ASYMMETRIC CROSS-VIEW DICTIONARY LEARNING FOR PERSON RE-IDENTIFICATION

Minyue Jiang, Yuan Yuan, Qi Wang

School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, Shaanxi, P. R. China, 710072.

ABSTRACT

Person re-identification is a critical yet challenging task in video surveillance which intends to match people over non-overlapping cameras. Most metric learning algorithms for person re-identification use symmetric matrix to project feature vectors into the same subspace to compute the similarity while ignoring the discrepancy between views. To solve this problem, we proposed an asymmetric cross-view matching algorithm with dictionary learning to alleviate the variations in human appearance across different views. Not only the views' dictionaries but also the persons' dictionary codes are constrained. Moreover, the 'between-class' and the 'within-class' distance are taken into consideration which makes the forming dictionary codes more robust and discriminative than the original feature vectors. The effectiveness of our approach is validated on the VIPeR and CUHK01 datasets. Experimental results show the proposed algorithm achieves compelling performance and asymmetric model plays an important role in the proposed approach.

Index Terms— Person re-identification, cross-view matching, dictionary learning

1. INTRODUCTION

Nowadays, public occasions such as shopping malls, airports, railway stations deploy a number of surveillance cameras to meet the growing requirements of security. Person re-identification is a fundamental task in video surveillance which aims to match people across surveillance cameras. In order to get a wide field of view, these cameras put in high position and camera views are not overlapping. Normally, person-reidentificaiton methods use appearance features to re-identify people across non-overlapping views from the captured videos. However, persons' appearance features differ in non-overlapping views due to great changes of illumination, pose or viewpoint and occlusion. These complicated environmental changes increase the difficulty of person re-identification.

This work is supported by National Natural Science Foundation of China under Grant 61379094, Natural Science Foundation Research Project of Shaanxi Province under Grant 2015JM6264.

Person re-identification has been paid more and more attention in the past five years . Many researches [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12] have been done to solve this challenging task. Feature representation [1, 2, 5, 9] and metric learning [4, 6, 7, 8, 9, 11, 12] are two fundamental problems in person re-identification. Approaches mainly focus on developing robust features against large changes across non-overlapping views or finding a distance metric learning algorithm which makes the same person's appearance feature vectors closer than different persons' vectors in the learned subspace.

For feature representation methods, Farenzena et al. [1] employed the weighted color histograms, Maximally Stable Color Regions (MSRC), Recurrent High-Structured Patches (RHSP) to capture image properties. Yang et al. [5] combined salient color names with color histograms to represent color distribution in appearance feature extraction. Local Maximal Occurrence(LOMO) [9] analyzed the horizontal occurrence of local features and maximized the occurence to make a stable representation against viewpoint changes.

However, it is extremely difficult to design an approach to extract stable features when undergoing large environmental changes. To solve this issue, an approriate distance metric or similarity function needs to be learned, making the distance between features of the same pedestrian smaller than different ones. Zheng et al. [6] proposed Probabilistic Relative Distance Comparison (PRDC) metric learning algorithm to maximize the probability of a pair of true match having a smaller distance than that of a wrong match pair. KISSME [4] learned a distance metric from equivalence constraints based on a statistical inference perspective without iteration. Pedagaid et al. [7] employed Local Fisher Discriminant Analysis to maximize the 'between-class' scatter while minimizing the 'within-class' scatter. Cross-view Quadratic Discriminant Analysis(XQDA) was proposed in [9] to learn a discriminant low dimensional subspace. Normally, metric learning algorithms project the original feature vector x_i into the same learned subspace by matrix L. Then the learned feature vector is $y_i = L^t x_i$, where superscript t represents the transpose of a matrix. Two samples y_i, y_j are evaluated the similarity by using Mahalanobis distance metric: $d(y_i, y_j) = \sqrt{(x_i - x_j)^t M (x_i - x_j)} = \|L^t(x_i - x_j)\|_2 = \|L^t x_i - L^t x_j\|_2$, where $M = L^t L$. This symmetric projec-

tion is applied to all camera views, ignoring the discrepancy between views.

In this work, we propose an asymmetric cross-view matching algorithm with dictionary learning to address the drawback of the symmteric models, i.e. we use a mapping matrix T to compensate the discrepancy between views and project dictionary codes into the same subspace to calculate the cosine similarity function. Besides, we also balance the impact of negative samples on the learned model and take the 'between-class' and the 'within-class' distance into account. Extensive experiments on the VIPeR and CUHK01 datasets show that our approach achieves state-of-the-art performance.

