2013 Seventh International Conference on Image and Graphics

Enhancing Person Re-identification by Robust Structural Metric Learning

Gang Yuan, Zhaoxiang Zhang, Yunhong Wang
SCSE, Beihang University, Beijing, China
zxzhang@buaa.edu.cn

***Abstract*—Person re-identification has become an important but also challenging task for video surveillance systems as it aims to match people across non-overlapping camera views. So far, most successful methods either focus on robust feature representation or sophisticated learners. Recently, metric learning has been applied in this task which aims to find a suitable feature subspace for matching samples from different cameras. However, most metric learning approaches rely on either pairwise or triplet-based distance comparison, which can be easily over-fitting in large scale and high dimension learning situation. Meanwhile, the performance of these methods can significantly decrease when the extracted features contain noisy information. In this paper, we propose a robust structural metric learning model for person re-identification with two main advantages: 1) it applies loss functions at the level of rankings rather than pairwise distances; 2) the proposed model is also robust to noisy information of the extracted features. The approach is verified on two available public datasets, and experimental results show that our method can get state-of-the-art performance.**

***Keywords*-person re-identification; structural metric learning; input sparsity; robust;**

I. Introduction

Matching people across non-overlapping camera views, known as person re-identification, has become one of the most important tasks in video surveillance systems. A successful person re-identification method needs to solve the following problem: when a target person disappears from one camera view, he or she can be identified in another view among a crowd of candidates. However, person re-identification is rather challenging due to several reasons. First, in a crowded public space monitored by cameras, it is difficult to obtain reliable biometric information, so traditional recognition techniques (e.g., face and gait recognition) are not suitable to deal with person re-identification. Second, due to changes in view angle, illumination, background clutter and occlusion, the appearance of target person from different camera views may be quite dissimilar while there may exist several non-target person whose visual appearance is quite similar to the target person (see Figure 1 (a)). Third, different from similar large scale search problems, its difficult to utilize accurate temporal and spatial constraints to ease the task. Thus, person re-identification has aroused an increased scientific interest recent years due to the need of practical applications and still unresolved difficulties.

So far, existing methods for person re-identification can be divided into three groups according to the difference of emphasis: 1) robust feature extraction [3], [4], [15], 2) discriminative learning models [6], [12], and 3) metric learning [17], [2]. The first group of methods tries to extract visual features that are distinctive, and simultaneously robust to large changes of appearance caused by different camera views. However, it is difficult to find a distinctive and relatively stable feature representation due to large intra-class variations(see Figure 1 (b)). In order to overcome above limitations, many researchers attempt to build discriminative models to utilize discriminative information from the training data. However, in real applications, this kind of methods may lose efficacy because discriminative information obtained from one camera view may be not discriminative anymore in another camera view (e.g., a bright-colored handbag in one view may be invisible in another view, see Figure 1 (c)). Recently, metric learning has been applied in person re-identification. In contrast to directly using Euclidean distance for matching, metric learning aims to find an optimal metric that can describe the transition in feature space between different camera views. However,

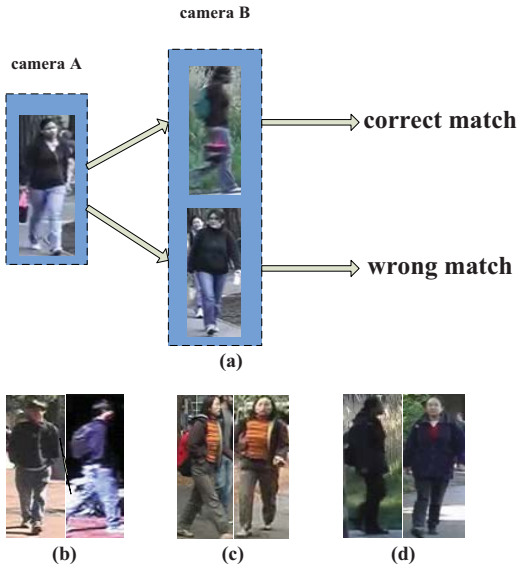


Figure 1. (a) shows an example for person re-identification problem. (b), (c) and (d) show image pairs from two different camera views which denote the challenges for person re-identification respectively.

