Learning Patch-Dependent Kernel Forest for Person Re-Identification

Wei Wang[1] Ali Taalimi[1] Kun Duan[2] Rui Guo[1] Hairong Qi[1]
[1]University of Tennessee, Knoxville [2]GE Research Center
{wwang34,ataalimi,rguo1,hqi}@utk.edu, {kun.duan}@ge.com

Abstract

In this paper, we propose a new approach for the person re-identification problem, discovering the correct matches for a query pedestrian image from a set of gallery images. It is well motivated by our observation that the overall complex inter-camera transformation, caused by the change of camera viewpoints, person poses and view illuminations, can be effectively modelled by a combination of many simple local transforms, which guides us to learn a set of more specific local metrics other than a fixed metric working on the feature vector of a whole image. Given training images in pair, we first align the local patches using spatially constrained dense matching. Then, we use a decision tree structure to partition the space of the aligned local patch-pairs into different configurations according to the similarity of the local cross-view transforms. Finally, a local metric kernel is learned for each configuration at the tree leaf nodes in a linear regression manner. The pairwise distance between a query image and a gallery image is summarized based on all the pairwise distance of local patches measured by different local metric kernels. Multiple decision trees form the proposed random kernel forest, which always discriminatively assign the optimal local metric kernel to the local image patches in re-identification. Experimental results over the public benchmarks demonstrate the effectiveness of our approach for achieving very competitive performances with a relatively simpler learning scheme.

1. Introduction

Person re-identification is to recognize the same person across a network of cameras with non-overlapping views. It is important for video surveillance by saving a lot of human effort on exhaustively searching for a person from large amounts of video sequences, *e.g.*, the large scale pedestrian retrieval [28] and the wide scale multi-camera tracking [41]. However, this is also a fairly challenging problem since the appearance of the same person may vary greatly in different camera views, due to the significant variations in camera viewpoints, illuminations, person poses, occlusions and backgrounds, *etc*. In addition, a surveillance camera usually observes hundreds of people in one day, many of which have similar appearances, therefore generating a lot of false alarms for the query image. See Figure 1 for some typical difficult examples.

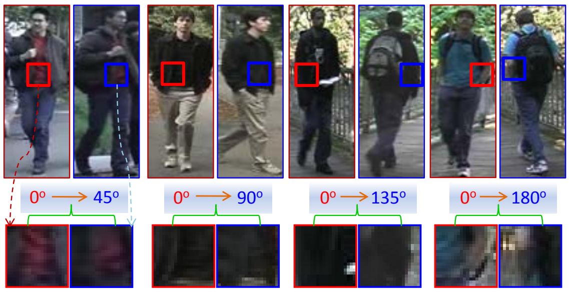


Figure 1. Samples of pedestrian images observed in different camera views in person re-identification. Each pedestrian has a different pose variation in the four examples between two cameras.

In literature, two lines of approaches have been proposed to tackle this problem. The *first* line concentrated on the development of viewpoint quasi-invariant local features, *e.g.*, color [11], texture [44] or gradient [47], as well as robust feature ensembles. However, these feature based methods still suffer from illumination changes, human shape deformations and difficulty of multi-feature ensembles. The *second* line is to learn a parametric distance metric to enforce features from the same individual to be closer than that from different individuals [22, 36, 44], also known as the metric learning (ML). However, ML usually deals with feature vectors of a complete image in learning of the metric. Although effective, this distance metric may not be the optimal to work well on certain local parts of each person image.

In the re-identification problem, image regions typically undergo both geometric transformation due to camera viewpoint changes and photometric transformation due to illumination variations. However, different regions suffer differently to these two transformations, *e.g.*, the smooth pure color regions suffer less while the texture or high gradient patches suffer more. In addition, the pose changes from one camera to another vary for different people, there is no fixed pattern, *e.g.*, $45^\circ \rightarrow 90^\circ$ or $0^\circ \rightarrow 45^\circ$, to describe the diverse pose changes, shown as in Figure 1. Therefore, the configuration of person images are multi-modal even if the people are observed in the same camera view. To fully formulate the overall inter-camera transformation $\mathcal{F}$, it must be a sophisticated non-linear function with a large number of unknown parameters. Obviously, single transform or unimodal metric function might not be the optimal to tackle the problem. Thus some of the recent works used kernel tricks applying ML in a non-linear kernel space [44, 3] or adopted nested formulations as in deep learning framework [1, 42], which is usually time-consuming in model training.

