2016 IEEE Conference on Computer Vision and Pattern Recognition

Joint Learning of Single-image and Cross-image Representations for Person Re-identification

Faqiang Wang[1,2], Wangmeng Zuo[1], Liang Lin[3], David Zhang[1,2], Lei Zhang[2]
[1]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
[2]Dept. of Computing, The Hong Kong Polytechnic University, Hong Kong, China
[3]Sun Yat-sen University, Guangzhou, China
tshfqw@163.com, cswmzuo@gmail.com, linliang@ieee.org, {csdzhang, cslzhang}@comp.polyu.edu.hk

Abstract

Person re-identification has been usually solved as either the matching of single-image representation (SIR) or the classification of cross-image representation (CIR). In this work, we exploit the connection between these two categories of methods, and propose a joint learning framework to unify SIR and CIR using convolutional neural network (CNN). Specifically, our deep architecture contains one shared sub-network together with two sub-networks that extract the SIRs of given images and the CIRs of given image pairs, respectively. The SIR sub-network is required to be computed once for each image (in both the probe and gallery sets), and the depth of the CIR sub-network is required to be minimal to reduce computational burden. Therefore, the two types of representation can be jointly optimized for pursuing better matching accuracy with moderate computational cost. Furthermore, the representations learned with pairwise comparison and triplet comparison objectives can be combined to improve matching performance. Experiments on the CUHK03, CUHK01 and VIPeR datasets show that the proposed method can achieve favorable accuracy while compared with state-of-the-arts.

1. Introduction

Person re-identification is the task of matching two pedestrian images from different viewpoints [11]. It has attracted increasing interests and encouraged considerable efforts in recent years due to its broad applications in video surveillance [34, 33]. This problem, however, is still very challenging and deserves further studies, because of the large variations in illumination, poses, viewpoints and background of pedestrian images.

The task of person re-identification can be accomplished by two categories of methods: (i) distance or similarity measures on single-image representation, which is the representation of a given image [13, 14, 16, 28, 20, 17, 26, 7] and (ii) classification on cross-image representation, which is the representation of an image pair [19, 1, 24]. For the first category of methods, single-image representation (SIR) is first obtained using either hand-crafted feature [13, 14, 16, 28, 20, 17, 40, 36, 21, 22] or deep convolutional neural network (CNN) approaches [7, 37, 38], and then a distance measure together with a threshold is utilized to predict whether two pedestrian images are matched or not. For the second category of methods, after obtaining the cross-image representation (CIR), person re-identification can be regarded as an ordinary binary classification task [1, 19, 4, 24].

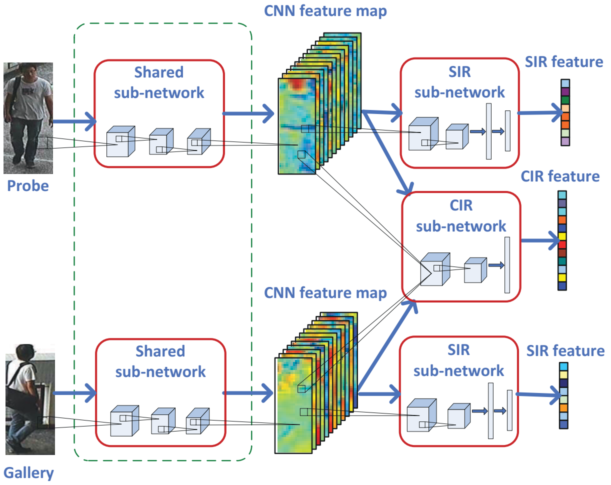

Figure 1. The sketch of the network for learning the single-image and cross-image representations.

These two categories of methods have their own advantages. The SIR has some outstanding advantages in terms of efficiency. Given a gallery set of N images, one can precompute their SIRs in advance. In the matching stage, we only need to extract the SIR of the probe image and compute

1063-6919/16 $31.00 © 2016 IEEE
DOI 10.1109/CVPR.2016.144

its distances to the SIRs of the gallery images, while for CIR classification method we need to extract the CIR between the probe image and each gallery image (*i.e.*, N times). On the other hand, compared with SIR, CIR is effective in capturing the relationships between the two images, and several approaches have been suggested to address horizonal displacement by local patch matching. Therefore, the SIR and CIR have their respective advantages and this finding inspires us to investigate a comprehensive way of combining these two representations in terms of both effectiveness and efficiency.

In this work, we study the connection between SIR and CIR, and propose a joint learning framework with deep CNN to exploit the advantages of these two categories of representation methods. Denote by $\mathbf{x}_i$ and $\mathbf{x}_j$ two pedestrian images. We adopt the following classifier based on the cross-image representation $g(\mathbf{x}_i, \mathbf{x}_j)$:

$$S_{\text{CIR}}\left(\mathbf{x}_i, \mathbf{x}_j\right) = \mathbf{w}^T g\left(\mathbf{x}_i, \mathbf{x}_j\right) - b \tag{1}$$

and use the Euclidean distance to measure the dissimilarity between the SIRs of $\mathbf{x}_i$ and $\mathbf{x}_j$:

$$S_{\text{SIR}}\left(\mathbf{x}_i, \mathbf{x}_j\right) = \left\| f\left(\mathbf{x}_i\right) - f\left(\mathbf{x}_j\right)\right\|_2^2 \tag{2}$$

where $(\mathbf{w}, b)$ is the parameter of the classifier, $f\left(\mathbf{x}_i\right)$ and $f\left(\mathbf{x}_j\right)$ are the SIRs of $\mathbf{x}_i$ and $\mathbf{x}_j$, respectively, and $\|\centerdot\|_2$ denotes the L_2 norm. With $S_{\text{SIR}}\left(\mathbf{x}_i, \mathbf{x}_j\right)$, a threshold t_S is introduced to predict whether the two pedestrian images are from the same person.

