Exploring Structural Information and Fusing Multiple Features for Person Re-identification

Yang Hu, Shengcai Liao, Zhen Lei, Dong Yi, Stan Z. Li *
Center for Biometrics and Security Research & National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
{yhu, scliao, zlei, dyi, szli}@nlpr.ia.ac.cn

Abstract

Recently, methods with learning procedure have been widely used to solve person re-identification (re-id) problem. However, most existing databases for re-id are small-scale, therefore, over-fitting is likely to occur. To further improve the performance, we propose a novel method by fusing multiple local features and exploring their structural information on different levels. The proposed method is called Structural Constraints Enhanced Feature Accumulation (SCEFA). Three local features (i.e., Hierarchical Weighted Histograms (HWH), Gabor Ternary Pattern HSV (GTP-HSV), Maximally Stable Color Regions (MSCR)) are used. Structural information of these features are deeply explored in three levels: pixel, blob, and part. The matching algorithms corresponding to the features are also discussed. Extensive experiments conducted on three datasets: VIPeR, ETHZ and our own challenging dataset MCSSH, show that our approach outperforms stat-of-the-art methods significantly.

1. Introduction

Person re-identification (re-id) is a challenging problem in video surveillance scenarios which has attracted more and more attention in recent years. It aims to associate identities of individuals across disjoint views in non-overlapping camera networks. The key issue is to measure the similarity between pedestrian images to estimate if they are from the same person. Due to the low resolution of images and the uncertainties of the views of different cameras, the appearance of images from the same person may change a lot. Moreover, there are illumination and background variations due to different environments and even occlusion when the person is in a crowd.

Most existing person re-id methods can be roughly classified into two groups: 1) methods only based on image (regardless of the topology of camera networks) [3, 5, 7, 8, 19, 18, 14, 17]; 2) methods utilizing spatial and temporal constraints within camera networks [1, 12, 13].

The image-based methods can be mainly divided into two types: 1) Extract visual features which are both distinctive and stable under various conditions between cameras. After feature extraction, an established distance measurement is applied to compare different person representations [3, 5, 8, 17]. The main difficulties of these methods are finding the applicable features under realistic conditions. 2) The other type aims to learn optimal distance measure for all features jointly via distance learning theory [14, 19]. These methods are less sensitive to feature selection, therefore they usually use very simple features, such as RGB, YCbCr, HSV color features and two types of texture features extracted by Schmid and Gabor filters which have been used in [7]. However, their results may be biased by the selection of the parameters, thus making these methods less flexible to different scenarios.

Methods combined with spatial and temporal information within camera networks can get much higher performance than those solely based on appearance [12]. Information like the traveling time across cameras, the expected entry/exit regions in the scene [13], people's locations and speed can be used as discriminative features. However, these methods are limited to the necessity of the knowledge of the environment such as where the cameras are deployed.

Traditional appearance-based methods usually make use of histogram features without paying enough attention to the structural information of the image. This paper highlights the structural constraints on local features, which facilitates the matching between two persons. The structural constraints are divided into three levels in bottom-up order: pixel-level constraints, blob-level constraints and part-level constraints. For the pixel-level, considering spatial locations of pixels, we extract color histograms by assigning weights according to the distance from each pixel to the central point of prior partitions. For the blob-level, two features are used. The first one is MSCR [4], by which not only the

*Corresponding author

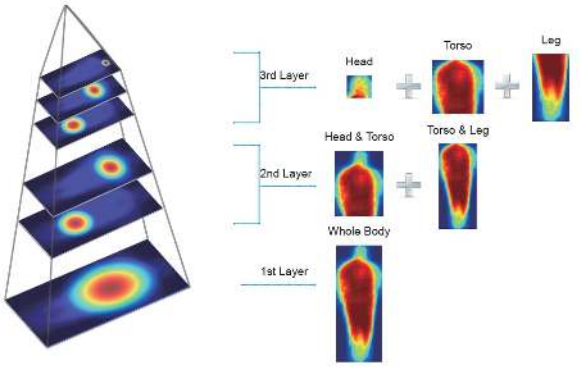

Figure 1. Hierarchical structure with 3 layers consisting of six components defined on the average image of the human body. The left is the two-dimensional Gaussian kernels applied on different layers. The right is the average image derived by accumulating all the human body images.

average colors are used, but also the locations and description capabilities of blobs are taken into account. The other is GTP-HSV which can describe massive keypoints detected on human body. Structural constraints are enhanced by graph matching approach. Finally, for the part-level, a hierarchical structure is formed with three layers to describe the targets, upon which the structural constraints of human body are utilized.

