Improving Person Re-identification by Viewpoint Cues

Sławomir Bąk Sofia Zaidenberg Bernard Boulay Francois Brémond
INRIA Sophia Antipolis, STARS/Neosensys
2004, route des Lucioles, BP93
06902 Sophia Antipolis Cedex - France
firstname.surname@inria.fr

Abstract

Re-identifying people in a network of cameras requires an invariant human representation. State of the art algorithms are likely to fail in real-world scenarios due to serious perspective changes. Most of existing approaches focus on invariant and discriminative features, while ignoring the body alignment issue. In this paper we propose 3 *methods for improving the performance of person re-identification. We focus on eliminating perspective distortions by using 3D scene information. Perspective changes are minimized by affine transformations of cropped images containing the target (1). Further we estimate the human pose for (2) clustering data from a video stream and (3) weighting image features. The pose is estimated using 3D scene information and motion of the target. We validated our approach on a publicly available dataset with a network of* 8 *cameras. The results demonstrated significant increase in the re-identification performance over the state of the art.*

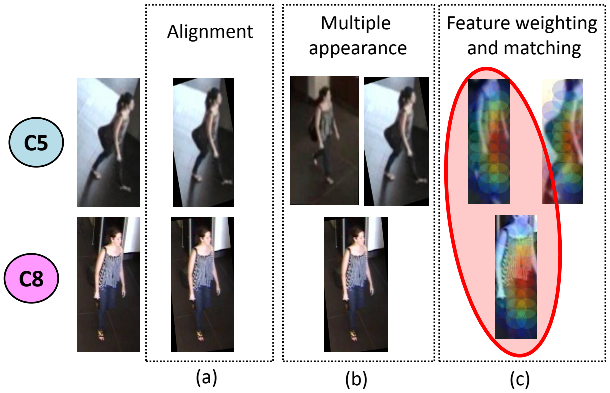


Figure 1. Improvements on re-identification using viewpoint cues: (a) target alignment (section 2); (b) multiple target appearance based on clustering (section 3.2); (c) pose orientation-driven weighting (section 3.3). The illustration shows an example of the same person viewed from two different cameras (C5 and C8).

1. Introduction

Person re-identification is a well defined task that requires finding a particular person in a network of cameras with non-overlapping fields of view. As video surveillance usually provides low resolution data, this task is approached by extracting a global appearance of the target using its clothing features. The clothing is not as distinctive as face or iris features, thus we often treat this problem as a retrieval task. However, the low recognition accuracy of existing approaches not only comes from insufficient discriminative power of clothing features, but also from significant appearance changes caused by variations in view angles, lighting conditions and person poses.

Current state of the art approaches concentrate either on *feature modeling* [2, 5, 15] producing descriptors invariant to camera changes or on *metric learning* [4, 9, 17] that should boost performance regardless of the representation choice. Designing a new descriptor, we need to look for a trade-off between its discriminative power and invariance through cameras. This task is particularly hard, especially, as this trade-off varies from data to data [14]. Metric learning approaches use training data to search for strategies that combine given features maximizing inter-class variation whilst minimizing intra-class variation. However, these approaches focus on learning a function which transfers the feature space from the first camera to the second one, introducing the requirement of training $\binom{c}{2}$ models for c cameras (*e.g.* a network of 8 cameras needs to train 56 metrics). Moreover, *metric learning* based approaches need large training data (hundreds of annotated image pairs with the same individual registered by different cameras) for every camera pair. This alone might make these approaches inapplicable in small camera networks where acquisition of labeled data is unattainable.

1.1. Perspective changes

Until now, most of appearance-based approaches were usually evaluated on cropped images that are manually extracted from a few video streams taken at eye level (no significant perspective changes). Slight pose changes can usually be approached by adopting perceptual principles of symmetry and asymmetry of the human body [5]. The extracted features are then weighted using the idea that features closer to the bodies' axes of symmetry are more robust against scene clutter.

