2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance

Reference-Based Person Re-Identification

Le An*, Mehran Kafai*, Songfan Yang, Bir Bhanu
Center for Research in Intelligent Systems
University of California, Riverside, CA 92521, USA
lan004@ucr.edu, mkafai@cs.ucr.edu, songfan.yang@email.ucr.edu, bhanu@cris.ucr.edu

Abstract

Person re-identification refers to recognizing people across non-overlapping cameras at different times and locations. Due to the variations in pose, illumination condition, background, and occlusion, person re-identification is inherently difficult. In this paper, we propose a reference-based method for across camera person re-identification. In the training, we learn a subspace in which the correlations of the reference data from different cameras are maximized using Regularized Canonical Correlation Analysis (RCCA). For re-identification, the gallery data and the probe data are projected into the RCCA subspace and the reference descriptors (RDs) of the gallery and probe are constructed by measuring the similarity between them and the reference data. The identity of the probe is determined by comparing the RD of the probe and the RDs of the gallery. Experiments on benchmark dataset show that the proposed method outperforms the state-of-the-art approaches.

1. Introduction

Person re-identification is a recognition task in which one matches the individuals across cameras in disjoint views. Accurate person re-identification facilitates the understanding of human behavior in areas covered by surveillance cameras. A direct application of re-identification is people tracking in multi-camera systems [12].

Recently there have been a lot of interests and efforts in person re-identification [6] [5] [18] [15] [10] [20]. However, the person re-identification still remains very challenging due to several reasons: (1) *Low resolution*. Most of the surveillance cameras are not able to capture high-resolution images due to hardware limitations, (2) *Arbitrary poses*. Since a subject is captured by surveillance cameras with non-overlapping field-of-views, the poses of the subject in each camera are usually not similar, (3) *Changing illumination*. The images are captured at different time and/or locations. As a consequence, the appearance of the person may change dramatically due to the illumination change, (4) *Occlusion*. A subject may carry accessories such as a suitcase which occludes the distinctive features of the subject from a certain view. Figure 1 shows some sample subjects captured in two different cameras. Due to large variations in pose, illumination and background, the appearance of the subjects differs significantly in the two views, which makes person re-identification inherently difficult.

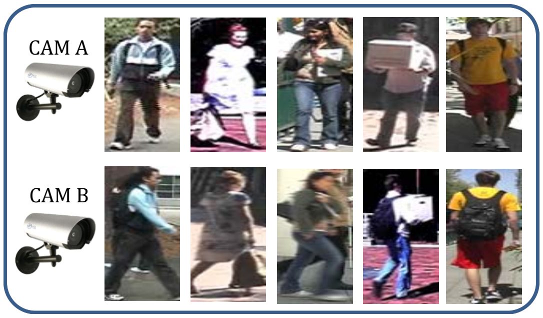

Figure 1. Samples from VIPeR dataset [6] in two camera views.

In order to recognize a given probe from a large gallery, the basic idea is to first extract robust feature representations for both probe and gallery images, and then perform the matching. This kind of approach is called *appearance-based* and only the visual cues are used.

Different appearance-based methods can be categorized into two groups. In the first group, the goal is to extract the feature representations that have low intra-class variation for the same subject and high inter-class variation among different subjects [6] [5] [18]. However, due to the significant appearance change across cameras, the intra-class variation is often larger than the inter-class variation. As a result, accurate classification is very difficult.

Another strategy is to learn the optimal distance measure for the image pairs [20] [10] [13]. The metric learning approaches train a transformation for the original feature representation by which the intra-class distances are minimized while the inter-class distances are maximized. The draw-

*indicates equal contribution. This work was supported in part by ONR grants N00014-12-1-1026, N00014-09-C-0388 and NSF grant 0905671. The contents and information do not reflect the position or policy of the U.S. Government.

978-1-4799-0703-8/13/$31.00 ©2013 IEEE

back of the metric learning approaches is that the learned model tends to overfit the training data. Also, some popular approaches [19] [8] [3] may require high computational costs due to complex optimization.