We call the proposed method *Structural Constraints Enhanced Feature Accumulation* (SCEFA), our model is unsupervised and does not require training samples. In contrast to most existing learning methods in this field, our method can avoid over-fitting caused by the learning procedure.

2. Features

In this section, we introduce three features to explore the structural information on three levels. In Sec. 2.1, HWH is discussed to enhance the structural constraints on both part-level and pixel-level; in Sec. 2.2 and Sec. 2.3 MSCR and GTP-HSV is discussed to enhance the structural constraints on blob-level, respectively.

2.1. Hierarchical Weighted Histograms

We present a hierarchical structure of human body which takes into account the inner connections of body parts. It consists of six components, upon which weighted color histograms are extracted (see Fig. 1). This feature is called *Hierarchical Weighted Histograms* (HWH). It should be noted that pixel-level and part-level structural constraints are enhanced in weighted color histograms and hierarchical structure, respectively.

We label the center and boundary of each basic component on the average pedestrian image computed from the overall sample images. A coarse-to-fine approach is applied to construct the hierarchical structure. The first layer corresponds to the bottom layer while the second and the third layer consist of two and three components, respectively, see Fig. 1. Partition of the first layer is the whole sketch of body. Finer partitions on the second layer are partial combinations of the three basic components.

HWH is built by applying a Gaussian kernel within each partition$\{P_1..., P_6\}$:

$$H(i) = \sum_{x,y} w(x,y)s(I(x,y) \in B(i)) \tag{1}$$

where H represents the histogram, $s(.)$ is a bool function, $B(i)$ is the value range of the i*th* bin, $w(x,y)$ is calculated as follows:

$$w(x,y) = \frac{1}{Z}\exp[-(\frac{(x-x_0)^2}{2\sigma_x^2} + \frac{(y-y_0)^2}{2\sigma_y^2})] \tag{2}$$

where (x_0, y_0) is the center of the partition, σ_x and σ_y is the deviation parameter (we set σ_x and σ_y three times of the width and height of the partition, respectively), and Z is the normalization coefficient.

2.2. Maximally Stable Color Regions

The Maximally Stable Color Regions (MSCR) operator [4] detects stable color blob regions by an agglomerative clustering step on each pixel from image, which shows invariance to scales and high repeatability. In this paper, after applying MSCR detector with all default parameters, we utilize the outputs of MSCR that consist of a second moment matrix and average color to match. It is noted that blob-level structural constraints are taken into account in the matching algorithm of MSCR (see Sec. 3.2 for details).

2.3. Gabor Ternary Pattern in Re-id

In this paper, we briefly consider person re-id as a problem seeking discriminative affine-invariant features in order to overcome the difficulty arising from viewpoint changes. The SIFT detector presented by Lowe [11] is one of the best detectors for keypoints as well as the most commonly used in person re-id [16]. However, the SIFT detector outputs insufficient number of keypoints due to the low resolution of image while describing a person. Thus, inspired by [10], we use a more proper keypoint detector CanAff and adopt Gabor Ternary Pattern (GTP) by combining it with HSV color information as a novel feature descriptor which we call GTP-HSV (see Fig. 2).

By using the GTP descriptor from the keypoint detected by CanAff, a 1296-dimensional feature vector is extracted. In this paper, we follow all the parameters in accordance with [10].