Our work is mainly motivated by the above observations. Suppose a specific metric can be learned from a small group of local image patch-pairs that share a consistent cross-view transform, not only the metric learning task becomes much easier, the combination of these specific metrics is also more effective to further ensure the pairwise distances of images from the same individual can be better minimized. Comparing to the deep learning architectures [1, 20], which approximates the overall transformation $\mathcal{F}$ as a series of nested functions with the distance metric defined as $\mathcal{D}(\mathbf{x}, \mathbf{y}) = d_K\left(...d_2\left(d_1(\mathbf{x}, \mathbf{y})\right)\right)$, where $\mathbf{x}, \mathbf{y}$ are the representations of two images from two different camera views, respectively, we try to partition out all the local transforms and decompose the overall transformation $\mathcal{F}$ into many independent sub-functions f_k, then our new distance metric is defined as $\mathcal{D}(\mathbf{x}, \mathbf{y}) = \sum_{x,y}\{d_1(x, y) + d_2(x, y) + ... + d_K(x, y)\}$, where x, y are features of local patches from the image pair $\mathbf{x}, \mathbf{y}$, respectively, *i.e.*, some segments of the concatenated feature vectors $\mathbf{x}, \mathbf{y}$. However, each d_k *only* works on a specific kind of the local patches from each image.

The main purpose of this paper is to learn specific metric kernels for different local image patches in measure of the pairwise distance. We propose a novel random kernel forest (RKF) based on the *consistent patch-to-patch transform* criteria for person re-identification. Our *main contribution* is the use of a highly efficient decision forest that is trained to discriminatively predict which kernel should be applied to measure the pairwise distance of any two given image patches. As shown in Figure 2, the tree structure jointly partitions the space of local patch-pairs from all the training image pairs into a set of sub-spaces at each tree leaf, where the transform of the local patches between cameras is simplified and consistent. *Furthermore*, a simple linear kernel can be learned at each leaf to describe the specific transform $f_{k,k=1,...,\mathrm{K}}$, such that the distance between any true patch-pair will be minimized in d_k. Combining with multiple decision trees in the forest, the model also effectively avoids over-fitting during training. *Finally*, since the decision tree recursively and jointly partitions the patch-pair space solely based on the thresholds on features, it is very fast in learning and prediction. Extensive experimental results demonstrate the effectiveness of our approach for achieving very competitive performance while adopting a relatively simple learning scheme.

2. Related Work

The existing person re-identification approaches can be broadly grouped into two categories: robust feature extraction and distance metric learning.

The existing works on feature design and selection can be further divided into unsupervised and supervised versions. Unsupervised approaches search for view invariant features via perceptual symmetry or certain prior assumptions [25, 2, 29, 30]. For example, Farenzena *et al.* [10] proposed the accumulation of local features by exploiting the symmetry property. Zhao *et al.* [47, 46] proposed a salience model for patches matching such that the reliable and discriminative matched patches can be identified for better performance. Liao *et al.* [23] proposed to maximize the occurrence of each local pattern among all the horizontal sub-windows to tackle the viewpoint changes. Supervised approaches select the most effective features by certain criteria [11, 31]. For example, Prosser *et al.* [37] formulated person re-identification as a ranking problem, and learned global feature weights based on an ensemble of RankSVM. Paisitkriangkrai *et al.* [35] improved the feature ensemble performance by learning the weights based on cumulated matching characteristics curve. Recently, Wu *et al.* [43] proposed an appearance model integrating the camera viewpoints and human pose information.