We show that classification on CIR is the generalization of conventional similarity measures based on SIR. Denote by $[\centerdot]_{\text{vec}}$ the vector form of a matrix. Using the Euclidean distance in (2) as an example, it is obvious to see that $S_{\text{SIR}}\left(\mathbf{x}_i, \mathbf{x}_j\right)$ is a special case of $S_{\text{CIR}}\left(\mathbf{x}_i, \mathbf{x}_j\right)$ with $\mathbf{w} = [\mathbf{I}]_{\text{vec}}$, $b = t_S$, and $g\left(\mathbf{x}_i, \mathbf{x}_j\right) = [\left(f(\mathbf{x}_i) - f(\mathbf{x}_j)\right)\left(f(\mathbf{x}_i) - f(\mathbf{x}_j)\right)^T]_{\text{vec}}$, where $\mathbf{I}$ is the identity matrix. As illustrated in Sect. 3.1, other distance or similarity measures [20, 3] are also special cases of CIR.

By using the deep CNN architecture, we propose a framework to jointly learn SIR and CIR for improving matching performance with the least increase of the computational cost. As illustrated in Fig. 1, our network consists of three sub-networks, *i.e.* first one shared sub-network and followed by two sub-networks for extracting SIR and CIR features, respectively. To save the computational cost, we can store the CNN feature maps from the shared sub-network of the gallery images in advance, and reduce the depth of the CIR sub-network to include only one convolutional layer and one fully-connected layer. In the test stage, the shared feature maps and SIR of each probe image are required to be computed one time, and only the CIR sub-network is used to compute the CIR between the probe image and each gallery image. Thus we can exploit the CIR to improve the matching accuracy, while exploiting the SIR and shared sub-network to reduce the computational cost.

Furthermore, we extend our model by utilizing two different deep CNNs for joint SIR and CIR learning based on either pairwise comparison objective or triplet comparison objective, respectively. For the pairwise comparison based network, we learn the CIR by standard support vector machine (SVM) [32]. For the triplet comparison based network, we learn the CIR by ranking SVM (RankSVM) [30]. Finally we combine the matching scores of these two networks together as the similarity of the image pair.

Experiments have been conducted on several public datasets for person re-identification, *i.e.* CUHK03 [19], CUHK01 [18] and VIPeR [12]. The results show that, joint SIR and CIR learning is effective in improving the person re-identification performance, and the matching accuracy can be further boosted by combining the learned models based on pairwise and triplet comparison objectives. Compared with the state-of-the-art approaches, the proposed methods perform favorably in person re-identification.

The rest of this paper is organized as follows. Section 2 reviews the related work. Section 3 describes the proposed model. Section 4 presents the deep network architecture. Section 5 reports the experimental results, and Section 6 concludes this paper.

2. Related Work

The existing person re-identification methods can be divided into two categories depending on whether they use the hand-crafted or deep CNN features. There have been many kinds of hand-crafted features used for person re-identification, including local binary patterns (LBP) [36, 16], color histogram [16] and local maximal occurrence (LOMO) [21, 22]. For the methods based on hand-crafted features, they usually focus on learning an effective distance/similarity metric to compare the features. For the methods based on deep CNN features, feature representation and classifier can be jointly optimized for learning either SIR or CIR features. This section will provide a brief review on these methods.

2.1. Metric Learning for Person Re-identification

Many distance metric learning methods have been developed for person re-identification. They aim to learn a distance metric to reduce the distance of the matched images, and enlarge the distance of the mismatched images. Among the existing distance metric learning methods, some of them are based on pairwise comparison constraint. Guillaumin *et al.* proposed a logistic discriminant metric learning (LDML) model by modeling the probability of a given sample pair $(\mathbf{x}_i, \mathbf{x}_j)$ and used the maximum log-likelihood as the objective function [13]. Following the keep-it-simple-and-straight forward (KISS) principle, Köstinger *et al.* proposed a KISS metric learning (KISSME) method to address the scalability issue of metric learning from equivalence constraints [16]. Li *et al.* developed a generalized similarity metric for person re-identification by introducing an adaptive threshold into Mahalanobis distance [20]. Li and Wang introduced the locally aligned feature transform to match the person images across camera views [17]. Liao *et al.* improved the KISSME method by learning a discriminant low dimensional subspace [21] based on the LOMO features. They also improved the LDML model by enforcing the positive semidefinite constraint and the asymmetric sample weighting strategy [22]. Some other works, including pairwise constrained component analysis (PCCA) [28], local Fisher discriminant analysis (LFDA) [29] and information-theoretic metric learning (ITML) [5], are also based on the pairwise comparison constraints.

Apart from the methods based on pairwise comparison constraints, some other methods are based on the triplet comparison constraints. Weinberger *et al.* proposed a large margin nearest neighbor (LMNN) model [35], where the distance metric is learned to separate the matched neighbors from the mismatched ones by a large margin. Dikmen *et al.* improved LMNN by adding the option of rejection [6]. Zheng *et al.* developed a person re-identification model by maximizing the likelihood that each sample is more closed to its matched sample than its mismatched sample [43].

2.2. Deep Learning for Person Re-identification

Due to the power of deep CNNs in learning discriminative features from large-scale image data, many methods have adopted the deep architecture to jointly learn the representation and the classifier [1, 4, 19, 37, 31]. Some of them focus on learning the SIR together with the similarity function. Schroff *et al.* proposed a FaceNet model for face verification [31], which adopts a deep CNN to learn the Euclidean embedding per image by using the triplet comparison loss. Online triplet generation is also developed to gradually increase the difficulty of the triplets in training. Ding *et al.* proposed a deep SIR learning model based on relative distance comparison for person re-identification [7]. It first presents an effective triplet generation strategy to construct triplets, which contains one image with a matched image and a mismatched image. For each triplet, this model learns the SIR by maximizing the relative distance between the matched pair and the mismatched pair.

