LETTER

Improving Person Re-Identification by Efficient Pairwise-Specific CRC Coding in the XQDA Subspace

Ying TIAN[†], **Mingyong ZENG**[††a)], **Aihong LU**[†], **Bin GAO**[††], ***Nonmembers***, ***and*** **Zhangkai LUO**[††], ***Student Member***

SUMMARY A novel and efficient coding method is proposed to improve person re-identification in the XQDA subspace. Traditional CRC (Collaborative Representation based Classification) conducts independent dictionary coding for each image and can not guarantee improved results over conventional euclidian distance. In this letter, however, a specific model is separately constructed for each probe image and each gallery image, i.e. in probe-galley pairwise manner. The proposed pairwise-specific CRC method can excavate extra discriminative information by enforcing a similarity item to pull similar sample-pairs closer. The approach has been evaluated against current methods on two benchmark datasets, achieving considerable improvement and outstanding performance.

key words: *person re-identification, collaborative representation based classification (CRC), XQDA*

1. Introduction

In the recent decade, person re-identification (re-id), which means recognizing a person-of-interest in different camera views, has gained consistent progress. Unlike face recognition, re-id has remained extremely challenging due to more uncontrolled variations in practical scenarios. To further improve the accuracy, current solutions mainly focus on designing or learning better descriptors [1] and(or) metrics [2].

Among existing re-id descriptors, SDALF [3] leverages the symmetry of pedestrians and fuses color with texture features. LOMO [2] combines HSV and texture histograms of overlapping patches in three-scale resolutions. Recently, GOG [4] proposes a hierarchical gaussian descriptor to model image regions in four different colorspaces. Besides, deep learning is becoming more and more popular for learning pedestrian representations. In this letter, we adopt GOG descriptor and focus on a better metric with coding based approach.

Metric learning can often boost re-id greatly with labeled or unlabeled training samples. KISSME (Keep It Simple and Straightforward MEtric) [5] is a famous inference based metric while XQDA learns a discriminant subspace before applying KISSME by cross-view quadratic discriminant analysis [2]. CNNA [6] proposes the common-near neighbor analysis to tackle with the deviation of badly-distributed samples. Furthermore, dictionary coding methods have also been applied, among which the researchers in Kyoto University (Wu, Minoh [7]) consider that Collaborative Representation based Classification (CRC) performs comparable but more efficient than sparse coding. The recent WLC [8] also proposes a weighted linear coding to learn multi-level descriptors. After learning sample-specific SVMs, LSSCDL [9] learns a pair of dictionaries and a mapping function to predict the similarity. With labeled training samples in two views as two dictionaries, KXCRC [10] proposes a supervised CRC extension considering both kernel tricks and cross-view coding.

Currently, thanks to the joint efforts in the re-id community, the recognition rate is becoming much higher than a decade ago. At the same time, improving re-id further is also becoming increasingly difficult. Considering that XQDA is already capable of mining most discriminative information in the training data, we derive an approximated XQDA subspace and adapt a CRC coding method to replace the previous Mahalanobis metric. Unlike conventional CRC where the coding of each sample on the dictionary is independent, each probe image and each gallery image are paired and coded together with a similarity constraint to form a pairwise-specific CRC model (PS-CRC). Our method is validated to achieve improved results on two public datasets.

2. XQDA Subspace and CRC Coding

As our approach is built upon the XQDA method, we briefly introduce its learned metric. In the XQDA (i.e. Cross-view Quadratic Discriminant Analysis), the distance between two d-dimensional pedestrian descriptors $\mathbf{x}_i$ and $\mathbf{x}_j$ can be calculated as

$$D\left(\mathbf{x}_i, \mathbf{x}_j\right) = \left(\mathbf{x}_{ij}\right)^T \mathbf{W}\left(\boldsymbol{\Sigma}_I^{-1} - \boldsymbol{\Sigma}_E^{-1}\right)\mathbf{W}^T\left(\mathbf{x}_{ij}\right) \tag{1}$$

