Attribute-Based Fused Feature for Person Re-identification

Qingqing Zhao, Changhong Chen *, Wanru Song and Feng Liu

School of Nanjing University of Posts and Telecommunications, Nanjing, China.

*Corresponding author e-mail: chenchh@njupt.edu.cn

Abstract. In this paper, we propose an attribute-based fused feature which combined low-level and attribute features for person re-identification. Pedestrian parsing and different low-light image enhancement methods are adopted before feature extraction and Random Forest (RF) classifier is used as the basic classifier. Firstly, each of the image is divided into eight strips, and the strips are composed into four parts: head, up-body, low-body and bag areas. Above these four parts, the gradient-LOMO (Local Maximal Occurrence) features are extracted separately to form sub-classifiers features. The sub-classifiers of the independent and different dimensional attributes are respectively trained and then fused to generate a classifier, which describes 21-dimensional attribute features and is combined with the correction mechanism. Then, this classifier can be used to predict attributes on unmarked datasets. In this way, the time complexity of manual marking can be reduced effectively. Finally, the low-level features, which extracted from original and foreground images, are combined with attributes as the fused feature. The results of experiments on three common person re-identification datasets (VIPeR, PRID450S and GRID), indicate that our attribute-based fused feature exhibits prominently performance which outperforms the state-of-the-art methods for person re-identification.

1. Introduction

Most re-identification methods have relied on low-level feature matching. However, low-level feature may not be robust to different view condition, illumination and background environments. Attribute features are sufficiently invariant to view condition and similar to a description offered verbally to a human eye-witness. Appearance-based feature has been widely used, even though no feature can fully represent pedestrians on Person Re-identification. Attribute-based fused feature that mainly used for dealing with various viewpoints, different illuminations and background environments.

Several effective features have been proposed such as LOMO [1], a three-scale pyramid of HSV color histograms and SILTP (scale-invariant LBP) features are extracted from the image processed by Retinex algorithms [2]. GOG proposes a novel region descriptor based on hierarchical Gaussian distribution of pixel features and uses mean color of local parts and covariance descriptors as the representation of the hierarchy [3].

Since 2016, deep learning has been applied in person re-identification [4, 5, 6], and has fused feature extraction and metric learning to a single framework. Deep learning requires a huge amount of data to train, but data in distinguish different persons [7] are small. However, some small person re-identification datasets such as VIPeR dataset also needed to research. This is because the accuracy is still lower than 55% and this datasets focus on background interference and multi-views problems. Low-

level features are mostly based on color histograms and texture information, they describe pixels content but no semantic details. Mid-level attribute features are defined to leveraging semantically for person re-identification [8, 9, 10, 11, 12].

Our contributions can be summarized in the following two aspects: (1) Pre-processing works on datasets, such as the light correction and the remove of background interference, performs robustly on different illuminations and complex background environments. The proposed gradient-LOMO feature describes gradient information and benefits for the attributes of silhouette corresponding features (i.e. bag's silhouette features). (2) The sub-classifiers that correspond to head, up-body, low-body and bag are trained separately. Then, the sub-classifiers are fused into a classifier with correction mechanism for predicting attributes on the unmarked dataset. In this way, the time complexity of manual marking can be reduced effectively. Finally, the low-level features, which extracted from original and foreground images, are combined with attributes as the fused feature.

2. Pre-processing works

Pre-processing works which include Retinex and pedestrian parsing are used in this paper. The pre-processing works are shown in Fig. 1. For each dataset, from left to right, indicates: original image, image after Retinex, image after pedestrian parsing and image with the remove of redundant information.

(a) (b) (c)

Figure 1. The pre-processing works of: (a) GRID dataset. (b) PRID450S dataset. (c) VIPeR dataset.