Despite learning SIR, some other methods are suggested to perform person re-identification based on CIR. Li *et al.* proposed a filter pairing neural network (FPNN) [19], which learns the CIRs by a patch matching layer followed by a maxout-grouping layer. In FPNN, the patch matching layer is used to model the displacement of each horizontal stripe in the images across views, the maxout-grouping layer improves the robustness of patch matching, and finally a softmax classifier is imposed on the learned CIR for person re-identification. The work in [1] shares the similar idea, but introduces a new layer to learn the cross-image representation by computing the neighborhood difference between two input images. The work in [4] learns the CIR by formulating the person re-identification task as a learning-to-rank problem. For each image pair, this model first stitchs its two images horizontally to form a holistic image, then feeds these images to a CNN to learn their representations. Finally the ranking loss is used to ensure that each sample is more similar to its positive matched image than its negative matched image. Liu *et al.* proposed a Matching CNN (M-CNN) architecture for human parsing [24], which learns the CIR of the image and a semantic region by a multi-layer cross image convolutional path to predict their matching confidence and displacements.

There are considerable differences between the proposed deep architecture and the previous networks. First, both SIR and CIR can be jointly learned with the proposed deep architecture, while only SIR is learned in [31, 7] and only CIR is learned in [24, 19, 4, 1]. Second, to improve the computational efficiency, we restrict the depth of the CIR sub-network with only a convolutional layer and a fully-connected layer. In contrast, multiple convolutional and fully-connected layers are adopted in [24, 19, 4, 1] for CIR learning. Besides, we present two deep CNN architectures for joint SIR and CIR learning based on pairwise and triplet comparison objectives, respectively, and the matching scores of these two networks can be combined to improve the matching accuracy.

3. Joint SIR and CIR Learning

In this section, we first discuss the connections between SIR and CIR, then propose two formulations (*i.e.* pairwise comparison formulation and triplet comparison formulation) for joint SIR and CIR learning, and finally introduce the matching scores for person re-identification.

3.1. Connection between SIR and CIR

With the SIR features, there are four commonly used distance/similarity measures for person re-identification, *i.e.* Euclidean distance, Mahalanobis distance, joint Bayesian [3], and LADF [20]. As explained in Sect. 1, Euclidean distance on SIRs can be regarded as a special case of CIR-based classification. In the following, we will show that the other measures are also special cases of CIR-based classification.

The Mahalanobis distance based on the SIR $\mathbf{z}_i = f(\mathbf{x}_i)$ can be formulated as $s(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{z}_i - \mathbf{z}_j)^T \mathbf{M} (\mathbf{z}_i - \mathbf{z}_j)$, where $\mathbf{M}$ is positive semi-definite. This formulation is equivalent to (1) when $\mathbf{w} = [\mathbf{M}]_{\text{vec}}$ and $g(\mathbf{x}_i, \mathbf{x}_j) = \left[(\mathbf{z}_i - \mathbf{z}_j)(\mathbf{z}_i - \mathbf{z}_j)^T\right]_{\text{vec}}$.

The joint Bayesian formulation [3] is defined as follows,

$$s\left(\mathbf{x}_i,\mathbf{x}_j\right)=\mathbf{z}_i^T\mathbf{A}\mathbf{z}_i+\mathbf{z}_j^T\mathbf{A}\mathbf{z}_j-2\mathbf{z}_i^T\mathbf{G}\mathbf{z}_j \tag{3}$$

which is the generalization of Mahalanobis distance. By setting $\mathbf{w}=\left([\mathbf{A}]_{\text{vec}}^T \quad [\mathbf{G}]_{\text{vec}}^T\right)^T$ and $g\left(\mathbf{x}_i,\mathbf{x}_j\right)=\left(\left[\mathbf{z}_i\mathbf{z}_i^T+\mathbf{z}_j\mathbf{z}_j^T\right]_{\text{vec}}^T \quad \left[-2\mathbf{z}_j\mathbf{z}_i^T\right]_{\text{vec}}^T\right)^T$ in (1), joint Bayesian can be regarded as a classifier $\mathbf{w}$ on the CIR $g(\mathbf{x}_i,\mathbf{x}_j)$.

The LADF [20] is defined as follows,

$$\begin{aligned} s\left(\mathbf{x}_i,\mathbf{x}_j\right)=&\frac{1}{2}\mathbf{z}_i^T\mathbf{A}\mathbf{z}_i+\frac{1}{2}\mathbf{z}_j^T\mathbf{A}\mathbf{z}_j+\mathbf{z}_i^T\mathbf{B}\mathbf{z}_j \\ &+\mathbf{c}^T\left(\mathbf{z}_i+\mathbf{z}_j\right)+b \end{aligned} \tag{4}$$

which is the generalization of Mahalanobis distance and joint Bayesian. It can also be viewed as a special case of (1) when $\mathbf{w}=\left([\mathbf{A}]_{\text{vec}}^T \quad [\mathbf{B}]_{\text{vec}}^T \quad \mathbf{c}^T \quad b\right)^T$ and $g(\mathbf{x}_i,\mathbf{x}_j)=\left(\frac{1}{2}\left[\mathbf{z}_i\mathbf{z}_i^T+\mathbf{z}_j\mathbf{z}_j^T\right]_{\text{vec}}^T \quad \left[\mathbf{z}_j\mathbf{z}_i^T\right]_{\text{vec}}^T \quad \left(\mathbf{z}_i+\mathbf{z}_j\right)^T \quad 1\right)^T$.

Despite the connections between SIR and CIR, they do have their own advantages and can be combined to improve the matching performance. For the SIR-based method, the SIR features of the gallery set can be precomputed in advance. For each probe image, we only require extract its SIR and compute its distance/similarity measure to the precomputed SIRs from the gallery images, making SIR computationally efficient for person re-identification. The CIR-based method can effectively model the complex relationships between the gallery and probe images, and is robust to spatial displacement and changed views. In the following, we will investigate the loss for joint SIR and CIR learning and design proper network architecture by considering both accuracy and efficiency factors.