where $\mathbf{x}_{ij} = \mathbf{x}_i - \mathbf{x}_j$, and $\mathbf{W} \in \mathbb{R}^{d\times r}$ is the learned subspace projection matrix from original d to lower r dimensions with generalized eigenvalue decomposition. $\mathbf{M} = \boldsymbol{\Sigma}_I^{-1} - \boldsymbol{\Sigma}_E^{-1}$ is the classical KISSME metric in the learned subspace, where $\boldsymbol{\Sigma}_I$ and $\boldsymbol{\Sigma}_E$ denote the covariance matrices of intra-personal and extra-personal classes in the training set, respectively.

The above metric is effective when computing the distance of two samples, but it is not convenient if we want to get a compact representation for each sample. As $\mathbf{M}$ also contains discriminative information, we define $\bar{\mathbf{W}} = \mathbf{W}\mathbf{M}^{1/2}$ as the XQDA subspace rather than using only $\mathbf{W}$. Then sample $\mathbf{x}$ can be projected to a lower subspace $\mathbf{y} = \bar{\mathbf{W}}^T\mathbf{x}$. An

Manuscript received September 30, 2017.
Manuscript revised December 2, 2017.
Manuscript publicized December 25, 2017.

[†]The authors are with Suzhou Institute of Trade and Commerce, Suzhou 215009, China.
[††]The authors are with College of Communications Engineering, Army Engineering University, Nanjing 210007, China.
a) E-mail: zengmingyong1987@gmail.com (Corresponding author)

DOI: 10.1587/transinf.2017EDL8216

additional benefit is that the mahalanobis based distance in Eq. (1) becomes the simple l_2 euclidean distance between $\mathbf{y}_i$ and $\mathbf{y}_j$, which is more computationally efficient.

However, adopting euclidean distance in above XQDA subspace may not be optimal. We thus study whether the coding strategy can perform better in this subspace. We focus on collaborative representation rather than sparse coding for its superior performance and efficiency on coding [7]. Using all gallery images as the dictionary $\mathbf{D}$, the CRC method represents each probe image $\mathbf{y}$ with a coding vector $\mathbf{z}$:

$$\min_{\mathbf{z}} \|\mathbf{y} - \mathbf{D}\mathbf{z}\|_2^2 + \lambda\|\mathbf{z}\|_2^2 \tag{2}$$

where λ is a scalar parameter to balance the representation residual and the regularization term. It assigns the probe image to the class that results in the smallest reconstruction error for classification.

3. Pairwise-Specific CRC Coding

Though successful in face recognition, the original CRC does not suit well for re-id as there exist very few images for each gallery person. A simple idea is to use training samples to construct $\mathbf{D}$ and code each test sample with the analytical solution in Eq. (2) $\mathbf{z}=\left(\mathbf{D}^T\mathbf{D}+\lambda\mathbf{I}\right)^{-1}\mathbf{D}^T\mathbf{y}$, then the cosine distance between coding vectors can be employed. However, above coding space does not always guarantee improved results when compared to the original XQDA subspace. Inspired by sample-specific models in LSSCDL [9] and KXCRC [10], we propose a pairwise-specific CRC coding method below.

On a shared dictionary $\mathbf{D}$ with k training samples, a probe sample $\mathbf{y}_p$ and a gallery sample $\mathbf{y}_g$ are proposed to be coded jointly as

$$\begin{aligned}\min_{\mathbf{z}_p,\mathbf{z}_g} &\left\|\mathbf{y}_p - \mathbf{D}\mathbf{z}_p\right\|_2^2 + \left\|\mathbf{y}_g - \mathbf{D}\mathbf{z}_g\right\|_2^2 \\ &+ \lambda\left\|\mathbf{z}_p\right\|_2^2 + \lambda\left\|\mathbf{z}_g\right\|_2^2 + \beta\left\|\mathbf{z}_p - \mathbf{z}_g\right\|_2^2\end{aligned} \tag{3}$$

where the main novelty lies in the last item which enforces a similarity constraint between two coding vectors. This item is expected to generate more similar coding vectors if $\mathbf{y}_p$ and $\mathbf{y}_g$ are from the same person than from different persons. The cosine distance between corresponding coding vectors thus becomes more appropriate for distance computing.

