PERSON RE-IDENTIFICATION USING SPARSE REPRESENTATION WITH MANIFOLD CONSTRAINTS

Behzad Mirmahboub[1], Hamed Kiani[1], Amran Bhuiyan[1], Alessandro Perina[1], Baochang Zhang[2], Alessio Del Bue[1], Vittorio Murino[1,3]

[1]Pattern Analysis and Computer Vision (PAVIS), Italian Institute of Technology, Genova, Italy
[2]School of Automation Science and Electrical Engineering, Beihang University, Beijing, China
[3]Department of Informatics, University of Verona, Italy

ABSTRACT

Human re-identification is still a challenging task due to the human pose and illumination variations. Nowadays, surveillance cameras with high frame rate are capable of capturing several consecutive frames from each person. Multi-shot images provide richer information of the target person compared to a single-shot image. They, however, produce a high cost of information redundancy which may degrade the performance of re-identification systems. In this paper, we propose a novel framework that combines sparse coding and manifold constraints to extract discriminative information from multi-shot images of one pedestrian for person re-identification across a set of non-overlapped surveillance cameras. The evaluation over two standard multi-shot datasets shows very competitive accuracy of our framework against the state-of-the-art.

Index Terms— Person Re-identification, Sparse Representation, Manifold Constraint

1. INTRODUCTION

Person re-identification is the problem of recognizing an individual in images captured by a camera (or cameras), given his/her images provided by other non-overlapping camera(s) [1]. This is a challenging task because of typical low resolution images, occlusions, varying poses, clutter background and uncontrolled varying illumination. In general, person re-identification is performed adopting two different scenarios, i.e., single-shot and multi-shot. Unlike single-shot re-identification scenarios [2, 3], in which only one image of each person is provided by each camera, multiple-shot scenarios employ multiple images of the same person [4, 5].

The assumption of multi-shot approaches is that multiple images of a same person provide richer information for extracting more discriminative and robust features/descriptors [6, 7, 8, 9, 10]. Multiple images of the same person taken over a very short period of time, however, are visually very similar and, as a result, share a huge amount of redundant information. Such redundancy, apart from inefficient complexity and memory usage, may degrade the discrimination and generalization of the re-identification system. For instance, it is known that incoherent dictionaries improve the performance of sparse representation [11]. If the images are all similar this assumption is no more valid.

Person re-identification methods are generally divided into two categories: direct and learning-based methods [4]. Direct methods extract descriptors such as color histogram and HOG from both probe and gallery images and examine the similarity between descriptors using statistical measures such as Bhattacharyya distance, l_2-norm or chi-square for image matching [6, 7, 8, 12]. Since direct matching is performed between a single probe image and a single gallery image, it is not capable of handling multi-shot images.

On the other hand, in learning-based method a set of training images for each individual is used to train a person-specific classifier. The idea is that the information in training images can be generalized to unseen test samples. Dictionary learning and sparse coding were successfully used for this purpose [9, 10]. Despite recent success of learning-based techniques, they still suffer from the lack of adequate training samples (for each person) and information redundancy of visually very similar training images.

In this paper, we propose a human re-identification framework based on sparse representation along with the manifold learning paradigm. The main objective is to take advantage from useful extra information that comes from the multi-shot scenario by learning a compact representation from the multiple images. To do so, we rely on manifold learning to perform dimensionality reduction [13]. We use this idea to summarize several frames of each person into a descriptor that is similar to all those frames and carries additional different information from all of them minimizing redundancies.

Given a test image, we first generate a set of augmented images by applying small translations and brightness variations [14]. Then, we extract image descriptor from the original and the augmented images and find their sparse representations. The sparse codes that are related to each person from each camera are used to build a manifold point. This point

is an approximation of the person in low-dimensional feature space that is used for human re-identification.