In this paper, we present a new framework for person re-identification using a reference-based scheme. During the training, a *reference set* containing images from different camera views are used to learn a subspace in which the data from different views are maximally correlated. Regularized Canonical Correlation Analysis (RCCA) is used for subspace learning. In the re-identification, given the probe and gallery images, the features are extracted and projected into the RCCA subspace using the learned projection matrices. Instead of matching the features of the probe and gallery directly, we generate a representation called reference descriptor (RD) using the reference set. The dimension of the RD is determined by the size of the reference set and is irrelevant to the size of the original image features. The matching is performed by measuring the similarity between the RDs of the probe and gallery. In this way, probe and gallery from different views are indirectly compared with respect to the reference set instead of being matched directly.

2. Related Work and Motivation

To extract stable feature representations from different camera views, various pursuits have been reported. In [2], the pictorial structures are adopted to localize the human parts and part-to-part correspondences are searched to match the subjects. Farenzena *et al.* [5] extract features that account for the overall chromatic content, the spatial arrangement and the presence of recurrent local motifs to match individuals with appearance variation. In [1], a model is learned in a covariance metric space to select features based on the idea that different regions for each subject should be matched specifically. Gray *et al.* [7] use AdaBoost to select the most discriminative features instead of using handcrafted features. The re-identification is formulated as a ranking problem with the development of an Ensemble RankSVM (ERSVM) in [17]. In [9] a two-step method is proposed by first using a descriptive model to obtain an initial ranking which is refined in the second step by a discriminative model with human feedback.

Recently, distance learning methods are gaining popularity for re-identification. In [10], a relaxed pairwise metric learning (RPLM) is proposed based on Mahalanobis distance learning which takes advantages of the structure of the data with reduced computational cost, achieving the *state-of-the-art* with simple feature descriptors. Köstinger *et al.* [13] propose a simple yet effective method to learn the distance metric based on a statistical inference perspective. Zheng *et al.* [20] formulate re-identification as a relative distance comparison (PRDC) problem which aims to maximize the likelihood that the distance between a pair of images of the same person is smaller than a pair of images of different people. In [15] a transferred metric learning is motivated by the insight that multiple metrics should be learned for visually different candidate sets. The standard metric learning techniques such as Large Margin Nearest Neighbor (LMNN) [19], Information Theoretic Metric Learning (ITML) [3], and Logistic Discriminant Metric Learning (LDML) [8] are also applicable to person re-identification. Dikmen *et al.* [4] develop a variant of LMNN by introducing a rejection option to the unfamiliar matches (LMNN-R) and achieve improved results.

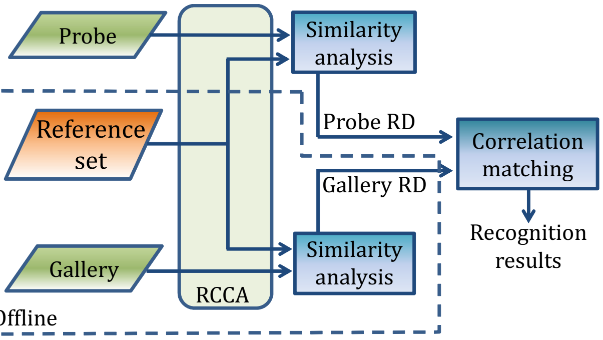

Figure 2. System diagram of our method. In the offline process, the reference set is first used to train a RCCA subspace in which the correlation between the data from two different views are maximized. After subspace projection, the gallery data are compared against all the reference data from the same view camera and the similarity scores form a new representation called reference descriptor (RD). The RD of the probe is generated in a similar manner and the correlation (cosine similarity) of the probe RD and the gallery RDs are compared to decide the final matching result.