The main contribution of this work are summarized as follows. (1) We use dictionary learning along with asymmetric mapping matrix to compensate the discrepancy between camera views. (2) The dictionaries and dictionary codes are constrained which makes the forming codes more robust and discriminative than the original features. (3) Feature extraction procedure is improved to better overcome viewpoint variation.

2. PROBLEM FORMULATION AND NOTATIONS

This work mainly focuses on dictionary learning algorithm for person re-identification. After extracting appearance features from pedestrians' images, feature vectors are combined into feature matrices. Given probe set and gallery set of pedestrians' feature matrix Y_1 and Y_2, our goal is that for every coloum feature vector $y_{1i} \in Y_1$, $i = 1, 2, \cdots, n$, the algorithm needs to find the same pedestrian $y_{2j} \in Y_2$, $j = 1, 2, \cdots, m$, where n and m are the number of pedestrians in probe set and gallery set respectively.

The proposed algorithm has 5 categories of matrix: feature matrix, dicitonary matrix, code matrix, projective matrix and mapping matrix. X_1 and X_2 are the feature matrices for view1 and view2 when training. B_1, B_2 are the dictionaries of view1, view2 respectively. A_1, A_2 are the codes under dictionaries B_1, B_2. T is the asymmetric mapping matrix to map A_2 into the subspace of A_1. P_1, P_2 are the reconstruction projective matrices. Using the aforementioned notations, we discuss the details of the proposed algorithm in Section 3.

3. ASYMMETRIC CROSS-VIEW DICTIONARY LEARNING

Figure 1 shows the summary of our re-identification procedure. We extract features from patches firstly (the details of features extraction are described in Section 5), and then use feature matrices to train the dicitonary algorithm. Finally, features are projected into the same subspace and then the probe images are compared with gallery images, forming similarity ranking list. According to the ranking list, we re-identify the probe pedestrian from gallery set.

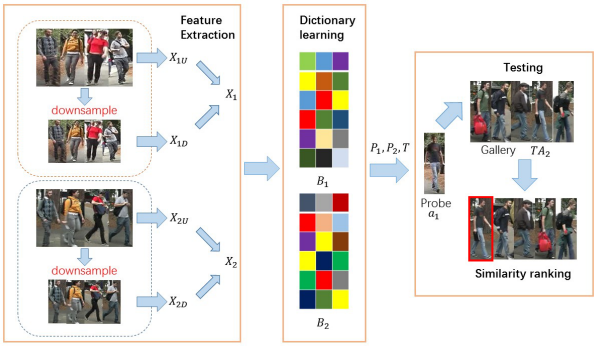

Fig. 1. Summary of Asymmetric Cross-View Dictionary Learning. The first two solid boxes are the training process and the third one is the testing process. X_{iU} and X_{iD}, $i = 1, 2$ indicate the feature matrix extracted on the original images and the downsampled images in view1 or view2. The red border image and probe image are the same pedestrian.

The main component of our algorithm is dictionary learning. Formally, the objective function is:

$$\begin{aligned} \min_{A_1,A_2,B_1,B_2,P_1,P_2,T} \quad & D(A_1, B_1, A_2, B_2) \\ & + E(A_1, P_1, A_2, P_2) \\ & + S(A_1, A_2, T) + R(B_1, B_2), \end{aligned} \tag{1}$$

which contains four parts.

$$D(A_1, B_1, A_2, B_2) = \frac{1}{2}(\|X_1 - B_1A_1\|_F^2 + \|X_2 - B_2A_2\|_F^2) \tag{2}$$

minimizes the dictionary coding reconstruction error. This is the traditional component in the dictionary learning algorithm. Specifically, we learn one dictionary for each view to ensure the dictionary can capture more details of different views.

$$E(A_1, P_1, A_2, P_2) = \frac{\alpha}{2}(\|P_1X_1 - A_1\|_F^2 + \|P_2X_2 - A_2\|_F^2) \tag{3}$$

minimizes the projection error. This projective dictionary learning algorithm is inspired by the projective dictionary learning [13] and Cross-view Projective Dictionary Learning (CPDL) [14]. Using this dictionary learning algorithm can save large computation and optimize very easily. Besides, we can use these projective matrices to get the probe set's and the gallery set's dictionary codes easily when testing performance.