In this paper we address the real world scenario, where pose orientation of a person might change due to serious perspective changes. Figure 1(a) illustrates that the change might be significant, thus having noticeable impact on recognition performance. We offer a simple but efficient affine transformation using 3D scene information to improve alignment of two images extracted from different cameras (section 2).

1.2. Multiple signature appearance

Further classification of appearance-based techniques distinguishes the *single-shot* and the *multiple-shot* approaches. The former class extracts appearance using a single image [8, 12, 15], while the latter employs multiple images of the same object to obtain a robust representation [2, 5, 6, 13, 17]. The main idea of multiple-shot approaches is to take advantage of several frames for extracting a more reliable representation [1, 6, 11]. Although *multiple-shot* approaches employ several frames for generating the signature, in the end they usually produce a single (averaged) representation ignoring the possible pose and view-angle changes that can occur while tracking the target in a single camera. Those approaches either select more frequent features or blend the appearance by averaging features. On the contrary, we propose to cluster the trajectory based on estimated pose cues and generate the signature for every significantly different appearance (see figure 1(b) where the appearance of the target is clustered into two appearances). Finally we produce a multiple signature consisting of several signatures ordered by the orientation of their pose. Having pose information, we determine weights of features w.r.t. the distance from the frontal viewpoint for every signature (see figure 1(c)). This improves the alignment and the matching accuracy.

This paper makes the following contributions:

- We eliminate perspective distortions by applying a simple affine transformation (rotation) on cropped images containing a person of interest. This transformation is based on 3D scene information (section 2).

- We offer to employ pose cues for clustering the trajectory with a target. Pose is estimated using 3D scene information and motion of the target. The proposed clustering allows to measure the reliability of the detected orientation in every frame (section 3.2).

- We propose an *orientation-driven* weighting strategy of image features. Our idea is that different regions of the target appearance ought to be aligned using pose cues, minimizing appearance change and increasing matching performance (section 3.3).

We validate all steps of our approach in section 4 before discussing perspectives and concluding.

2. Target alignment

The changing viewpoint in a network of cameras is an important issue and might significantly distort the visual appearance of a person. This problem has a direct impact on the re-identification accuracy. Eliminating perspective distortions in an image region containing the target is often called *image rectification* [10]. Although employing rectification methods gives satisfactory results in pedestrian detection tasks, we observed that the extracted homography between the image and the ground planes can still produce significant distortions in the texture inside the target appearance. As a result, instead of employing rectification, we propose a simple method that only rotates the cropped image with of target by an angle α. This angle is extracted using 3D scene information (see figure 2(b)), by mapping the vector orthogonal to the ground plane living in real world coordinates to the vertical of a given image. This mapping is achieved by employing the calibration information of the camera (*i.e.* we employ Tsai calibration).

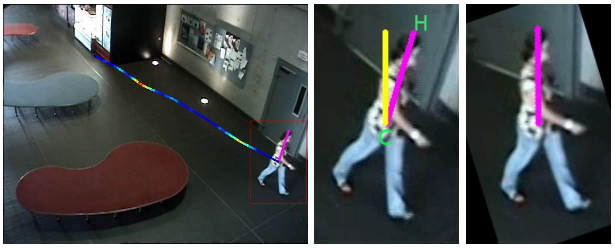
Rotation angle α: Given a detected person (a rough bounding box around the person, see figure 2(b)), we select the center point of the detection (point C). Having a pixel location of this point in the image plane (x_c^i, y_c^i), we compute its corresponding point in the world coordinate system (x_c^r, y_c^r) using calibration information and a fixed height of a person $h = 1.7$ m. From this point, we can easily compute the orthogonal to the ground plane in the world coordinate system meeting the head point (x_h^r, y_h^r) that has its corresponding location (x_h^i, y_h^i) in the image plane (point H). The rotation angle can be computed by

$$\alpha = \arctan\left(\frac{x_h^i - x_c^i}{y_c^i - y_h^i}\right). \quad (1)$$

Figure 2(c) illustrates the result of the rotation.

