UNSUPERVISED PERSON RE-IDENTIFICATION VIA RE-RANKING ENHANCED SAMPLE-SPECIFIC METRIC LEARNING

Heng Zhao[1] *Zhenjun Han*[*1,2] *Zhaoju Li*[1] *Fei Qin*[1,2]

[1]School of Electronics, Electrical and Communication Engineering
University of Chinese Academy of Sciences, Beijing, China.
[2]Insititute of Electrics, Chinese Academy of Sciences.
hanzhj@ucas.ac.cn

ABSTRACT

Despite of the great progress of image-based person re-identification, most existing methods use supervised metric learning to build re-identification models and thus require repeated human effort to annotate sample pairs from non-overlapping cameras. In this paper, we propose an unsupervised sample-specific metric learning approach (SSML) to alleviate this problem. Specifically, using samples those are negatives (with a high probability) to the query samples, we train a local metric for each query sample following the max-margin learning theory. Moreover, a KNN intersection re-ranking (KIRR) method is used to further decrease the ambiguity of samples and aggregate the re-identification performance. With experiments on three widely used person re-identification datasets: VIPeR, CUHK01, and PRID, we demonstrate that the proposed approach is simple but effective.

***Index Terms*—** Person Re-identification, Sample-specific Metric Learning, Re-ranking

1. INTRODUCTION

Person re-identification is to associate the same person across non-overlapping surveillance scenes. Researchers have made great progress on feature representation and metric learning models to address the challenges in this area, such as the variations of posture, illumination, view angle, and scale.

Effective feature representations including the Ensemble of Local Features (ELF) [1], Symmetry-Driven Accumulation of Local Features (SDALF) [2], Fisher Vectors (LDFV) [3], mid-level filter [4], Local Maximal Occurrence (LOMO) [5], and Hierarchical Gaussian Descriptor [6] have been proposed. With extracted features, supervised metric-learning approaches [7, 8, 4, 9, 10, 11, 12, 5, 13] including KISSME [7], RDC [8], kLFDA [9], and XQDA [5] have been developed to construct person-identification models.

Despite of the effectiveness of supervised metric learning methods, their performance usually relies on well annotated sample pairs, which require costly human effort for sample annotation in various surveillance scenes. This has severely limited the scalability and usage of person re-identification in practical scenarios.

To tackle this problem, unsupervised methods [2, 14, 15, 16, 17] have been proposed. These methods usually utilize unlabeled data which are abundant in related surveillance scenes, and the scalability and the practicability make them receive great attention. Wang et al. [15] utilized localised saliency feature without cross-view discriminative information. Lisanti et al. [14] designed hand-crafted appearance features and sparse reconstruction. Kodirov et al. [16] proposed a dictionary learning based model which is intrinsically suitable for unsupervised learning. Peng et al. [17] used the idea of transfer learning and thus requires partially annotated samples, without minimizing human effort on sample annotation. These methods aim finding feature spaces where matching is effective but lack the way to learn effective metrics, which makes them less discriminative than supervised approaches.

Different from existing unsupervised methods, we propose a method aiming learning discriminative metric and only using unlabeled data from the same scene of the query sample. Unlike many other works aiming to train a single global distance metric for all datasets, that suffers limitation while handling the data with diversity, we propose an unsupervised sample-specific metric learning method (SSML) for person re-identification. Based on the max-margin theory, we train a local metric for each query sample using itself as positive and all other unlabeled data from the same scene as negatives. The learned metric in this camera view would perform the same discriminative ability to the corresponding query sample in any other view, e.g. the gallery view. Specifically, the metric is learned via an optimization problem with margin constraint like SVM. Furthermore, a KNN intersection re-ranking (KIRR) method is designed to further decrease the ambiguity of unrelated gallery samples and improve the performance during the matching process. It is effective especially in real scenario that the gallery set is very large. Our approach is competitive to the state-of-the-art unsupervised methods and