Based on the number of images used, the aforementioned approaches can be divided into two groups. The *single-shot* approaches (*e.g.*, [7]) use a single image to describe a subject from one view while the *multiple-shot* approaches (*e.g.*, [9]) extract features from multiple images. In this paper, we propose a novel framework for *single-shot* person re-identification. Instead of designing complex feature representations or learning the distance metric, we generate new feature representations called reference descriptors (RDs) for the probe and gallery data using a *reference set*. The reference set is a set of images of the subjects from different views. The subjects in the probe and gallery are disjoint to the subjects in the reference set. By using the reference set, we bypass the necessity to match the image pairs directly and the image feature space is transformed to the reference space. Experiments on standard benchmark dataset show that the proposed method outperforms current methods.

3. Technical Approach

The system diagram is illustrated in Figure 2. In the offline process, the image pairs of the reference set are used to learn a subspace by Regularized Canonical Correlation

Analysis (RCCA) to maximize the correlation between the data from different views. The features of the galley and probe are extracted and projected into the learned RCCA subspace and then their RDs are generated. The matching is performed by measuring the similarity between the RDs of the gallery and probe images.

3.1. Canonical Correlation Analysis

First introduced in [11], CCA aims to explore the relationship between two sets of random variables from the different observations on the same data (*e.g.*, images of subjects from different views). CCA finds projections such that the correlation between these two sets of random variables is maximized after projection. Figure 3 shows a pictorial example of the CCA principle.

Given two sets of data observations, $D^A = \{d_i^A \in \mathbb{R}^m, i = 1, 2, ..., N\}$ and $D^B = \{d_i^B \in \mathbb{R}^n, i = 1, 2, ..., N\}$, CCA aims at obtaining two sets of basis vectors $W_A \in \mathbb{R}^m$ and $W_B \in \mathbb{R}^n$ such that the correlation coefficient ρ of $W_A^T D^A$ and $W_B^T D^B$ is maximized. The objective function to be maximized is given by

$$\begin{aligned} \rho &= \frac{Cov(W_A^T D^A, W_B^T D^B)}{\sqrt{Var(W_A^T D^A)}\sqrt{Var(W_B^T D^B)}} \\ &= \frac{W_A^T C_{AB} W_B}{\sqrt{W_A^T C_{AA} W_A W_B^T C_{BB} W_B}} \end{aligned} \tag{1}$$

where C_{AA} and C_{BB} are the covariance matrices of D^A and D^B. C_{AB} is the covariance matrix of D^A and D^B.

Equivalently, the CCA can be formulated as a constrained optimization problem by

$$\underset{W_A, W_B}{\operatorname{argmax}} W_A^T C_{AB} W_B \tag{2}$$

subject to $W_A^T C_{AA} W_A = 1$ and $W_B^T C_{BB} W_B = 1$.

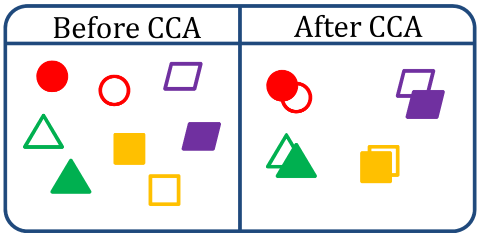

Figure 3. An illustration of the CCA principle. The filled and unfilled shapes indicate data from two views. Before CCA projection, the data from different views are scattered in the original feature space (left). After CCA projection, the features of the same data from different views are better coupled (right).

Using Lagrange multiplier, the solution of the optimization problem for CCA is equivalent to the solution of the following generalized eigenvalue problems

$$\begin{aligned} C_{AB} W_B &= \lambda C_{AA} W_A \\ C_{BA} W_A &= \lambda C_{BB} W_B \end{aligned} \tag{3}$$

where $C_{BA} = C_{AB}^T$. CCA is performed in an unsupervised manner and both correlation maximization and dimension reduction can be achieved simultaneously by choosing the number of basis vectors to use.