3. Multiple target appearance

This section introduces the method for extracting the pose of a person using 3D scene information and the motion of the target (section 3.1). Using pose, we cluster the trajectory into clusters with reliable pose detection



(a) trajectory (b) cropped (c) rotated

Figure 2. Affine transformation of the target image: (a) trajectory of the target (color of the trajectory illustrates the reliability of the detected pose, see section 3.2 for details); (b) the cropped image obtained by the detection algorithm ; (c) the rotated image.

(section 3.2) and generate multiple signatures for the single trajectory. For each such generated signature, we determine orientation-driven weights for signature matching (section 3.3).

3.1. Pose orientation θ

Given detection results for n frames, we compute a set of central points $\mathcal{C}^r = \{(x^r_{c,1}, y^r_{c,1}), \ldots, (x^r_{c,n}, y^r_{c,n})\}$ that lie in the real world coordinates system and correspond to the center of the detections in the image plane. Using calibration information, we extract the position of the camera projected on the ground plane (x^r_{cam}, y^r_{cam}). For each position $k \in [2, n]$ on the trajectory, let $\mathbf{m}_k$ be the motion vector defined as

$$\mathbf{m}_k = \left[x^r_{c,k} - x^r_{c,k-1}, y^r_{c,k} - y^r_{c,k-1}\right], \qquad (2)$$

and $\mathbf{v}_k$ the viewpoint vector defined as

$$\mathbf{v}_k = \left[x^r_{cam} - x^r_{c,k-1}, y^r_{cam} - y^r_{c,k-1}\right]. \qquad (3)$$

We define the pose orientation θ_k using dot product between these two vectors

$$\theta_k = \arccos\left(\frac{\mathbf{v}_k \cdot \mathbf{m}_k}{|\mathbf{v}_k||\mathbf{m}_k|}\right). \qquad (4)$$

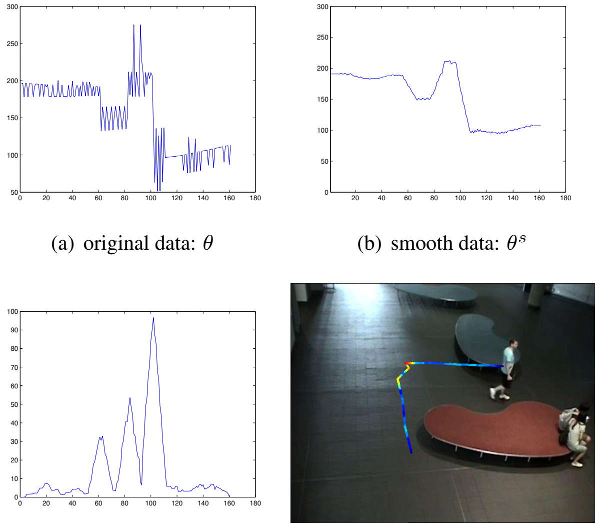
Figure 3(a) presents θ values for the given trajectory (figure 3(d)) and figure 5 shows an illustration of orientation angle θ.

3.2. Orientation-driven clustering

In figure 3(a), we can notice that the θ estimation might be noisy. We minimize noise by smoothing the data:

$$\theta^s_k = \sum_{l=k-z}^{k+z} \frac{\theta_l}{2z+1}, \qquad (5)$$

where z is a smoothing parameter (we set $z = 5$ in experiments). This operation provides us more reliable pose orientation cues (figure 3(b)).



(a) original data: θ (b) smooth data: θ^s

(c) derivative: δ (d) orientation reliability

Figure 3. Orientation driven trajectory clustering: (a) original pose orientation θ estimated from the video; (b) the result of the smoothing operation; (c) our control derivative function; (d) the reliability of the trajectory (red color indicates low reliability, while blue stands for the highest).