978-1-5090-2175-8/17/$31.00 ©2017 IEEE

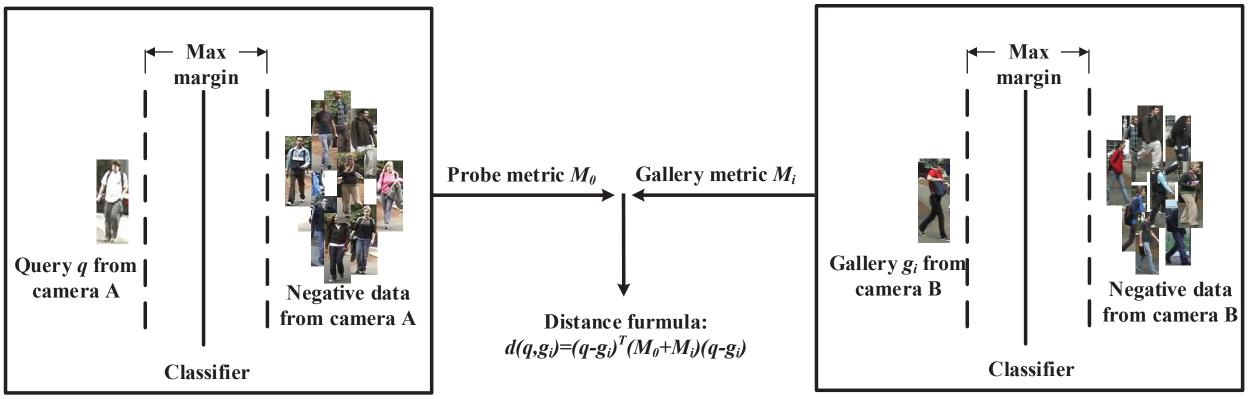


Fig. 1. Instead of training a single global distance metric, we train a local metric for each query sample with a single positive sample and a large number of negative samples. Negatives come from the same camera view with a query sample.

suitable for real scene.

2. UNSUPERVISED PERSON RE-ID APPROACH

We describe the proposed model in two parts: the unsupervised sample-specific metric learning (SSML) algorithm and the KNN intersection re-ranking (KIRR) algorithm.

2.1. Sample-specific Metric Learning

We denote the feature vector of a query (probe) sample as q and a gallery set as $G = \{g_i\}_{i=1}^{n}$, where n is the size of the gallery. A common metric learning is to learn a global Mahalanobis metric M in distance equation:

$$d(q, g_i)_M^2 = (q - g_i)^T M (q - g_i). \tag{1}$$

The proposed algorithm, Fig. 1, looks for the local matrix M [18] that maximizes the margin between the query example q and the closest negative example in g_1,...,g_n,

$$M = \underset{M \succeq 0}{\text{argmax}}(\min_{1 \leq i \leq n} (g_i - q)^T M (g_i - q)). \tag{2}$$

Here negative samples are captured from the same camera view but have different ID from the query sample. It is easy to get a large number of unlabeled negative samples via some pedestrian detection algorithms like DPM [19]. It happens with high probability that negative samples have different ID from the query sample especially using some strategies like similarity comparing and tracking algorithm.

Equation (2) does not have an unique solution in the current form. So we rewrite the equation as an optimization problem with constraint:

$$\begin{aligned} &M(q) = \underset{M \succeq 0}{\text{argmin}} \frac{1}{2} ||M||^2 \\ &s.t.\ (g_i - q)^T M (g_i - q) \geq c \ \forall i \in \{1, \dots, n\}. \end{aligned} \tag{3}$$

The constant c is set as 2 for convenience later on. We define $\tilde{g}_i = g_i - q$, $M = y^T y$. So the constraint in equation (3) can be rewritten as the form of a quadratic kernel:

$$\begin{aligned} &(g_i - q)^T M (g_i - q) = \tilde{g}_i y^T y \tilde{g}_i = \tilde{g}_i \tilde{g}_i^T \cdot y y^T \\ &= \varphi(\tilde{g}_i) \cdot \varphi(y) = k(\tilde{g}_i, y). \end{aligned} \tag{4}$$