Due to illumination imbalance on different views and the importance of color information, Retinex algorithm has been applied for datasets. Pedestrian parsing of Deep De-compositional Network, is applied to obtain the foregrounds for each image [13]. For the architecture DDN, the input of which is a feature vector x^c, and the output is a set of label maps of body parts. The first occlusion estimation layer employs the rectified linear function [14]. $\rho(x)$ as the activation function. On the top of DNN, x^c is transformed into several label maps $y_1,...,y_M$ with the corresponding weight matrices $W^{t1},W_1^{t2},...,W_M^{t2}$ and biases $b^{t1},b_1^{t2},...,b_M^{t2}$, label map $y_i \in [0,1]^n$ is formulated by

$$y_i = \tau(W_i^{t2}\rho(W^{t1}+b^{t1})+b_i^{t2}) \tag{1}$$

Where, $y_{ij}=0$ indicates the pixel belongs to the background, otherwise the pixel belongs to body.

3. The proposed method

Inspired by [1, 3], our feature describes gradient information and simplifies the structure. We formulate the fused features as

$$\begin{aligned} F_{fusion} &= w_1 \times W_{pre-acc} \times f_{attr} + w_2 \times (f_{low-ori} + f_{low-processed}) \\ s.t. \quad & w_1 + w_2 = 1 \end{aligned} \tag{2}$$

Where, w_1 is the weight of attribute feature, and w_2 is the weight of low-level feature which called gradient-LOMO, $f_{attr}, f_{low-ori}$ are separately attribute feature and the low-level feature from original or

processed foreground images, $W_{pre-acc}$ means the weight from Random Forest (RF) classifier on annotated dataset.

The gradient-LOMO feature contains of HSV color space, SILTP texture information and gradient information M_θ. Gradient orientation $O = \arctan(I_y / I_x)$ is calculated from x and y derivatives I_x, I_y of intensity I, the orientation is quantized into four bins: $M_{\theta \in \{0^\circ,\ldots,270^\circ\}}$.

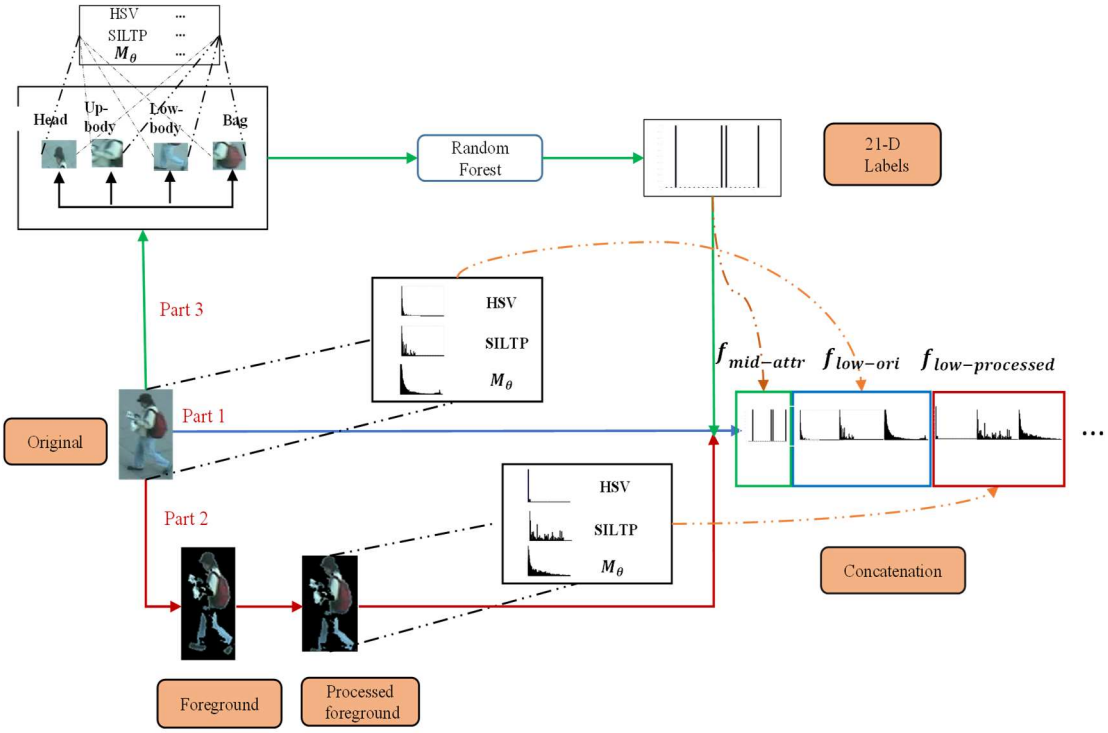

Figure 2. The entire structure of fused features.

