Person Re-identification by Multiple Feature Representations and Metric Learning

Meibin Qi, Jingxian Han(✉), and Jianguo Jiang

School of Computer and Information, Hefei University of Technology,
Hefei 230009, China
jingxhan@163.com

Abstract. Person re-identification is the problem of matching pedestrian images captured from multiple cameras. Feature representation and metric designing are two critical aspects in person re-identification. In this paper, we first propose an effective Convolutional Neural Network and learn it with mixed datasets as a general deep feature extractor. Secondly, we extract the hand-crafted feature of images as a supplement, then we learn the independent distance metrics for deep feature representation and hand-crafted feature representation, respectively. Finally, we validate our method on three challenging person re-identification datasets, experimental results show the effectiveness of our approach, and we achieve the best rank-1 matching rates on all the three datasets compare with the state-of-the-art methods.

1 Introduction

Person re-identification aims to identify whether two pedestrian images observed from disjoint camera views belong to the same person or not, which has great significance in video surveillance systems. Large variations in viewpoint, illumination and body posture across different camera views can cause a great appearance variance, which makes the re-identification still a challenging problem. Typically, methods for re-identification include two vital steps: (1) developing robust feature representations to handle the variations in pedestrian images; (2) designing discriminative distance metrics to measure the similarity between pedestrian images.

Representative feature descriptors include [1–8], which mostly come from color and texture. Gray *et al.* [1] used boosting to select a subset of optimal features composed by texture and color features; Farenzena *et al.* [2]proposed Symmetry-Driven Accumulation of Local Features (SDALF) consisted of both symmetry and asymmetry color and texture information; Zhao *et al.* [3] learned the mid-level filter (Mid-Level) from patch clusters with coherent appearance obtained by pruning hierarchical clustering trees to get view-invariant and discriminative features; SalMatch [5] was proposed to exploit both patch matching and salience matching for person re-identification, and in [7], Local Maximal Occurrence (LOMO) was proposed, which was extracted from the local HSV

Z. Zhang and K. Huang (Eds.): IVS 2016, CCIS 664, pp. 83–90, 2016.
DOI: 10.1007/978-981-10-3476-3_10

histograms and SILTP features with sliding windows to make a stable representation against viewpoint changes. However, due to the limitations of hand-crafted feature descriptors, it is hard to extract abstract and intrinsic features of the images, which makes these appearance-based features are highly susceptible and difficult to achieve a balance between discriminative power and robustness.

In recent years, many metric learning approaches have been proposed [5,7,9–14] and achieved remarkable performance for person re-identification. Representative methods include Cross-view Quadratic Discriminant Analysis (XQDA) [7], Large Scale Metric Learning from Equivalence Constraint (KISSME) [9], Metric learning to Rank (MLR) [10], Pairwise Constrained Component Analysis (PCCA) [11] and Large Margin Nearest Neighbor (LMNN) [14]. These methods extracted the hand-crafted features first to learn the transformation matrix of the initial feature space, which makes the distance become smaller between the same individuals and larger between different individuals in transformed feature space, some of them achieved impressive improvements for person re-identification.

Compared with the hand-crafted features based methods aforementioned, there are several deep learning based methods have been proposed [15–19]. More abstract and internal features can be learned automatically with the deep architecture, which makes the feature representation rather robust compared with those hand-crafted features. Li *et al.* [15] proposed a novel filter pairing neural network (FPNN) to jointly optimize feature learning, geometric transforms, photometric transforms, misalignment, occlusions and classification. Yi *et al.* [16] used a siamese deep convolutional architecture to learn the texture feature, color feature and metric together in fully cross dataset setting. Ahmed *et al.* [17] presented a deep neural network with layers specially designed for capturing relationships between different camera views. Wu *et al.* [18] used very small convolution filters and increased the depth of the network to improve the performance of re-identification. Xiao *et al.* [19] learned deep feature representation from multiple domains with Convolutional Neural Networks (CNNs). However, these deep neural network need to learn a large number of parameters, small datasets usually can not get remarkable results.