3.2. Regularized Canonical Correlation Analysis

Often in practice, the feature dimension of the data is significantly larger than the number of data samples. In this case the covariance matrices C_{AA} and C_{BB} may be singular and their inverse would be ill-conditioned. Regularized CCA (RCCA) has been proposed to solve this problem and it prevents overfitting [14]. In the solution of RCCA, the generalized eigenvalue problem becomes

$$\begin{aligned} C_{AB} W_B &= \lambda (C_{AA} + \lambda_1 I_A) W_A \\ C_{BA} W_A &= \lambda (C_{BB} + \lambda_2 I_B) W_B \end{aligned} \tag{4}$$

where λ_1 and λ_2 are the two non-negative regularization parameters. I_A and I_B are two identity matrices. Usually λ_1 and λ_2 are determined by cross-validation.

3.3. Offline Processing

In the offline training stage, images $\{I_i^A, i = 1, 2, ..., N\}$ and $\{I_i^B, i = 1, 2, ..., N\}$ of N subjects from two different cameras A and B are available as a *reference set*. The features (*e.g.*, color and texture) from each image are extracted and two feature sets $\{F_i^A, i = 1, 2, ..., N\}$ and $\{F_i^B, i = 1, 2, ..., N\}$ are obtained. Since the features are from images in different views, we first learn a RCCA subspace in which the correlations between these two sets of features $\{F_i^A\}$ and $\{F_i^B\}$ are maximized. The RCCA projection matrices W_A and W_B are learned from Eq. 4.

By projecting the original features into the RCCA subspace, we obtain the projected features of the reference set $\{f_i^A, i = 1, 2, ..., N\}$ and $\{f_i^B, i = 1, 2, ..., N\}$ with reduced dimension and enhanced correlation.

Suppose a gallery of M subjects are from camera A, the features of the gallery subjects are extracted and projected using W_A. The representation for the j^{th} subject in the gallery set is f_j^g. From f_j^g a new representation RD R_j^g is calculated by

$$R_j^g = [s(f_j^g, f_1^A), s(f_j^g, f_2^A), \ldots, s(f_j^g, f_N^A)]^T \tag{5}$$

where $s(a, b)$ denotes the similarity between the features a and b. We use the inverse of the Euclidean distance as the similarity measure. In this process, the representation of the

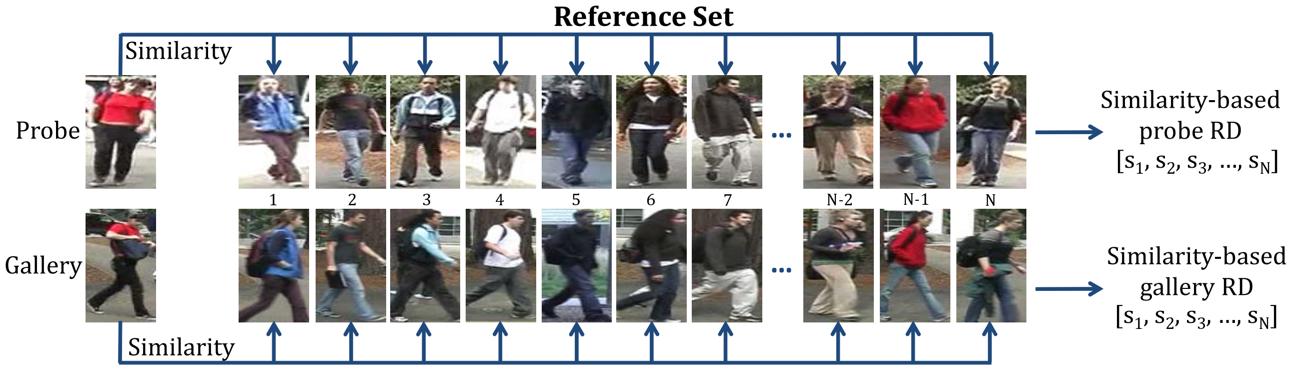

Figure 4. The reference descriptor (RD) for a probe/gallery subject is generated by computing the similarity between the probe/gallery data and each of the subjects in the reference set. N is the number of subjects in the reference set and s_i is the similarity between the probe/gallery and the i^{th} subject in the reference set.

gallery subject is transformed to a descriptor of length N regardless of the original feature dimension and each element in R_j^g indicates the similarity between this gallery subject and a reference subject. The projected features of the reference set from camera A $\{f_i^A, i = 1, 2, ..., N\}$ are acting like basis functions and they jointly describe the appearance of a gallery subject by similarity measures. Figure 4 shows the basic idea of how the RDs are generated.