The key idea of clustering the trajectory is to obtain multiple appearances of the target w.r.t. its orientation. By detecting significant changes in orientation, we believe that we can detect significant changes in the appearance. We estimate the pose changes using the control derivative function δ_k defined as

$$\delta_k = \max_{l=k-z \ldots k+z} \frac{d\theta^s_k(t+l)}{dt}. \qquad (6)$$

We use this function to measure the reliability of the orientation θ. We assume that *peaks* in δ (see figure 3(c)) and their neighborhoods might provide unreliable information. Figure 3(d) illustrates the trajectory and its reliability. We can notice that the trajectory is unreliable during the pose change (the color red indicates low reliability, while blue – the highest). Frames with estimated unreliable orientation ($\delta > 10$) are removed from the trajectory, determining gaps in the trajectory, thus clustering it into the multiple appearances. Each appearance cluster is labeled with its *mean* orientation (*e.g.* the person in figure 3(d) was separated into two clusters, labeled with orientation 191° and 99°). The output multiple signature for the target consists of signatures computed from each cluster.

3.3. Orientation-driven feature weighting and matching

Each trajectory cluster consists of a set of cropped and rotated images with estimated poses. From such images we

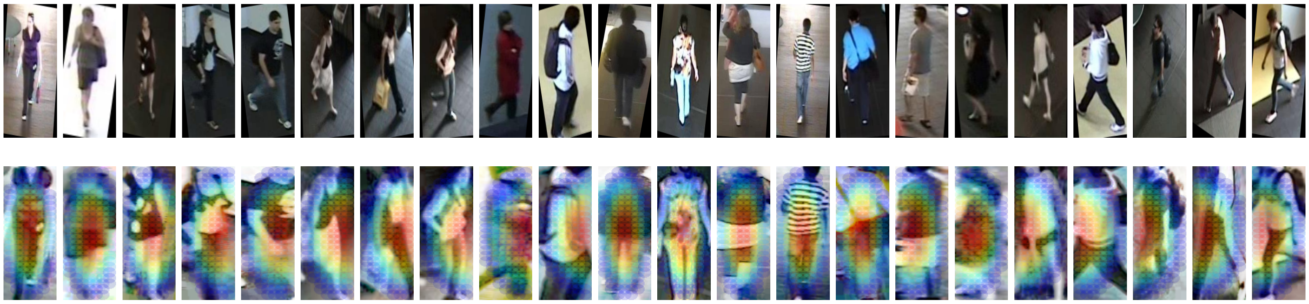


Figure 4. Sample aligned images and their weight distributions for different orientations. The first row shows aligned images, rotated by α. The second row illustrates weight distributions (looking from the left to the right the orientation angle θ in these examples varies from 16° to 309° with a $\sim$ 13° step).

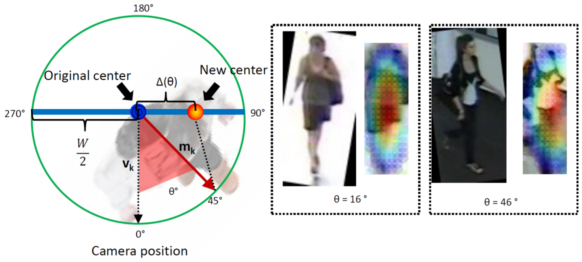


Figure 5. Computation of the new center of the Epanechnikov kernel. The original kernel is shifted based on the detected orientation of the person w.r.t. to the camera. We shift the kernel by $\Delta(\theta^s)$ in the horizontal direction. On the right we can see the change of weight distributions w.r.t. the detected orientation angle for $\theta^s = 16°$ and $\theta^s = 46°$.

extract a dense grid of overlapping color features (patches). We propose a weighting strategy of those patches based on the estimated pose. We eliminate background information by employing an Epanechnikov kernel, while determining the feature weights.

