Riemannian Set-level Common-Near-Neighbor Analysis for Multiple-shot Person Re-identification

Wei Li
Graduate School of Informatics
Kyoto University, Kyoto 606-8501, Japan
liwei@mm.media.kyoto-u.ac.jp

Yang Wu, Yasutomo Kawanishi, Masayuki Mukunoki, Michihiko Minoh
Academic Center for Computing and Media Studies
Kyoto University, Kyoto 606-8501, Japan

Abstract

Multiple-shot person re-identification deals with the problem to build the correspondence between the images of the same person appearing at different sites captured by onsite deployed cameras. The difficulty stems from large within-class but small between-class variations caused by the change of person appearance and environment. Traditional methods on feature/signature design and/or distance/dissimilarity exploration have been largely investigated, resulting in a quick slowing down of the performance improvement. This paper proposes a novel solution called "Riemannian Set-level Common-Near-Neighbor Analysis" by absorbing the essence of two distinctive and effective state-of-the-art models. More concretely, it generates the discriminative covariance-based representation for each set of images following the Mean Riemannian Covariance Grid approach, while at the same time creatively realizes the set-level neighborhood information based ranking inheriting the key idea of sample-level Common-Near-Neighbor Analysis. Experiments have been conducted on widely-used benchmark datasets, showing significant performance improvement over state-of-the-art methods.

1 Introduction

Multiple-shot person re-identification is one of the most important but challenging issues in visual surveillance. Solutions to it have generally focused on either or both of two important aspects: feature/signature design and distance/dissimilarity exploration. For the first aspect, one typical method is "Histogram Plus Epitome" [5]. It focuses on the overall chromatic contents via histogram representation and recurrent local patches via epitomic analysis to extract the global and local features of the human appearance. Another representative method is "Haar-based and DCD-based Signature" [7]. It can build a satisfactorily invariant and discriminative signature from the existed haar-like features and dominant color descriptors relying on the Adaboost scheme. For the second aspect, one well-known method is "Minimum Point-wise Distance (MPD)" [3]. It measures the minimum distance between any pair of points from the two sets, which is simple and effective when the features are discriminative enough. Since MPD only depends on the nearest points between the sets, such a local point distribution based method is sensitive to outliers. Another exemplary method is "Set Based Discriminative Ranking" [9]. It iteratively constructs convex hulls for set-to-set distance measurements and optimizes the metric for ranking relying on these measurements. Although this method considers the distance between convex hulls of the set pair, which is called "Convex Hull based Image Set Distance (CHISD)" [1] and supposed to be more robust than MPD, yet it is more or less influenced by local parts determined by the layout of nearest points between the sets which support the convex hulls.

Recently, two methods have pursued novel solutions on these two aspects and have achieved remarkable results. One of them is "Mean Riemannian Covariance Grid (MRCG)" [8]. It exploits spatio-temporal information of person appearance by dense overlapping covariance grids with Karcher mean based signature generation in Riemannian space, and achieves encouraging results. The other method is "Common-Near-Neighbor Analysis (CNNA)" [4]. It utilizes neighborhood information of each sample pair to model a new dissimilarity, which has shown leading performance on single-shot person re-identification. Even though, directly applying the same idea to the multiple-shot problem is not a good choice. It will result in a huge increase of computational cost caused by the combinatorial explosion of sample pairs, while at the same time the performance is likely to be limited due to its ignorance of the correlation among images of the same set.

Inspired by MRCG and CNNA, this paper proposes a novel method called "Riemannian Set-level Common-Near-Neighbor Analysis (RSCNNA)". It generates a covariance grids based signature for each set of images and then builds a set-level neighborhood information based dissimilarity measurement for effective set-based ranking. All of these are done in Riemannian space rather than Euclidean space to cooperate with the covariance-based representation. There are two major points that distinguish RSCNNA from its analogues: (1) it extends traditional single sample based CNNA to the set level, to effectively and efficiently solve the multiple-shot person re-identification problem; (2) it enhances the power of covariance-based representation in Riemannian space by exploring the set-level neighborhood information. Following sections will detail the model of the proposed RSCNNA approach and the experimental justification of its superiority, with a short conclusion for the whole paper.