The entire structure of fused features is shown in Fig. 2. As shown in Part 1 and Part 2, gradient-LOMO feature is extracted separately from original images and processed images. During the process of Part 2, pedestrian parsing and the remove of redundancy information have been applied to the fusion feature.

Images are resized into 128×48 and divided into 8 strips. Then, the strips are composed into four parts: head (the first two strips), up-body (the next two strips), low-body (the last four strips) and bag (the middle four strips) areas. The head sub-classifier contains attributes of male, mid-hair, dark-hair and bald. Red-shirt, light-shirt, dark-shirt, green-shirt, no-coats and patterned are the attributes of up-body. Not-light-jeans-color, dark-bottoms, light-bottoms, bare legs, shorts, jeans and skirts are belongs to the low-body attributes. Has-hand-bag-carrier-bag, has-backpack, has-satchel are attributes of bag. The above procedures are shown in Fig. 3. The gradient-LOMO feature is extracted from four parts shown in Part3. Four sub-classifiers that based on Random Forest (RF) classifier are trained on annotated VIPeR and PRID2011 datasets and then combined into a classifier. The classifier can be used to predict the attributes of the unmarked datasets (PRID450S and the GRID). The parameters of RF classifier are set to $ntrees = 100$ and the feature selection equals to the square of gradient-LOMO.

Figure 3. The generation details of strips and four areas.

Constrains of attributes can be formulated in (3), where $l=1$ indicates that the attribute exits.

$$\begin{aligned} s.t. & \; l_{red-shirt} + l_{blue-shirt} + l_{green-shirt} = 1 \\ & l_{jeans} + l_{not-light-dark-jeans-color} = 1 \\ & l_{mid-hair} + l_{dark-hair} + l_{bald} = 1 \\ & l_{light-bottoms} + l_{dark-bottoms} = 1 \\ & l_{jeans} + l_{skirts} + l_{shorts} = 1 \\ & l_{light-shirt} + l_{dark-shirt} = 1 \\ & l_{jeans} + l_{bare-legs} = 1 \\ & if \quad l_{male} = 1, l_{skirts} \neq 1 \end{aligned} \tag{3}$$

4. Experimental results

VIPeR [1], PRID450S [15], and GRID [16] are considered in this paper and shown in Fig. 4. Cross-view quadratic discriminant analysis (XQDA), which is adopted as the metric learning method.

(a) (b) (c)

Figure 4. Some example image pairs from VIPeR, PRID450S and GRID datasets. Images in the same column are captured from the same person.

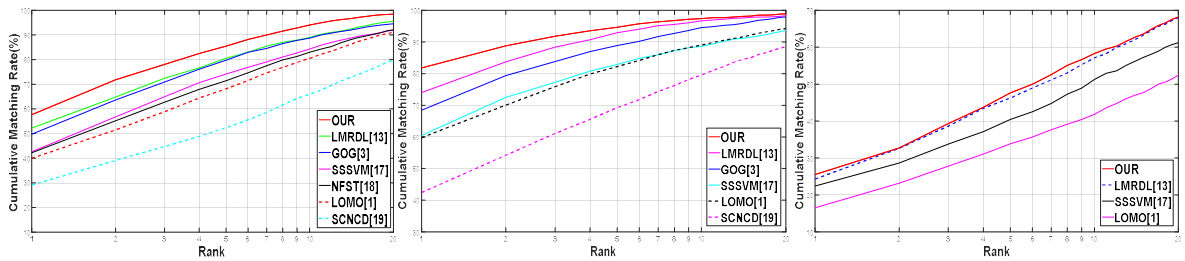

Figure 5. Comparison of CMC curves and rank1 matching rates on datasets.

Table 1-3 shows the comparison of rank-r matching rate on three datasets.

4.1. Experiments on VIPeR

VIPeR is a challenging person re-identification database containing 632 person image pairs from two camera views. The sets on this dataset is to divide the 632 pairs of images into half for training and the other half for testing.

Table 1. Comparison of top-r cumulative matching rates (%) on VIPeR dataset.