In contrast, the approaches that focus on metric learning usually extract image features in a more straightforward manner, *e.g.*, color or texture histograms from predefined image regions. A lot of metric learning algorithms have been proposed recently [50, 16, 32, 44]. For example, Mahalanobis (M-distance) learning has been proposed for re-identification problem [14, 34], as M-distance can implicitly model the transition in feature space between two camera views. Pedagadi *et al.* [36] applied FDA (fisher discriminant analysis) together with PCA and LPP (locality preserving projections) to derive a low-dimensional yet discriminant subspace. Li *et al.* [22] developed a locally-adaptive decision function (LADF) that jointly models a distance metric and a locally adaptive thresholding rule to achieve good performance. Dictionary learning [26, 15, 40] is also proposed to bridge the appearance across two cameras with the assumption that the manifold of local patches in spaces of two camera views are similar. Recently, Chen *et al.* [3] proposed an explicit polynomial kernel approach that learns a similarity function to maximize the difference between the similarity score of true and false image pairs.

Other than these two main research lines, some other interesting and novel approaches have also been proposed for the re-identification problem. For example, the deep learn-

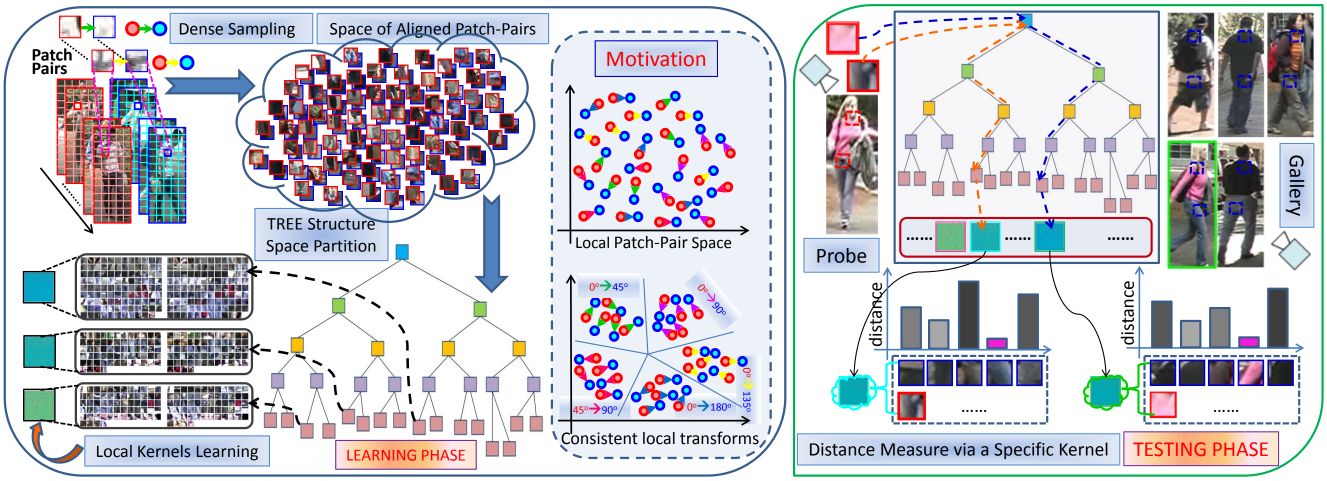


Figure 2. Illustration of the main idea. Left: learning phase, the aligned patch pairs of the same person from different cameras are separated in a tree structure based on the *consistent patch-to-patch transform criteria*. At each tree leaf, a simple but effective kernel is learned to describe the simplified transform. Right: testing phase, given a probe image, a suitable kernel will be selected based on the decision tree for each of its local image patch. With the optimal local kernel, the distance between the true patch pairs will be well minimized.

ing framework was applied to exploit the information of the cross-input difference features by multiple layers of the neural network [20, 45, 1]. The mid-level features, *e.g*., filters and semantic attributes [18, 8], were also explored. Zhao *et al*. [48] proposed to learn mid-level filters by mining the cross-view invariance in subsets of local patch features. Shi *et al*. [39] proposed a new approach for learning a semantic attribute model from existing fashion datasets, and adapted the resultant model to facilitate person re-identification.