978-0-7695-5050-3/13 $26.00 © 2013 IEEE
DOI 10.1109/ICIG.2013.99

most existing metric learning methods [1], [9] use direct or relative distance comparison by sampling pairs or triplets from the training set. This strategy can easily lead to over-fitting especially in large scale and high dimension learning situation. Moreover, existing metric learning methods are not robust to noisy information from the extracted features. This may result in obvious effectiveness decrease especially in person re-identification (e.g., features contain much noisy information due to variations of illumination, background clutter and occlusion, see Figure 1 (d)).

In this paper, we build a structural metric learning [10], [11] framework to solve the person re-identification problem. Compared to traditional metric learning approaches, structural metric learning has the following advantages: 1) structural metric learning allow us to apply loss functions at the level of rankings, rather than pairwise distances; 2) the learned metric can be robust to noisy information of features by imposing a group sparsity penalty on the learned metric to promote input sparsity, and a trace penalty to promote output sparsity. Experimental results show that state-of-the-art performance can be obtained by formulating person re-identification as a structural metric learning problem.

II. Related work

There is a lot of work to solve person re-identification by designing a feature representation that can be both distinctive and robust to large appearance variations caused by changes of camera views. For instance, Gheissari et al. [4] present a novel application of triangulated model fitting to people that can deal with pose variations. Wang et al. [15] divide the image into several regions and capture their color spatial structure by a co-occurrence matrix. Farenzena et al. [3] try to utilize a strategy to extract distinctive and stable features. This strategy is based on the localization of perceptual relevant human parts, driven by asymmetry/symmetry principles.

Besides above approaches, there are also several methods trying to learn a discriminative learning model. For instance, Gray et al. [6] propose to select a subset of local features to match people by boosting. Prosser et al. [12] formulate person re-identification as a ranking problem. The authors use RankSVM to learn a feature subspace that the target person can get the highest rank.

Recent years, metric learning has been applied in this area. Dikmen et al. [2] apply a metric learning framework to obtain a efficient distance metric for large margin nearest neighbor classification. Zheng et al. [17] treat person re-identification as a relative distance comparison problem. The proposed method tries to maximize the probability that a true match pair have smaller distance than a wrong match pair. However, these methods are built on pairwise distances and they don't take the noisy information of features into consideration.

III. Robust Structural Metric Learning for Person Re-identification

Person re-identification can be formulated as the following metric learning problem: given an image x of person A (x is represented by a feature vector, i.e. $x \in R^d$), x' is another image of person A, x'' is an image of any other person. We want that the distance between x and x' is smaller than that between x and x''. So a distance function $d(\cdot,\cdot)$ should be learned to satisfy $d(x,x') < d(x,x'')$. One popular approach for metric learning is to learn a Mahalanobis distance metric:

$$d_M(x,x') = (x-x')^T M(x-x') \tag{1}$$

M is a positive semi-definite matrix, i.e., $M \succeq 0$. Once a distance metric M has been learned, we can use nearest neighbor search to identify the target person by calculating the distances between probe image(i.e., target person from one camera view) and gallery images(i.e., all the candidates from another camera view).

A. Structural Metric Learning

Before we introduce structural metric learning, some preliminary definitions need to be made: Let $\mathbb{S}^d$ and $\mathbb{S}^d_+$ denote the sets of $d \times d$, real-valued, symmetric and positive semi-definite matrices. For matrices A and B, the Frobenius inner product is denoted by $\langle A,B\rangle_F := \Sigma_{i,j}A_{ij}B_{ij}$, and is normed by $\|A\|_F := \sqrt{\langle A,A\rangle_F}$.

Structural metric learning is based on structural SVM [14] and aims to optimize $M \in \mathbb{S}^d_+$ to minimize a ranking loss function $\Delta : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ over permutations $\mathcal{Y}$ induced by distance, which can be selected according to different ranking measures.