As different probe-gallery pairs have specific coding vectors, multiple pairwise optimizations in Eq. (3) seem to bring much computation burden at the first sight. In fact, there still exist closed-form analytical solutions and the computation can be accelerated efficiently through precomputing and storing some shared variables. By setting the derivatives with respect to $\mathbf{z}_p$ and $\mathbf{z}_g$ to zero in Eq. (3), we can obtain interdependent coding vectors:

$$\mathbf{z}_p = \mathbf{P}\left(\mathbf{D}^T\mathbf{y}_p + \beta\mathbf{z}_g\right), \quad \mathbf{z}_g = \mathbf{P}\left(\mathbf{D}^T\mathbf{y}_g + \beta\mathbf{z}_p\right) \tag{4}$$

where $\mathbf{P} = \left(\mathbf{D}^T\mathbf{D} + (\lambda+\beta)\mathbf{I}\right)^{-1}$ and $\mathbf{I}$ is the identity matrix. Note that if $\beta = 0$, above PS-CRC model degrades to original CRC method where each coding vector is independent. By substituting $\mathbf{z}_g$ and $\mathbf{z}_p$ in (4) and denoting $\mathbf{Q} = \left(\mathbf{I} - \beta^2\mathbf{P}^2\right)^{-1}$, the analytical solutions can be re-written as

$$\mathbf{z}_p = \mathbf{Q}\mathbf{P}\mathbf{D}^T\mathbf{y}_p + \beta\mathbf{Q}\mathbf{P}^2\mathbf{D}^T\mathbf{y}_g \tag{5}$$

$$\mathbf{z}_g = \mathbf{Q}\mathbf{P}\mathbf{D}^T\mathbf{y}_g + \beta\mathbf{Q}\mathbf{P}^2\mathbf{D}^T\mathbf{y}_p \tag{6}$$

Denoting $\mathbf{A} = \mathbf{Q}\mathbf{P}\mathbf{D}^T$ and $\mathbf{B} = \beta\mathbf{Q}\mathbf{P}^2\mathbf{D}^T$, it is obvious that all pairwise coding models share the same projection matrices $\mathbf{A}$ and $\mathbf{B}$. Besides, we can pre-compute the intermediate coding vectors $\mathbf{z}_p^A = \mathbf{A}\mathbf{y}_p$, $\mathbf{z}_p^B = \mathbf{B}\mathbf{y}_p$ for all $\mathbf{y}_p \in \mathbf{Y}_p$, and pre-compute $\mathbf{z}_g^A = \mathbf{A}\mathbf{y}_g$, $\mathbf{z}_g^B = \mathbf{B}\mathbf{y}_g$ for all $\mathbf{y}_g \in \mathbf{Y}_g$. Then either $\mathbf{z}_p$ or $\mathbf{z}_g$ can be calculated with just one addition of two pre-computed vectors, followed by the cosine distance computing. The detailed procedures are summarized in Algorithm 1.

Unlike traditional CRC method which constructs $\mathbf{D}$ with all gallery samples, we propose to pick part of the training samples to form the dictionary via unsupervised k-means or simple random sampling. An alternative idea is to learn a more distinctive $\mathbf{D}$ by supervised dictionary learning. However, it turns supervised and the results may not be improved consistently in the XQDA subspace as most of the discrimination in the training labels has already been exploited by XQDA. We further highlight the unsupervised nature in PS-CRC when compared with KXCRC [10], which uses all probe and gallery training samples to form two supervised dictionaries where respective samples in the same dictionary column represent the same person.