3.2. Pairwise Comparison Formulation

Denote by $\{((\mathbf{x}_i,\mathbf{x}_j),h_{ij})\}$ the doublet training set, where $\mathbf{x}_i$ and $\mathbf{x}_j$ are the ith and jth training samples, respectively. h_{ij} is the label assigned to the doublet $(\mathbf{x}_i,\mathbf{x}_j)$. If $\mathbf{x}_i$ and $\mathbf{x}_j$ are from the same class, then $h_{ij}=1$, otherwise $h_{ij}=-1$. Let $f\left(\mathbf{x}_i\right)$ be the SIR of $\mathbf{x}_i$ and b_{SIR} be a distance threshold. In the pairwise comparison formulation, the similarity of the positive pair is expected to be higher than a given threshold, while the similarity of the negative pairs is expected to be lower than the threshold. The Euclidean distance of the SIRs for any doublet $(\mathbf{x}_i,\mathbf{x}_j)$ should satisfy the constraints as follows:

$$\begin{aligned} &\left\|f\left(\mathbf{x}_i\right)-f\left(\mathbf{x}_j\right)\right\|_2^2\le b_{\text{SIR}}-1+\xi_{ij}^P \quad \text{if } h_{ij}=1 \\ &\left\|f\left(\mathbf{x}_i\right)-f\left(\mathbf{x}_j\right)\right\|_2^2\ge b_{\text{SIR}}+1-\xi_{ij}^P \quad \text{if } h_{ij}=-1 \end{aligned} \tag{5}$$

where ξ_{ij}^P is a nonnegative slack variable. Then the loss function of SIR learning is

$$L_{\text{SIR}}^P=\sum_{i,j}\left[1+h_{ij}\left(\left\|f\left(\mathbf{x}_i\right)-f\left(\mathbf{x}_j\right)\right\|_2^2-b_{\text{SIR}}\right)\right]_+ \tag{6}$$

where $[z]_+=\max(z,0)$.

The CIR learning can be formulated as a binary classification problem, where the CIR for any doublet $(\mathbf{x}_i,\mathbf{x}_j)$ should satisfy the constraints:

$$\begin{aligned} &\mathbf{w}^Tg\left(\mathbf{x}_i,\mathbf{x}_j\right)\le b_{\text{CIR}}-1+\zeta_{ij}^P \quad \text{if } h_{ij}=1 \\ &\mathbf{w}^Tg\left(\mathbf{x}_i,\mathbf{x}_j\right)\ge b_{\text{CIR}}+1-\zeta_{ij}^P \quad \text{if } h_{ij}=-1 \end{aligned} \tag{7}$$

where b_{CIR} is the threshold and ζ_{ij}^P is a nonnegative slack variable. We use the loss function of the standard SVM [32] to learn CIR:

$$L_{\text{CIR}}^P=\frac{\alpha_P}{2}\left\|\mathbf{w}\right\|_2^2+\sum_{i,j}\left[1+h_{ij}\left(\mathbf{w}^Tg\left(\mathbf{x}_i,\mathbf{x}_j\right)-b_{\text{CIR}}\right)\right]_+. \tag{8}$$

where α_P is a trade-off parameter, and we set $\alpha_P=0.0005$ in the experiments.

The overall loss function of pairwise comparison based representation learning method is the combination of (6) and (8):

$$L^P=L_{\text{SIR}}^P+\eta_PL_{\text{CIR}}^P \tag{9}$$

where η_P is a trade-off parameter and we set $\eta_P=1$ in our experiments.

3.3. Triplet Comparison Formulation

The triplet comparison formulation is trained on a series of triplets $(\mathbf{x}_i,\mathbf{x}_j,\mathbf{x}_k)$, where $\mathbf{x}_i$ and $\mathbf{x}_j$ are from the same class, while $\mathbf{x}_i$ and $\mathbf{x}_k$ are from different classes. To make the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ smaller than the one between $\mathbf{x}_i$ and $\mathbf{x}_k$, for any triplet $(\mathbf{x}_i,\mathbf{x}_j,\mathbf{x}_k)$ the SIR should satisfy the following constraint:

$$\left\|f\left(\mathbf{x}_i\right)-f\left(\mathbf{x}_k\right)\right\|_2^2-\left\|f\left(\mathbf{x}_i\right)-f\left(\mathbf{x}_j\right)\right\|_2^2\ge 1-\xi_{ijk}^T \tag{10}$$

where ξ_{ijk}^T is a nonnegative slack variable. Then the loss function of SIR learning is

$$L_{\text{SIR}}^T=\sum_{i,j,k}\left[1-\left\|f\left(\mathbf{x}_i\right)-f\left(\mathbf{x}_k\right)\right\|_2^2+\left\|f\left(\mathbf{x}_i\right)-f\left(\mathbf{x}_j\right)\right\|_2^2\right]_+ \tag{11}$$

The CIR learning can be formulated as a learning-to-rank problem, where the CIRs should satisfy the following constraint:

$$\mathbf{w}^Tg\left(\mathbf{x}_i,\mathbf{x}_k\right)-\mathbf{w}^Tg\left(\mathbf{x}_i,\mathbf{x}_j\right)\ge 1-\zeta_{ijk}^T \tag{12}$$

where ζ_{ijk}^T is a nonnegative slack variable. We use the loss function of the RankSVM [30] to learn CIR:

$$L_{\text{CIR}}^T=\frac{\alpha_T}{2}\left\|\mathbf{w}\right\|_2^2+\sum_{i,j,k}\left[1+\mathbf{w}^Tg\left(\mathbf{x}_i,\mathbf{x}_k\right)-\mathbf{w}^Tg\left(\mathbf{x}_i,\mathbf{x}_j\right)\right]_+ \tag{13}$$

where α_T is a trade-off parameter, and we set $\alpha_T=0.0005$ in the experiments.

The overall loss function of triplet comparison based learning method is the combination of (11) and (13):

$$L^T = L^T_{\text{SIR}} + \eta_T L^T_{\text{CIR}} \tag{14}$$

where η_T is a trade-off parameter and we set $\eta_T = 1$ in our experiments.