In general, Structural metric learning can be expressed as the following convex optimization problem:

$$\begin{aligned} \min_{M \in \mathbb{S}^d_+} & \; tr(M) + \frac{C}{n}\sum_{q\in\mathcal{X}} \xi_q \\ & s.t. \forall q \in \mathcal{X}, y \in \mathcal{Y}: \\ \langle M, \psi(q,y_q) - \psi(q,y)\rangle_F & \geq \Delta(y_q,y) - \xi_q \end{aligned} \tag{2}$$

here, $\mathcal{X} \subset \mathbb{R}^d$ denotes the whole training set of n data points; $\mathcal{Y}$ is the set of all permutations over $\mathcal{X}$; $C > 0$ is a slack trade-off parameter; $\psi : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{S}^d$ denotes a feature map which encodes an input-output pair(q,y); $\Delta(y_q,y) \in [0,1]$ is a desired margin indicating the similarity between two rankings, i.e., loss incurred for predicting a ranking y rather than the ground truth ranking y_q. The feature map ψ [7] is designed to make sure $\langle M, \psi(q,y)\rangle_F$ is large when the ranking of $\mathcal{X}$ induced by distance from q agrees with y, and small otherwise. For a query $q \in \mathbb{R}^d$, $\mathcal{X}^+_q \subseteq \mathcal{X}$ denotes the relevant set of q, and $\mathcal{X}^-_q \subseteq \mathcal{X}$ denotes the irrelevant set. ψ

is called partial order feature, and defined by

$$\psi(q,y) := \sum_{i\in\mathcal{X}_q^+}\sum_{j\in\mathcal{X}_q^-} y_{ij}\frac{\phi(q,x_i)-\phi(q,x_j)}{\mid \mathcal{X}_q^+ \mid \cdot \mid \mathcal{X}_q^- \mid}$$
$$\phi(q,x) := -(q-x)(q-x)^T, \quad (3)$$
$$y_{ij} := \begin{cases} +1 & \text{if } i \prec_y j \\ -1 & \text{otherwise} \end{cases}$$

and $\phi(q,i)$ is a feature map which encodes the relationship between a query q and a data point i. The regularization term $tr(M)$ in (2) is used as a convex surrogate for $rank(M)$ to promote sparsity and low-rank solutions. However, the objective function ignores the off-diagonal elements of M, so it can't necessarily promote input sparsity, and performance can be influenced by noisy information of features.

B. Robust Structural Metric Learning

As mentioned in the introduction section, for person re-identification problem, the extracted low-level features may contain much noise due to background, occlusion and variation of illumination. So we want the learning algorithm to provide a distance metric M which only relies on informative features. Furthermore, if some input dimension i is non-informative, then the corresponding rows and columns of M should suppress the effectiveness of feature, i.e., $M_{i\cdot} = M_{\cdot i} = 0$. However, it is not reasonable to enforce sparsity for rows corresponding to informative features, as the ability of the learning algorithm to exploit correlation between informative features will be limited and output sparsity will not be promoted. This suggests that we can make a natural row (or column) grouping of the entries of M when enforcing sparsity. As a result, rows corresponding to informative features may be dense, and sparsity is still enforced over rows to avoid relying upon too many features.

As in the group lasso, row-sparsity can be promoted by mixed-norm regularization [16]:

$$\|M\|_{2,1} := \sum_{i=1}^{d}\|M_{i\cdot}\|_2 \quad (4)$$

so the robust structural metric learning formulation is finally defined by

$$\min_{M\in\mathbb{S}_+^d} tr(M) + \lambda\|M\|_{2,1} + \frac{C}{n}\sum_{q\in\mathcal{X}}\xi_q$$
$$s.t. \forall q\in\mathcal{X}, y\in\mathcal{Y}: \quad (5)$$
$$\langle M, \psi(q,y_q)-\psi(q,y)\rangle_F \geq \Delta(y_q,y) - \xi_q$$

The robust structural metric learning balances the trade-off between input and output sparsity through a hyper-parameter $\lambda > 0$ which can be tuned by cross-validation.