As we know, color-based descriptors are very discriminative and robust to the viewpoint changes. Thus, with

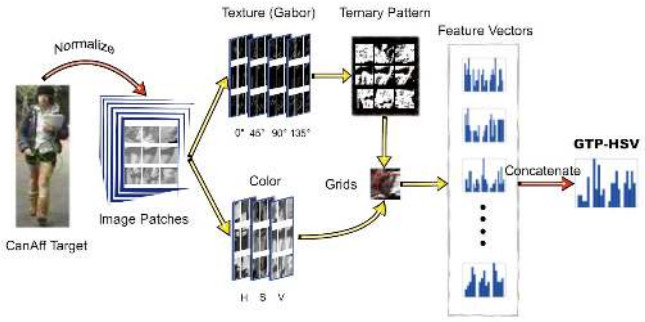

Figure 2. Flowchart of the GTP-HSV extraction. In this case only nine keypoints are processed.

the purpose of obtaining a more representative local feature descriptor, we not only extract texture information, but also color information ($16 \times 8 \times 4$ HSV used in this paper) from normalized image patches. After concatenating with the feature vector of Gabor filters, the dimensionality of combined feature vector will be very high. In order to reduce the computation complexity, we apply PCA to getting a M-dimensional feature vector, where $M = 128$ in this paper.

3. Multiple Feature Matching

3.1. HWH Matching

Each partition of HWH has already considered pixel-level structural constraints (by applying Gaussian weight) while the hierarchical structure of HWH reflects the part-level structural information. It is inappropriate to compute the similarity of each partition independently, because underlying structural relation may exist between partitions. Thus, HWH is extracted and concatenated to a final representative histogram taken as a whole, therefore the matching between HWHs becomes structure constrained matching both on pixel-level and part-level. Furthermore, matching between HWHs is calculated by the Hellinger distance [15] instead of the Battacharrya distance, from which we found 1% improvement of the recognition rate.

3.2. Blob Matching for MSCR

Due to the uncertainty of the size of detected MSCR area, it is inadequate by computing the distance of average color and circle centers as in [3]. Therefore it is better to compare the blobs which have similar size of area. The distance between two blobs d_{blob} can be measured as:

$$d_{blob} = \alpha d_{color} + \beta d_y + \mu d_{area} \tag{3}$$

where d_{color} is the difference of average color value (Lab color space), d_y is the difference of vertical coordinate and d_{area} is the difference of the size of area. α, β and μ are combination coefficients for these three distances, respectively. In this paper, we take $\alpha = 0.4$, $\beta = 0.6$ and $\mu = 0.4$.

In addition, to have a better corresponding map between blobs, we present a bidirectional matching strategy: given two sets of blobs $B^p = \{b_1^p, b_2^p, ..., b_{N_p}^p\}$ and $B^g = \{b_1^g, b_2^g, ..., b_{N_g}^g\}$ from probe and gallery respectively, where N_p and N_g are total number of the blobs in probe and gallery. First, for each blob t in B^p, using Equ. 3 to find a matched pair (t, q) which has the minimal distance from blob t to set B^g. By repeating this step for all the blobs in B^p, a set of matched pairs (also called assignment in Sec. 3.3) $A_{p\to g}$ can be set up. Second, for each blob t in B^g, we calculate a matching pair (t, q) which has the minimal distance from blob t to set B^p and then go through all the blobs in B^g, and a set of assignments $A_{g\to p}$ can be obtained. We consider the intersection of $A_{p\to g}$ and $A_{g\to p}$, denoted as $A_{p\leftrightarrow g} = A_{p\to g} \cap A_{A\leftarrow g} = \{(a_1, b_1), (a_2, b_2), ..., (a_n, b_n)\}$, where n is the total number of best matched pairs. The distance between the MSCR features can be finally obtained by:

$$d_{MSCR}(B_p, B_g) = \frac{1}{n}\sum_{i=1}^{n} d_{blob}(a_i, b_i) \tag{4}$$

3.3. Graph Matching for GTP-HSV

The GTP-HSV feature exhibits the following desirable properties for feature matching: (1) Invariant to the scale changes and affine transformations. (2) High repeatability of keypoints in two views of a same person. In this paper, we apply graph matching which takes into account the structural constraints among keypoints and uses spectral technique to seek the best matches. To the best of our knowledge, this is the first time applying graph matching method in re-id.