3.4. Prediction

We use both of SIR and CIR for matching. For the given image pair $(\mathbf{x}_i, \mathbf{x}_j)$, we take the Euclidean distance $\|f(\mathbf{x}_i) - f(\mathbf{x}_j)\|_2^2$ as the indicator of the SIRs, and take $\mathbf{w}^T g(\mathbf{x}_i, \mathbf{x}_j)$ as the indicator of the CIR. In this view, we use the combination of these indicators, which is as follows,

$$S(\mathbf{x}_i, \mathbf{x}_j) = \|f(\mathbf{x}_i) - f(\mathbf{x}_j)\|_2^2 + \lambda \mathbf{w}^T g(\mathbf{x}_i, \mathbf{x}_j) \tag{15}$$

where λ is the trade-off parameter. This parameter can be selected by cross validation. In the experiments, we set it as $\lambda = 0.7$ in the pairwise comparison model, and $\lambda = 1$ in the triplet comparison model. We compare $S(\mathbf{x}_i, \mathbf{x}_j)$ with a threshold t to decide whether these two images $\mathbf{x}_i$ and $\mathbf{x}_j$ are matched or not. If $S(\mathbf{x}_i, \mathbf{x}_j) < t$, then $\mathbf{x}_i$ and $\mathbf{x}_j$ are matched, otherwise they are not matched.

We also combine the matching scores of the learning models based on pairwise and triplet comparison formulations, which are denoted by $S_P(\mathbf{x}_i, \mathbf{x}_j)$ and $S_T(\mathbf{x}_i, \mathbf{x}_j)$, respectively. The combined matching score is $S_{P\&T}(\mathbf{x}_i, \mathbf{x}_j) = S_P(\mathbf{x}_i, \mathbf{x}_j) + \mu S_T(\mathbf{x}_i, \mathbf{x}_j)$, where μ is a trade-off parameter and we set it as $\mu = 0.5$ in the experiments.

4. Deep Convolutional Neural Network

4.1. Network Architecture

Instead of using the hand-crafted image features, we jointly learn the SIRs and CIRs using a deep CNN. For the pairwise comparison formulation, we learn the SIRs ($f(\mathbf{x}_i)$ and $f(\mathbf{x}_j)$) and CIR $g(\mathbf{x}_i, \mathbf{x}_j)$ for the image pair $(\mathbf{x}_i, \mathbf{x}_j)$. For the triplet comparison formulation, we learn the SIRs ($f(\mathbf{x}_i)$, $f(\mathbf{x}_j)$ and $f(\mathbf{x}_k)$) and the CIRs ($g(\mathbf{x}_i, \mathbf{x}_j)$ and $g(\mathbf{x}_i, \mathbf{x}_k)$) for the image triplet $(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k)$. The deep architectures of the pairwise and triplet comparison models are illustrated in Fig. 2 and Fig. 3, respectively. Each of these two networks consists of a SIR learning sub-network (green part), a CIR learning sub-network (red part), and a sub-network shared by SIR and CIR learning (blue part). For each of the probe and gallery images, its CNN feature maps (yellow part) from the shared sub-network and the SIR feature are computed once. Only the CIR learning sub-network is used to extract the CIR features for each image pair of probe image and gallery image.

Shared sub-network. The sub-network in the blue part of Figs. 2 and 3 is shared by SIR learning and CIR learning. It consists of two convolutional layers with rectified linear unit (ReLU) activation. Each of them is followed by a pooling layer. The kernel sizes of the first and second convolutional layers are 5×5 and 3×3, respectively. The stride of the convolutional layers is 1 pixel. The kernel sizes of the first and second pooling layers are set to 3×3 and 2×2, respectively.

SIR sub-network. We use the sub-network in the green part of Figs. 2 and 3 to learn the SIR $f(\mathbf{x}_i)$ for the input image $\mathbf{x}_i$. This sub-network contains one convolutional layers with ReLU activation, a pooling layer and two fully-connected layers. The kernel sizes of the convolutional layer and the pooling layer are 3×3 and 2×2. The output dimensions of these two fully-connected layers are 1000 and 500, respectively. For the pairwise and triplet comparison model, there are two and three sub-networks, which share the same parameter, to learn the SIR, respectively.

CIR sub-network. We use the sub-network in the red part of Figs. 2 and 3 to learn the CIR $g(\mathbf{x}_i, \mathbf{x}_j)$ for the input image pair $(\mathbf{x}_i, \mathbf{x}_j)$. This sub-network contains one convolutional layer with ReLU activation followed by one pooling layer and one fully-connected layer. The kernel sizes of the convolutional layer and the pooling layer are 3×3 and 2×2. The output dimension of the fully-connected layer is 1000. Denote by $\phi_p(\mathbf{x}_i)$ the pth channel of the CNN feature map of $\mathbf{x}_i$ from the shared sub-network. When we extract the CIR of $(\mathbf{x}_i, \mathbf{x}_j)$, the CIR sub-network is feeded by the CNN feature maps of $\mathbf{x}_i$ and $\mathbf{x}_j$ from the shared sub-network. The first convolutional layer of CIR sub-network is used to compute the cross-image feature map as follows

$$\begin{aligned} \varphi_r(\mathbf{x}_i, \mathbf{x}_j) = \max\Big(0, b_r + \sum_q \mathbf{k}_{q,r} * \phi_q(\mathbf{x}_i) \\ + \mathbf{l}_{q,r} * \phi_q(\mathbf{x}_j)\Big), \end{aligned} \tag{16}$$

where $\varphi_r(\mathbf{x}_i, \mathbf{x}_j)$ is the rth channel of cross-image feature map, $\mathbf{k}_{q,r}$ and $\mathbf{l}_{q,r}$ are different convolutional kernels of the qth channel of the shared sub-network feature map and the rth channel of cross-image feature map. The similar operation has also been used in [24].

4.2. Network Training

There are three main steps in the training process, including data preprocessing, doublet/triplet generation and network training. Like most of the deep models, back propagation (BP) is utilized to train the proposed network. The details of the first two steps are described as follows.