The rationale for first projecting the features into the RCCA subspace is to better couple the features $\{f_i^A, i = 1, 2, ..., N\}$ and $\{f_i^B, i = 1, 2, ..., N\}$. In the re-identification, a probe image is described using $\{f_i^B, i = 1, 2, ..., N\}$. Since $\{f_i^A, i = 1, 2, ..., N\}$ and $\{f_i^B, i = 1, 2, ..., N\}$ are maximally correlated after RCCA projection, the matching between the RD of the probe and the RD of the gallery becomes meaningful and reliable.

3.4. Re-Identification

The goal of re-identification is to accurately recognize a probe subject in one camera view from a gallery of identified subjects in a different camera view. Suppose the detection of a subject from camera B (I_p) is given as probe, the features F^p are extracted. The projected feature f^p in the RCCA subspace is given by

$$f^p = W_B^T F^p \tag{6}$$

The RD of the probe, R^p, is computed using the projected features of the reference set from camera B $\{f_i^B, i = 1, 2, ..., N\}$ by

$$R^p = [s(f^p, f_1^B), s(f^p, f_2^B), \ldots, s(f^p, f_N^B)]^T \tag{7}$$

where f_i^B is the projected features of the reference subject i in camera B.

The identity of the subject is determined by finding the subject k that is most similar to the probe using cosine similarity between R^p and R_k^g among M gallery subjects

$$\underset{k}{\operatorname{argmax}} \frac{(R^p)^T \cdot R_k^g}{\|R^p\| \, \|R_k^g\|} \tag{8}$$

Compared to other similarity/distance measures (*e.g.*, Euclidean, Chi-square), cosine similarity is computationally efficient especially for high dimensional feature descriptors and/or large datasets.

4. Experiments

4.1. Dataset

We evaluate our method on the VIPeR dataset [1], which is one of the most popular benchmark datasets for person re-identification [6]. The VIPeR dataset is designed in a *single-shot* scenario. It contains image pairs of 632 pedestrians. The images were taken by two cameras with significant view change. In addition, the illumination may also change dramatically. Other aspects such as cluttered background and occlusions further make this dataset more challenging. It is considered as the most challenging dataset currently available for pedestrian re-identification. For each person, a single image is available from each camera view. All of the images in the VIPeR dataset are normalized to 128 × 48. Some sample image pairs are shown in Figure 1.

In our experiments we follow the experimental protocols in the previous work [5] [13]. The image pairs are randomly divided into two sets of 316 pairs each. One set is used as the reference set and the other is used for testing. In the testing, the images from one camera are used as gallery data and images from the other camera are the probes. The experiments are performed 10 times and the average results are reported.

[1] Available at http://vision.soe.ucsc.edu/?q=node/178

4.2. Feature Extraction and Parameters

For feature descriptors we follow the feature extraction scheme in [10]. The HSV and Lab color features are used to describe the color appearance of the subject. For the texture descriptor we use Local Binary Patterns (LBP) [16]. The image is divided into blocks of size 8×16. The blocks are overlapping by 50% in horizontal and vertical directions. Thus, the total number blocks for one image of size 128×48 would be $31 \times 5 = 155$. For each block, the quantized mean values of the HSV and Lab color channels are computed. The 8-bit LBP histogram is extracted from the block and the final feature representation is the concatenation of the means of the color channels and the LBP histogram. In the RCCA projection the first 50 eigenvectors in the projection matrices W_A and W_B are used (*i.e.*, the RCCA reduces the dimensions of the original features to 50). λ_1 and λ_2 are set to $10^{-1.6}$. The number of RCCA dimensions and the regularization parameters are chosen by cross-validation for optimal re-identification accuracy.

