Person Re-Identification via Region-of-Interest based Features

Jianlou Si, Honggang Zhang, Chun-Guang Li
School of Communication and Information Engineering, Beijing University of Posts and Telecommunications
Beijing 100876, PR China
{sijianlou, zhhg, lichunguang}@bupt.edu.cn

***Abstract*—Person re-identification is still a challenging task due to large visual appearance variations caused by illumination, background, viewpoints and poses in multi-camera surveillance. To address these challenges, many methods have been proposed. In this paper, we present an efficient method, called Region-of-Interest based Features (ROIF), via combining textural and chromatic features. It consists of two main phases — region-of-interest exploration from image and features extraction from ROI. Experimental results on the database VIPeR show that our method can yield promising accuracy with a quite cheap time cost.**

***Index Terms*—Color, Multi-camera, Person re-identification, Region-of-Interest, Texture**

I. Introduction

With multi-camera surveillance application becoming more and more popular, person detection and tracking has extended from single camera to multiple cameras [1], [2], [3], [4], [5], [6], [7]. In this context, person re-identification across cameras at different locations and different time has received a lot of attention in recent years [8], [9], [10], [11], [12], [13], [14].

In surveillance scenario, images are captured from a complicated environment. For the sake of real-time monitoring and long-time running, the installed cameras have a low resolution (see Fig. 1). In general, it is assumed that the same people wear the same clothes between different cameras. Consequently, color features can be used to identify the individual. However, under the uncontrolled illumination and low resolution condition, individuals with similar clothes cannot be matched exactly if using color only. If we consider their texture features in addition, these counterparts can be distinguished further. In this paper, we explore interesting regions and extract region-of-interest based features (ROIF for short) to describe the individual.

II. Related Work

Based on different applications, person re-identification problem can be described with two ways, i.e., multi-camera multi-object tracking and individual matching.

In the pedestrian detection and tracking scenario, person re-identification focuses on matching the individuals by the consideration of temporal constraints and ground plane homography, e.g., [1], [2], [3], [4], [5], [6], [7]. Yu et al. [4] use color and face features as the representation, and measure the distance with spatio-temporal constraints. Hamdoun et al. [14] use signatures based on SURF collected on the video sequences.

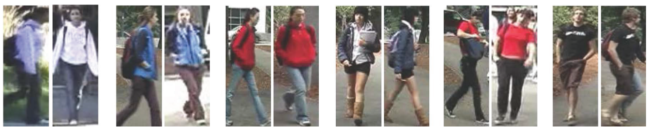

Fig. 1. Image pairs captured from different cameras.

In the individual identification or matching scenario, person re-identification is solved with the appearance model only, without temporal information, e.g., [8], [9], [10], [11], [12]. To improve the identification accuracy, Pedagadi et al. [13] propose to slide a window to extract a high-dimensional color histogram and then use the LFDA to reduce its high dimensionality. Kviatkovsky et al. [11] propose a color invariant signature in log-RGB color space to overcome the illumination variation. Farenzena at al. [8] exploit symmetry and asymmetry perceptual principles to weight features. Zheng et al. [12] formulate the re-identification task as a relative distance comparison (RDC) learning problem and propose to learn an optimal similarity measure between image pairs.

All the aforementioned works share the same protocol: extracting proper features and designing an effective distance measure. The main focus of this paper is feature extraction. Generally, there are two aspects to describe the appearance, i.e., global representation and local features integration. The global representation describes an image by a low-dimensional vector, and it may result the loss of details, e.g., [1]. Instead, the local features integration can keep details of very well, but it may generate high dimensionality and computation, e.g., [9], [13]. To overcome their disadvantages, we propose to reduce dimensionality and improve the performance via features from ROIs.

In this paper, we propose an efficient scheme for extracting ROIF, and implement ROIF based on HSV histogram and chromatic content respectively. In addition, we integrate ROIF with other features.

978-1-4799-6139-9/14/$31.00 ©2014 IEEE

III. Region-of-Interest based Features

ROIF scheme consists of two steps: a) finding interest regions by key point detection or texture analysis, and b) extracting features from ROIs using HSV histogram or chromatic content.

A. Finding Region-of-Interest

1) Silhouette Partition: We assume that the silhouette of an individual is present. In the video sequence case, it is obvious that a silhouette S can be obtained from the background using human detection and Gaussian Mixture Model (GMM) [15]. In the one-shot case, a silhouette S is extracted by using STEL component analysis [16].