$$S(A_1, A_2, T) = \varphi \sum_{i,j=1}^{n} w_{ij}(a_{1i} - Ta_{2j})^2 \tag{4}$$

considers the 'between-class' and 'within-class' distance in two views. $w_{ij} = \begin{cases} 1 & id_i = id_j \\ -\frac{0.01}{n_{neg}} & id_i \neq id_j \end{cases}$ is the coefficient

between a_{1i} and Ta_{2j} when training. This item has three roles in this approach. Firstly, the value 0.01 is to reduce the impact of negative samples. Secondely, this item can reduce the intra-class variance and enlarge the inter-class variance when optimizing the objective function. Thirdly, an asymmetric mapping matrix is introduced into this algorithm to compensate the discrepancy between views.

$$R(B_1, B_2) = \frac{\theta}{2} \|B_1 - B_2\|_F^2 + \lambda(\|B_1\|_F^2 + \|B_2\|_F^2) \quad (5)$$

regulates the dictionaries and constrains them to be similar to alleviate the misalignments among views.

$\alpha, \varphi, \theta, \lambda$ is the hyperparamters in the algorithm. Combining all these four items, we can get the asymmetric cross-view dictionary learning for person re-identification. The forming codes are more compact and discriminant than the original feature vectors.

CPDL [14] is also a cross-view dictionary learning algorithm for person re-identification. The main difference between our work and CPDL [14] is that we add a constraint on the dictionaries' codes which considers the 'between-class' and 'within-class' distance. We map the view2's codes into view1's subspace by matrix T to compensate the discrepancy between two views. Secondly, CPDL [14] has two stages of dictionary learning combining image level and patch level. Such procedure makes the model more complicated while the performance doesn't improve dramatically. Thirdly, feature extraction strategies are different. Although the proposed approach uses more color and texture features, the forming feature vector's dimension is lower than CPDL [14] due to a larger patch area and fewer downsampling operations.

4. OPTIMIZATION

We optimize the equation over $A_1, A_2, P_1, P_2, T, B_1, B_2$ one at a time while fixing other matrices. This one variable function is a convex optimizaiton problem and has closed solution. We can simplify S into $S(A_1, A_2, T) = \varphi tr(A_1 L(TA_2)^t)$, where $L_{ij} = \begin{cases} \sum_j w_{ij} & id_i = id_j \\ w_{ij} & id_i \neq id_j \end{cases}$ is the Laplacian matrix, $tr(\cdot)$ indicates the trace of the matrix. Optimization process is described next.

Update steps for A_1, A_2. We fix other variables and optimize over A_1. The objective function can be simplified into:

$$J(A_1) = \frac{1}{2} \|X_1 - B_1 A_1\|_{\mathrm{F}}^2 + \frac{\alpha}{2} \|P_1 X_1 - A_1\|_{\mathrm{F}}^2 + \varphi \mathrm{tr}(A_1 L (TA_2)^t). \quad (6)$$

Setting $\frac{\delta J(A_1)}{\delta A_1} = 0$, we get the solution

$$A_1 = (B_1^t B_1 + \alpha I)^{-1}(B_1^t X_1 + \alpha P_1 X_1 - \varphi T A_2 L), \quad (7)$$

where I is the identity matrix. The optimization for A_2 is similar and the solution is

$$A_2 = (B_2^t B_2 + \alpha I)^{-1}(B_2^t X_2 + \alpha P_2 X_2 - \varphi T^t A_1 L). \quad (8)$$

Update steps for P_1, P_2. Similar to the optimization process for A_1,A_2, we can get the simplified objective function $P_1^* = \arg\min_{P_1} \|P_1 X_1 - A_1\|_{\mathrm{F}}^2$. The solution is

$$P_1 = A_1 {X_1}^t (X_1^t X_1 + \gamma I)^{-1}, \quad (9)$$

where γ is a regularization parameter. Accordingly, the solution for P_2 is

$$P_2 = A_2 {X_2}^t (X_2^t X_2 + \gamma I)^{-1}. \quad (10)$$

Update steps for T.

$$T^* = \arg\min_T tr(A_1 L (T A_2)^t) = A_1 L A_2^t. \quad (11)$$

Update steps for B_1,B_2. The objective function is

$$J(B_1) = \frac{1}{2} \|X_1 - B_1 A_1\|_{\mathrm{F}}^2 + \frac{\theta}{2} \|B_1 - B_2\|_{\mathrm{F}}^2 + \lambda \|B_1\|_{\mathrm{F}}^2. \quad (12)$$