We define $y_0 = -1$ for q and $y_i = 1$ for g_i where $1 \leq i \leq n$. Equation (3) can be rewritten as the form of SVM,

$$\begin{aligned} &M(q) = \underset{M \succeq 0}{\text{argmin}} \frac{1}{2} ||M||^2 \\ &s.t.\ (< M, \varphi(\tilde{g}_i) > -1) \geq 1 \ \forall i \in \{1, \dots, n\}. \end{aligned} \tag{5}$$

It can be seen that this is equivalent to a kernel SVM problem, and it is easy to solve by quadratic programming method. Since $\varphi(g) \succeq 0$ for any g, $\varphi(\tilde{q}) = \varphi(0) = 0$, and $y_i = 1$ for $i \geq 1$, the final solution M has the form:

$$M = \sum_{i=1}^{n} \alpha_i y_i \varphi(\tilde{g}_i) \succeq 0,\ \alpha_i \geq 0. \tag{6}$$

So the matrix M in (2) is indeed positive semidefinite.

After obtaining the metric M_0 for a query q, the distance can be calculated between q and the gallery samples $g_1, \dots, g_n$. The same strategy can be utilized on each gallery sample to generate n metrics $\{M_i\}_{i=1}^{n}$ as shown in Fig. 1. The final distance of two samples is:

$$d(q, g_i)^2 = (q - g_i)^T (M_0 + M_i)(q - g_i). \tag{7}$$

Like most other metric learning methods, SSML would suffer performance loss in open (practical) gallery set due to the ambiguity of many unrelated samples in it. So we propose a re-ranking strategy to ease this problem.

2.2. KNN Intersection Re-Ranking

In this section an unsupervised re-ranking strategy is introduced. This method is utilized based on the initial gallery rank of the query sample. The initial rank is generated based on the distance between the query and every gallery sample.

The core idea of re-ranking is that: if the query sample's k-nearest neighbor and the gallery sample's k-nearest neighbor have some same gallery samples. Then the query sample is often similar with this gallery sample in some degree. Based on this phenomenon, we can re-rank the initial result of each query sample to overcome the open gallery problem.

We denote the feature vector of a query sample as q and a gallery set as $G = \{g_i\}_{i=1}^{n}$ as section 2.1. After calculating the distances $d(q, g_i)$, an initial query-gallery rank can be obtained as $R_q(G) = \{g_i^0\}_{i=1}^{n}$ where g_i^0 represents i-th sample in the rank list which satisfy $d(q, g_1^0) < d(q, g_2^0) <, \ldots, < d(q, g_n^0)$.

First, a rank score $S_r(q, g_i^0) = 1/i$ is defined to present the similarity between q and g_i^0. Next, we calculate the number of common k-nearest neighbors of q and g_i^0. Specifically, we donate $n_k(q)$ as the k-nearest neighbors of the query q and $n_k(g_i^0)$ as the k-nearest neighbors of the gallery sample g_i^0. So we can define:

$$S_{kn}(q, g_i^0) = |n_k(q) \cap n_k(g_i^0)|. \tag{8}$$

The final new similarity between q and g_i^0 is defined as:

$$S_f(q, g_i^0) = S_{kn}(q, g_i^0) \times S_r(q, g_i^0) = \frac{|n_k(q) \cap n_k(g_i^0)|}{i}. \tag{9}$$

Aided by equation (9), a new similarity score between query and gallery can be obtained, which has better performance with open gallery problem.