If viewed from the perspective of motivation, our work is most close to the LAFT approach [19], which jointly partitions the image spaces of two camera views into different subspaces according to the similarity of inter-camera transforms. However, the main difference between our works are that: (i) LAFT partitions the image space of each camera view instead of the more fine local patches space, where the problem of local regions suffering from different transforms can be better tackled; (ii) LAFT uses a gating network to softly assign the given image pair to a configuration type and requires feature selection with sparsity and log-determinant divergence regularization. In contrast, we assign the optimal kernel to the given local patch much more efficiently in the tree structure and do not require post feature selection. If viewed from the perspective of methodology, our work is also relevant with [47, 21]. Though [47] also plays with local patches, it assumes the salient parts on each person appearance can be captured by different cameras. [21] also uses random forest, which however was used as a traditional classifier to recognize two image features are from the same person or not. In contrast, the random forest in our work is utilized to discover different patch-level inter-camera transforms in a tree structure.

3. Method

Random forests [6] is a well-known decision tree based classifier ensemble. It has been widely used in many computer vision problems recently, such as image denoising [9], edge detection [7], image classification [38], human pose estimation [17], *etc*. In our work, random forest has been strategically designed to **decompose** the multi-modal inter-camera transformation into multiple simple and independent uni-modal transforms.

3.1. Model

Traditional machine learning problems try to learn a category specific probability distribution or a decision boundary to answer which category a given sample belongs to. In contrast, the person re-identification problem deals with image pairs and tries to determine whether a pair of samples are from the same category or not. Formally, for a pair of image samples represented by $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, respectively, each of which corresponds to a class label $\mathcal{C}(\mathbf{x})$ and $\mathcal{C}(\mathbf{y})$, we need to decide whether they are from the same category, *i.e*., $\mathcal{C}(\mathbf{x}) = \mathcal{C}(\mathbf{y})$, or not. The ability of dealing with unseen categories is the key for person re-identification, since most of the testing samples are from unseen persons which do not exist in the training set. The proposed approach still follows the distance metric learning framework. Given a set of N training pedestrian image pairs $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$, $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N\}$, which are observed by two disjoint camera $\mathcal{X}$ (cam$\mathcal{X}$) and camera $\mathcal{Y}$ (cam$\mathcal{Y}$), our goal is to learn a distance metric $\mathcal{D}(\mathbf{x}_i, \mathbf{y}_i)$ that any pair of two samples from the same person generates the smallest distance.

Mathematically, the gist of metric learning is to learn a projection $\mathbf{P}$ and find a common subspace to measure the pairwise distance, *e.g.*, $||\mathbf{Px}-\mathbf{Py}||^2 = (\mathbf{x}-\mathbf{y})^\top\mathbf{P}^\top \cdot \mathbf{P}(\mathbf{x}-\mathbf{y}) = (\mathbf{x}-\mathbf{y})^\top\mathbf{W}(\mathbf{x}-\mathbf{y})$, where $\mathbf{W}$ is a semi-definite matrix. However, as explained in §1, the complex transformation between cam$\mathcal{X}$ and cam$\mathcal{Y}$ is multi-modal, hence it cannot be well learned with a single fixed metric $\mathbf{W}$. In our work, we learn a more comprehensive overall mapping function $\mathcal{F}_\mathbf{M} : \mathcal{X} \to \mathcal{Y}$ which is parameterized by $\mathbf{M} = \{\mathbf{m}_1, \mathbf{m}_2, ..., \mathbf{m}_K\}$, where each $\mathbf{m}_k$ represents a simple local transform that learned from a small set of ***automatically*** selected local patch-pairs in group $G_k = \{(x_i, y_i)^n_{i=1,2,...}\}$ from the training set of image pairs $\{\mathbf{X}, \mathbf{Y}\}$, where i is the subscript of each local patch in group $G_{k,k=1,2,...,K}$ and n denotes which image it comes from. Hereinafter, each independent local transform f_k parameterized by kernel $\mathbf{m_k}$ is denoted as $f_{\mathbf{m}_k}$. Learning such a kernel $\mathbf{m}_k$ is generally formulated using the empirical risk minimization:

$$\mathbf{m}^*_k = \arg\min_{\mathbf{m}_k} \frac{1}{|G_k|} \sum_{i \in G_k} \mathcal{L}(y_i, f_{\mathbf{m}_k}(x_i)) \tag{1}$$