Given a set of N_p candidate keypoints $\{K_i\}$ from a probe image and a set of N_g target keypoints $\{K_i^{'}\}$ from a gallery image, the corresponding map of assignments becomes a binary value set. Considering the speed of calculation, only keypoints which have similar feature representation are calculated:

$$\begin{aligned} x = (&x_1^1, x_1^2, ..., x_1^{l_1}, ..., x_{n_p}^1, x_{n_p}^2, ..., \\ &x_{n_p}^{l_{n_p}}, ..., x_{N_p}^1, x_{N_p}^2, ..., x_{N_p}^{l_{N_p}}) \end{aligned} \tag{5}$$

where $x_{n_p}^l \in \{0, 1\}$, $l = 1, 2, ..., l_{n_p}$, $n_p = 1, 2, ..., N_p$, and l is the number of candidate assignments for a keypoint which is determined by a feature similarity threshold $T_{GTP-HSV}$. For each candidate assignment $a = (i, i^{'})$ (use i represent keypoint K_i for short in this section), there is an associating affinity that measures how well keypoints i matches $i^{'}$. In addition, for each pair of assignments (a, b), where $a = (i, i^{'})$ and $b = (j, j^{'})$, there is an affinity that

measures how compatible the features (f_i, f_j) are with the features $(f_{i'}, f_{j'})$. Therefore, the candidate assignments $a = (i, i^{'})$ can be seen as nodes forming an undirected graph which is weighted by the individual scores w_a at node a and $w_{a,b}$ on the edge connecting to the node a and b. By this way, the matching problem can be formulated as the following optimization form:

$$S(x) = \sum_{a \in A} w_a x_a + \lambda \sum_{a,b \in N} w_{a,b} x_a x_b \tag{6}$$

where A is the set of assignments and N is the compatible pairs of assignments, x_a represent $x^l_{n_p}$ in Equ. 5, and λ is the weight factor, which is set to 0.1 in our experiment. (a, b) is defined as a compatible assignment pair when keypoints (i, j) and $(i^{'}, j^{'})$ are geometrical neighbor at the same time. Only one-to-one matching is allowed in our model, therefore, Equ. 6 has to obey the following constraints:

$$\sum_{i=n_p, i^{'}} x_a \leq 1, \quad \sum_{i, i^{'}=n_g} x_a \leq 1 \tag{7}$$

With proper matrix manipulation, Equ. 6 can be reformulated as

$$S(x) = X^T M X \tag{8}$$

where M is called the affinity matrix, in which $M(a, a) = \exp\{-d(f_i, f_{i'})\}$, $M(a, b) = \exp\{-\frac{1}{\sigma}\|d_{i,j} - d_{i',j'}\|_2\}$. If the two assignments are not compatible, we set $M(a, b) = 0$; otherwise, $M(a, b) = \exp\{-\frac{1}{\sigma}\|d_{i,j} - d_{i',j'}\|_2\}$, where $d_{i,j}$ is distance from keypoint i to j and σ is the average radius of the area describing keypoints. Therefore, we can get:

$$x^* = \arg\max(X^T M X) \tag{9}$$

By applying the spectral approach, the above optimal problem becomes finding the main cluster from the assignments graph and can thus be solved by eigenvector technique [9]. After the eigenvalue decomposition of affinity matrix M, the values of main eigenvector are interpreted as the confidence of corresponding assignments. In addition, we reject those assignments conflicted with constraints described in Equ. 7, as well as the ones with low confidence. Finally, the optimal assignments of the keypoints between probe and gallery can be obtained by quantizing the solution into a binary value vector. In proposed method, however, it is preferable to utilize the original solution (not quantized into a binary value vector) $d_{GTP-HSV} = x^{*T} M x^*$ as the similarity score.