2 Riemannian Set-level Common-Near-Neighbor Analysis

2.1 Set-level Common-Near-Neighbor Analysis

CNNA has been proven to be effective for the issue of person re-identification [4]. Since samples within the same class will share more common near neighbors than those from different classes, CNNA explores this kind of neighborhood information to model a novel dissimilarity, which can further make intra-class dissimilarities smaller than inter-class dissimilarities for all samples.

However, this method operates on the sample level and is designed for the target of single-shot person re-identification, which is much different from the multiple-shot case [3]. Though it's possible to directly apply it to multiple-shot problems, it is undesirable to do so. If we transform the multiple-shot problem into a single-shot problem to solve, the efficiency will be low. Because the dissimilarity between each pair of samples is required to measure in this case, when the sample number increases in each set, the computation will be combinatorial explosion. Furthermore, CNNA explores the common-near-neighbor information for every sample pair, so it cannot model the correlations among the multiple images within the same set which is important for robustness to within-set variations and representation ability to cover as much information as possible from individual images. and the outliers which may be noise will also be emphasized during the operation on the sample level by CNNA. Thus, the effectiveness will be low as well.

Inspired by CNNA, we propose a new model called "Set-level Common-Near-Neighbor Analysis (SCNNA)" to explore the neighborhood information among sets instead of samples towards the multiple-shot person re-identification problem. When most sets of the same class stay closer to each other than those from different classes, the sets within the same class will share more similar set-level neighborhood structure than those from different classes. SCNNA utilizes this information to further ensure that inter-class dissimilarities are larger than intra-class dissimilarities for all sets.

Technically, SCNNA incorporates the set-level neighborhood information into a novel dissimilarity, named "Set-level Common-Near-Neighbor (SCNN)" dissimilarity, which is composed of a symmetric term and an asymmetric term combined by the following function:

$$D^{\text{SCNN}}(a,b) = D^{\text{Symmetric}}(a,b) + 2\lambda n D^{\text{Asymmetric}}(a,b). \quad (1)$$

In Equation (1), a and b denote two arbitrary sets; λ is the trade-off parameter between $D^{\text{Symmetric}}(a,b)$ and $D^{\text{Asymmetric}}(a,b)$; the "Fixed-number" for the neighborhood size, denoted by n, is suggested to be half of the average number of sets per class. Considering symmetry, $D^{\text{Symmetric}}(a,b)$ is given by:

$$D^{\text{Symmetric}}(a,b) = D_{\text{Fixed-number}}(a,b) + D_{\text{Fixed-number}}(b,a), \quad (2)$$

where

$$D_{\text{Fixed-number}}(a,b) = \sum_{i=0}^{n-1} O_b(f_a(i)). \quad (3)$$

In Equation (2), $D_{\text{Fixed-number}}(a,b)$ sums the rank orders of a's list's top elements in b's Rank-Order list under the setting of n, as shown in Figure 1, and $D_{\text{Fixed-number}}(b,a)$ is calculated in the similar way. In Equation (3), $f_a(i)$ is the i^{th} element in a's Rank-Order list; $O_b(f_a(i))$ returns the rank order of $f_a(i)$ in b's list. Here, Rank-Order list of an assigned set is formed by the ranking of all the other sets according to their dissimilarities to this set.

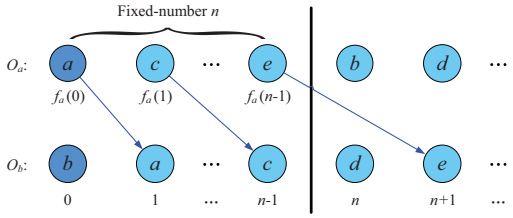

Figure 1. Sets are denoted by a, b, c, d, e and so on. $D_{\text{Fixed-number}}(a,b)$ is calculated from the 0^{th} to the $(n-1)^{th}$ nearest neighbor set in a's Rank-Order list.

As the asymmetric term, $D^{\text{Asymmetric}}(a,b)$ is given by:

$$D^{\text{Asymmetric}}(a,b) = min(O_a(b), O_b(a)). \quad (4)$$

In Equation (4), $O_a(b)$ is the rank order of b in a's Rank-Order list, and $O_b(a)$ is defined in the similar way.