Setting $\frac{\delta J(B_1)}{\delta B_1} = 0$, the solution is

$$B_1 = (X_1 A_1^t + \theta B_2)(A_1 A_1^t + (\theta + 2\lambda) I)^{-1}. \quad (13)$$

Similarly,

$$B_2 = (X_2 A_2^t + \theta B_1)(A_2 A_2^t + (\theta + 2\lambda) I)^{-1}. \quad (14)$$

Repeating the optimization procedure, the algorithm can converge very quickly. After training, we can get the probe codes $C^1 = P_1 Y_1$ and mapping gallery codes $C^2 = T P_2 Y_2$ and then compute the cosine similarity between column vector c_i^1 and c_j^2, forming similarity ranking list and re-identify the probe pedestrian.

5. EXPERIMENTAL RESULTS

We evaluate our approach on the public person re-identification VIPeR [15] and CUHK01 [16] datasets.

5.1. Experimental Settings

Feature extraction. We extract appearance features on local patches with dense grids. The patch size is 16×16 and the grid step is 8 pixels. 16-bin histogram in each color channel of RGB, HSV, LAB, YCbCr color space is extracted and dense SIFT features are extracted in OPPONENT color space [17]. Besides, HOG texture features are also extracted.

Then we divide the images simply into right and left halvies. For every half of the image, we use maxpool to

Table 1. Comparison of state-of-the-art results on the VIPeR dataset. The cumulative matching scores (%) at rank 1,5,10, and 20 are listed

Methods	rank=1	rank=5	rank=10	rank=20
KISSME [4]	19.60	-	62.20	77.00
SDALF [1]	19.87	38.89	49.37	65.73
SalMatch [18]	30.16	-	-	-
LADF [8]	30.22	64.70	78.92	90.44
kLFDA [11]	32.2	65.8	79.7	90.9
CPDL [14]	33.99	64.21	77.53	88.58
IDLA [19]	34.81	-	75.63	84.49
PolyMap [20]	36.80		83.70	91.70
SCNCD [5]	37.80	68.50	81.20	90.40
XQDA(LOMO) [9]	40.00	68.13	80.51	91.08
MLAPG [12]	40.73	69.97	82.34	92.37
SSSVM [21]	42.66	-	84.27	91.93
Ours(no$T$)	38.04	67.69	79.18	89.15
Ours	**42.94**	**73.58**	**84.63**	**93.70**

maximize the local occurence of pattern in the same horizontal position to overcome significant viewpoint variation. This feature extraction strategy is different from LOMO's [9]. LOMO [9] didn't divide image. The reason for this division is that we find it is more robust under complicated environment in which the right half's pattern and left half's may not be symmetric. To extract more general features, we also downsample the original 128×48 (160×60) images to 64×24 (80×30) by 2×2 average pooling for VIPeR(CUHK01) dataset, then repeat the feature extraction procedure.

Hyperparameters. In this algorithm there are four hyperparameters, including $\alpha, \varphi, \theta, \lambda$. They are all set to 1.This algorithm is not sensitive to these hyperparameters. Components of the dicitonary learning algorithm play an equally important role in this approach.

Evaluation Metric. We report the rank-k matching rates as our evaluation metric. A rank-k matching rate indicates the percentage of the probe image with correct matches found in the top k rank against the gallery set.

5.2. Results on VIPeR Dataset

VIPeR[15] is a challenging person re-identification dataset. It contains of 632 pedestrians pairs in two different outdoor views. The images captured in two cameras undergo significant variations in illumination, pose ,viewpoint. 632 pairs of images are randomly divided into half, one half for training and the other half for testing. Images from view1 are used as probe and other images from view2 as gallery. We evaluate the performance of the algorithm by repeating training and testing procedure 10 times and getting an average cumulative matching scores. The experimental results show the effectiveness of our algorithm.

From **Table 1**, We can see that our method gets the best performance compared with other methods. The rank-1 accuracy is higher than CPDL[14] by 8.95%, indicating that considering the 'between-class' and 'within-class' distance is necessary and using a more straightforward way training the dictionary can perform better.

Table 2. Comparison of state-of-the-art results on the CUHK01 dataset. The cumulative matching scores (%) at rank 1,5,10, and 20 are listed

Methods	rank=1	rank=5	rank=10	rank=20
CPDL [14]	59.47	81.26	89.72	93.10
XQDA(LOMO) [9]	63.24	-	90.04	94.16
MLAPG [12]	**64.24**	-	**90.84**	**94.92**
Ours	63.96	**84.77**	90.12	94.24

5.3. Results on CUHK01 Dataset

CUHK01 [16] dataset consists of 3884 pedestrian images captured by 2 different views from 971 persons. Each person has two images in each view. Images from view1 are used as probe and other images from view2 as gallery. 485 image pairs are for training while the remaining image pairs are for testing.