3.4. Score Fusion

Finally, a weighted score fusion is conducted via:

$$\begin{aligned} d(I_p, I_g) = {} & \phi_1 d_{HWH}(H_p, H_g) + \phi_2 d_{MSCR}(B_p, B_g) \\ & + \phi_3 d_{GTP-HSV}(G_p, G_g) \end{aligned} \tag{10}$$

where weighted coefficients ϕ_1, ϕ_2 and ϕ_3 are set as 0.4, 0.2, and 0.4, respectively.

4. Experiment

We validate the proposed approach on two publicly available datasets: VIPeR [6] and ETHZ [2]. To have a further understanding of the performance of proposed method, we conduct the experiment on our own dataset MCSSH (see Fig. 5), which is a challenging multi-shots dataset.

It is noted that in order to verify the flexibility and robustness of our method, we do not take any specific steps for the multi-shots datasets. Even without any learning process for spatio-temporal information of the multi-shots datasets, the proposed approach can still achieve state-of-the-art results.

Evaluation Settings. In our experiments, as for single-shot dataset, we randomly split each pair of persons and then put them into Cam A and Cam B respectively. After that, we randomly take all images of p people (classes) from Cam A to set up probe set and left the remainders unused. Gallery set is also set at the same time by finding the corresponding persons in Cam B. Then we repeat this procedure for 10 times and use the average cumulative match characteristic (CMC) curves [6] to show the recognition performance.

As for multi-shots dataset, we put all images together and randomly select one image from all persons (classes) to set up the gallery set and the rest are used as probe set. This procedure is repeated for 100 times to get the average CMC curves.

VIPeR. The Viewpoint Invariant Pedestrian Recognition database is the most popular and largest public dataset of re-id. As a single-shot dataset, it contains 632 pairs of pedestrians and images in VIPeR suffer greatly from illumination and viewpoint changes, making it a very challenging dataset.

Fig. 3 represents our results as well as those obtained by other state-of-the-art methods, namely PRDC [19] and SDALF [3], with the probe set $p = 316$, $p = 432$, $p = 532$. Results for Adaboost [7], again lower than others, are depicted for comparison purpose only. Besides, CMC scores obtained by these four approaches are reported in Tab. 1. It can be seen from both Fig. 3 and Tab. 1 that SCEFA always outperforms the other approaches especially on top ranks. Generally, SCEFA stands best on rank-1 which achieves great improvement (around 12% with three different number of probe set) for correct pair matching. With the increasing number of probe set (which means decrease number of the train set), the performance of SCEFA, as expected, is not impaired as much as the performances obtained by learning methods like PRDC and Adaboost. Both of the appearance-based approaches SCEFA and SDALF, outperform PRDC when $p = 532$, while SDALF is lower than PRDC with $p = 316$.

ETHZ. The ETHZ dataset consists of 146 persons with

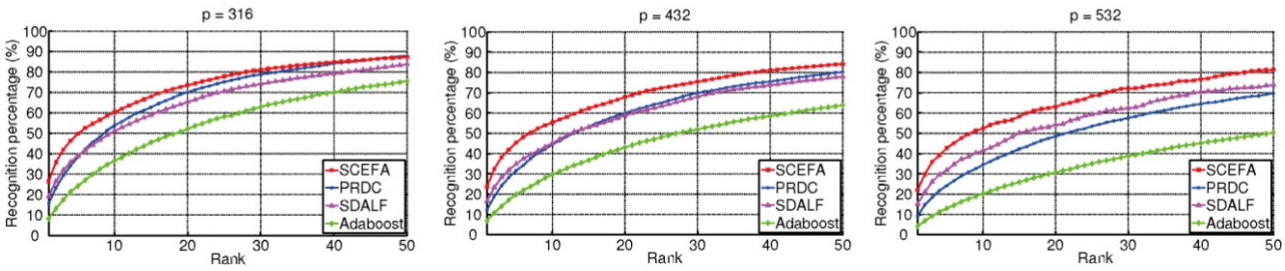

Figure 3. VIPeR dataset: CMC curves for SCEFA and other state of the art approaches.