2. SPARSE REPRESENTATION CLASSIFICATION

Sparse representation emerged as a powerful tool for signal and image processing [15]. In this method every n-dimensional vector b is represented as:

$$b = Ax \tag{1}$$

where A is a full-rank $n \times m$ matrix called dictionary. The columns of A are representative training samples called basis vectors or atoms. The idea of sparsity is that vector b is given by a linear combination of only few atoms. Therefore, m-dimensional vector x is sparse and only few of its elements are non-zero. The number of training samples is assumed to be high enough to span the entire sample space. This means that the dictionary A is over-complete ($m > n$) with the number of basis vectors greater than the dimensionality of input vector b. Given the dictionary A and input vector b, the optimization problem estimating a sparse vector x is called Lasso regression, is represented as:

$$x^* = arg\min_x \|b - Ax\|_2^2 + \lambda \|x\|_1 \tag{2}$$

where the l_1-norm forces the sparsity on x, l_2-norm minimizes the reconstruction error and λ controls the tradeoff between these two terms.

Sparse representation is naturally discriminative. It selects only basis vectors that most compactly represent a signal and therefore is useful for classification. However, if data is strongly redundant, sparsity assumption is weaker because more atoms can describe a single test sample. Decision rule to classify a new test vector b is obtained based on reconstruction error of b. Eq. 1 can be rewritten as $b = \sum_i (A_i x_i)$ where A_i is a part of dictionary A whose columns are only samples from class i and x_i is a part of vector x whose elements are related only to class i. So, the normalized reconstruction error of b for class i is computed as:

$$e_i = \frac{\|b - A_i x_i\|_2}{\|b\|_2}, \quad i \in \{1, ..., C\} \tag{3}$$

where C is the total number of classes. The class (label) i with minimum error is assigned to the input sample [10].

3. MANIFOLD CONSTRAINT

A manifold is a topological space that locally resembles Euclidean space near each point [13]. The main technique to learn the manifold is Locally Linear Embedding which states that each data point is expected to be a linear combination of its neighboring points. If we assume that points/samples are located on a manifold, the optimization problem in Eq. 2 is

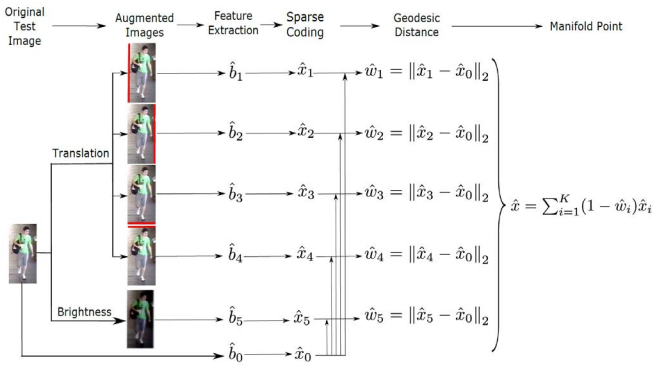

Fig. 1. Schematic diagram for estimating the manifold point

extended by an additional constraint imposing that input sample b should be on a manifold. This is formulated as:

$$x^* = arg\min_{x,\hat{b}} \|b - Ax\|_2^2 + \lambda \|x\|_1 + \sigma \left\|b - \hat{b}\right\|_1 \tag{4}$$

where $\hat{b}$ is the point on manifold and σ assigns an importance factor to the manifold constraint. This objective function is not easy to be optimized due to the additional manifold constraint. Zhang *et al.* [16] proved that the constraint on input sample can be transferred to a constraint on sparse representation of the sample. Therefore the objective function in Eq. 4 can be written as:

$$x^* = arg\min_x \|b - Ax\|_2^2 + \lambda \|x\|_1 + \sigma' \|x - \hat{x}\|_1 \tag{5}$$

where A is a matrix whose columns are training samples, b is a test sample, x is the sparse representation of b and $\hat{x}$ is the sparse representation of a point on the manifold.