Data preprocessing. To make the model robust to the image translation variance, we randomly crop the input images before the training process. The original image size in our experiment is 180×80 pixels. We randomly select the cropped image center from $[80, 100] \times [30, 50]$ and crop the original image to 160×60 pixels. We also enlarge the

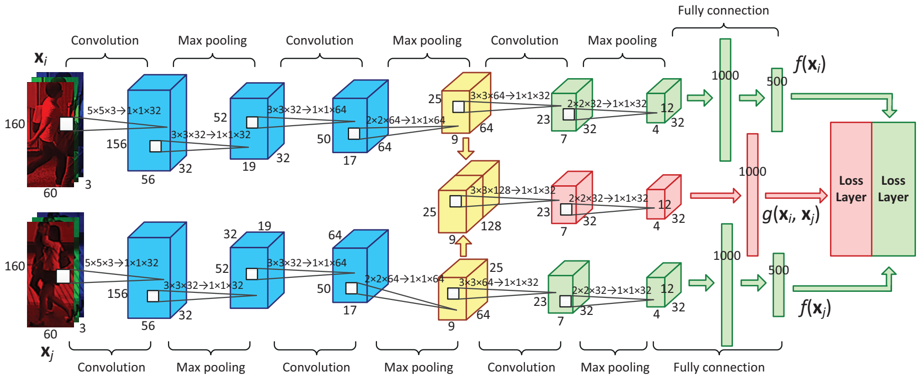

Figure 2. The proposed deep architecture of the pairwise comparison model (best viewed in color)

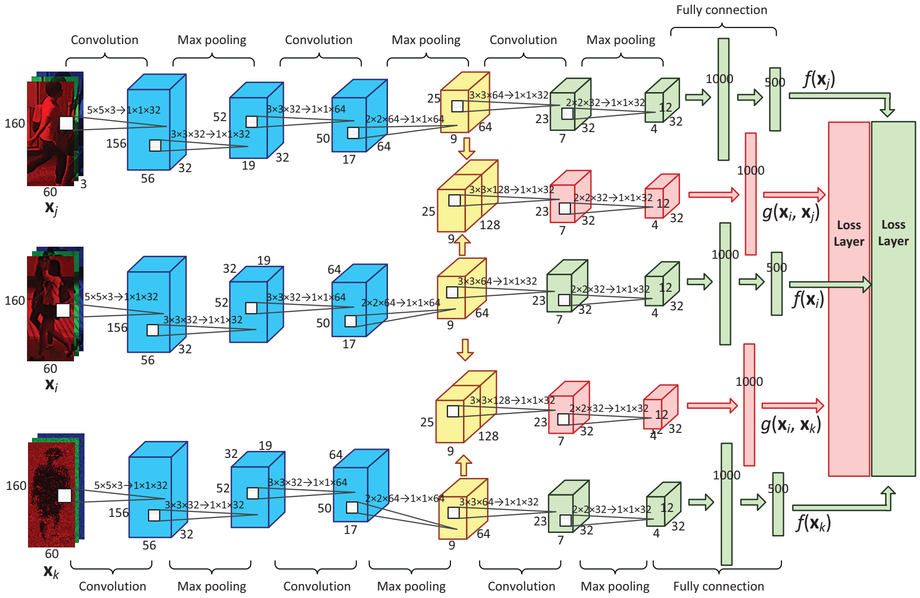

Figure 3. The proposed deep architecture of the triplet comparison model (best viewed in color)

training set by creating the horizontal mirror of each training images.

Doublet/triplet generation based on mini-batch strategy. Since the training set may be too large to be loaded into the memory, we divide the training set into multiple mini-batches. Following the strategy in [7], for each iteration, we randomly select 80 classes from the training set, and construct 60 doublets or triplets for each class. Using this strategy, we can generate 4,800 doublets or triplets in each round of training. For the SIR learning, we use all of the 4,800 doublets or triplets in training. For the CIR learning, we randomly select 100 doublets or triplets for training.

5. Experiments

In this section, we evaluate the proposed method using three person re-identification datasets, *i.e.* CUHK03 [19][1], CUHK01 [18][1] and VIPeR [12][2]. The proposed method is implemented based on the Caffe framework [15]. We set the momentum as $\gamma = 0.5$ and set the weight decay as $\mu = 0.0005$. We train the network for 150,000 iterations. It takes about 28–34 hours in training with a NVIDIA Tesla K40 GPU. The learning rates of pairwise and triplet comparison models are 1×10^{-3} and 3×10^{-4} before the 100,000th iteration, respectively. After that their learning rates reduce to 1×10^{-4} and 3×10^{-5}.

5.1. CUHK03 Dataset

The CUHK03 dataset contains 14,096 pedestrian images, which were taken from 1,467 persons by two surveillance cameras [19]. Each person has 4.8 images on average. All of the images are collected from five video clips. The dataset provides both the manually cropped bounding box

[1] http://www.ee.cuhk.edu.hk/~rzhao/

[2] http://vision.soe.ucsc.edu/projects

Model	SIR	CIR	Combined
Pairwise	37.15	35.70	43.36
Triplet	43.23	43.46	51.33
Combined	44.35	45.40	52.17

Table 1. The rank-1 accuracies (%) of the proposed pairwise and triplet comparison models

Model	SIR	SIR&CIR
Pairwise	24h18m	28h25m
Triplet	29h33m	33h27m

Table 2. The training times of the proposed pairwise and triplet comparison models

and the automatically cropped bounding box with a pedestrian detector [9]. Here we use the images cropped by the pedestrian detector in our experiments. Following the testing protocol in [19], the identities in this dataset are randomly divided into non-overlapping training and test set. The training set consists of 1,367 persons and the test set consists of 100 persons. By this strategy, 20 partitions of training and test set are constructed. The reported cumulative matching characteristic (CMC) curve and accuracy are averaged by these 20 groups. For each person in the test set, we randomly select one camera view to construct the probe set, and use one image from another camera view as the gallery set. By this way we construct 10 pairs of probe and gallery sets for testing. The result is averaged by these 10 groups. The reported results of CUHK03 dataset are based on single-shot setting.

First, we report the accuracies of different settings of the proposed pairwise and triplet comparison models in Table 1. For each of the pairwise and triplet comparison models, we report the matching accuracies by SIR and CIR, respectively. The CMC curves of these settings are in the supplementary material. From the results, we can see that the SIR and CIR based matching have comparable results. However, their combination achieves a higher accuracy than either of them. The accuracy of triplet comparison model is higher than pairwise comparison model, and their combination also outperforms either of them. We also report the training time of the proposed model in Table 2. Compared with SIR learning, the proposed joint SIR and CIR learning model can achieve substantial improvement of matching accuracy with slight increase of training time.