Method	Reference	r = 1	r = 5	r = 10	r = 20
Our	Proposed	57.72	85.41	93.96	98.45
LMRDL	2017[13]	52.18	80.54	88.92	95.60
MPCNN	2016[16]	47.80	74.70	84.80	91.10
SSSVM	2016[17]	42.66	74.21	84.27	91.93
NFST	2016[18]	42.28	71.46	82.94	92.06
GOG	2016[3]	49.70	79.70	88.70	94.50
LOMO	2015[1]	40.00	68.13	80.50	91.10
SCNCD	2014[19]	37.80	68.50	81.20	90.40

Our method contains more pre-processing procedure and attribute information than methods in Table 1. Compared with our method, LOMO [1] contains less gradient information and GOG [3] is too complicated. Differs from the classifier of SSSVM [19], RF classifier is superior in accuracy among current algorithms.

4.2. Experiments on PRID450S.

The PRID450S dataset contains 450 persons with 900 image pairs captured from two disjoint camera views, each image for each person in one camera view. Serious viewpoint changes, partial occlusion and background interference are the obviously challenges for person re-identification. Pedestrians both from probe and gallery datasets are all un-annotated.

Table 2. Comparison of top-r cumulative matching rates (%) on PRID450s dataset.

Method	Reference	r = 1	r = 5	r = 10	r = 20
Our	Proposed	82.18	95.02	98.36	99.02
LMRDL	2017[13]	74.04	92.93	96.62	98.22
GOG	2016[3]	68.40	88.80	94.50	97.80
SSSVM	2016[17]	60.49	82.93	88.58	93.60
LOMO	2015[1]	59.78	82.22	89.02	94.22
SCNCD	2014[19]	42.44	69.22	79.56	88.44

Five state-of-the-art approaches, including XQDA [1], GOG [3], LMRDL [17], SSSVM [19] and SCNCD [21] are compared here. SCNCD describes more color information, however overlooks gradient information. Unlike the other two datasets, this dataset uses its own segmentation template.

4.3. *Experiments on GRID.*

The GRID dataset was captured in a crowded underground station with eight disjoint cameras. There are 250 persons that each has one pair images captured from different views. Besides, there are 775 extra

images that do not belong to any of the 250 persons. Pedestrians both from probe and gallery datasets are all un-annotated.

Table 3. Comparison of top-r cumulative matching rates (%) on GRID dataset.

Method	Reference	r = 1	r = 5	r = 10	r = 20
Our	Proposed	25.52	47.60	58.32	68.16
LMRDL	2017[13]	24.32	46.24	57.12	68.08
SSSVM	2016[17]	22.40	40.40	51.28	61.20
LOMO	2015[1]	16.56	33.84	41.86	52.40

Compared with five state-of-the-art approaches, including XQDA [1], LMRDL [17], and SSSVM [19], it is obviously that our method outperforms all the previous results, achieving 25.52% rank-1 matching rate.

5. Conclusion

In this paper, we propose an effective attribute-based fused feature for person re-identification, which is proved to be robust against viewpoint changes, background clutter and illumination variations. Considering the characteristics of datasets, the decreasing of non-uniform illumination and the parsing of foregrounds methods have been adopted in pre-processing work. Four sub-classifiers that based on Random Forest (RF) classifier are trained on annotated datasets and then combined into a 21-dimension classifier with correction mechanism. The classifier can be used to predict the attributes of the unmarked datasets. The proposed attributed-based fused descriptor contains attributes features and low-level features that obtained from original and processed foreground images. Experiments indicate that the proposed method improves the state-of-the-art rank-1 identification rates on the three datasets respectively.

Acknowledgments

This work was financially supported by the Natural Science Foundation of China No. 61471201.

References

[1] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, Person re-identification by Local Maximal Occurrence representation and metric learning, in Proc. of Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, pp. 2197–2206, January,2015.

[2] D. J. Jobson, Z.-u. Rahman, and G. A. Woodell, A multiscale retinex for bridging the gap between color images and the human observation of scenes, in Proc. of IEEE Transactions on Image Processing, vol. 6, no. 7, pp. 965–976, July, 1997.

[3] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, Hierarchical Gaussian Descriptor for Person Re-identification, in Proc. of Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, pp.1363-1372, January 27-30,2016.

