Manifold Ranking Weighted Local Maximal Occurrence Descriptor for Person Re-identification

Foqin Wang[1], Xuehan Zhang[1], Jinxin Ma[2], Jin Tang[1], and Aihua Zheng* [1]

[1]*Anhui University, Hefei, China*
[2]*University of Greenwich, London, United Kingdom*

***Abstract*—Person re-identification is an important task of matching pedestrians across non-overlapping camera views. In this paper, we exploit a weighted feature descriptor for person re-identification. We firstly compute the weights on the superpixel level via graph-based manifold ranking algorithm, then integrate the computed weights into a patch-based feature descriptor, named local maximal occurrence. Finally, the weighted descriptors are fed into a top-push distance learning to mitigate the cross-view gaps. We evaluate the proposed method on three benchmark datasets iLIDS-VID, PRID 450S and VIPeR. The promising experimental results demonstrate the effectiveness of the proposed method comparing with the state-of-the-arts.**

keywords: Person re-identification; Manifold ranking; Local maximal occurrence; Weighted descriptor.

I. Introduction

Person re-identification is defined as the process of determining whether a given individual under one camera has already appeared under other cameras. It is a very important fundamental process in smart surveillance. In recent years, although a host of researchers have proposed many state-of-the-art methods for person re-identification, it's still a challenging task due to various problems, such as illumination changes [23], object scale differences, imaging angle difference and partial occlusions.

Generally speaking, there are two steps for person re-identification: 1) Appearance modeling, to describe the person image with the features that maintaining strong invariance for the same person, and clear distinction among the different person. 2) Learning method, usually using the metric learning to train a distance measurement or classifier, then solving the person re-identification problem by minimizing/maxmizing the distance among the samples of same person/different person.

Although the metric learning scheme can mitigate the cross view gaps somehow, the appearance model is the preliminary problem which brings the way toward the success of person re-identification. Retrospectively, Gray et al. [4] proposed the ensemble of localized features to solve viewpoint invariant for person re-identification. Farenzena et al. [7] proposed the symmetry-driven accumulation of local features with the weighted HSV histogram, MSCR and RHSP. Prosser et al. [9] learned the feature weights in view of an entirety with RankSVM specific metric for pedestrian query settings. D Kouno et al. [22] proposed utilizing depth information for the problem based on an image from an overhead camera by decreasing the influence of occluded images. Fergnani et al. [13] exploited the unartificial proportions of person body composed by a division of body parts, which only relies on a real-time estimation of facial symbol and not requires complex transmutable part models.

Apparently, the weights of the pixels with person body are supposed to be larger than the background region, Zhao et al. [10] exploited the patch-based saliency indication to match persons with similar feature. In addition, they further proposed an unsupervised learning approach [11] to build reliable corresponding relationship between image pairs by learning localized saliency. Rui Zhao etc. [12] combined human salience feature and SDALF method to match identical individual. Ma C et al. [14] proposed a method pretreating samples for person re-identification task, based on the saliency map improved using a bottom-up saliency approach.

However, most of previous appearance-based methods are greatly limited for lack of foreground object priori information. In this paper, we proposed a novel framework using the manifold ranking algorithm to rank the pixels in the bounding boxes of person image with both foreground and background cues but in a different manner. The main contribution of this paper can be summarised as:

- we propose a novel weighted feature descriptor via graph based manifold ranking. Specifically, a close-loop graph is constructed on the superpixel nodes, which is further ranked by the weights based on their distance to the background and foreground cues. We further integrate the superpixel weights into a patch based feature descriptor, local maximal occurrence, to construct the weighted feature descriptor for person re-identification.
- Extensive experiments on benchmark datasets demonstrate the promising performance of proposed weighted feature descriptor. It can also be integrated to all the existing pixel, superpixel or patch based appearance models.

II. Manifold Ranking Weighted Descriptor

In order to preserve the boundary of the person in the images, we firstly generate the superpixels of the image via the state-of-the-art SLIC method [1]. The manifold ranking algorithm is then designed based on the superpixel level.