In order to estimate the manifold point $\hat{x}$, we generate some additional images by applying spatial translation and brightness transfer function [14] on the original image. Then, we extract feature vectors from these augmented images, $\hat{b}_i$, and calculate their sparse representations $\hat{x}_i$ using Eq. 2, with $i = 1, ..., K$, where K is the number of augmented images. The sparse representation of the original image can also be denoted as $\hat{x}_0$. The geodesic distance between the original image and the i^{th} augmented image is defined as $\hat{w}_i = \|\hat{x}_i - \hat{x}_0\|_2$ which is used to estimate the manifold point as $\hat{x} = \sum_{i=1}^{K}(1 - \hat{w}_i)\hat{x}_i$. This means that similar images, in terms of their sparse representations, contribute more to the generation of the manifold point. This process is depicted in Fig. 1. Parameters λ and σ' in Eq. 5 are empirically set to 0.1 and 0.2, respectively, for all experiments.

4. EXPERIMENTAL RESULTS

In this section, we evaluate our method for the task of multi-shot person re-identification comparing with state-of-the-art

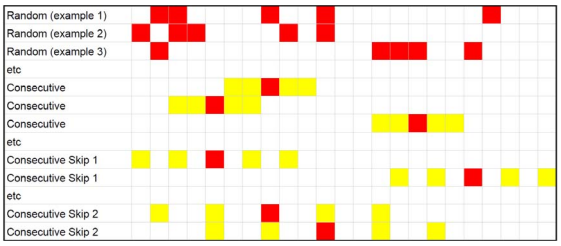

Fig. 2. Multi-shot frames selection. Each square is a frame. Dark squares are random selection and light squares are consecutive selections after one random selection.

methods. We use the feature extraction proposed by [10] including color histograms (i.e., RGB and HS) and Histogram of Oriented Gradients (HOG) for image representation. Also we use Cumulative Weighted Brightness Transfer Function (CWBTF) as proposed in [14] to generate augmented images to form manifold points.

4.1. Datasets

We evaluate our method on two multi-shot datasets: SAIVT-SOFTBIO [17] and PERSON RE-ID 2011 [18].

SAIVT-SOFTBIO contains images of 152 subjects. Each person is captured with eight different static cameras in indoor environment under varying pose, illumination and background. The number of images for each person in each camera varies from 0 to 240 consecutive frames. Training and testing are done using the images from two different cameras.

PERSON RE-ID 2011 consists of pedestrians images from two different static surveillance cameras in outdoor environment with extreme view-point and illumination changes. Camera A captures 385 persons with clear background. Camera B captures 749 persons walking over a zebra crossing in a street. The first 200 persons appear in both cameras. Images for both single-shot and multi-shot scenarios are provided. We use the multi-shot scenario, where the number of images for each person changes from 5 to 675 consecutive frames.

4.2. Manifold Generation

We use multiple shots of each person and their augmented images (generated by translations and WCBTF) to form manifold points. We apply our proposed method on 5 Vs. 5 person re-identification and we use four scenarios for selecting frames from train set (for dictionary learning) and test set (to form manifold points) as depicted in Fig. 2.

Each manifold point is built using all 5 frames of one person and their augmented images. In the random scenario, 5 frames are randomly selected to capture diversity in the dataset. We also try three more frame selection strategies based on consecutive shots: one random frame and its two consecutive previous and next frames (Consecutive in Fig 2), one random frame and its two consecutive previous and next frames with one skipped frame in between (Consecutive Skip 1 in Fig 2), one random frame and its two consecutive previous and next frames with two skipped frames in between (Consecutive Skip 2 in Fig 2). Since the number of images for each person in dataset is different, we choose only those persons with enough images for our scenarios.

Given five shots of each person, we investigate three different image transformation strategies to generate augmented images from the original test image. These augmented images are used to construct the manifold point as represented in Fig. 1. These strategies are:

1) WCBTF that generates one augmented image from each test image. Since we select 5 test shots for each person, each manifold point is build based on 5 images with augmented brightness.