3. EXPERIMENTS

3.1. Datasets and Settings

Datasets. Three widely used public datasets are used for the experiments. **VIPeR** [20]. It contains 632 pedestrians with resolution 128*48. Each pair of person images is taken from two distinct camera views. It is very challenging due to the low resolution and the variations of illumination and posture. We randomly selected half samples of the set for training and the rest for testing. **CUHK01** [21]. It is collected in a campus environment with a resolution of 160*60. It contains 971 persons. Each person contains two images in each camera view where one camera captures the front or back view of persons while another camera captures the side view. We just select the first of two images from each person in each camera. The dataset was randomly divided to 485 for training and 486 for testing. **PRID** [22]. It is a more realistic dataset with two camera views. Camera view A contains 385 persons, camera view B contains 749 persons, with 200 of them appearing in both views. For fair comparison with published results [16, 17], we randomly select 100 persons from the 200 persons in both camera views for the training set, while the remaining 100 of view A are used as the probe set, and remaining 649 of view B are used as gallery.

Settings. For the person appearance representation, we use a 5138-D feature vector [10] concatenating with colour histogram, HOG [23] and LBP [24], which is widely utilized in unsupervised person re-id methods. We reduce the feature dimension to 400 using PCA to reduce the computation. We set the parameter k in re-ranking close to 10 percent of the size of test data. All final results are obtained by averaging with 10 trials.

3.2. Result and Analysis

3.2.1. *Combination of probe and gallery metric*

Table 1. Results of SSML on VIPeR, CUHK01, and PRID.

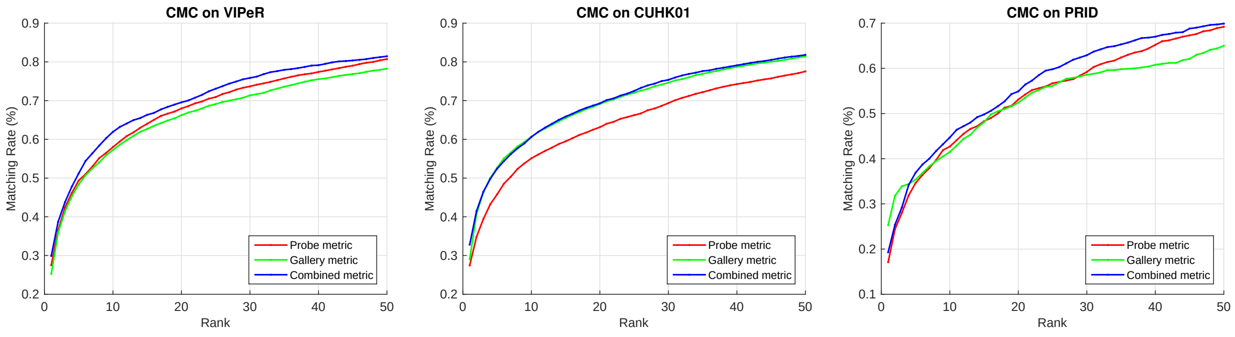
Rank-1	VIPeR	CUHK01	PRID (gal: 100)	PRID (gal: 649)
Probe metric	27.53	27.44	35.90	17.10
Gallery metric	25.25	29.05	34.80	**25.30**
Combined	**29.91**	**32.82**	**40.90**	19.30

Tab. 1 and Fig. 2 show that the combined metric is usually better than the single probe metric or gallery metric. But if the gap between the performance of the probe metric and the gallery metric is too large, the combined result would have no additional gain. It is due to the fact that the gallery set (649) in PRID is much larger than the probe set (100) due to containing a lot of unrelated persons. If we just use the gallery set (100) with the persons corresponding to the probe set (100) in PRID, then the rank-1 performance would be: probe metric 35.90%, Gallery metric 34.80%, Combined metric 40.90%.

In order to fill the gap between the performances 25.30% and 40.90%, we apply a KNN intersection re-ranking (KIRR) which is effective in the condition that the gallery set containing a large number of unrelated persons.