Please note that each small group G_k is discovered automatically, we will illustrate how to partition the space of aligned local patch-pairs into different sub-spaces and get the resultant groups G_k in the next subsection via the decision tree structure. In this work, each $f_{\mathbf{m}_k}$ is defined as a linear mapping function describing the decomposed *uni-modal* local transform. The loss function $\mathcal{L}$ is simply defined as $(y_i - \mathbf{m}_k x_i)$, and $\mathbf{m}_k$ is just a linear mapping kernel that can be efficiently solved in closed form as $\hat{\mathbf{y}}\hat{\mathbf{x}}^\top(\hat{\mathbf{x}}\hat{\mathbf{x}}^\top + \lambda\mathbf{I})^{-1}$, where λ is a regularizing parameter being small value, and $\hat{\mathbf{x}} = [x_1, x_2, ..., x_i, ..]_{x_i \in G_k}$ and $\hat{\mathbf{y}} = [y_1, y_2, ..., y_i, ..]_{y_i \in G_k}$. Finally, the overall inter-camera transformation from $\mathcal{X}$ to $\mathcal{Y}$ can be formulated as $\mathcal{F}_\mathbf{M} = \sum_1^K f_{\mathbf{m}_k}$, with each of the $f_{\mathbf{m}_k}$ representing one uni-modal transform that works on certain specific kind of image local patches.

Finally, our *local distance metric* is defined as $d_k(x_i, y_j) = ||y_j - f_{\mathbf{m}_k}(x_i)||^2$, where the optimal kernel $\mathbf{m}_k$ for each image patch x_i is *automatically* and discriminatively assigned by the tree structure. Then, the *overall distance metric* is defined as:

$$\begin{aligned} \mathcal{D}(\mathbf{x}, \mathbf{y}) &= ||\mathbf{y} - \mathcal{F}(\mathbf{x})||^2 \\ &= \textstyle\sum_{r,c} ||y_{[r,c]} - \frac{1}{Q}\sum_q f^q_{\mathbf{m}_k}(x_{[r',c']})||^2 \end{aligned} \tag{2}$$

where the subscript $[r, c]$ denotes the coordinates of each local patch in images $\mathbf{x}, \mathbf{y}$. Notice that the $\mathbf{x}, \mathbf{y}$ are normalized images of pedestrian appearance cropped out from surveillance data. We use a greedy distance measure, which will be detailed in §3.3, to compute the pairwise patches distance, thus the $[r, c]$ in $\mathbf{y}$ and $[r', c']$ in $\mathbf{x}$ do not have to be identical. As to be introduced in §3.2, we formulate the uni-modal transform $f_{\mathbf{m}_k}(x)$ Q times in Q decision trees to avoid over-fitting, the final output thus is a mean value of the Q predictions.

3.2. Random Kernel Forest

A forest is an ensemble of Q decision trees T_q [6]. Given a sample x, the prediction of $T_q(x)$ from each tree is combined using an ensemble model, *e.g.*, an average value, into a single output. Each decision tree consists of non-terminal (split) and terminal (leaf) nodes. A tree T_q classifies a sample $x \in \mathcal{X}$ by recursively branching left or right child node down the tree structure until reaching a leaf node. Each non-terminal node z in the tree is associated with a binary split function h with parameters θ_z:

$$h(\phi(x), \theta_z) = \begin{cases} 0 & \text{for} \quad \phi(x) < \tau_z \\ 1 & \text{for} \quad \phi(x) \geq \tau_z \end{cases} \tag{3}$$

Then, sample x will be sent to left if $h(\phi(x), \theta_z) = 0$, otherwise, right. The split function $h(\phi(x), \theta_z)$ can be arbitrarily complex, but a typical choice is just a threshold that a single entry on the feature vector x is compared to, *e.g.*, $\theta_z = (k_z, \tau_z)$, then $h(\phi(x), \theta_z) = [x(k_z) < \tau_z]$, where $[\cdot]$ is an indicator function, $\phi(x) = x$, and $x(k_z)$ is the k_z-th entry on the feature vector x. Other than setting $\phi(x) = x$, the function $\phi(x)$ also can be of other forms, for example, we use the "pairwise" difference of two entries on the feature vector x, *i.e.*, $\phi(x) = x(k_1) - x(k_2)$. Both the two entries k_1, k_2 are randomly selected from feature vector x.