Table 1. VIPeR dataset: top ranked matching rate (%) with $p = 316$, $p = 432$ and $p = 532$ in the probe set, r is the rank.

Methods	p=316				p=432				p=532			
	r = 1	r = 5	r = 10	r = 20	r = 1	r = 5	r = 10	r = 20	r = 1	r = 5	r = 10	r = 20
SCEAF	**26.49**	**49.80**	**60.29**	**73.54**	**23.71**	**45.39**	**55.39**	**67.89**	**22.13**	**42.72**	**52.03**	**63.19**
PRDC [19]	15.66	38.42	53.86	70.09	12.29	31.55	44.49	59.91	9.12	24.19	34.40	48.55
SDALF [3]	19.11	38.97	51.07	65.29	16.58	34.80	45.09	58.75	15.19	31.72	41.45	54.15
Adaboost [7]	8.17	24.15	36.58	52.12	6.83	19.81	29.75	43.06	4.19	12.95	20.21	30.73

8555 images in total, where images of people are taken from a moving camera in a busy street. Viewpoint and illumination changes as well as occlusions are the main challenges in ETHZ. The ETHZ dataset is structured with 3 sequences, in this paper we call these sequences ETHZ1, ETHZ2 and ETHZ3 respectively.

The comparative results of our experiments carried out on ETHZ are shown in Fig. 4. We compare SCEFA with state of the art appearance-based method SDALF. It can be seen that SCEFA outperforms SDALF on the first two sequences of ETHZ: ETHZ1 and ETHZ2, while the CMC curves almost overlap on ETHZ3. It is, however, obvious that SCEFA outperforms SDALF at rank-1 on ETHZ3 with around 5% improvement. The reason of the decreasing performance of SCEFA on ETHZ3 is that the images in ETHZ3 are different from the other two: the pedestrian contours do not fill the bounding box, therefore, when we segment the pedestrian at a fixed ratio, the three basic components we get are not corresponded to the body parts accurately which will affect the features we extract.

Multiple Camera Scenario in Station Hall (MCSSH). This dataset is captured in a busy station hall. It contains 345 pedestrians with totally 1561 images. This dataset captures images in real monitoring scenario, we can see some sample images in Fig. 5.

To give more insight on how our proposed method performs on multi-shots dataset, we carry out the experiments on MCSSH. Our experiments on the MCSSH dataset follow the protocol pre-mentioned in multi-shots settings. Fig. 6

Figure 5. Examples of the MCSSH dataset.

shows our results as well as the one obtained by SDALF. We can see that our proposed method outperforms SDALF completely by 17% at rank-1 in addition to the average 13% improvement over all rank scores. Note that with structural constraints enhanced, local features can provide notable benefit in re-id application. Even by using the same features, significant improvement can be achieved by exploring structural information.

5. Conclusions

In this paper, we combine three suitable features for high performance person re-identification and introduce the structural constraints enhancement on three levels into the feature extraction and matching process to improve the performance further. As demonstrated by the experimental results, our approach outperforms state-of-the-art methods in-

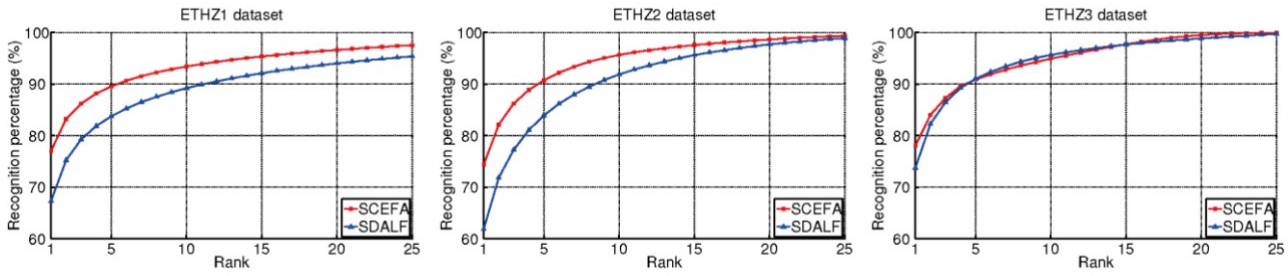

Figure 4. ETHZ dataset: CMC curves for SCEFA and SDALF. Left to right: results on SEQ.#1, on SEQ.#2, and on SEQ.#3.