To address these problems, firstly, we learn a general Convolutional Neural Network with the mixture of various datasets as our deep feature extractor, which increases the scale of training set to make small datasets are applicable and enables us learn better features from multiple datasets. Then we extract the appearance-based features of pedestrian images as a supplement. Finally, we learn different metrics for the deep feature representation and hand-crafted feature representation, respectively, which makes the distance metrics more discriminative. Experiments show the superior performance of our proposed approach when compared with the state-of-the-art works.

2 Proposed Approach

In this paper, we extract both deep features and hand-crafted features to represent pedestrian images, and then learn the distance metrics respectively for the two types of feature representations to measure the similarity between different images in a more discriminative way. Section 2.1 introduces the Convolutional Neural Network we proposed to extract the deep feature representation of the images. Section 2.2 introduces the multiple feature representations and our independent metric learning.

2.1 Our Deep Architecture

Inspired by [19,20], we build a CNN model described in Table 1, and mix the various datasets together to train a general CNN as our deep feature extractor for all the datasets. Specifically, three benchmark datasets include VIPER, CUHK01 and CUHK03 are used to validate our method, and all the images are scaled to 144 × 56 pixels.

Table 1. The Architecture of Our Proposed CNN

Name	Patch size/stride	Input size
conv1	$3 \times 3/1$	$144 \times 56 \times 3$
conv2-conv3	$3 \times 3/1$	$144 \times 56 \times 32$
pool3	$2 \times 2/2$	$144 \times 56 \times 32$
Inception 4a,4b	As in Fig. 1(a)	$78 \times 28 \times 32$
Inception 5a,5b	As in Fig. 1(a)	$36 \times 14 \times 384$
Inception 6a,6b	As in Fig. 1(b)	$18 \times 7 \times 786$
Global pool	$9 \times 4/1$	$9 \times 4 \times 1536$
fc7	Logits	$1 \times 1 \times 1536$
fc8	Logits	$1 \times 4 \times 2048$
Softmax	Classifier	$9 \times 4 \times 2168$

The structure of our CNN is the same with [19] expect the last two Inception modules and the two fully connected layers. Figure 1(b) shows the structure of our last two Inception modules, which was applied to image classification in [20], it expanded the filter bank outputs of the original Inception modules in Fig. 1(a) to promote high dimensional representation. After this, two fully connected layers were applied, the first has 2048 channels and the second contains the channels are equaled with the number of the individuals in training set which is set to 2168 in our model.

2.2 Multi-features Fusion and Independent Metric Learning

After trained the proposed CNN, we extract the fc7 layers output as the deep feature representation for the training and testing set, and exploit the handcrafted feature LOMO [7] consisted of local HSV histograms and SILTP features

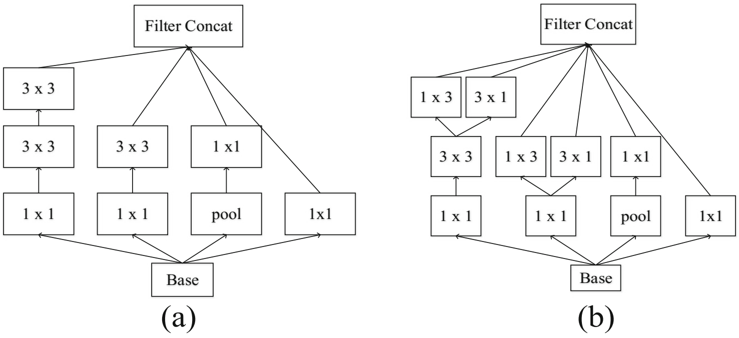


Fig. 1. Inception modules used in our CNN structure, which were all proposed in [20] for image classification, module in (b) is an expanding of (a) to promote high dimensional representations on the coarsest grid

as a complement. Then we learn the independent distance metric with XQDA [7] for the two types of feature representations, respectively.