2) Image translation by one pixel in four different directions (top, down, left and right) that generates four augmented images from each test image. Therefore, each manifold point is built using $5 \times 4 + 5 = 25$ images.

3) The combination of WCBTF and image translation that first generates five images using WCBTF of the 5 original shots, and then generates again four additional images by translating each WCBTF augmented image. As a result, there are 25 images in total to form the manifold point.

The common tool to evaluate human re-identification performance is the Cumulative Matching Characteristic (CMC) curve and its Area Under Curve (AUC). This curve represents correct matching rate as a function of top n ranks. Fig. 3 shows the result of the proposed approach on SAIVT-SOFTBIO dataset. CMC curves are plotted for three different types of manifolds that we mentioned above. In each diagram, we plot CMC curves for four frame selection scenarios. As we can see, random frame selection strategy outperforms the consecutive strategies. Moreover, we observe that the combination of image translation (Fig. 3 (a)) and brightness transfer function (Fig. 3 (b)) improves the matching rate and obtains higher AUC (Fig. 3 (c)).

4.3. Comparison with state-of-the-art

According to the previous experiment, we choose to generate manifold point using the combination of brightness transfer function and image translation. We run the proposed method with random frame selection for ten times and compare the average results with state-of-the-art methods. On PRID dataset we compare with 8 methods including Custom Pictorial Structures (CPS) [8], Symmetry-Driven Accumulation of Local Features (SDALF) [7], Weighted Brightness Transfer Function (WBTF) [12], Kernel Canonical Correlation Analysis (KCCA) [22], kernel Local Fisher Discriminant Analysis (kLFDA) [23], Learning Midlevel Filters (LMF) [24], regularized Pairwise Constrained Component Analysis (rPCCA) [23] and RankSVM (RSMV) [3]. The CMC curves and AUCs respect to different ranks are depicted in Fig. 4. As

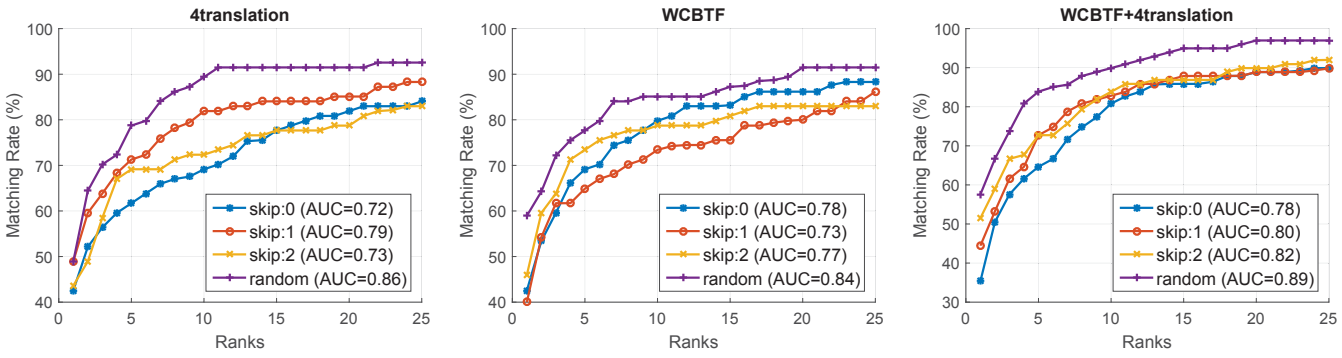

Fig. 3. CMC curves for different scenarios of frame selection based on manifold generation using (a) translation (b) WCBTF (c) WCBTF with translation on SAIVT-SOFTBIO (similar views) dataset.