4.3. Evaluation Criteria

The top rank recognition rates and the Cumulative Matching Characteristic (CMC) curves are reported. The CMC curve represents the expectation of finding the correct match in the top r matches. In other words, a rank-r recognition rate shows the percentage of the probes that are correctly recognized from the top r matches in the gallery.

4.4. Comparisons to Current Methods

Table 1 shows the comparisons of the proposed method and the *state-of-the-art* approaches that either focus on feature extraction and selection or distance metric learning. As compared to all the current methods, our approach achieves the highest recognition rates in the top ranks listed in Table 1. Figure 5 shows the CMC curves comparison up to the rank 30 since the recognition rates in the top ranks are of most importance. Compared to the second best results by RPLM [10], the proposed method performs better at different ranks and at rank-1 the performance gain by our method is over 11%. Moreover, compared to the recognition results using RCCA projected features only, the combination of RCCA and RDs boosts the results. Note that even when only RCCA projected features are used for direct matching without RDs, the results are still competitive compared to the current methods in Table 1.

Rank→	$r = 1$	10	20	50	100
Proposed	**30**	**75**	**87**	**96**	**99**
RPLM [10]	27	69	83	95	99
PS [2]	22	57	71	87	N/A
SDALF [5]	20	50	65	85	N/A
KISSME [13]	20	62	77	92	98
DDC [9]	19	52	65	80	91
LMNN [19]	18	59	75	91	97
PRDC [20]	16	54	70	87	97
ITML [3]	14	52	71	90	98
ERSVM [17]	13	50	67	85	94
ELF [7]	12	43	60	81	93
LDML [8]	5	21	30	51	71
LMNN-R* [4]	20	68	80	93	99

Table 1. The comparison of the top ranked recognition rates (in %) on the VIPeR dataset. (* indicates the best run reported.)

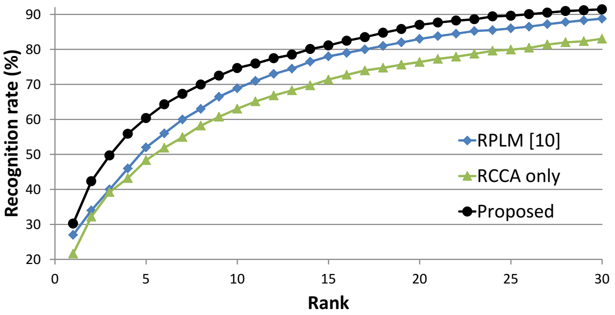

Figure 5. The comparison of the CMC curves for the proposed method and RPLM [10]. The CMC curve by only using RCCA is also provided.

4.5. Effects of Reference Set Size

In Table 2 we evaluate the performance of our method with reduced reference set (referred as training set in [10] and [20]). In this case all the data from the VIPeR dataset are used. That means, as the size of the reference set decreases, the number of subjects in the gallery and probe data increases, which raises the difficulty of the re-identification task. As can be seen in Table 2, even with smaller reference set, the proposed approach still achieves the best results compared to RPLM [10] and PRDC [20]. In addition, the CMC curves of the proposed method up to rank 30 with different reference set sizes are shown in Figure 6.

Figure 7 shows the impact of the size of the reference set on the same testing data. In this case the sizes of the gallery and the probe data remain the same (316, half of the VIPeR dataset). As shown in Figure 7, the recognition rates remain low with a very small reference set. The reason is that a small reference set is not able to effectively describe and

Ref. set size→	N=200			N=100		
Rank→	$r = 1$	10	20	$r = 1$	10	20
RPLM [10]	20	56	71	11	38	52
PRDC [20]	13	44	60	9	34	49
Proposed	**22**	**59**	**75**	**15**	**47**	**60**

Table 2. The comparison of the recognition rates (in %) with different reference set sizes.