The general optimization procedure is listed as Algorithm 1. In order to make compactness, we define

$$\delta\psi(q,y^*,y) = \psi(q,y^*) - \psi(q,y) \quad (6)$$

Algorithm 1 Robust Structural Metric Learning for Person Re-identification

Input: data $\mathcal{X}$ (n data points), ground truth rankings $y_1^*,\cdots,y_n^*$, slack trade-off $C>0$, hyper-parameter $\lambda > 0$, accuracy threshold $\epsilon > 0$

Output: metric $M \succeq 0$, slack variable $\varepsilon \geq 0$

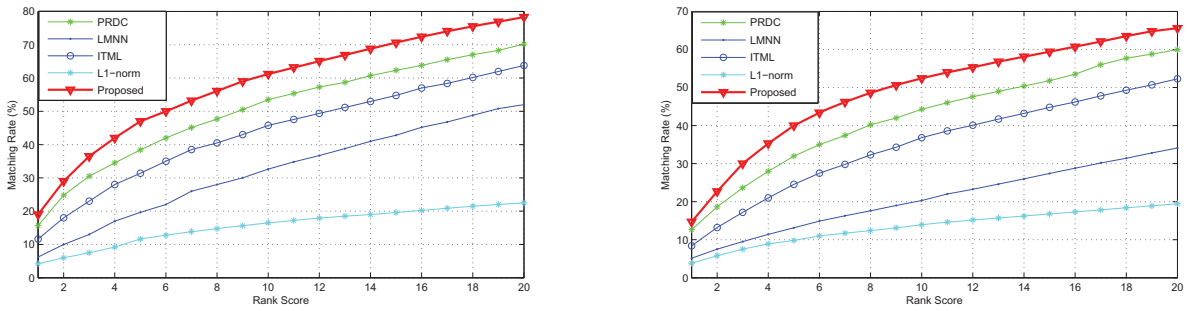
1: $\mathcal{C} \leftarrow \varnothing$
2: **repeat**
3: Solve for the optimal metric and slack:
$(M,\varepsilon) \leftarrow argmin_{M,\varepsilon} f(M,\varepsilon) = tr(M) + \lambda\|M\|_{2,1} + C\varepsilon$
s.t. $M\succeq 0, \varepsilon \geq 0$
$\forall(y_1,y_2,\cdots,y_n)\in\mathcal{C}$:
$\frac{1}{n}\sum_{i=1}^n\langle M, \delta\psi(q_i,y_i^*,y_i)\rangle_F \geq \frac{1}{n}\sum_{i=1}^n\Delta(y_i^*,y_i) - \varepsilon$
4: **for** $i=1$ to n **do**
5: $y_i \leftarrow argmax_{y\in\mathcal{Y}}\Delta(y_i^*,y) + \langle M, \psi(q_i,y)\rangle_F$
6: **end for**
7: $\mathcal{C} \leftarrow \mathcal{C}\cup(y_1,\cdots,y_n)$
8: **until** $\frac{1}{n}\sum_{i=1}^n\Delta(y_i^*,y_i) - \langle M,\delta\psi(q_i,y_i^*,y_i)\rangle_F \leq \varepsilon+\epsilon$

Note that $|\mathcal{Y}|$ (number of constraints in (5)) is super-exponential in the size of the training set, this makes the optimization procedure unpractical with current techniques. Here, we can use 1-Slack margin-rescaling cutting-plane algorithm [8] to make the optimization process feasible. The algorithm tries to alternate between optimizing the metric M, and updating the constraint set $\mathcal{C}$ with a new batch of rankings $(y_1,y_2,\cdots,y_n)$, which is most violated by the resulting M. The optimization process will terminate when the most-violated constraint is satisfied (ϵ defines). The implementation of step (3) for finding the optimal M in each iteration can be realized through a gradient decent solver. After each gradient step, the updated M is projected back onto $\mathbb{S}_+^d$ by spectral decomposition. However, it can be computationally expensive when projecting M back onto $\mathbb{S}_+^d$ after each gradient step. So the alternating direction method of multipliers (ADMM) has been proposed to make the optimization process more efficient. Due to space limitation, we would not show the completely optimization process, details can be found in [10].