The XQDA aims to learn a discriminant subspace and an effective distance metric at the same time. Given a pair of images (i, j) captured from different views, $\mathbf{x}_i$ and $\mathbf{y}_j$ are the original features of the images. The distance between image i and j is formulated as:

$$f(\mathbf{x}_i, \mathbf{y}_j) = (\mathbf{x}_i - \mathbf{y}_j)^{\mathrm{T}} \mathbf{W} \mathbf{M} \mathbf{W}^{\mathrm{T}} (\mathbf{x}_i - \mathbf{y}_j) \; . \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{d \times r}$ is the subspace projection matrix, $\mathbf{M} \in \mathbb{R}^{r \times r}$ is the learned metric kernel, d is the dimension of the original feature space, and $r(r < d)$ is the dimension of the transformed feature space.

In this paper, we suppose $\mathbf{x}_i^{dr}$ and $\mathbf{x}_i^{hr}$ are the deep feature representation and hand-crafted feature representation of image i, respectively, $\mathbf{y}_j^{dr}$ and $\mathbf{y}_j^{hr}$ have the similar meanings. The distance between image i and j can be re-formulated as:

$$d(i, j) = d_n(\mathbf{x}_i^{dr}, \mathbf{y}_j^{dr}) + d_n(\mathbf{x}_i^{hr}, \mathbf{y}_j^{hr}) \; . \tag{2}$$

where $d_n(\mathbf{x}_i^{dr}, \mathbf{y}_j^{dr})$ is the normalization of $d(\mathbf{x}_i^{dr}, \mathbf{y}_j^{dr})$, $d_n(\mathbf{x}_i^{hr}, \mathbf{y}_j^{hr})$ is the normalization of $d(\mathbf{x}_i^{hr}, \mathbf{y}_j^{hr})$, which are all calculated by Eq. (1).

3 Experiments

3.1 Datasets and Experiment Protocols

We validate the proposed approach on three widely-used person re-identification datasets include VIPER [21], CUHK01 [22], and CUHK03 [15].

VIPER is one of the most challenging dataset for person re-identification, it contains 632 pairs of person images taken from two camera views with various poses, viewpoints and illumination conditions. The CUHK01 dataset is larger in scale than VIPER, it contains 971 persons captured from two disjoint views and each person has two images in each camera view, camera A captured the frontal or back view of the individuals while camera B captured the side views. And the CUHK03 dataset is one of the most largest published person re-identification datasets, it includes five different pairs of camera views with more than 14,000 images of 1467 pedestrians, in addition, both manually cropped pedestrian images and images automatically detected by the pedestrian detector of [23] are all provided, this is a more realistic setting considering misalignment, occlusions, body part missing and detector errors.

We follow the widely adopted experimental protocols for VIPER and CUHK01 datasets, the individuals in these dataset are randomly divided into half for training and the other half for testing. And for CUHK03, we follow the settings in [19,24], using both manually cropped pedestrian images and images automatically detected together and then randomly select 100 individuals for testing, the other 1367 individuals are used for training. We mix the three selected training sets together to train a general CNN which is employed to extract the deep feature representations of images for various datasets, we use Caffe [25] deep learning framework implement our network. And then we exploit the same individuals used for training our CNN in the three datasets to learn distance metrics for different datasets, respectively. The result is evaluated by cumulative matching characteristic (CMC) curve [26], which is also known as rank-n, an estimate of finding the correct match in the top n match. This procedure is repeated 10 times and the average of rank-n is reported for different dataset.

