identification problem. We notice that traditional sparse representation may not work well because of the view discrepancy problem. To solve this problem, we propose to impose a cross-view transformation to transform the features into an enhanced space where samples in one view are representative to samples in the other. The sparse representation is performed in this latent space, aiming to reduce the reconstruction error and obtain a more reliable sparse coefficient. Experiments on three datasets show the superiority of our algorithm. We notice that the efficiency of our algorithm is low if the training set is large, we will focus on the problem of increasing the efficiency of our method in the future work.

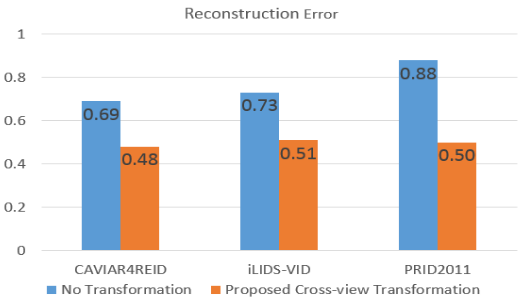


Fig. 6. Mean reconstruction error on CAVIAR4REID, iLIDS-VID and PRID 2011.

Acknowledgment

This project is supported by the NSFC (61573387), the GuangDong Program (Grant No.2015B010105005), the Guangdong Natural Science Foundation (No. 2015A030311047) and Shenzhen Innovation Program (No. JCYJ20150401145529008).

References

[1] T. Avraham, I. G. M. Lindenbaum, and S. Markovitch. Learning implicit transfer for person re-identification. In ECCV Work- Shop, 2012.

[2] O. Chapelle and S. S. Keerthi. Efficient algorithms for ranking with svms. Information Retrieval, pages 201 C 215, 2010.

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

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

[5] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In BMVC, 2011.

[6] M. Grant and S. Boyd. Graph implementations for nonsmooth convex programs. In V. Blondel, S. Boyd, and H. Kimura, editors, Recent Advances in Learning and Control, Lecture Notes in Control and Information Sciences, pages 95 - 110, 2008.

[7] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In Image Analysis, pages 91 - 102, 2011.

[8] S. Karanam, Y. Li, and R. J. Radke. Person re-identification with discriminatively trained viewpoint invariant dictionaries. In ICCV, pages 4516 - 4524, 2015.

[9] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth and H. Bischof. Large scale metric learning from equivalence constraints. In CVPR

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

[11] G. Lisanti, I. Masi, A. D. Bagdanov, and A. D. Bimbo. Person re-identification by iterative re-weighted sparse ranking. T-PAMI, 2015.

[12] K. Liu, B. Ma, W. Zhang, and R. Huang. A spatio- temporal appearance representation for video-based pedestrian re-identification. In ICCV, 2015

[13] C. C. Loy, C. Liu, and S. Gong. Person re-identification by manifold ranking. In Image Processing (ICIP), 2013 20th IEEE International Conference on.

[14] N. Martinel, A. Das, C. Micheloni, and A. Roy-Chowdhury. Re-identification in the function space of feature warps. T-PAMI, pages 1656 - 1669, 2014.

[15] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In CVPR, pages 3318 - 3325, 2013.

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

[17] A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. T-PAMI, pages 210 - 227, 2008.

[18] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In CVPR, pages 3586 - 3593, 2013.

[19] S. Chen, C. Guo, and J. Lai. Deep Ranking for Person Re-Identification via Joint Representation Learning. In TIP, pages 2353 - 2367, 2016.