Algorithm 1 : Pairwise-Specific CRC (PS-CRC)

Input: Dictionary $\mathbf{D}$, probe set $\mathbf{Y}_p$, gallery set $\mathbf{Y}_g$, parameters λ, β
Step 1: Compute projection matrices $\mathbf{P}$, $\mathbf{Q}$, $\mathbf{A}$ and $\mathbf{B}$
Step 2: Pre-compute intermediate coding matrices
$\mathbf{Z}_p^A = \mathbf{A}\mathbf{Y}_p, \mathbf{Z}_p^B = \mathbf{B}\mathbf{Y}_p$
$\mathbf{Z}_g^A = \mathbf{A}\mathbf{Y}_g, \mathbf{Z}_g^B = \mathbf{B}\mathbf{Y}_g$
Step 3: Pairwise CRC coding and cosine distance computing
for $\mathbf{y}_p \in \mathbf{Y}_p$ **do**
  Get corresponding coding vectors $\mathbf{z}_p^A$, $\mathbf{z}_p^B$ from $\mathbf{Z}_p^A$, $\mathbf{Z}_p^B$
  for $\mathbf{y}_g \in \mathbf{Y}_g$ **do**
    Get corresponding coding vectors $\mathbf{z}_g^A$, $\mathbf{z}_g^B$ from $\mathbf{Z}_g^A$, $\mathbf{Z}_g^B$
    $\mathbf{z}_p = \mathbf{z}_p^A + \mathbf{z}_g^B, \mathbf{z}_g = \mathbf{z}_g^A + \mathbf{z}_p^B$
    Calculate the cosine distance
    $Dist(\mathbf{y}_p, \mathbf{y}_g) = 1 - \mathbf{z}_p^T\mathbf{z}_g \Big/ \left(\|\mathbf{z}_p\|_2^2 \|\mathbf{z}_g\|_2^2\right)$
  end for
end for
Output: Distance matrix between $\mathbf{Y}_p$ and $\mathbf{Y}_g$: $Dist(\mathbf{Y}_p, \mathbf{Y}_g)$

4. Experimental Results

The proposed PS-CRC method in the XQDA subspace is validated on two public datasets, i.e. VIPeR [3] and CUHK01 [2]. VIPeR contains 632 pedestrians with exactly 2 images for each person while CUHK01 contains 3884 images of 971 persons (i.e. four samples for each identity). The Cumulative Matching Characteristic (CMC) curve is used as

Table 1 Comparison with baseline and other PS-CRC variants.

Method	R1	R5	R10	R20	time(s)
Baseline	48.58	79.24	88.89	94.05	**0.04**
PS-CRC	**52.44**	81.77	90.09	95.63	0.61
PSCRC-M	50.13	80.28	89.18	94.56	0.61
SI-CRC	49.53	80.54	89.49	94.62	0.07
KPS-CRC	52.15	**82.09**	**90.51**	**95.73**	2.14

the evaluation tool which represents the probability of finding the correct match in the first n matches. As in KXCRC, we adopt the same GOG descriptor and repeat the same 10 folds random test for an average result. In each fold, a half of the dataset are used for training and the rest are for testing. The original 27622-d GOG feature is projected into XQDA subspace and then normalized with unit l_2 length. A half of the training samples are picked to form the dictionary $\mathbf{D}$ and the parameters are set to achieve best result for each method. For PS-CRC, we set $\lambda = 0.3$ and $\beta = 5$ on VIPeR.