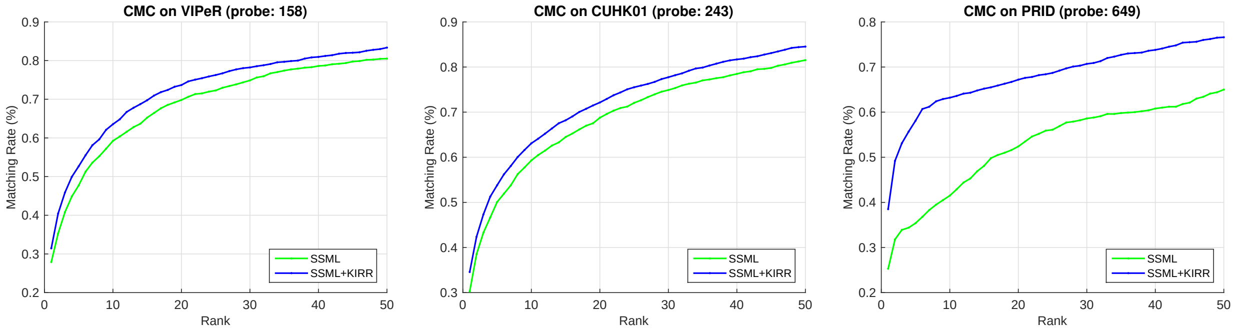
3.2.2. *Combined metric with re-ranking*

The re-ranking experiment on three datasets is presented in Fig. 3 in settings:VIPeR (probe:158, gallery:316); CUHK01 (probe:243, gallery:486); PRID (probe:100, gallery:649). It shows that by applying KIRR, the performance would be improved especially in PRID (from 25.30% to 38.50%), almost close to the performance (40.90%) using the gallery set (100). It validates that KIRR can effectively use the discriminative information from large number of the unrelated persons in the gallery set and reduce the negative impact of the unrelated samples.



(a) (b) (c)

Fig. 2. Experimental results on VIPeR, CUHK01, and PRID in form of cumulative matching characteristic (CMC) curve corresponding to Table 1.



(a) (b) (c)

Fig. 3. Experimental results on VIPeR, CUHK01, and PRID in form of CMC.

3.2.3. Comparison with the state-of-the-art

Table 2. Rank-1 comparisons with state-of-the-arts on VIPeR, CUHK01, and PRID.

Rank-1	VIPeR	CUHK01	PRID(gal: 649)
ISR [14]	27.0	-	17.0
GTS [15]	25.2	-	-
DLILR [16]	29.6	28.4	21.1
UCDTL [17]	**31.5**	27.1	24.2
Ours-SSML	29.91	32.82	25.30
Ours-SSML+KIRR	29.94	**32.82**	**38.50**

Some state-of-the-art unsupervised methods are selected for comparison in Tab. 2. These methods include graphical model based GTS [15], sparse representation classification based ISR [14], dictionary learning with iterative laplacian regularisation based DLILR [16], and unsupervised cross dataset transfer learning UCDTL [17]. Comparing with recent state-of-the-art UCDTL [17], our approach achieves competitive result on VIPeR and better result by a large margin on CUHK01 and PRID. Noting that the proposed approach only uses unlabeled target dataset for training, without other source data involved.

4. CONCLUSION

In this paper, a novel unsupervised approach called sample-specific metric learning (SSML) is proposed, which trains a local metric for the query sample using unlabeled data from the same scene. To enhance the unsupervised metric learning, a KNN intersection re-ranking (KIRR) strategy is utilized to alleviate the open gallery set problem and further improve the performance. Experiment shows that our approach is comparable to the state-of-the-art methods and has great potential be to applied in various real-world scenes.

5. ACKNOWLEDGEMENT

This research is supported in partial by National Nature Science Foundation of China (NSFC) under Grant No. 61401426, 61671427, Beijing Municipal Science and Technology Commission under Grant Z161100001616005, and Science and Technology Innovation Foundation of Chinese Academy of Sciences under Grant CXJJ-16Q218.