* Corresponding author

978-1-5090-5756-6/17/$31.00 ©2017 IEEE
SERA 2017, June 7-9, 2017, London, UK

A. Weight Computation via manifold ranking

Given an image $\mathbf{S} = \{s^1, s^2, ..., s^n\}$, where s^i denotes the i-th superpixel of the image, we construct a graph $G = (V, E)$, where $\mathbf{V}$ is the node (superpixel) set and $\mathbf{E}$ is the set of undirected edges. Specifically, the superpixel node is connected to both neighbor superpixels and the secondary neighbor superpixels sharing the common superpixel border. Furthermore, the superpixel nodes on the boundary of image $\mathbf{S}$ are also connected. Based on this k-regular graph, the affinity matrix $\mathbf{W^e} = [\mathbf{w^e_{ij}}]$ which reflects the weights between each pair of the nodes can be defined as:

$$w^e_{ij} = \begin{cases} e^{-\frac{\|c_i - c_j\|}{\sigma^2}}, & if\ s^i\ and\ s^j\ are\ connected, \\ 0, & otherwise. \end{cases} \quad (1)$$

where c_i and c_j are the mean color values of superpixels s^i and s^j respectively. The constant σ, which is set as 0.85 in this paper, controls the intensity of $\mathbf{W^e}$.

The task of graph-based manifold ranking is to rank the nodes on the basis of their correlations to the given query node. According to [24], the ranking function with unormalized Laplacian matrix can be defined as:

$$\mathbf{F} = (\mathbf{D} - \beta \mathbf{W^e})^{-1} y \quad (2)$$

where $\mathbf{D} = diag\{d_{11}, ..., d_{ii}, ..., d_{nn}\}$ is the degree matrix described as $d_{ii} = \Sigma_i w^e_{ij}$. $\mathbf{y} = [y_1, y_2, ..., y_n]^\top$ is a indicator vector, defined as:

$$y_i = \begin{cases} 1, & s^i\ is\ a\ query, \\ 0, & otherwise. \end{cases} \quad (3)$$

By given some superpixels as the queries, the objective of Eq. 2 is to rank the remaining superpixels according to their similarities to the queries. Since the position of foreground is difficult to confirm, two-step bottom-up strategy is designed to rank the superpixels.

Step 1. Obtaining foreground queries by boundary background queries. It's noted that the superpixels on the boundary of the image are with much higher possibility as the background [3], [2]. Therefore, we use the four partial boundaries (top, bottom, left and right) as queries to rank the rest of the superpixels in the image. Specifically, given y^t as the indicator matrix of the top boundary, the ranks of remnant unlabelled superpixels F^t can be calculated according to their similarities to the top boundary queries via Eq. 2. Then the corresponding foreground queries can be obtained by:

$$R^t = A - \tilde{F}^t \qquad A = [1, 1, ..., 1]^\top \in R^{n \times 1} \quad (4)$$

where $\tilde{F}^t$ is the normalization of F^t. R^b, R^l and R^r are obtained in the same manner based on the bottom, left and right boundaries respectively. The finally foreground queries can be obtained via:

$$R_{bq} = R^t \odot R^b \odot R^l \odot R^r \quad (5)$$

where $\odot$ denotes dot product symbol.

Step 2. Ranking computation via foreground queries. In order to produce the binary indicator vector y^{fg} for the obtained foreground queries R_{bq}, we use the mean of R_{bq} as the threshold to binarize R_{bq}, which can also guarantee the number of foreground queries. The final foreground ranking F^{fg} is obtained through substituting y by y^{fg} in Eq. 2. The normalized foreground ranking $\tilde{F}^{fg}$ indicate the weights of the superpixels. The larger ranking/weight, the higher probability of the corresponding superpixel to be the foreground/person body.

Fig. 1 demonstrates several visualized examples of the weight computation via manifold ranking. From which we can see, the weights can further enhance the human bodies on the person images.

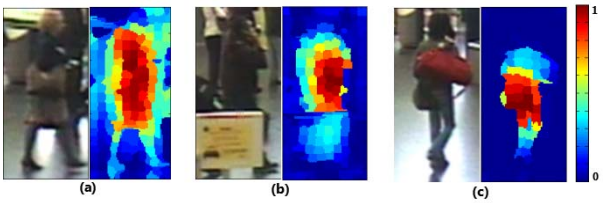

Fig. 1. The superpixels weight distribution of the manifold ranking.