Second, we investigate the sensitivity of rank-1 accuracy to the trade-off parameter λ in (15). Fig. 4 shows the curves of rank-1 accuracy on the test set versus λ. It can be observed that the pairwise and triplet comparison models reach the highest accuracies when $\lambda = 0.7$ and 1, respectively.

We also compare the performances of the proposed method and some other state-of-the-art methods, including Euclidean distance, ITML [5], LMNN [35], metric learning to rank (RANK) [27], LDML [13], symmetry-driven accumulation of local features (SDALF) [8], eSDC [41], KISSME [16], FPNN [19], the work by Ahmed *et al.* [1], and LOMO+XQDA [21]. Fig. 5 illustrates the CMC curves and the rank-1 accuracies of these methods. We can see that the rank-1 accuracy of the proposed method can reach 52.17%, which is 5.92% higher than the second best performance method (LOMO+XQDA).

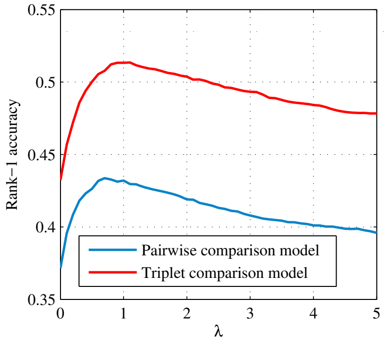

Figure 4. Rank-1 accuracy versus λ in the CUHK03 dataset (best viewed in color)

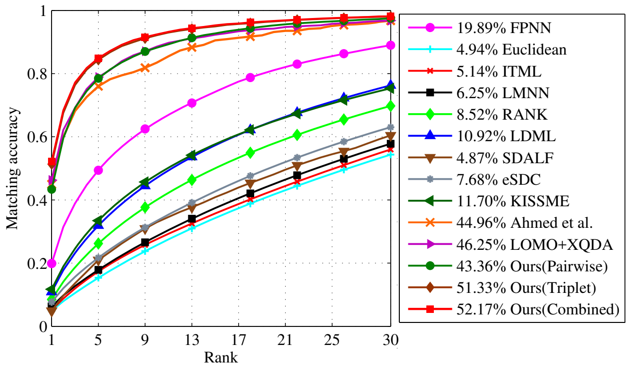

Figure 5. The rank-1 accuracies and CMC curves of different methods on the CUHK03 dataset [1] (best viewed in color)

5.2. CUHK01 Dataset

The CUHK01 dataset consists of 3,884 pedestrian images taken by two surveillance cameras from 971 persons. Each person has 4 images. This dataset has been randomly divided into 10 partitions of training and test sets, and the reported CMC curves and rank-1 accuracies are averaged on these 10 groups.

Following the protocol in [1], we use 871 persons for training and 100 persons for testing. We pretrain the deep network using CUHK03 dataset for 100,000 iterations, and fine-tune the CNN using the training set of CUHK01 for 50,000 iterations. On the basis of the single-shot setting, we report the CMC curves and rank-1 ac-

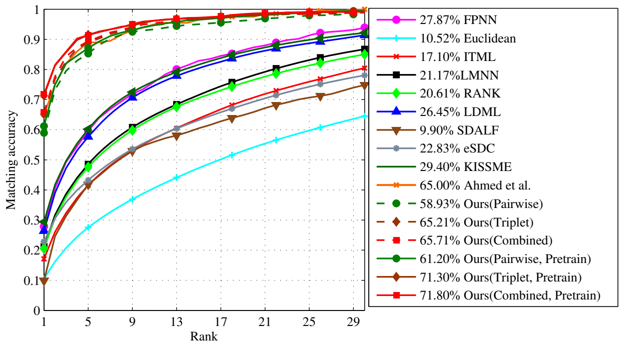

Figure 6. The rank-1 accuracies and CMC curves of different methods on the CUHK01 dataset [1] (best viewed in color)

curacies of the proposed model (marked as "*Ours (Pairwise/Triplet/Combined, Pretrain)*") and the other state-of-the-art person re-identification methods, including FPNN [19], Euclidean distance, ITML [5], LMNN [35], RANK [27], LDML [13], SDALF [8], eSDC [41], KISSME [16], and the work by Ahmed *et al.* [1] in Fig. 6. The rank-1 accuracy of the proposed method is much higher than the other competing methods. We also report the result using the same setting in [1] without pre-training (marked as "*Ours (Pairwise/Triplet/Combined)*"). In this setting, the rank-1 accuracy of the proposed method is much higher than most of the competing methods and is comparable to [1].

5.3. VIPeR Dataset

The VIPeR dataset consists of 1,264 images from 632 persons [12]. These images are taken by two camera views. We randomly select 316 persons for training, and use the rest 316 persons for testing. For each person in the test set, we randomly select one camera view as the probe set, and use the other camera view as the gallery set. Following the testing protocol in [1], we pretrain the CNN using CUHK03 and CUHK01 datasets, and fine-tune the network on the training set of VIPeR. We report the CMC curves and rank-1 accuracies of local Fisher discriminant analysis (LF) [29], pairwise constrained component analysis (PCCA) [28], aPRDC [23], PRDC [43], enriched BiCov (eBiCov) [25], PRSVM [2], and ELF [10], saliency matching (SalMatch) [40], patch matching (PatMatch) [40], locally-adaptive decision function (LADF) [20], mid-level filters (mFilter) [42], visWord [39], the work by Ahmed *et al.* [1], the proposed model, *etc.* The proposed method performs better than most of the other competing methods except mFilter [42]+LADF [20], which is the combination of two methods.