The steps of SCNNA are displayed in Algorithm 1.

Algorithm 1 Set-level Common-Near-Neighbor Analysis

Require: Query image sets X_qs and corpus image sets X_cs; feature space mapping $\mathcal{P}$.

Ensure: Ranking Y_q of all X_cs for each X_q.

1: All X_qs and all X_cs are mapped by $\mathcal{P}$ into a new feature space, denoted by $\mathcal{X}_q$s and $\mathcal{X}_c$s, respectively.
2: All $\mathcal{X}_q$s and all $\mathcal{X}_c$s are listed together into the set group $\mathcal{X}$.
3: For a pair of $\mathcal{X}_q$ and $\mathcal{X}_c$, $\mathcal{X}$ are sorted by set-to-set dissimilarity measurements to each $\mathcal{X}_q$ and each $\mathcal{X}_c$, respectively, to obtain the set-level Rank-Order lists.
4: SCNN dissimilarity $\mathcal{D}$ is measured for each pair of $\mathcal{X}_q$ and $\mathcal{X}_c$ taking advantage of the set-level Rank-Order lists.
5: For each $\mathcal{X}_q$, according to $\mathcal{D}$s of all $\mathcal{X}_c$s to it calculated in step 4, all $\mathcal{X}_c$s are re-ranked to return the result Y_q.

From the efficiency perspective, SCNNA treats the samples in the same class as one whole set, so it is much faster than CNNA when there are multiple-shot images in each set. Suppose the average number of samples belonging to the same class (i.e., a set) is k,

and there are totally C classes, then the computational cost for re-identifying a query set (k samples) in CNNA will be $o(k^3C^2\log(kC))$, given that the fast sorting algorithm is adopted (i.e., $o(n\log n)$ complexity for n items). However, re-identifying cost for the same task in SCNNA will be only $o(kC\log k)$. Therefore, SCNNA is more than k^2C times faster than CNNA.

From the effectiveness perspective, firstly, SCNN dissimilarity creatively uses rank orders to compute the set-to-set dissimilarity. This can be regarded as a kind of quantized set-to-set dissimilarity measurement. Usually, non-uniform set distribution may impair the effectiveness of ranking which relies on the traditional set-to-set dissimilarity measurement, like MPD or CHISD. The quantized set-to-set dissimilarity measurement can overcome the non-uniform set distribution problem; secondly, the symmetric term concerns the "Fixed-number" of nearest neighbors for each set other than the set itself by means of the sum of the "Fixed-number" of rank orders in both Rank-Order lists for each set pair. Sum has the statistical averaging effect, which offers more robustness; last but not least, robustness of the symmetric term is influenced by the "Fixed-number". When there are only two sets in each class, the recommended "Fixed-number" will be half class size as "$n = 1$". On this condition, the asymmetric ranking problem is more obvious, which means a given pair of sets usually don't have the same rank order for each other in their own Rank-Order lists. It is heuristic and unfair to judge the rank order by randomly considering one side of them or simply averaging them. This problem can be tackled with the asymmetric term, which makes up the limitation of the symmetric term. The cooperation of the symmetric term and the asymmetric term makes SCNN dissimilarity more flexible and reliable.

2.2 Dissimilarity Measurement in Riemannian Space

SCNNA is based on set-level Rank-Order lists, which are formed by set-to-set dissimilarity measurements, thus the capability of it highly depends on the effectiveness of these measurements.

MRCG [8] is one state-of-the-art method for the multiple-shot person re-identification problem. It uses the covariance descriptors computed from dense overlapping grids on the images, which effectively capture the discriminative information of appearance details for each person. This covariance-based representation naturally leads to a dissimilarity measurement in Riemannian space. By condensing the information of all the samples within the set into a Karcher mean based signature, MRCG can deal with the large intra-class variations caused by pose changing, illumination varying, and occlusions. The algorithm of MRCG is detailed in [8].