The first experiment is to evaluate the components in our method on the VIPeR dataset. The XQDA subspace with l_2 euclidean distance is used as the baseline method. One module is modified from our PS-CRC method while others are kept still to justify the corresponding design. These variants include: (1) PSCRC-M, $\mathbf{M}$ in the XQDA subspace is removed; (2) SI-CRC, sample independent CRC, i.e. $\beta = 0$; (3) K-PSCRC, similar kernel techniques are applied in PS-CRC as KXCRC has reported improved results over its non-kernel models. Table 1 lists the CMC values at some top ranks (e.g. R5 denotes cumulative matching accuracy at rank 5) of these methods and the average time for distance computing after tuning their parameters to the best results.

From Table 1, PS-CRC has improved more than 3% Rank-1 rate over the baseline XQDA subspace method, which can be considered a great enhancement as recent methods seem to reach saturated results on this challenging VIPeR dataset. PSCRC-M and SI-CRC perform worse than PS-CRC, indicating that $\mathbf{M}$ in the XQDA subspace is discriminative and the pairwise-specific coding strategy is crucial to mine extra discrimination. KPS-CRC achieves slightly better results on the later ranks but its top ranks are relatively weaker. It suggests the pairwise-specific coding plays the most important role rather than the kernel tricks and the benefit brought by kernels in the discriminant XQDA subspace is not as great as that in the supervised KXCRC. Though not as efficient as the simple l_2 distance, the computation of all the distances $Dist(\mathbf{Y}_p, \mathbf{Y}_g)$ in PS-CRC takes less than 1 second, which is still very efficient due to the analytical solutions and the pre-computed matrices. As KPS-CRC needs to tune additional kernel parameters and achieves similar results with more computing time, we highlight only PS-CRC in this letter.

Then we continue to compare with current related methods listed in Table 2 on VIPeR. Among these methods, LOMO and GOG are two recent excellent re-id descriptors with more than 25000 dimensions, which are both reduced to lower dimensions for evaluation with XQDA. LSSCDL further improves LOMO's result by sample-specific SVM learning and semi-coupled dictionary learning. WLC is a recent feature learning method which also employs unsupervised dictionary coding on a dictionary constructed by k-means. KXCRC is a supervised cross-view CRC coding method with kernel extensions. Table 2 reveals that PS-CRC outperforms these methods at all ranks. It should be noted that KXCRC tunes λ but does not tune the cross-view codings (i.e. with fixed $\beta = 1$), which may hamper its result as β may play a more vital role (e.g. $\beta > \lambda$ in our PS-CRC).

Table 2 Comparison with state-of-the-art methods on VIPeR.

Method	Reference	R1	R5	R10	R20
LOMO	CVPR15 [2]	40.0	68.0	80.5	91.1
LSSCDL	CVPR16 [9]	42.7	-	84.3	91.9
WLC	AAAI17 [8]	51.4	76.4	84.8	-
GOG	CVPR16 [4]	49.7	79.7	88.7	94.5
KXCRC	ArXiv16 [10]	51.6	80.8	89.4	95.3
PS-CRC	Ours	**52.4**	**81.8**	**90.1**	**95.6**

Table 3 Comparison with current methods on CUHK01.

Method	Reference	R1	R5	R10	R20
LOMO*	CVPR15 [2]	63.2	81.0	90.1	93.5
LSSCDL	CVPR16 [9]	**65.9**	88.1	92.1	96.0
WLC	AAAI17 [8]	65.8	81.1	85.9	-
GOG	CVPR16 [4]	57.8	79.1	86.2	92.1
KXCRC	ArXiv16 [10]	61.2	80.9	87.3	93.2
PS-CRC	Ours	65.8	**88.1**	**92.2**	**96.2**

Next we conduct further validation experiments on the larger CUHK01 dataset. The results are listed in Table 3, where the compared methods are the same with that in Table 2. Similar with VIPeR, we adopt GOG descriptor and repeat 10 folds tests on CUHK01 for an average result. For each test, 486 persons are randomly sampled from 971 persons for training and the rest 485 persons are used for testing. Note that in CUHK01, each person has two images in each camera view, only one image is selected for a person under each camera view in testing. For the compared LOMO method, the multi-shot result is reported since its authors only employed the more advantageous multi-shot setting. For our PS-CRC, a half of the training samples are picked to form the dictionary $\mathbf{D}$ and the parameters are set as $\lambda = 0.5$, $\beta = 9$. From Table 3, it is clear that PS-CRC improves greatly from the baseline GOG method, resulting in 8% leap in Rank-1 rate from 57.8% to 65.8%. Besides, it outperforms the other three dictionary learning/coding based methods WLC, LSSCDL and KXCRC.