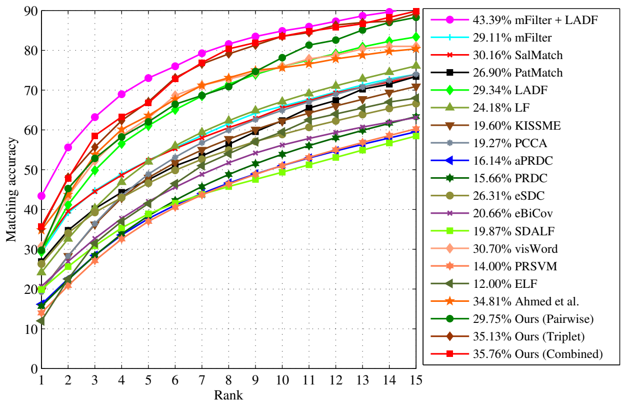

Figure 7. The rank-1 accuracies and CMC curves of different methods on the VIPeR dataset [1] (best viewed in color)

6. Conclusion

In this work, we propose an approach for person re-identification by joint SIR and CIR learning. Since SIR is efficient in matching, while CIR is effective in modeling the relationship between probe and gallery images, we fuse their losses together to utilize the advantages of both these representations. We present a pairwise comparison formulation and a triplet comparison formulation for joint SIR and CIR learning. For each of these two models, we formulate a deep neural network to jointly learn the SIR and CIR. Experimental results validate the efficacy of joint SIR and CIR learning, and the proposed method outperforms most of the state-of-the-art models in the CUHK03, CUHK01 and VIPeR datasets. In the future, we will investigate other ways to integrate SIR and CIR learning (*e.g.*, explicit modeling on patch correspondence), and study model-level fusion from pairwise and triplet comparisons.

Acknowledgments

This work is supported by NSFC grant (61271093), the program of ministry of education for new century excellent talents (NCET-12-0150), and the Hong Kong RGC General Research Fund (PolyU 152212/14E). This work is also supported in part by Special Program for Applied Research on Super Computation of the NSFC-Guangdong Joint Fund (the second phase). The authors would like to thank NVIDIA Corporation for the donation of Tesla K40 GPU.

References

[1] E. Ahmed, M. Jones, and T. Marks. An improved deep learning architecture for person re-identification. In *CVPR*, 2015.

[2] L. Bazzani, M. Cristani, A. Perina, and V. Murino. Multiple-shot person re-identification by chromatic and epitomic analyses. *Pattern Recognit. Letters*, 33(7):898–903, 2012.

[3] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun. Bayesian face revisited: A joint formulation. In *ECCV*, 2012.

[4] S. Chen, C. Guo, and J. Lai. Deep ranking for person re-identification via joint representation learning. *arXiv: 1505.0682*, 2015.

[5] J. Davis, B. Kulis, P. Jain, S. Sra, and I. Dhillon. Information-theoretic metric learning. In *ICML*, 2007.

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

[7] S. Ding, L. Lin, G. Wang, and H. Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognit.*, 48(10):2993–3003, 2015.

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

[9] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1627–1645, 2010.

[10] N. Gheissari, T. B. Sebastian, and R. Hartley. Person reidentification using spatiotemporal appearance. In *CVPR*, 2006.

[11] S. Gong, M. Cristani, S. Yan, and C. C. Loy, editors. *Person Re-Identification.* Springer, 2014.

[12] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008.

[13] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? metric learning approaches for face identification. In *ICCV*, 2009.

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

[15] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM International Conference on Multimedia*, 2014.

[16] M. Köstinger, M. Hirzer, P. Wohlhart, P. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012.

[17] W. Li and X. Wang. Locally aligned feature transforms across views. In *CVPR*, 2013.

[18] W. Li, R. Zhao, and X. Wang. Human re-identification with transferred metric learning. In *ACCV*, 2012.

[19] W. Li, R. Zhao, T. Xiao, and X. Wang. DeepReID: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014.

[20] Z. Li, S. Chang, F. Liang, T. Huang, L. Cao, and J. Smith. Learning locally-adaptive decision functions for person verification. In *CVPR*, 2013.

[21] S. Liao, Y. Hu, X. Zhu, and S. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015.

[22] S. Liao and S. Li. Efficient PSD constrained asymmetric metric learning for person re-identification. In *ICCV*, 2015.

[23] C. Liu, S. Gong, C. C. Loy, and X. Lin. Person re-identification: What features are important? In *ECCV Workshops and Demonstrations*, 2012.

[24] S. Liu, X. Liang, L. Liu, X. Shen, J. Yang, C. Xu, L. Lin, X. Cao, and S. Yan. Matching-CNN meets KNN: Quasi-parametric human parsing. In *CVPR*, 2015.

[25] B. Ma, Y. Su, and F. Jurie. BiCov: a novel image representation for person re-identification and face verification. In *BMVC*, 2012.

[26] N. Martinel, C. Micheloni, and G. L. Foresti. Saliency weighted features for person re-identification. In *ECCV Workshop on Visual Surveillance and Re-identification*, 2014.

[27] B. Mcfee and G. Lanckriet. Metric learning to rank. In *ICML*, 2010.

[28] A. Mignon and F. Jurie. PCCA: A new approach for distance learning from sparse pairwise constraints. In *CVPR*, 2012.

[29] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *CVPR*, 2013.

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

[31] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.

[32] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 2nd edition, 2000.

[33] R. Vezzani, D. Baltieri, and R. Cucchiara. People reidentification in surveillance and forensics: A survey. *ACM Computing Surveys (CSUR)*, 46(2):29, 2013.

[34] X. Wang. Intelligent multi-camera video surveillance: A review. *Pattern Recognit. Letters*, 34:3–19, 2013.

[35] K. Weinberger, J. Blitzer, and L. Saul. Distance metric learning for large margin nearest neighbor classification. In *NIPS*, 2005.

[36] F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In *ECCV*, 2014.

[37] D. Yi, Z. Lei, S. Liao, and S. Li. Deep metric learning for person re-identification. In *ICPR*, 2014.

[38] R. Zhang, L. Lin, R. Zhang, W. Zuo, and L. Zhang. Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification. *IEEE Trans. Image Process.*, 24(12):4766–4779, 2015.

[39] Z. Zhang, Y. Chen, and V. Saligrama. A novel visual word co-occurrence model for person re-identification. In *ECCV Workshop on Visual Surveillance and Re-identification*, 2014.

[40] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In *ICCV*, 2013.

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

[42] R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identification. In *CVPR*, 2014.

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