The proposed method is compared to recent methods. Experimental results are shown in **Table 2**. The rank-1 accuracy of our algorithm is 0.28% inferior to MLAPG [12]. But the proposed method's training time is much shorter than MLAPG [12] due to the closed solution in each iteration.

5.4. Disscussion

We evaluate T in the dictionary learning. Asymmetric mapping matrix T is eliminated, so the third part of objective function becomes $S(A_1, A_2) = \varphi tr(A_1 L A_2^t)$. The same procedure is repeated on the VIPeR dataset. Experimental result (ours(no$T$)) is shown in the Table 1. The performance decreases about 4% which means the mapping matrix can compensate the dicrepancy between two views and asymmetric model can handle more complicated environment than symmetric model.

6. CONCLUSIONS

In this paper, we propose an asymmetric cross-view dictionary learning algorithm. This algorithm uses asymmetric mapping matrix to compensate the discrepancy between views. The constraints on dictionaries and codes are necessary for person re-identification. The effectiveness of our algorithm is evaluated on the public VIPeR and CUHK01 datasets. Our method demonstrates state-of-the-art performance on two datasets which means the asymmetric model can handle more complicated environment than the symmetric models.

7. REFERENCES

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

[2] Douglas Gray and Hai Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *European Conference on Computer Vision*, 2008.

[3] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang, "Towards open-world person re-identification by one-shot group-based verification," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. DOI: 10.1109/TPAMI.2015.2453984, 2015.

[4] Martin Kostinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof, "Large scale metric learning from equivalence constraints," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2288–2295.

[5] Yang Yang, Jimei Yang, Junjie Yan, Shengcai Liao, Dong Yi, and Stan Z. Li, "Salient color names for person re-identification," in *European Conference on Computer Vision*, 2014, pp. 536–551.

[6] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang, "Person re-identification by probabilistic relative distance comparison," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 649–656.

[7] Sateesh Pedagadi, James Orwell, Sergio Velastin, and Boghos Boghossian, "Local fisher discriminant analysis for pedestrian re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3318–3325.

[8] Wei Li and Xiaogang Wang, "Locally aligned feature transforms across views," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3594–3601.

[9] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2197–2206.

[10] Wei-Shi Zheng, Xiang Li, Tao Xiang, Shengcai Liao, Jianhuang Lai, and Shaogang Gong, "Partial person re-identification," in *IEEE Conference on Computer Vision*, 2015.

[11] Fei Xiong, Mengran Gou, Octavia Camps, and Mario Sznaier, "Person re-identification using kernel-based metric learning methods," in *European Conference on Computer Vision*, 2014.

[12] Shengcai Liao and Stan Z. Li, "Efficient psd constrained asymmetric metric learning for person re-identification," in *IEEE Conference on Computer Vision*, 2015.

[13] Shuhang Gu, Lei Zhang, Wangmeng Zuo, and Xiangchu Feng, "Projective dictionary pair learning for pattern classification," in *Advances in Neural Information Processing Systems 27*, 2014, pp. 793–801.

[14] Sheng Li, Ming Shao, and Yun Fu, "Cross-view projective dictionary learning for person re-identification," in *Proceedings of International Joint Conference on Artificial Intelligence*, 2015, pp. 2155–2161.

[15] Douglas Gray, Shane Brennan, and Hai Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, 2007, vol. 3, pp. 2155–2161.

[16] Rui Zhao, Wanli Ouyang, and Xiaogang Wang, "Learning mid-level filters for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2528–2535.

[17] Koen E.A. van de Sande, Theo Gevers, and Cees G.M.Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1582–1596, 2010.

[18] Rui Zhao, Wanli Ouyang, and Xiaogang Wang, "Person re-identification by salience matching," in *IEEE Conference on Computer Vision*, 2014, pp. 144–151.

[19] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3908–3916.

[20] Dapeng Chen, Zejian Yuan, Gang Hua, Nanning Zheng, and Jingdong Wang, "Similarity learning on an explicit polynomial kernel feature map for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1565–1573.

[21] Ying Zhang, Baohua Li, Huchuan Lu, Atshushi Irie, and Xiang Ruan, "Sample-specific svm learning for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.