6. REFERENCES

[1] Douglas Gray and Hai Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *ECCV*. 2008, pp. 262–275, Springer.

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

[3] Bingpeng Ma, Yu Su, and Frédéric Jurie, "Local descriptors encoded by fisher vectors for person reidentification," in *ECCV, Workshops and Demonstrations*. Springer, 2012, pp. 413–422.

[4] Rui Zhao, Wanli Ouyang, and Xiaogang Wang, "Learning mid-level filters for person re-identification," in *CVPR*. IEEE, 2014, pp. 144–151.

[5] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li, "Person re-identification by local maximal occurrence representation and metric learning," in *CVPR*, 2015, pp. 2197–2206.

[6] Tetsu Matsukawa, Takahiro Okabe, Einoshin Suzuki, and Yoichi Sato, "Hierarchical gaussian descriptor for person re-identification," in *CVPR*, 2016, pp. 1363–1372.

[7] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof, "Large scale metric learning from equivalence constraints," in *CVPR*. IEEE, 2012, pp. 2288–2295.

[8] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang, "Reidentification by relative distance comparison," *PAMI, IEEE Transactions on*, vol. 35, no. 3, pp. 653–668, 2013.

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

[10] Giuseppe Lisanti, Iacopo Masi, and Alberto Del Bimbo, "Matching people across camera views using kernel canonical correlation analysis," in *ICDSC*. ACM, 2014, p. 10.

[11] Martin Hirzer, Peter M Roth, Martin Köstinger, and Horst Bischof, "Relaxed pairwise learned metric for person re-identification," in *ECCV*. 2012, pp. 780–793, Springer.

[12] Yang Yang, Jimei Yang, Junjie Yan, Shengcai Liao, Dong Yi, and Stan Z Li, "Salient color names for person re-identification," in *ECCV*. 2014, pp. 536–551, Springer.

[13] Li Zhang, Tao Xiang, and Shaogang Gong, "Learning a discriminative null space for person re-identification," in *CVPR*, 2016, pp. 1239–1248.

[14] Giuseppe Lisanti, Iacopo Masi, Andrew D Bagdanov, and Alberto Del Bimbo, "Person re-identification by iterative re-weighted sparse ranking," *PAMI, IEEE transactions on*, vol. 37, no. 8, pp. 1629–1642, 2015.

[15] Hanxiao Wang, Shaogang Gong, and Tao Xiang, "Unsupervised learning of generative topic saliency for person re-identification," in *BMVC*, 2014.

[16] Elyor Kodirov, Tao Xiang, and Shaogang Gong, "Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification," in *BMVC*, 2015, vol. 3, p. 8.

[17] Peixi Peng, Tao Xiang, Yaowei Wang, Massimiliano Pontil, Shaogang Gong, Tiejun Huang, and Yonghong Tian, "Unsupervised cross-dataset transfer learning for person re-identification," in *CVPR*, 2016, pp. 1306–1315.

[18] Ethan Fetaya and Shimon Ullman, "Learning local invariant mahalanobis distances.," in *ICML*, 2015, pp. 162–168.

[19] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan, "Object detection with discriminatively trained part-based models," *PAMI, IEEE Transactions on*, vol. 32, no. 9, pp. 1627–1645, 2010.

[20] Douglas Gray, Shane Brennan, and Hai Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *PETS*. Citeseer, 2007, vol. 3.

[21] Wei Li, Rui Zhao, and Xiaogang Wang, "Human reidentification with transferred metric learning.," in *ACCV*, 2012, pp. 31–44.

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

[23] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *CVPR*. IEEE, 2005, vol. 1, pp. 886–893.

[24] Timo Ahonen, Abdenour Hadid, and Matti Pietikainen, "Face description with local binary patterns: Application to face recognition," *PAMI, IEEE transactions on*, vol. 28, no. 12, pp. 2037–2041, 2006.