Being impressed by the effectiveness of MRCG on the issue of multiple-shot person re-identification, we embed it into the SCNNA model, resulting in our proposed "RSCNNA" concept which makes use of the merits of both methods. The general steps of RSCNNA are in the following: since multiple-shot images for the same person within the same camera are treated to stay in one set, firstly, the covariance-based representation is extracted by MRCG for each set; then, with these representations, the SCNN dissimilarities between query sets and corpus sets are measured in Riemannian space; after that, set-based ranking is carried out according to these dissimilarities.

3 Experimental Results

Figure 2. Illustration of ETHZ, i-LIDS-MA, and i-LIDS-AA datasets.

We demonstrate our proposed method RSCNNA on public benchmark datasets: ETHZ [2], i-LIDS-MA [8], and i-LIDS-AA [8]. Typical samples of them are illustrated inX Figure 2. The ETHZ dataset is composed of three video sequences of crowded street scenes captured by two moving cameras mounted on a baby carrier. We utilize three subsets of it extracted by Schwartz and Davis for person re-identification [6]. There are 83 pedestrians within 4857 images for ETHZ1, 35 pedestrians within 1936 images for ETHZ2, and 28 pedestrians within 1762 images for ETHZ3. The iLIDS MCTS dataset is a video dataset captured at an airport arrival hall in the busy time under a multi-camera CCTV network [10]. We use the datasets i-LIDS-MA and i-LIDS-AA extracted by Bak *et al.* for testing their MRCG method for multiple-shot person re-identification [8]. The i-LIDS-MA dataset contains 40 individuals extracted from two cameras. From both cameras, 46 frames are annotated manually for each person. The i-LIDS-AA dataset is made of 100 individuals seen from both cameras, obtained by the HOG-based human detector and tracker. Undoubtedly, for i-LIDS-AA, the noisy detection and tracking results make the task of person re-identification more challenging.

We use all the persons in each mentioned dataset. We normalize all the images into 192×64 pixels. We repeat it for 10 times cross-validation and average the results for evaluation. Experimental results are illustrated by the CMC (Cumulative Matching Characteristic) curves, which represent the expectation of finding the correct match in the top several matches. For each person, we randomly select 10 images each time. And for fairness, in each time, we use the same selected data for method comparison. For each image, the covariance grid size is set to 16×16. The "Fixed-number" n and the trade-off parameter λ of SCNNA are suggested as "$n = 1$" and "$\lambda = 1$", because each person only has two sets (one query set and one corpus set from two different cameras).

We compare our proposed method RSCNNA with MPD [3], CHISD [1], and MRCG [8]. For MPD and CHISD, we adopt the discriminative vector descriptor concatenated of Dense-Sampled-Color-Histograms and Schmid-Filter-Bank in Euclidean space, which has

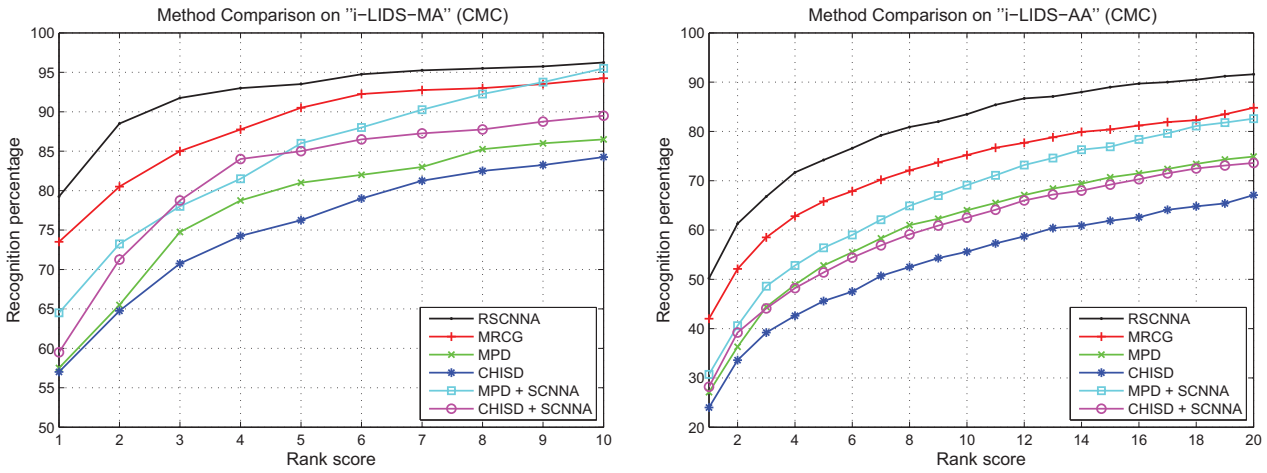

Figure 3. CMC performance comparison on the i-LIDS-MA and i-LIDS-AA datasets.