Using symmetry of the body to weight features has been proved effective in [8]. Here we utilize the symmetry axis of an image to approximate the symmetry axis of a body (see Fig. 2). After that, we can assign different weight to each interest region and utilize the characteristic of torso and legs separately.

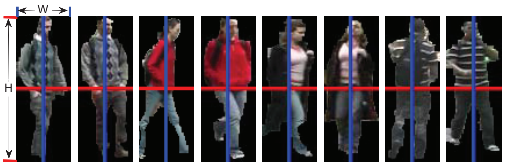

Fig. 2. Silhouette partition. The blue line is vertical symmetry axis, and the red line is horizontal symmetry axis.

2) Finding ROI: After silhouette partition, we have shrunk the interest area to the body region. Further, in order to obtain more sophisticated regions which can describe individual particularly, we find ROIs by their textural and structural properties.

ROI is the region with rich textures, so it can be detected with high repeatability and distinctness. Since SURF is designed as an invariant feature and MSER represents stable pattern instinctively, they can be used to find ROIs. To confirm this point, we compared the performance of five different feature detectors in Fig. 3. Obviously, the key points detected by SURF and MSER are more stable than others. Notice that, most of these points based on corners, e.g., FAST, MinEigen, Harris, are situated on the margin of the body, so they are not able to represent the whole appearance exactly. However, the points based on SURF and MSER are locating within the body, and hence including the most information of the appearance. Based on these observations, we can find ROIs in the positions which are detected by SURF and MSER.

B. Extracting Feature from ROI

1) ROIF based on HSV histogram: **Finding key patch.** SURF [17] is a scale- and rotation-invariant key point detector and descriptor. Unfortunately, because the original SURF is a point based descriptor, the same key point is always missing when there is a great change of the perspective. Therefore, in the case of low resolution of images and large variation of viewpoints, SURF descriptor cannot be matched very well. To remedy the drawback of SURF, we relax this key point as a 7×7 key-point-centred patch P_k, where k denotes the k-th key-point-centred patch of image I and calculate a histogram in each patch as descriptor $\boldsymbol{h}_k$.

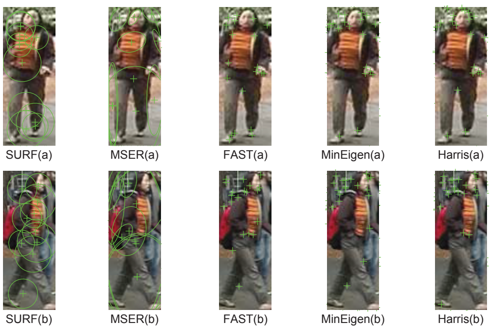

Fig. 3. Key points extracted by different detectors. The first row is from camera A, and the second is from camera B.

Weighting on patch. The features are extracted from HSV space. In order to eliminate the variance induced by the light condition, we adopt histogram equalization at first. Similar to [8], the significance of a point is inversely proportion to the distance to symmetry axis. We weight each patch P_k with a one-dimensional Gaussian kernel denoted as follows

$$w_k = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\{-\frac{(x_k - \mu_0)^2}{{\sigma_0}^2}\}, \tag{1}$$

where x_k is the x-coordinate of k-th key point, μ_0 is equal to the x-coordinate of vertical symmetry axis, and σ_0 is a prior set $W/2$. The HSV histogram is accumulated from all patches found in one image as $w_k \cdot \boldsymbol{h}_k$, where H, S and V channels are divided into 16, 16 and 8 equal bins respectively. The reason we divided the V space with less bins is that it contains the information about illumination and we reduce the effect of illumination variation.

Weighting on pixel. To show the different significance of every pixel $P(m,n)$, we give each pixel a weight as follows

$$\alpha(m,n) = \exp\{-\frac{(m-m_0)^2 + (n-n_0)^2}{2\sigma_1^2}\}, \tag{2}$$

where (m_0, n_0) is the index of the central pixel, (m,n) is the index of each element in the same patch, and σ_1 is equal to 3. Then every pixel's contribution to histogram can be calculated as $\alpha(m,n) \times P(m,n)$ (See Fig. 4).