B. Weighted feature Description

We integrate the obtained weights into the patch based local maximal occurrence [16] to construct the weighted feature descriptor. Specifically, we equalize all the pixel weights in the same superpixel. A subwindow with size of 10×10, overlapping step of 5 pixels is slided horizontally and vertically on the Retinexed image [17]. Then the $8 \times 8 \times 8$ HSV histogram and two scales of SILTP histograms [18] ($SILTP^{0.3}_{4,3}$ and $SILTP^{0.3}_{4,5}$) are extracted for each subwindow, where the weight of the subwindow is calculated by averaging the weights of all the pixels covered by the subwindow. To deal with the viewpoint changes, the maximal local occurrence of the same histogram bin is obtained as the horizontal histogram/feature and the feature of each image is obtained by aligning all the horizontal groups(the number of vertical subwindows). Furthermore, the multi-scale information was considered by three-scale pyramid representation with two 2×2 local average pooling operations. The final feature vector for each image has ($8 \times 8 \times 8$ color bins + $3^4 \times 2$ SILTP bins) $\times$ (24+11+5 horizontal groups) = 26,960 dimensions for a 128×64 image. The log transform is employed to suppress large bin values, and normalize both HSV and SILTP features to unit length.

III. Top-push Distance Learning

After obtaining the weighted feature description, a metric learning is required to mitigate the cross view gap. In this paper, we employ the top-push distance learning [15].For the training set $\mathbf{X} = \{(x^a_i, x^b_i)\}^m_{i=1}$, where $x^a_i \in \mathbb{R}^d$ denotes the feature vector of the ith person in camera a, we denote $\mathcal{D}(x^a_i, x^b_j)$ as the distance between feature vectors x^a_i and x^b_j. In person re-id, we always expect that the distance between image pairs of the same person should be smaller than that of

the different person. Hence, for each example x_i^a, our objective is to optimize following program:

$$\mathcal{D}(x_i^a, x_i^b) + \rho < \min \mathcal{D}(x_i^a, x_k^b), \tag{6}$$

where ρ is a relaxing parameter. To quantify the above program, we aim to minimize a hinge loss function:

$$\min \sum_{x_i^a, x_j^b} \max\{\mathcal{D}(x_i^a, x_i^b) - \min \mathcal{D}(x_i^a, x_k^b) + \rho, 0\}. \tag{7}$$

In order to strengthen the correlation of samples of positive pairs, the distance between samples of the same class is further integrated into the objective function:

$$\begin{aligned} f(D) &= (1-\alpha) \sum_{x_i^a, x_i^b} \mathcal{D}(x_i^a, x_i^b) \\ &+ \alpha \sum_{x_i^a, x_i^b} \max\{\mathcal{D}(x_i^a, x_i^b) - \min \mathcal{D}(x_i^a, x_k^b) + \rho, 0\}, \end{aligned} \tag{8}$$

where $\alpha \in [0, 1]$ to balance the penalizes of the large distances between positive pairs and the small distances between closest samples. Besides, we specially consider the optimization of Mahalanobis distance under Criterion 4:

$$\mathcal{D}(x_i^a, x_i^b) = (x_i^a - x_i^b)^T \mathbf{M} (x_i^a - x_i^b), \tag{9}$$

where $\mathbf{M} \succeq \mathbf{0}$ is a positive semi-definite matrix. The detailed optimization please refer to [15]. The final distance between person i in camera a and person j in camera b is obtained by:

$$D(x_i^a, x_j^b) = \min_j \{\mathcal{D}(x_i^a, x_j^b)\}, \tag{10}$$

IV. Experimental Results

We evaluate our method on the benchmark datasets iLIDS-VID [5], Prid 450S [6] and VIPeR [4]. In our experiment, one half of image pairs are randomly selected for training and testing. The performance is evaluated by using the widely-use Cumulative Matching Characteristics (CMC) curve, which represents the expectation of finding the correct match pair in the top k matches. All the experimental results are based on 10 random trials.

iLIDS-VID [5]. The i-LIDS dataset contains 300 indoor pedestrians images from two different camera views in a busy airport arrival hall. It is very challenging due to clothing similarities among people, lighting and viewpoint variations across camera views, cluttered background and random occlusions.