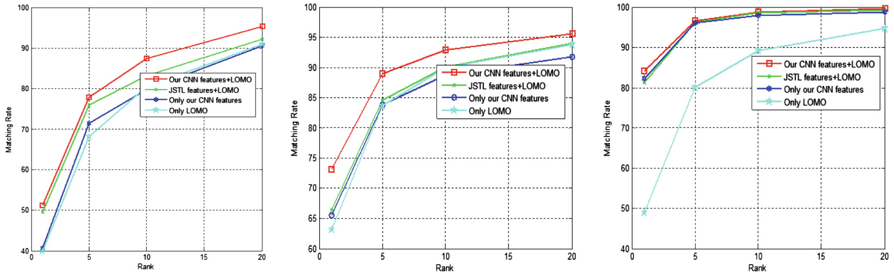
3.2 Evaluations of Proposed Method

In order to validate the effectiveness of the proposed method, here we conduct a series of experiments with different settings to evaluate the effectiveness of our approach, which include: (i) use our proposed method; (ii) replace our CNN with JSTL [19]; (iii) without hand-crafted feature representation; (iv) without our deep feature representation.

Figure 2 shows the rank-$n(n = 1, 5, 10, 20)$ matching rates for different experiments and datasets. Experimental results show the effectiveness of the proposed method, our method achieves the better performance than other compared methods on all the three datasets. The first two experiments validate the effectiveness of our proposed CNN, by expanding filter bank outputs to promote higher dimensional representation, we can achieve a better performance. And the last two experiments validate that the two types of feature representations can complement each other well.

3.3 Comparison with State-of-the-Arts

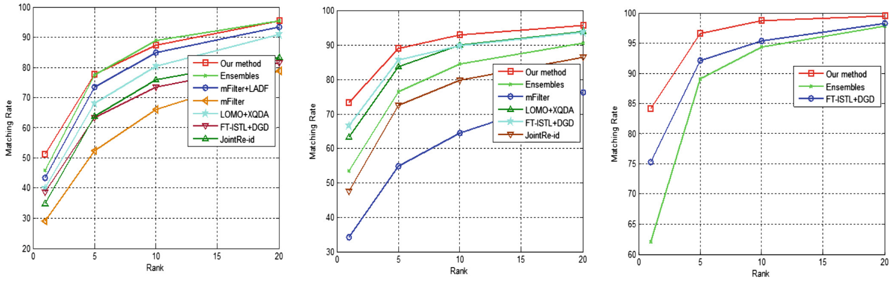
We compare our approach with the following state-of-the-art methods: Metric Ensembles (Ensembles) [24], mFilter+LADF [3], mFilter [3], LOMO+XQDA [7], FT-JSTL+DGD [19] and JointRe-id [17]. Figure 3 shows the results on VIPER,



(a) VIPER (b) CUHK01 (c) CUHK03

Fig. 2. The experimental results of different methods on the three datasets, measured by rank-1, rank-5, rank-10 and rank-20 matching rates. The red curves represent our proposed method, which achieve the best rank-1 matching rates for all the three datasets (Color figure online)

CUHK01 and CUHK03 datasets. Our method improves rank-1 recognition rates by 5.4%, 7.6% and 8.7% on the three datasets compare with the state-of-the-arts.



(a) VIPER (b) CUHK01 (c) CUHK03

Fig. 3. Performance comparison of the proposed method with the state-of-the-arts for VIPER, CUHK01 and CUHK03 datasets. Our approach outperforms all the state-of-the-art methods in most cases, especially on rank-1 matching rate

4 Conclusion

In this paper, we present an effective deep architecture trained with a mixture of various datasets to extract deep features of pedestrian images, then we use the deep feature representation and hand-crafted feature representation to learn different metrics, respectively. By using both deep feature representation and hand-crafted feature representation, we can gain more robust and comprehensive features, and learning independent distance metrics for the two types feature representation can realize a higher discriminative power. We conduct extensive experiments on three widely used person re-identification datasets to validate our approach. Experimental results demonstrate that our method achieves a better performance than other state-of-the-art methods in most cases.