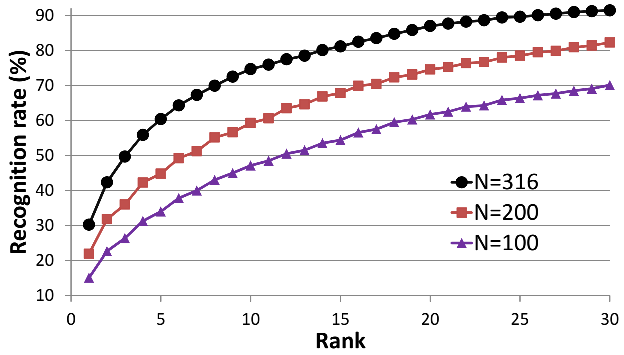

Figure 6. The comparison of the CMC curves of the proposed method with different reference set of size N.

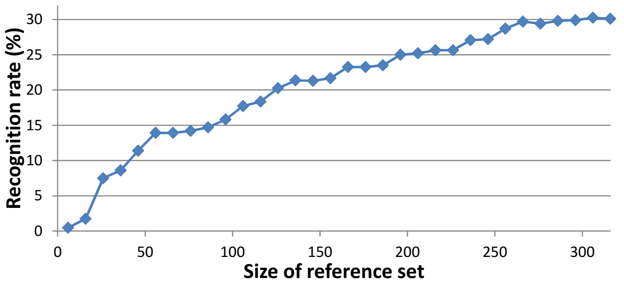

Figure 7. The recognition rates with different sizes of the reference set on the same testing data.

distinguish a much larger gallery and probe set of different subjects. The performance keeps increasing as the reference set expands. With more than 260 subjects in the reference set, the performance tends to stabilize.

5. Conclusions

In this paper, a referenced-based approach for *single-shot* person re-identification is proposed. In contrast to the previous methods in which either sophisticated features are developed or the distance metric is learned, a *reference set* with images from different camera views is utilized. We first project the features of the reference set into a common subspace where their correlation is maximized using Regularized Canonical Correlation Analysis (RCCA). The projected features of the reference set are then used to generate the reference descriptors (RDs) for the gallery and the probe data. The re-identification is performed by comparing the RDs of the probe and gallery subjects. The advantage of the proposed method is that the direct comparison between the gallery and probe in the original feature space is bypassed and the RDs are more distinct among different subjects and more consistent for the same subject despite of the large appearance variation in the original images, as a result of correlation maximization using RCCA. In addition, the dimension of the RDs is irrelevant to the original feature representation. Therefore, different features can be extracted from different camera views for better discrimination. The experiments on a challenging benchmark dataset show that the proposed method outperforms the *state-of-the-art* methods.

References

[1] S. Bak, G. Charpiat, E. Corvee, F. Bremond, and M. Thonnat. Learning to match appearances by correlations in a covariance metric space. In *ECCV*, 2012.

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

[3] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *ICML*, 2007.

[4] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja. Pedestrian recognition with a learned metric. In *ACCV*, 2011.

[5] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010.

[6] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *PETS*, 2007.

[7] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008.

[8] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? metric learning approaches for face identification. In *ICCV*, 2009.

[9] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *SCIA*, 2011.

[10] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *ECCV*, 2012.

[11] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):pp. 321–377, 1936.

[12] W. Hu, M. Hu, X. Zhou, T. Tan, J. Lou, and S. Maybank. Principal axis-based correspondence between multiple cameras for people tracking. *IEEE T-PAMI*, 2006.

[13] M. Köstinger, M. Hirzer, P. Wohlhart, P. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012.

[14] S. E. Leurgans, R. A. Moyeed, and B. W. Silverman. Canonical correlation analysis when the data are curves. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 1993.

[15] W. Li, R. Zhao, and X. Wang. Human reidentification with transferred metric learning. In *ACCV*, 2012.

[16] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE T-PAMI*, 2002.

[17] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang. Person reidentification by support vector ranking. In *BMVC*, 2010.

[18] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu. Shape and appearance context modeling. In *ICCV*, 2007.

[19] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *JMLR*, 2009.

[20] W.-S. Zheng, S. Gong, and T. Xiang. Reidentification by relative distance comparison. *IEEE T-PAMI*, 2013.