We define the Epanechnikov kernel as a function of the orientation θ^s with variable width. Modifying the center of the kernel w.r.t. θ^s, we address the problem of appearance changes. Let every image be scaled to a size of $W \times H$. Defining the center of the Epanechnikov kernel, we shift it in the horizontal direction by

$$\Delta(\theta^s) = sgn(sin(\theta^s))\frac{W sin^2(\theta^s)}{2} \tag{7}$$

(see figure 5). As a result, we define kernel parameters as

$$h = \frac{H}{2} \tag{8}$$

$$w(\theta^s) = \frac{W}{2} + \Delta(\theta^s). \tag{9}$$

Given patch p of the dense feature grid, we define its weight as $\mathbf{K}_{x,y}(\theta^s)$ expressed by

$$\begin{cases} \frac{3}{4}\left(1 - \left(\frac{x}{w(\theta^s)}\right)^2 - \left(\frac{y}{h}\right)^2\right), & |\left(\frac{x}{w(\theta^s)}\right)^2 + \left(\frac{y}{h}\right)^2| \leq 1; \\ 0, & \text{otherwise}, \end{cases} \tag{10}$$

where (x, y) are coordinates of patch p computed relatively to the shifted center of Epanechnikov kernel. The weight represents the patch's contribution to the signature. Figure 4 illustrates sample images and determined weights of image regions w.r.t. to orientation θ^s.

Signature matching for retrieval Let $\mathcal{A}$ and $\mathcal{B}$ be multiple signatures of two individuals whose similarity we want to measure. Each signature consists of a set of signatures extracted w.r.t. θ^s. Before computing the similarity between corresponding features in a dense grid, we address the alignment problem. We align two signatures by minimizing the difference between kernels:

$$\min_{a \in \mathcal{A}, b \in \mathcal{B}} |\mathbf{K}(\theta^s_a) - \mathbf{K}(\theta^s_b)|, \tag{11}$$

where a and b are sub-signatures of $\mathcal{A}$ and $\mathcal{B}$, expressed by weight matrices $\mathbf{K}$. Having aligned signatures, we compute similarity of corresponding features using averaged Epanechnikov weights.

4. Experimental results

This section focuses on evaluating all steps of our approach in a real world scenario. We illustrate the improvements at each step of the processing chain, discussing its importance. Finally, we show that our approach significantly outperforms state of the art results.

4.1. Database and experimental setup

We carry out experiments on the SAIVT-SOFTBIO database [3]. This dataset consists of 152 people moving through a network of 8 cameras. Subjects travel in an uncontrolled manner, thus most of subjects appear only in a subset of the camera network. This provides a highly unconstrained environment reflecting a real-world scenario. In average, each subject is registered by 400 frames spanning up to 8 camera views in challenging surveillance conditions (significant illumination, pose and viewpoint

STEP	$r = 1$	$r = 5$	$r = 10$	$r = 25$
BASELINE	7.11%	21.09%	35.79%	59.60%
TA	11.12%	28.09%	42.79%	63.60%
TA+MS	20.30%	39.62%	52.92%	71.91%
TA+MS+W	**23.02%**	**42.73%**	**54.64%**	**72.89%**

Table 1. Validation of the proposed contributions on the SAIVT-SOFTBIO dataset. Values correspond to the recognition accuracy averaged among all 56 pairs of cameras at different ranks r.

changes). Each camera captures data at 25 frames per second and at a resolution of 704×576 pixels. Although some cameras overlap, we do not use this information while testing re-identification algorithms. The database provides XML files with annotations given by coarse bounding boxes indicating the location of the subjects in each frame. Using the centers of these bounding boxes, we obtain trajectories of the subjects. Subject images are cropped and aligned using the method described in section 2. After clustering the trajectory, we randomly select $N = 10$ subsequent images from each cluster to compute the sub-signature of the multiple signature appearance.

4.1.1 Appearance model

Every cropped image is scaled into a fixed size window of 64×192 pixels. A set $\mathbf{P}$ of rectangular sub-regions is produced by shifting 32×32 regions with a 16 pixels step. This operation results in $|\mathbf{P}| = 33$ overlapping rectangular sub-regions. From each sub-region, we extract RGB color histograms. We minimize color dissimilarities caused by camera illumination changes by applying *histogram equalization* to each color channel.