In general, the appearance of the torso and legs are completely different, so we calculate the histogram separately. The HSV histogram based ROIF is defined as follows

$$\boldsymbol{h}_{up} = \sum_{k:P_k \in torso} w_k \cdot \boldsymbol{h}_k, \tag{3}$$

$$\boldsymbol{h}_{down} = \sum_{k:P_k \in legs} w_k \cdot \boldsymbol{h}_k, \tag{4}$$

$$\boldsymbol{h} = [\boldsymbol{h}_{up}, \boldsymbol{h}_{down}]. \tag{5}$$

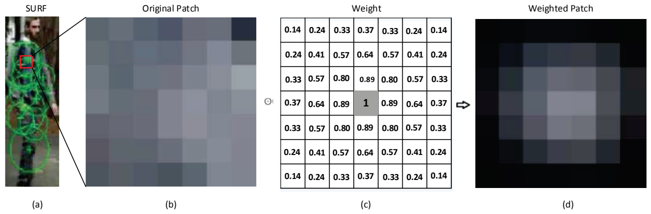

Fig. 4. The illustration of weighting on pixel. (a) The original image. (b) The original patch. (c) The weight matrix with a Gaussian kernel, where the element is $\alpha(m,n)$ defined in Eq. (2). (d) The weighted patch.

2) ROIF based on chromatic content: MSER [18] is the region having the property of invariance to affine transformation of image intensities, which is likely to keep the characteristic from pictures of different perspectives. In order to avoid redundancy of the representation, motivated by MSCR [8], we describe this region as a 9-dimensional chromatic content vector $\boldsymbol{c}_k$, which consist of its area, centroid, second moment matrix and average color.

The same as P_k, MSER based R_k is extracted from the torso and legs separately. And each region is also weighted as w_k. The whole chromatic content is defined as follows

$$\boldsymbol{c}_{up} = \sum_{k:R_k \in torso} w_k \cdot \boldsymbol{c}_k, \tag{6}$$

$$\boldsymbol{c}_{down} = \sum_{k:R_k \in legs} w_k \cdot \boldsymbol{c}_k, \tag{7}$$

$$\boldsymbol{c} = [\boldsymbol{c}_{up}, \boldsymbol{c}_{down}]. \tag{8}$$

3) Distance Measure: We extract ROIF from two aspects, i.e., SURF and MSER, and using histogram and chromatic content vector as descriptor respectively. In general, in order to combine two kinds of features together, we may concatenate them first, and then reduce the dimensionality by PCA or LFDA [13]. But for the sake of enhancing the integratability of our method, we combine them by the distance measure. So that, the other features can be integrated in the same way conveniently.

In our method, the distance d_{ij} between two images $I(i)$ and $I(j)$ is calculated as follows

$$d_{ij} = \gamma_1 \times d(\boldsymbol{h}_i, \boldsymbol{h}_j) + \gamma_2 \times d(\boldsymbol{c}_i, \boldsymbol{c}_j), \tag{9}$$

where γ_1 and γ_2 are the weights of ROIF based on histogram and chromatic content, $\boldsymbol{h}_i$ and $\boldsymbol{h}_j$ are the HSV histogram of image $I(i)$ and $I(j)$, and the similar with $\boldsymbol{c}_i$ and $\boldsymbol{c}_j$. We use the L_2 norm of the disparity between two feature vectors to calculate the distance measurement $d(.,.)$, that is to say, $d(\boldsymbol{h}_i, \boldsymbol{h}_j) = \|\boldsymbol{h}_i - \boldsymbol{h}_j\|_2$ and $d(\boldsymbol{c}_i, \boldsymbol{c}_j) = \|\boldsymbol{c}_i - \boldsymbol{c}_j\|_2$. Due to that the histogram $\boldsymbol{h}$ and the content $\boldsymbol{c}$ are incomparable, their distance cannot add directly. We normalize their distance as follows

$$d(\boldsymbol{h}_i, \boldsymbol{h}_j) = \frac{d(\boldsymbol{h}_i, \boldsymbol{h}_j)}{\max_{j \in \mathcal{I}} d(\boldsymbol{h}_i, \boldsymbol{h}_j)}, i \in \mathcal{I}, \tag{10}$$

$$d(\boldsymbol{c}_i, \boldsymbol{c}_j) = \frac{d(\boldsymbol{c}_i, \boldsymbol{c}_j)}{\max_{j \in \mathcal{I}} d(\boldsymbol{c}_i, \boldsymbol{c}_j)}, i \in \mathcal{I}, \tag{11}$$

where $\mathcal{I}$ is the index set for images in the database. After that, the distance can be calculated easily by the weighted sum. In this paper, we set γ_1 and γ_2 equal to 0.5.