been widely used by state-of-the-art methods [9, 11]. In RSCNNA, we use the same feature representation designed in MRCG to demonstrate the advantage of using SCNNA, which can tackle the non-uniform distribution problem and the asymmetric ranking problem of sets in Riemannian space. Meanwhile, for conviction, we test the reliability and robustness of SCNNA by its collaboration with MPD and CHISD, denoted by "MPD + SCNNA" and "CHISD + SCNNA", respectively.

As a result, on ETHZ1, RSCNNA obtains 99.04% recognition rate on Rank-1; on ETHZ2 and ETHZ3, RSCNNA gets perfect results on Rank-1. These results are better than any other reported ones. The experimental results on iLIDS-MA and iLIDS-AA are drawn in Figure 3. Obviously, our proposed method outperforms the state-of-the-art methods. In greater detail, SCNNA has satisfactory collaboration with not only MRCG in Riemannian space, but also MPD and CHISD in Euclidean space. Although MRCG still performs better than MPD, CHISD, MPD + SCNNA, and CHISD + SCNNA, by performing SCNNA in Riemannian space, our proposed RSCNNA greatly prevails over MRCG, showing that it enhances the power of covariance-based representation by integrating the set-level neighborhood information.

4 Conclusion

This paper has proposed a novel approach called RSCNNA for the issue of multiple-shot person re-identification. It integrates the essence of two powerful models and further extends their abilities. Extensive experiments on standard datasets have shown its significant superiority to not only its two analogues but also other state-of-the-art competitors. Possible future work includes applying RSCNNA to the issue of across-camera human tracking.

Acknowledgments This work was supported by "R&D Program for Implementation of Anti-Crime and Anti-Terrorism Technologies for a Safe and Secure Society", Special Coordination Fund for Promoting Science and Technology of the Ministry of Education, Culture, Sports, Science and Technology, the Japanese Government.

References

[1] H. Cevikalp and B. Triggs. Face recognition based on image sets. In *CVPR*, pages 2567–2573, Jun. 2010.

[2] A. Ess, B. Leibe, and L. V. Gool. Depth and appearance for mobile scene analysis. In *ICCV*, pages 1–8, Oct. 2007.

[3] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, pages 2360–2367, Jun. 2010.

[4] W. Li, Y. Wu, M. Mukunoki, and M. Minoh. Common-near-neighbor analsis for person re-identification. In *ICIP*, Oct. 2012.

[5] B. Loris, C. Marco, P. Alessandro, F. Michela, and M. Vittorio. Multiple-shot person re-identification by hpe signature. In *ICPR*, pages 1413–1416, Aug. 2010.

[6] W. R. Schwartz and L. S. Davis. Learning discriminative appearance-based models using partial least squares. In *Proceedings of the XXII Brazilian Symposium on Computer Graphics and Image Processing, SIBGRAPI*, pages 322–329, Oct. 2009.

[7] B. Slawomir, C. Etienne, B. Francois, and T. Monique. Person re-identification using haar-based and dcd-based signature. *AVSS*, pages 1–8, Aug. 2010.

[8] B. Slawomir, C. Etienne, B. Francois, and T. Monique. Multiple-shot human re-identification by mean riemannian covariance grid. In *AVSS*, pages 179–184, Aug. 2011.

[9] Y. Wu, M. Minoh, M. Mukunoki, and S. Lao. Set based discriminative ranking for recognition. In *ECCV*, Oct. 2012.

[10] W.-S. Zheng, S. Gong, and T. Xiang. Associating groups of people. In *BMVC*, pages 23.1–23.11, Sep. 2009.

[11] W.-S. Zheng, S. Gong, and T. Xiang. Re-identification by relative distance comparison. *In press with PAMI*, 2012.