IV. Experiments

A. Datasets

We evaluate our method on two available public datasets: the VIPeR dataset [5] and the ETHZ dataset [13]. The VIPeR dataset contains 632 pairs of images taken from two camera views(i.e., each person has 2 images). This dataset is challenging due to large variations in viewpoint, illumination and pose, but with less occlusion. The ETHZ dataset contains several video sequences of urban scenes captured by moving cameras which is originally designed



(a) p=316 (b) p=432

Figure 2. Performance comparison using CMC curves on VIPeR dataset

Table I
TOP RANKED MATCHING RATE (%) ON VIPER DATASET. p IS THE NUMBER OF PERSON IN THE TESTING SET; r IS THE RANK.

Methods	p=316				p=432			
	r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20
Proposed	**19.1**	**46.9**	**60.8**	**78.2**	**14.2**	**39.9**	**52.1**	**65.6**
PRDC	15.6	38.4	53.8	70.1	12.6	31.9	44.3	59.9
LMNN	6.2	19.6	32.6	52.3	5.1	13.1	20.3	33.9
ITML	11.6	31.4	45.8	63.9	8.4	24.5	36.8	52.3
L1-norm	4.2	11.6	16.5	22.4	3.8	9.8	13.9	19.4

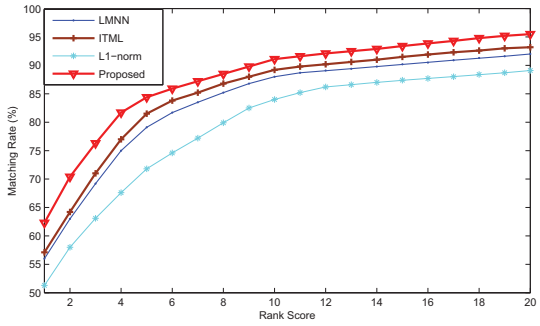


Figure 3. Performance comparison using CMC curves on ETHZ dataset

for pedestrian tracking. Later Schwartz and Davis [13] extract images of a set of people from the video sequences to build a new dataset for person re-identification. This dataset contains 3 video sequences, 146 people and 8555 images in total. Sequence 1 contains 83 persons (4857 images), sequence 2 contains 35 persons (1961 images), and sequence 3 contains 28 persons (1762 images). This dataset contains large illumination changes and occlusion but with less viewpoint changes, as the images for one person are captured by a single moving camera. As two datasets have different characteristic, it is suitable to use them to evaluate a person re-identification method.

B. Feature Representation

Color and texture features has been proved to be effective for person re-identification. In order to compare with other state-of-the-art methods, the feature representation is same to that used in [17], [12]. This feature representation contains color features (RGB, YCbCr, HSV) and texture features (Schmidt and Gabor filters), and all these are put together to form a feature vector of 2784 dimensions. Note that all these features are low-level and widely used, so this representation is both generic and representative.

C. Experiment Setup

For robust structural metric learning, the loss function Δ are fixed to mean average precision (MAP), and trade-off $\mathcal{C}$ is varied over $\{1, 10, \cdots, 10^5\}$, and the hyper-parameter λ is chosen from $\{10^{-2}, 10^{-1}, 1\}$, and the accuracy parameter ϵ is set as 10^{-5}.

The setting for VIPeR dataset is similar to [17], we random select p people of images for testing and the rest for training. The test images are divided into a gallery set and a probe set. The gallery set consists of one image for each person, and the remaining images form the probe set. In this setting, larger p means smaller training set. For ETHZ dataset, we randomly select 5 images for each person and use half of the people of images for training and the rest for testing. All the experiment procedure is repeated 10 times and the average results are reported.

For evaluation, we use the average cumulative match characteristic (CMC) curves to show the ranked matching

rates which is widely used in person re-identification. A rank r matching rate indicates the percentage of the probe images with correct matches found in the top r ranks against the p gallery images. Rank 1 matching rate is actually the correct matching rate. Note that in practice, although a high rank 1 matching rate is critical, it is also important to get the top r ranked matching rate with a relatively small r value because the top matched images can be verified using temporal reasoning or human operator.

D. Experimental Results on VIPeR Dataset

We compare our model with the following methods: LMNN [9], ITML [1] and PRDC [17]. Here, we also report the results of L1-norm as a baseline. LMNN and ITML are both state-of-the-art metric learning methods which have been proved effective in many application areas. PRDC is a metric learning method specially designed for person re-identification which aims to maximise the probability of a pair of true match having a smaller distance than that of a wrong match pair. This makes PRDC less over-fitting than other metric learning methods for person re-identification task.