IV. Experiment Results

In this section, we evaluate our method on the benchmark database VIPeR [10] compared with some other well designed ones. This database contains two views image sets captured from camera A and camera B. The two sets consist of 632 individual pairs, and there is a one-to-one correlation between them with conspicuous change caused by illumination and viewpoint. All images are normalized to 128 × 48 pixels. In this paper, we consider the images from camera A as probe set and images from camera B as gallery set.

The evaluating procedure is to select a probe from probe set and compare against all the individuals in gallery set, then return a matching rank in terms of the distance between probe and gallery. The results are shown by the Cumulative Matching Characteristic (CMC) curve, which represents the recognition rate in the top n matches [10].

To evaluate the performance of our method, we reimplement the algorithms proposed in [8] and [11]. We run our algorithm on 10 different random subsets, which include 316 image pairs every time, and calculate the final result as the average. Fig. 5(a)(b) show the results obtained by six single features. It is seen that our method is always as good as others, and even outperforms some of them. Specially speaking, rank 20 re-identification rate is about 50% for ROIF, versus 39% to 50% for others. Fig. 5(a) also shows the improvement of performance by combine SURF and MSER together.

Another advantage of our method is that it can be integrated easily with other approaches. To confirmed this point, we combine ROIF with PartsSC [11] just by normalizing the distance matrixes and summing them up. Fig. 5(c) illustrates the performance of our integration. It is obvious that our method catches up with the top level easily. In particular, rank 10 re-identification rate is around 51% for ROIF, versus 46% for WHSV+MSCR. To compare our method with state-of-the-art, we also re-implement algorithms in [12], [13] and [10], and list the results in Table I. And the average consuming time is 11.65ms on a common PC, which is much faster than other learning based methods, e.g., LF in [13] is 347.4ms.

TABLE I
Re-Identification Rate on VIPeR

	Rank1	Rank5	Rank10	Rank20
ELF	8.1%	24.1%	36.6%	52.1%
RDC	15.7%	38.4%	53.9%	70.1%
SDALF	19.9%	38.9%	49.4%	65.7%
ROIF+PartsSC	21.9%	40.4%	52.5%	63.3%
LF	24.2%	55.4%	69.5%	80.9%

V. Conclusion

In this paper, we addressed the person re-identification problem via combining texture and color features in ROI. To be specific, we proposed an efficient approach, called ROIF,

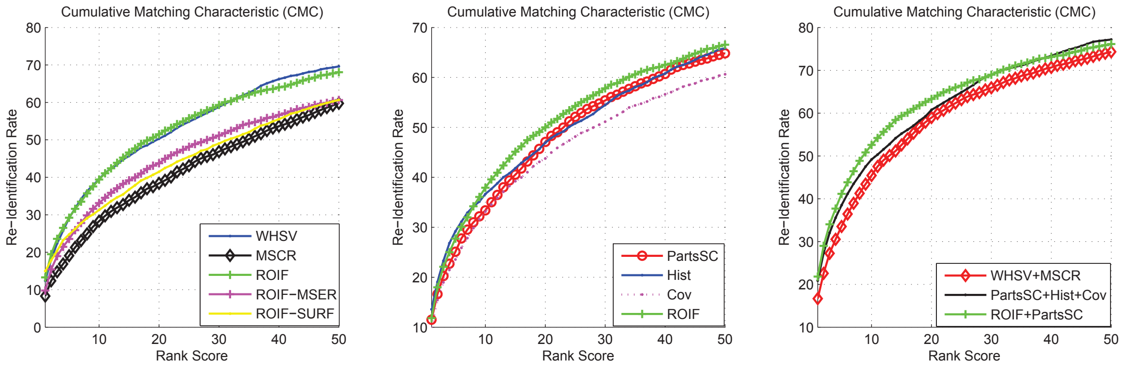

Fig. 5. Performance on VIPeR. (a) ROIF vs. WHSV and MSCR [19]. (b) ROIF vs. color invariant [11]. (c) Integrated ROIF vs. other methods.

which determines the region-of-interest by texture and extracts the color signature from these regions as the representation of an individual. We compared the proposed ROIF method on VIPeR database with ELF, RDC, SDALF, PartsSC and LF. Experimental results shown that our approach could achieve comparable accuracy with a lower computational complexity. In addition, we improved the performance further by integrating with other features. In future we will investigate the effect of metric learning and explore the optimal combination strategy for more sophisticated feature fusion.