In fact, PS-CRC can be thought as a second stage or a re-rank technique after the initial XQDA metric. And above experiments have demonstrated its potentials for performance improvement and outperforming results on both datasets. However, it would be irresponsible if we do not point out its two drawbacks. One is that the parameters should be tuned on different datasets and we argue that the specific parameters may reflect the distribution bias of certain datasets. The other drawback is that the whole method is supervised because it is built on the supervised XQDA.

We claim to use an unsupervised dictionary in this letter because we find the learned dictionary does not necessarily improve on the test datasets. Thus it may not guarantee to gather a reasonably good dictionary if the training set has insufficient images or identities on other small datasets. Though with above shortcomings explained, we tend to consider that it is still worthwhile to tune the parameters or try to build a good dictionary for the great performance gain PS-CRC can bring.

5. Conclusion

A novel dictionary coding method PS-CRC has been proposed to improve person re-identification. As the original feature space is often high dimensional, PS-CRC is built upon the XQDA subspace. Different from traditional coding methods that do not consider the specialty for every probe-gallery sample-pair, we propose pairwise-specific models by putting together respective CRC models and adding another similarity item between them. Though introducing multiple models, it is still very efficient as the intermediate coding vectors can be computed in advance. On the VIPeR and CUHK01 datasets, the proposed method has improved considerably over the baseline method and achieves outstanding performance. Future directions include evaluating on more datasets, learning better dictionary by more sophisticated supervised or unsupervised methods and deeper study on kernel techniques.

References

[1] M. Zeng, Z. Wu, C. Tian, L. Zhang, and L. Hu, "Efficient person re-identification by hybrid spatiogram and covariance descriptor," IEEE Conf. Computer Vision and Pattern Recognition Workshops, pp.48–56, 2015.

[2] S. Liao, Y. Hu, X. Zhu, and S.Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp.2197–2206, 2015.

[3] L. Bazzani, M. Cristani, and V. Murino, "Symmetry-driven accumulation of local features for human characterization and re-identification," Comput. Vis. Image Und., vol.117, no.2, pp.130–144, 2013.

[4] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical Gaussian descriptor for person re-identification," IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp.1363–1372, 2016.

[5] M. Kostinger, M. Hirzer, P. Wohlhart, P.M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp.2288–2295, 2012.

[6] W. Li, M. Mukunoki, Y. Kuang, Y. Wu, and M. Minoh, "Person re-identification by common-near-neighbor analysis," IEICE Trans. Inf. & Syst., vol.E97-D, no.11, pp.2935–2946, Nov. 2014.

[7] W. Li, Y. Wu, M. Mukunoki, and M. Minoh, "Bi-level relative information analysis for multiple-shot person re-identification," IEICE Trans. Inf. & Syst., vol.E96-D, no.11, pp.2450–2461, Nov. 2013.

[8] Y. Yang, L. Wen, S. Lyu, and S.Z. Li, "Unsupervised learning of multi-level descriptors for person re-identification," 31st AAAI Conf. Artificial Intelligence, pp.4306–4312, 2017.

[9] Y. Zhang, B. Li, H. Lu, A. Irie, and X. Ruan, "Sample-specific SVM learning for person re-identification," IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp.1278–1287, 2016.

[10] R. Prates and W.R. Schwartz, "Kernel cross-view collaborative representation based classification for person re-identification," arXiv, 1611.06969, 2016.