[4] D. Yi, Z. Lei, S. Liao, and S. Li, Deep Metric Learning for Person Re-identification, in Proc. of 22nd International Conference on Pattern Recognition, Swedish Soc Automated Image Anal, Stockholm, SWEDEN, pp.34-39, August 24-28,2014.

[5] W. Li, R. Zhao, T. Xiao, and X. Wang, Deep Re-ID: Deep Filter Pairing Neural Network for Person Re-identification, in Proc. of Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, pp.152-159, June 23-28, 2014.

[6] E. Ahmed, M. Jones, and T. K. Marks, An improved deep learning architecture for person re-identification, in Proc. of Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, pp.3908-3916, June 07-12,2015.

[7] Lin Y, Zheng L, Zheng Z, et al, Improving Person Re-identification by Attribute and Identity Learning, in Proc. of Computer Vision and Pattern Recognition, Puerto Rico, April, 2017.

[8] Layne, R., Hospedales, T.M., Gong, S, Person re-identification by attributes, in Proc. of 23rd British Machine Vision Conference, University of Surrey, Guildford, England, September 03-07, 2012. DOI: 10.5244/C.26.24.
[9] Layne, R., Hospedales, T.M., Gong, S, Towards person identification and re-identification with attributes, in Proc. of 12th European Conference on Computer Vision (ECCV), Florence, Italy, vol.7538, pp.402-412, October 07-13, 2012.
[10] Layne R, Hospedales T M, Gong S, Attributes-Based Person Re-Identification, in Proc. of Advances in Computer Vision and Pattern Recognition, London, January, pp.93-117, 2014.
[11] Jie S, Person re-identification based on bag of words and attributes, Dalian: Dalian University of Technology, 2016.
[12] Prosser, B., Zheng, W.S., Gong, S., Xiang, T, Person re-identification by support vector ranking, in Proc. of British Machine Vision Conference (BMVC), Aberystwyth, UK, vol.42,no.7, pp.1-11, August 31–September 3, 2010.
[13] Luo, P., Wang, X., Tang, X, Pedestrian Parsing via Deep Decompositional Network, in Proc. of IEEE International Conference on Computer Vision (ICCV), pp. 2648–2655, 2013.
[14] V. Nair and G. E. Hinton, Rectified linear units improve restricted boltzmann machines, in Proc. of International Conference on International Conference on Machine Learning (ICML), Omnipress, pp.807-814, 2010.
[15] P. M. Roth, M. Hirzer, M. Kostinger, C. Beleznai, and H. Bischof, Mahalanobis distance learning for person reidentification, in Proc. of Person Re-Identification Advanced in Computer Vision and Pattern Recognition, Spring, Heidelberg, pp.247–267, 2014.
[16] C. C. Loy, T. Xiang, and S. Gong, Time-Delayed Correlation Analysis for Multi-Camera Activity Understanding, in Proc. of International Journal of Computer Vision(IJCV), United States, vol.90, no.1, pp.106–129, 2010.
[17] Dong H, Gong S, Liu C, et al, Large margin relative distance learning for person re-identification, in IET Computer Vision, vol.11, no.6, pp.455-462, September 11, 2017.
[18] Cheng, D., Gong, Y., Zhou, S., et al. Person Re-identification by Multi-channel Parts-Based CNN with Improved Triplet Loss Function, in Proc. Of Computer Vision and Pattern Recognition, Las Vegas, NV, USA, pp.1335–1344, December 12, 2016.
[19] Zhang, Y., Li, B., Lu, H., et al. Sample-Specific SVM Learning for Person Re-identification, in Proc. of Computer Vision and Pattern Recognition, Las Vegas, NV, USA, pp.1278–1287, December 12, 2016.
[20] Zhang, L., Xiang, T., Gong, S, Learning a Discriminative Null Space for Person Re-identification, in Proc. of Computer Vision and Pattern Recognition, Las Vegas, NV, USA, pp.1239–1248, December 12, 2016.
[21] Yang, Y., Yang, J., Yan, J., et al, Salient color names for Person Re-identification, in Proc. of European Conf. on Computer Vision Spring Cham, vol. 8689, pp. 536–551, 2014.