Acknowledgment

This work was partially supported by the Next-Generation Broadband Wireless Mobile Communications Network Technology key project of the Ministry of Science and Technology of China under grant no. 2012ZX03002019-002, National Natural Science Foundation of China under Grant nos. 61273217, 61175011 and 61171193.

References

[1] C.-Y. Lin, L.-W. Kang, J.-H. Kao, C.-S. Lu, and Y.-T. Wu, “Multi-camera invariant appearance modeling for non-rigid object identification in a real-time environment,” *Journal of Visual Communication and Image Representation*, vol. 24, no. 6, pp. 717 – 728, 2013.

[2] V. I. Morariu and O. I. Camps, “Modeling correspondences for multi-camera tracking using nonlinear manifold learning and target dynamics,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006, pp. 545–552.

[3] S. Sternig, T. Mauthner, A. Irschara, P. M. Roth, and H. Bischof, “Multi-camera multi-object tracking by robust hough-based homography projections,” in *IEEE International Conference on Computer Vision Workshop*, 2011, pp. 1689–1696.

[4] S. Yu, Y. Yang, and A. G. Hauptmann, “Harry potter’s marauder’s map: Localizing and tracking multiple persons-of-interest by nonnegative discretization,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3714–3720.

[5] X. Chen, K. Huang, and T. Tan, “Object tracking across non-overlapping cameras using adaptive models,” in *Asian Conference on Computer Vision Workshop*, 2012, pp. 464–477.

[6] K. Kim and L. S. Davis, “Multi-camera tracking and segmentation of occluded people on ground plane using search-guided particle filtering,” in *European Conference on Computer Vision*, 2006, pp. 98–109.

[7] X. Wang, “Intelligent multi-camera video surveillance: A review,” *Pattern Recognition Letters*, vol. 34, no. 1, pp. 3–19, 2013.

[8] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, “Person re-identification by symmetry-driven accumulation of local features,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2360–2367.

[9] D. Gray and H. Tao, “Viewpoint invariant pedestrian recognition with an ensemble of localized features,” in *European Conference on Computer Vision*, 2008, pp. 262–275.

[10] D. Gray, S. Brennan, and H. Tao, “Evaluating appearance models for recognition, reacquisition, and tracking,” in *IEEE International Workshop on Performance Evaluation for Tracking and Surveillance*, 2007.

[11] I. Kviatkovsky, A. Adam, and E. Rivlin, “Color invariants for person reidentification,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1622–1634, 2013.

[12] W. Zheng, S. Gong, and T. Xiang, “Reidentification by relative distance comparison,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 653–668, 2013.

[13] S. Pedagadi, J. Orwell, S. A. Velastin, and B. A. Boghossian, “Local fisher discriminant analysis for pedestrian re-identification,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3318–3325.

[14] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux, “Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences,” in *IEEE International Conference on Distributed Smart Cameras*, 2008, pp. 1–6.

[15] J. Qu and Z. Liu, “Non-background HOG for pedestrian video detection,” in *International Conference on Natural Computation*, 2012, pp. 535–539.

[16] N. Jojic, A. Perina, M. Cristani, V. Murino, and B. J. Frey, “Stel component analysis: Modeling spatial correlations in image class structure,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 2044–2051.

[17] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, “Speeded-up robust features (surf),” *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346 – 359, 2008.

[18] J. Matas, O. Chum, M. Urban, and T. Pajdla, “Robust wide-baseline stereo from maximally stable extremal regions,” *Image and Vision Computing*, vol. 22, no. 10, pp. 761 – 767, 2004.

[19] L. Bazzani, M. Cristani, and V. Murino, “Symmetry-driven accumulation of local features for human characterization and re-identification,” *Computer Vision and Image Understanding*, vol. 117, no. 2, pp. 130 – 144, 2013.