Table 1. Comparison on SAIVT-SoftBio dataset

Dataset	Camera 3/8				Camera 5/8			
Ranks	Rank 1	Rank 5	Rank 10	Rank 20	Rank 1	Rank 5	Rank 10	Rank 20
Fused [17]	36.4	60.3	76.0	87.6	20.0	33.0	50.4	67.8
PFDS [19]	33.2	60.5	74.0	87.2	18.6	32.9	**53.0**	**85.3**
RankSVM [20]	32.4	68.4	82.0	92.9	14.9	40.5	57.9	75.0
LFDA [21]	12.2	36.8	54.6	74.9	9.3	27.1	41.2	60.6
Ours	**56.9**	**83.0**	**87.6**	**94.7**	**22.7**	**41.6**	48.6	66.9

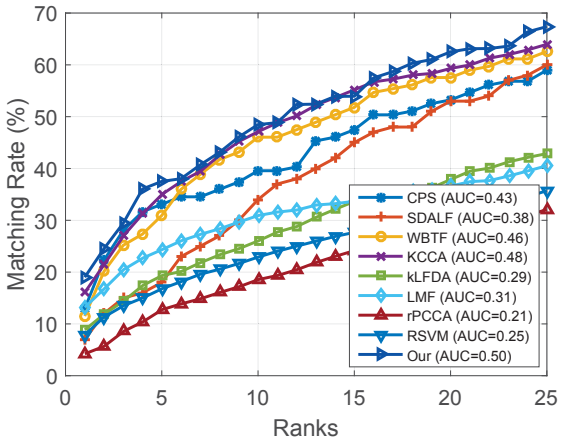

Fig. 4. CMC curves of our method based on translation and WCBTF manifold in comparison with state-of-the-art methods on Person Re-Id dataset

we see our proposed method outperforms all other methods. Only in rank 15, KCCA method shows higher matching rate than us. But the AUC of our method is higher.

On SAIVT dataset we compare with 4 methods including Fused [17], Pairwise Feature Dissimilarities Space (PFDS) [19], RankSVM [20] and Local Fisher Discriminant Analysis (LFDA) [21]. The images in SAIVT dataset are captured with eight different cameras. Following [17], we select two cameras for training (cameras 3 and 5) and one camera for testing (camera 8). Camera 3 and 8 capture images from similar views, while the views of camera 5 and 8 are dissimilar and, as a result, images taken by cameras 5 and 8 are affected by severe posing, illumination and appearance changes [17]. This leads to lower matching rate for camera 5 and 8. The results of person re-identification for different ranks are shown in Table 1. In case of camera 3/8 our proposed method performs better than all other methods. Between camera 5 and 8 although our method does not show good performance for rank 10 and 20, but it outperforms other methods in rank 1 and 5 that are more important for person re-identification.

5. CONCLUSION

We introduced a novel method for human re-identification using sparse representation with manifold constraints. Our method is categorized as a multi-shot re-identification approach and aims to preserve the discriminative characteristics of multiple images of a person by applying a manifold constraint over augmented sparse representations. We empirically found out that random frames selection and augmenting them using spatial translations and brightness transfer function gives the best performance to generate manifold point. The evaluation over SAIVT and PRID datasets shows the superiority of our approach. Future work can be using tracking method to improve the re-identification performance [25].

6. REFERENCES

[1] Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen Change Loy, *Person re-identification*, vol. 1, Springer, 2014.

[2] Zhe Lin and Larry S Davis, "Learning pairwise dissimilarity profiles for appearance recognition in visual surveillance," in *Advances in Visual Computing*, pp. 23–34. Springer, 2008.

[3] Bryan Prosser, Wei-Shi Zheng, Shaogang Gong, Tao Xiang, and Q Mary, "Person re-identification by support vector ranking.," in *BMVC*, 2010, vol. 2, p. 6.

[4] Loris Bazzani, Marco Cristani, and Vittorio Murino, "Symmetry-driven accumulation of local features for human characterization and re-identification," *Computer Vision and Image Understanding*, vol. 117, no. 2, pp. 130–144, 2013.