References

1. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5302, pp. 262–275. Springer, Heidelberg (2008). doi:10.1007/978-3-540-88682-2_21
2. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person reidentification by symmetry-driven accumulation of local features. In: Computer Vision and Pattern Recognition (CVPR), vol. 23, pp. 2360–2367 (2010)
3. Zhao, R., Ouyang, W., Wang, X.: Learning mid-level filters for person reidentification. In: CVPR, pp. 144–151 (2014)
4. Hu, Y., Liao, S., Lei, Z., Yi, D., Li, S.Z.: Exploring structural information and fusing multiple features for person re-identification. In: Computer Vision and Pattern Recognition Workshops (CVPRW), vol. 13, pp. 794–799 (2013)
5. Zhao, R., Ouyang, W., Wang, X.: Person re-identification by salience matching. In: Computer Vision (ICCV), pp. 2528–2535 (2013)
6. Ma, B., Su, Y., Jurie, F.: Local descriptors encoded by fisher vectors for person re-identification. In: Fusiello, A., Murino, V., Cucchiara, R. (eds.) ECCV 2012. LNCS, vol. 7583, pp. 413–422. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33863-2_41
7. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: CVPR, vol. 8, pp. 2197–2206 (2015)
8. Ma, B., Su, Y., Jurie, F.: Covariance descriptor based on bio-inspired features for person re-identification and face verification. Image Vis. Comput. **32**, 379–390 (2014)
9. Koestinger, M., Hirzer, M., Wohlhart, P.: Large scale metric learning from equivalence constraints. In: CVPR, pp. 2288–2295 (2012)
10. McFee, B., Lanckriet, G.R.G.: Metric learning to rank. In: International Conference on Machine Learning, pp. 775–782 (2010)
11. Mignon, A., Jurie, F.: PCCA: a new approach for distance learning from sparse pairwise constraints. In: CVPR, vol. 157, pp. 2666–2672 (2012)
12. Hirzer, M., Roth, P.M., Köstinger, M., Bischof, H.: Relaxed pairwise learned metric for person re-identification. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7577, pp. 780–793. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33783-3_56
13. Zheng, W.S., Gong, S., Xiang, T.: Person re-identification by probabilistic relative distance comparison. In: CVPR, vol. 42, pp. 649–656 (2011)
14. Weinberger, K.Q., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. J. Mach. Learn. Res. **10**, 207–244 (2009)
15. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: deep filter pairing neural network for person re-identification. In: CVPR, pp. 152–159 (2014)
16. Yi, D., Lei, Z., Li, S.Z.: Deep metric learning for practical person re-identification. In: ICPR, pp. 34–39 (2014)
17. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: CVPR, pp. 3908–3916 (2015)
18. Wu, L., Shen, C., Hengel, A.V.D.: Personnet: person reidentification with deep convolutional neural networks. arXiv preprint arXiv:1601.07255 (2016)
19. Xiao, T., Li, H., Ouyang, W., Wang, X.: Learning deep feature representations with domain guided dropout for person re-identification. arXiv preprint arXiv:1604.07528 (2016)

20. Szegedy, C., Vanhoucke, V., IoffeSzegedy, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. arXiv preprint arXiv:1512.0056 (2015)
21. Gray, D., Brennan, S., Tao, H.: Evaluating appearance models for recognition, reacquisition, and tracking. In: IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS), vol. 3 (2007)
22. Li, W., Zhao, R., Wang, X.: Human reidentification with transferred metric learning. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012. LNCS, vol. 7724, pp. 31–44. Springer, Heidelberg (2013). doi:10.1007/978-3-642-37331-2_3
23. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. Pattern Anal. Mach. Intell. **32**, 1627–1645 (2010)
24. Paisitkriangkrai, S., Shen, C., Hengel, A.V.D.: Learning to rank in person reidentification with metric ensembles. arXiv preprint arXiv:1503.01543 (2015)
25. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: convolutional architecture for fast feature embedding. In: ACM, pp. 675–678 (2014)
26. Moon, H., Phillips, P.J.: Evaluating appearance models for recognition, reacquisition, and tracking. Perception **30**, 303–321 (2001)