PRID 450S [6]. It contains 450 image pairs observed from two disjoint cameras with viewpoint changes, background interference and partial occlusion. It is also a challenging person re-identification dataset due to the background interference, partial occlusion and viewpoint changes.

VIPeR [4]. The VIPeR dataset is a challenging person re-identification database that has been widely used for benchmark evaluation. It contains of 632 observed from two different non-overlapping camera views. Each person has one image pair. Images in VIPeR contains large variations in background, illumination, and viewpoint.

A. Evaluation on Benchmarks

We evaluate our manifold ranking weighted descriptor (M-RWD) with three state-of-the-art feature descriptors ELF (the Ensemble of Localized Features) [4], HistLBP (Local Binary Patterns Histograms) [20] and CN_color (sRGB values to probabilities over Color Names) [19]. The parameter setting of our algorithm is : $\sigma = 10$ in Eq. 1, $\beta = 0.75$ in Eq. 2 and $\alpha = 0.2$ in Eq. 8. The comparison results are reported in Table I. As we can see, our proposed descriptor significantly ourperforms the other three descriptors. Specifically, the Rank 1 matching rates of ours can achieve 20.33%, 58.47% and 38.76% for iLIDS-VID, PRID 450S and VIPeR respectively, which demonstrate the promising performance of the manifold ranking weighted descriptor.

B. Component Analysis

In order to demonstrate the weight contribution of the manifold ranking, we further evaluate the local maximal occurrence descriptor without the weight integration. The comparison results on iLIDS-VID is demonstrated in Fig. 2. From which we can see, the matching rate can be further improved by integrating the weights.

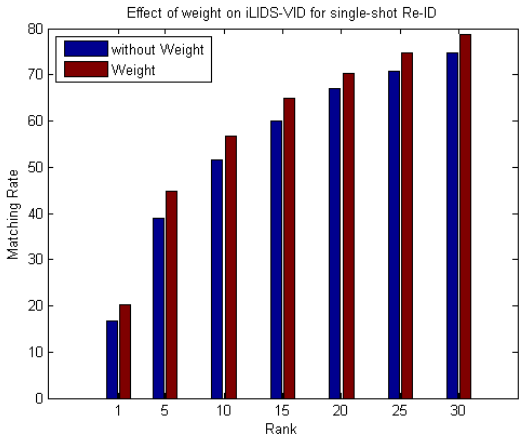

Fig. 2. The comparison results on iLIDS-VID dataset.

V. Conclusion

In this paper, we have proposed a weighted descriptor for person re-identification via manifold ranking on a graph for person images. It incorporates both background and foreground cues to generate the weight maps on superpixel level. We have further integrated the weight maps into the patch-based local maximal occurrence feature to construct the weighted feature descriptor. The evaluations on three benchmark datasets demonstrated the performance of our method. Our future work will focus on extend our method into multi-shot person re-identification.

VI. Acknowledgement

This study was funded by the National Nature Science Foundation of China (61502006, 61602006, 61472002) and the Natural Science Foundation of Anhui Province(1508085QF127).

TABLE I
EXPERIMENTAL RESULTS COMPARISON ON THE THREE DATASETS

	iLIDS-VID				PRID 450S				VIPeR			
Feature	Rank1	Rank5	Rank10	Rank20	Rank1	Rank5	Rank10	Rank20	Rank1	Rank5	Rank10	Rank20
ELF [4]	8.1	20.9	31.7	47.1	17.2	38.3	50.7	62.4	21.0	45.5	60.2	75.4
HistLBP [20]	7.9	21.0	30.0	43.9	18.2	40.0	52.1	65.6	23.0	53.2	67.0	82.1
CN_color [19]	5.0	17.3	26.2	38.0	10.2	25.3	35.0	45.1	24.1	47.2	61.3	77.8
MRWD(**OURS**)	**20.3**	**44.9**	**56.7**	**70.4**	**57.2**	**80.9**	**88.4**	**93.7**	**38.5**	**71.7**	**84.9**	**95.3**

REFERENCES

[1] Achanta R, Shaji A, Smith K, et al. SLIC superpixels[J]. Epfl, 2010.