[5] Slawomir Bak, Etienne Corvee, François Bremond, and Monique Thonnat, "Multiple-shot human re-identification by mean riemannian covariance grid," in *AVSS*. IEEE, 2011, pp. 179–184.

[6] Niloofar Gheissari, Thomas B Sebastian, and Richard Hartley, "Person reidentification using spatiotemporal appearance," in *CVPR*. IEEE, 2006, vol. 2, pp. 1528–1535.

[7] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *CVPR*. IEEE, 2010, pp. 2360–2367.

[8] Dong Seon Cheng and Marco Cristani, "Person re-identification by articulated appearance matching," in *Person Re-Identification*, pp. 139–160. Springer, 2014.

[9] Xiao Liu, Mingli Song, Dacheng Tao, Xingchen Zhou, Chun Chen, and Jiajun Bu, "Semi-supervised coupled dictionary learning for person re-identification," in *CVPR*. IEEE, 2014, pp. 3550–3557.

[10] Giuseppe Lisanti, Iacopo Masi, Andrew Bagdanov, and Alberto Del Bimbo, "Person re-identification by iterative re-weighted sparse ranking," *PAMI*, 2014.

[11] Tong Lin, Shi Liu, and Hongbin Zha, "Incoherent dictionary learning for sparse representation," in *ICPR*. IEEE, 2012, pp. 1237–1240.

[12] Amitava Datta, Lisa M Brown, Rogerio Feris, and Sharath Pankanti, "Appearance modeling for person re-identification using weighted brightness transfer functions," in *ICPR*. IEEE, 2012, pp. 2367–2370.

[13] Tong Lin and Hongbin Zha, "Riemannian manifold learning," *PAMI*, vol. 30, no. 5, pp. 796–809, 2008.

[14] Amran Bhuiyan, Alessandro Perina, and Vittorio Murino, "Exploiting multiple detections to learn robust brightness transfer functions in re-identification system," in *ICIP*. 2015.

[15] Michael Elad, *Sparse and Redundant Representations, From Theory to Applications in signal and Image Processing*, Springer, 2010.

[16] Baochang Zhang, Alessandro Perina, Vittorio Murino, and Alessio Del Bue, "Sparse representation classification with manifold constraints transfer," *CVPR*, 2015.

[17] Alina Bialkowski, Simon Denman, Sridha Sridharan, Clinton Fookes, and Patrick Lucey, "A database for person re-identification in multi-camera surveillance networks," in *DICTA*. IEEE, 2012, pp. 1–8.

[18] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof, "Person re-identification by descriptive and discriminative classification," in *Image Analysis*, pp. 91–102. Springer, 2011.

[19] Jorge Garcia, Niki Martinel, Gian Luca Foresti, Alfredo Gardel, and Christian Micheloni, "Person orientation and feature distances boost re-identification," in *ICPR*. IEEE, 2014, pp. 4618–4623.

[20] Thorsten Joachims, "Optimizing search engines using clickthrough data," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002, pp. 133–142.

[21] Masashi Sugiyama, "Local fisher discriminant analysis for supervised dimensionality reduction," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 905–912.

[22] Giuseppe Lisanti, Iacopo Masi, and Alberto Del Bimbo, "Matching people across camera views using kernel canonical correlation analysis," in *Proceedings of the International Conference on Distributed Smart Cameras*. ACM, 2014, p. 10.

[23] Fei Xiong, Mengran Gou, Octavia Camps, and Mario Sznaier, "Person re-identification using kernel-based metric learning methods," in *ECCV*, pp. 1–16. Springer, 2014.

[24] Rui Zhao, Wanli Ouyang, and Xiaogang Wang, "Learning mid-level filters for person re-identfiation," in *CVPR*, Columbus, USA, June 2014.

[25] Baochang Zhang, Alessandro Perina, Zhigang Li, Vittorio Murino, Jianzhuang Liu, and Rongrong Ji, "Bounding multiple gaussians uncertainty with application to object tracking," *IJCV*, pp. 1–16, 2016.