4.1.2 Evaluation metrics

Re-identification performance is analyzed in terms of recognition rate, using the averaged *cumulative matching characteristic* (CMC) curve [7]. The CMC represents the expectation of finding the correct match in the top n matches. nAUC is a quantitative scalar of the CMC curve computed by normalizing the area under the CMC curve. Every multiple signature is used as a query to the gallery set of multiple signatures from the other cameras. This procedure has been repeated 10 times to obtain averaged CMC results for each camera pair. Having 8 cameras, we evaluate our approach on 56 pairs. While validating the steps of our approach, we provide CMC curves and nAUC values corresponding to the recognition accuracy averaged among all 56 pairs of cameras at different ranks r.

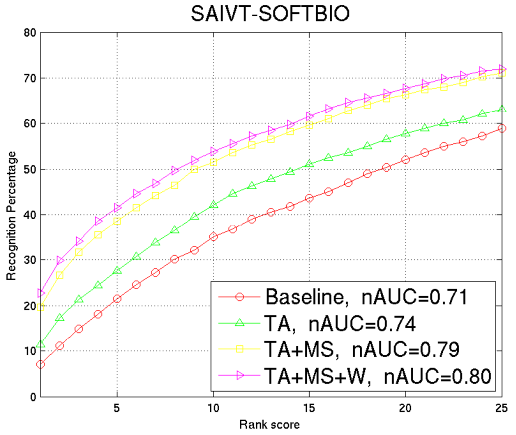


Figure 6. CMC curves obtained on the SAIVT-SOFTBIO dataset. CMC and nAUC values correspond to the recognition accuracy averaged among all 56 pairs of cameras at different ranks r.

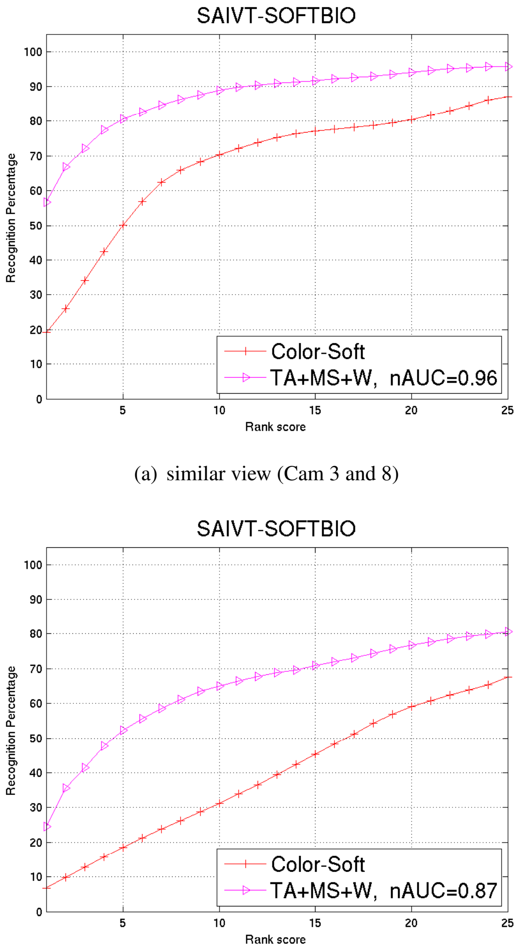
4.2. Results

Figure 6 and table 1 illustrate the impact of each step of our algorithm on the re-identification accuracy. BASELINE corresponds to signatures extracted using randomly selected $N = 10$ subsequent frames. Labels TA, MS and W correspond respectively to the given contributions: **T**arget **A**lignment (section 2), clustering into **M**ultiple target **S**ignature (section 3.2) and the orientation-driven **W**eighting (section 3.3). From the results it is apparent that each step of the algorithm has a significant impact on the performance. We consistently increase the recognition for all ranks employing the proposed steps. We can notice that the most important step is related to computation of the multiple signature appearance (*i.e.* for $r = 1$ we can notice an increase of about 9% in the recognition accuracy). This step is achieved by our clustering method. Although employing Epanechnikov kernel (step W) improves the recognition, we were expecting more significant improvement. Weighting the features or finding the salience regions [16] in images is still a challenging task and we will address this problem in the future work.