[2] Lee D Y, Sim J Y, Kim C S. Visual Tracking Using Pertinent Patch Selection and Masking[C]// Computer Vision and Pattern Recognition. IEEE, 2014:3486-3493.

[3] Itti L, Koch C, Niebur E. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1998, 20(11):1254-1259.

[4] Gray D, Tao H. Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features[C]// Computer Vision - ECCV 2008, European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings. DBLP, 2008:262-275.

[5] Wang T, Gong S, Zhu X, et al. Person Re-Identification by Discriminative Selection in Video Ranking.[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2016, 38(12):1-1.

[6] Hirzer M, Beleznai C, Roth P M, et al. Person re-identification by descriptive and discriminative classification[C]// Scandinavian Conference on Image Analysis. Springer-Verlag, 2011:91-102.

[7] Farenzena M, Bazzani L, Perina A, et al. Person re-identification by symmetry-driven accumulation of local features[C]// Computer Vision and Pattern Recognition. IEEE, 2010:2360-2367.

[8] Li W, Zhao R, Wang X. Human reidentification with transferred metric learning[C]// Asian Conference on Computer Vision. Springer-Verlag, 2012:31-44.

[9] Engel C, Baumgartner P, Holzmann M, et al. Person Re-Identification by Support Vector Ranking.[C]// British Machine Vision Conference, BMVC 2010, Aberystwyth, UK, August 31 - September 3, 2010. Proceedings. DBLP, 2010:1-11.

[10] Zhao R, Ouyang W, Wang X. Person Re-identification by Salience Matching[C]// IEEE International Conference on Computer Vision. IEEE, 2013:2528-2535.

[11] Wang H, Gong S, Xiang T. Unsupervised Learning of Generative Topic Saliency for Person Re-identification[J]. British Machine Vision Association Bmva, 2014.

[12] Zhao R, Ouyang W, Wang X. Unsupervised Salience Learning for Person Re-identification[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2013:3586-3593.

[13] Fergnani F, Alletto S, Serra G, et al. Body Part Based Re-Identification from an Egocentric Perspective[C]// The Workshop on Egocentric. 2016:355-360.

[14] Ma C, Miao Z, Li M. Saliency preprocessing for person re-identification images[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2016:1941-1945.

[15] You J, Wu A, Li X, et al. Top-push video-based person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 1345-1353.

[16] Liao S, Hu Y, Zhu X, et al. Person re-identification by local maximal occurrence representation and metric learning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 2197-2206.

[17] Jobson D J, Rahman Z, Woodell G A. A multiscale retinex for bridging the gap between color images and the human observation of scenes[J]. IEEE Transactions on Image processing, 1997, 6(7): 965-976.

[18] Liao S, Zhao G, Kellokumpu V, et al. Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes[C]//Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010: 1301-1306.

[19] Zheng L, Shen L, Tian L, et al. Scalable Person Re-identification: A Benchmark[C]// IEEE International Conference on Computer Vision. IEEE, 2015:1116-1124.

[20] Xiong F, Gou M, Camps O, et al. Person Re-Identification Using Kernel-Based Metric Learning Methods[J]. Lecture Notes in Computer Science, 2014, 8695:1-16.

[21] Li R, Fang L. Cluster Sensing Superpixel and Grouping[C]// IEEE Conference on Computer Vision and Pattern Recognition Workshops. IEEE Computer Society, 2016:1350-1358.

[22] Kouno D, Shimada K, Endo T. Person Identification Using Top-View Image with Depth Information[C]// Acis International Conference on Software Engineering, Artificial Intelligence, NETWORKING and Parallel & Distributed Computing. IEEE, 2012:140-145.

[23] Ishida S, Fukui S, Iwahori Y, et al. Construction of Shadow Model by Robust Features to Illumination Changes[J]. International Journal of Software Innovation, 2015, 1(4):45-55.

[24] Yang C, Zhang L, Lu H, et al. Saliency detection via graph-based manifold ranking[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2013: 3166-3173.