Experimental results on VIPeR dataset (see Figure 2) clearly show that our model can significantly outperform other methods in all ranks. This infer that it is more efficient to optimize a ranking list than directly utilizing a pairwise distance comparison. Besides, as our model can ensure input and output sparsity, this infer that noisy information removal can also benefit our model as the extracted low-level features may contain some noisy information. As PRDC is claimed to particularly effective when a training set is small, we also implement our method in a smaller size ($p = 432$) of training set. A more specific experimental results (see Table 1) show the robustness of our model as it can still outperform other methods. Our model can still get state-of-the-art performance even if there are less training samples. This infer that our model is also robust to the under-sampling of data .

E. Experimental Results on ETHZ Dataset

For ETHZ dataset, as the implementation of PRDC is not available, we just compare our model with LMNN and ITML, and also make the results of L1-norm as a baseline. As mentioned above, all the person images are captured from a single moving camera which may not reflect the practical situation, so there are less viewpoint changes in this dataset. The effectiveness of metric learning may decrease as it aims to capture the transitions in inter-camera feature space. However, experimental results (see Figure 3) show that our model can still outperform other metric learning methods and state-of-the-art performance can be obtained. As the proposed model optimizes the loss function at the level of rankings, this infers that it can take full advantage of multiple images for one person. Besides, the robustness to noisy features also benefits our model in another way.

V. Conclusion

In this paper, we proposed a novel structural metric learning framework to tackle person re-identification. Our model aims to learn an optimal distance metric by applying loss function at the level of rankings, rather than pairwise distances. Meanwhile, our model is also robust to noisy information of the extracted features, and this is realized by promoting both input and output sparsity. Experimental results clearly show that our method can outperform both state-of-the-art metric learning methods and methods specially designed for person re-identification. It is also demonstrated that our method can still get state-of-the-art performance even if the size of the training set is small and this infers the robustness of our model to under-sampling of data.

VI. Acknowledgement

This work is funded by the National Basic Research Program of China (No. 2010CB327902), the National Natural Science Foundation of China (No. 61005016), the Open Projects Program of National Laboratory of Pattern Recognition, and the Fundamental Research Funds for the Central Universities.

References

[1] J. Davis, B. Kulis, P. Jain, S. Sra, and I. Dhillon. Information theoretic metric learning. *In ICML*, 2007.
[2] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja. Pedestrian recognition with a learned metric. *In ACCV*, 2010.
[3] M. Farenzena, L. Bazzan, A. Perina, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. *In CVPR*, 2010.
[4] N. Gheissari, T. Sebastian, and R. Hartley. Person reidentification using spatiotemporal appearance. *In CVPR*, 2006.
[5] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. *In IEEE International workshop on performance evaluation of tracking and surveillance*, 2007.
[6] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. *In ECCV*, 2008.
[7] T. Joachims. A support vector method for multivariate performance measures. *In ICML*, 2005.
[8] T. Joachims, T. Finley, and C.-N. J. Yu. Cutting-plane training of structural svms. *Machine Learning*, 77(1):27–59, 2009.
[9] K.Weinberger, J. Blitzer, and L. Saul. Distance metric learning for large margin nearest neighbor classification. *In NIPS*, 2006.
[10] D. K. H. Lim, B. McFee, and G. Lanckriet. Robust structural metric learning. *In ICML*, 2013.
[11] B. McFee and G. Lanckriet. Metric learning to rank. *In ICML*, 2010.
[12] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang. Person reidentification by support vector ranking. *In BMVC*, 2010.
[13] W. Schwartz and L. Davis. Learning discriminative appearance-based models using partial least squares. *In Brazilian Symposium on Computer Graphics and Image Processing*, 2009.

[14] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6:1453–1484, 2005.
[15] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu. Shape and appearance context modeling. *In ICCV*, 2007.
[16] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68(1):49–67, 2006.
[17] W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. *In CVPR*, 2011.