For comparing with state of the art approach, we illustrate the performance of Color-soft [3] in figure 7. We followed the experimental setup in accordance with [3], thus focusing only either on cameras with similar view (figure 7(a)) or on cameras with a significant viewpoint change (figure 7(b)). The results show clearly that our approach significantly improves state of the art performance.

5. Conclusion and perspectives

This paper proposed 3 improvements for person re-identification in a real-world scenario. We focused on eliminating perspective distortions by using 3D scene informa-



(a) similar view (Cam 3 and 8)

(b) dissimilar view (Cam 5 and 8)

Figure 7. Performance comparison using CMC curves, our (TA+MS+W) *vs*. Color-Soft [3]. We can notice the significant improvement, especially for dissimilar views (b). The example of images from camera 5 and 8 are given in figure 1.

tion. Perspective changes are minimized by rotating the cropped images w.r.t. the orthogonal to the ground plane. Using different pose orientations, we divided the trajectory into clusters capturing multiple target appearances, thus significantly improving the recognition accuracy. Further, we proposed a pose-driven weighting strategy to eliminate background information and improve signature alignment. We demonstrated that all steps of our approach consistently improve the re-identification performance, outperforming state of the art. Future work will focus on improving the weighting strategy. We plan to learn *a priori* weights for matching different poses using a depth sensor. This offline learned weights could be then applied while matching different poses.

Acknowledgements

This work has been supported by PANORAMA European project.

References

[1] S. Bak, G. Charpiat, E. Corvee, F. Bremond, and M. Thonnat. Learning to match appearances by correlations in a covariance metric space. In *ECCV*, 2012. 2

[2] L. Bazzani, M. Cristani, A. Perina, M. Farenzena, and V. Murino. Multiple-shot person re-identification by hpe signature. In *ICPR*, 2010. 1, 2

[3] A. Bialkowski, S. Denman, S. Sridharan, C. Fookes, and P. Lucey. A database for person re-identification in multi-camera surveillance networks. In *DICTA*, 2012. 4, 5, 6

[4] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja. Pedestrian recognition with a learned metric. In *ACCV*, 2010. 1

[5] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010. 1, 2

[6] N. Gheissari, T. B. Sebastian, and R. Hartley. Person reidentification using spatiotemporal appearance. In *CVPR '06*, pages 1528–1535, Washington, DC, USA, 2006. IEEE Computer Society. 2

[7] D. Gray, S. Brennan, and H. Tao. Evaluating Appearance Models for Recognition, Reacquisition, and Tracking. *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, 2007. 5

[8] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV '08*, pages 262–275, Berlin, Heidelberg, 2008. Springer-Verlag. 2

[9] M. Hirzer, P. Roth, M. KÃűstinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *ECCV*, 2012. 1

[10] Y. Li, B. Wu, and R. Nevatia. Human detection by searching in 3d space using camera and scene knowledge. In *ICPR*, 2008. 2

[11] O. Oreifej, R. Mehran, and M. Shah. Human identity recognition in aerial images. In *CVPR*, 2010. 2

[12] U. Park, A. Jain, I. Kitahara, K. Kogure, and N. Hagita. Vise: Visual search engine using multiple networked cameras. In *ICPR '06*, volume 3, pages 1204–1207, Aug. 2006. 2

[13] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by support vector ranking. In *BMVC*, pages 21.1–21.11, 2010. 2

[14] M. Varma and D. Ray. Learning the discriminative power-invariance trade-off. In *ICCV*, 2007. 1

[15] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu. Shape and appearance context modeling. In *ICCV*, 2007. 1, 2

[16] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013. 5

[17] W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. In *CVPR*, 2011. 1, 2