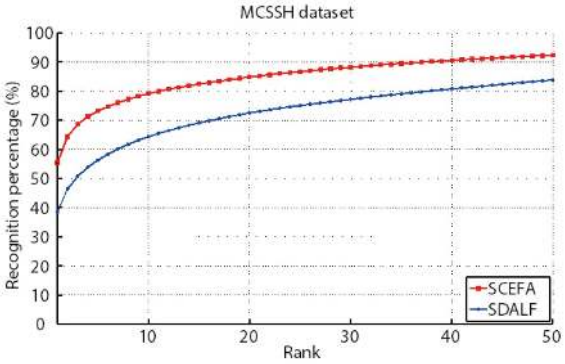

Figure 6. MCSSH dataset: CMC curves for SCEAF and SDALF.

cluding learning based and feature selection based method significantly. As an unsupervised method, our approach is robust on most re-identification datasets.

Acknowledgement

This work was supported by the Chinese National Natural Science Foundation Project #61070146, #61105023, #61103156, #61105037, #61203267, National IoT R&D Project #2150510, National Science and Technology Support Program Project #2013BAK02B01, Chinese Academy of Sciences Project No. KGZD-EW-102-2, European Union FP7 Project #257289 (TABULA RASA), and AuthenMetric R&D Funds.

References

[1] I. O. de Oliveira and J. L. de Souza Pio. People reidentification in a camera network. In *DASC*, pages 461–466, 2009. 1

[2] A. Ess, B. Leibe, and L. J. V. Gool. Depth and appearance for mobile scene analysis. In *ICCV*, pages 1–8, 2007. 4

[3] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, pages 2360–2367, 2010. 1, 3, 4, 5

[4] P.-E. Forssén. Maximally stable colour regions for recognition and matching. In *CVPR*, 2007. 1, 2

[5] N. Gheissari, T. B. Sebastian, and R. Hartley. Person reidentification using spatiotemporal appearance. In *CVPR (2)*, pages 1528–1535, 2006. 1

[6] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, volume 3, page 5, 2007. 4

[7] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV (1)*, pages 262–275, 2008. 1, 4, 5

[8] 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 *ICDSC*, pages 1–6, 2008. 1

[9] M. Leordeanu and M. Hebert. A spectral technique for correspondence problems using pairwise constraints. In *ICCV*, pages 1482–1489, 2005. 4

[10] S. Liao, A. K. Jain, and S. Z. Li. Partial face recognition: Alignment-free approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99(PrePrints):1, 2012. 2

[11] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, pages 1150–1157, 1999. 2

[12] R. Mazzon, S. F. Tahir, and A. Cavallaro. Person re-identification in crowd. *Pattern Recognition Letters*, 33(14):1828–1837, 2012. 1

[13] B. Prosser, S. Gong, and T. Xiang. Multi-camera matching using bi-directional cumulative brightness transfer functions. In *BMVC*, 2008. 1

[14] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by support vector ranking. In *BMVC*, pages 1–11, 2010. 1

[15] D. G. Simpson. Minimum hellinger distance estimation for the analysis of count data. *Journal of the American statistical Association*, 82(399):802–807, 1987. 3

[16] L. F. Teixeira and L. Corte-Real. Video object matching across multiple independent views using local descriptors and adaptive learning. *Pattern Recognition Letters*, 30(2):157–167, 2009. 2

[17] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. H. Tu. Shape and appearance context modeling. In *ICCV*, pages 1–8, 2007. 1

[18] W.-S. Zheng, S. Gong, and T. Xiang. Associating groups of people. In *BMVC*, 2009. 1

[19] W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. In *CVPR*, pages 649–656, 2011. 1, 4, 5