Suppose the training set, *i.e.* the aligned local patch-pairs $\mathcal{S} = \{(x_i, y_i)^{n=1,...,N}_{i=1,...}\}$, are extracted from the training image pairs $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$ and $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N\}$. Training of the decision tree for joint spaces partition involves searching for the parameter θ_z of each split function $h(\phi(x), \theta_z)$, which can well split the training data to maximize an objective function, *i.e.*, *information gain.*

$$\mathcal{I}_z = \mathcal{I}(\mathcal{S}_z, \mathcal{S}^L_z, \mathcal{S}^R_z) = E(\mathcal{S}_z) - \sum_{v \in L,R} \frac{|\mathcal{S}^v_z|}{|\mathcal{S}_z|} E(\mathcal{S}^v_z) \tag{4}$$

where $\mathcal{S}^L_z = \{(x, y) \in \mathcal{S}_z | h(\phi(x), \theta_z) = 0\}$, $\mathcal{S}^R_z = \mathcal{S}_z \backslash \mathcal{S}^L_z$, and the term E is an index function. Then, learning of the parameter θ_z is guided as to maximize $\mathcal{I}_z$. The same learning will be executed on each non-terminal nodes recursively until it reaches a leaf node or the gain falls below a threshold.

For typical classification problems, the term E is defined as the Shannon entropy, *i.e.*, $E(\mathcal{S}) = -\sum_c s_c \log(s_c)$, where s_c is the fraction of elements in $\mathcal{S}$ with label c [6, 7]. In contrast, the index function E of our task is defined as the regression error $||\hat{\mathbf{y}}_z - \mathbf{m}_z\hat{\mathbf{x}}_z||^2$. Therefore, training for classification tasks partitions training samples into successive homogeneous sub-clusters, while training for our task jointly partitions the local patch-pairs from two spaces into successive sub-spaces where their inter-camera transforms become consistent and easier to formulate, layer by layer in the tree structure. Finally, we are able to define our local metric at each **leaf node** with a specific kernel, and the combination of those local kernels can approximate any complicated multi-modal inter-camera transformations.

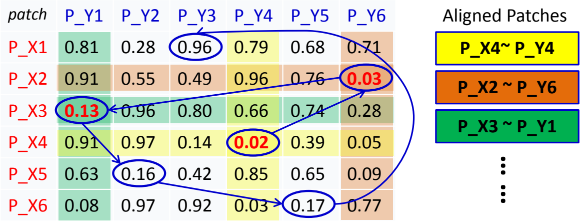
3.3. Patch Features and Alignment

Features of local patches: features of local patches on an overlapping dense grid are extracted, as shown in Figure 2. The features used for patch representation include: 10-bin color histogram extracted from each of the 3 channels of HSV color space and each of the 3 channels of LAB color space, 9-bin gradient histogram extracted from the intensity space, and 59-bin LBP features also extracted from the intensity space. The 8 channels of features are finally concatenated to form a final 128-dimensional feature vector for each local patch.

Constrained patches alignment: Suppose the images $\mathbf{x}_n, \mathbf{y}_n$ of person appearance are cropped out [13] and normalized from surveillance data in advance. In each image $\mathbf{x}_n$, the appearance of human body is usually segmented into several horizontal stripes [36, 47] to incorporate certain spatial constraint in patches matching and alignment. The feature of a local patch is denoted as $\{x^n_{r,c}\}$, indicating it's from the r-th row and c-th column on the dense grid. Since the two images from cam$\mathcal{X}$ and cam$\mathcal{Y}$ might be taken with different viewpoints, as shown in Figure 1, we need to roughly align the local patches in measure of the distance between the two images $\mathbf{x}$ and $\mathbf{y}$. Therefore, when a local patch $\{x^n_{r,c}\}$ is matched to a corresponding one in the image $\mathbf{y}_u$: $\{y^u_{r,c}\}$, its search is constrained to the set of $\{y^u_{[r-1,r+1],c=1,\ldots,C}\}$. With searching in a small range $[r-1, r+1]$, we can relieve the neg-effect in patches matching caused by the vertical misalignment. We perform the patches matching in a greedy way, in both the extraction of training patch-pairs and the testing of images re-identification. Each patch $x^n_{r,c}$ is matched to its nearest neighbor in its searching set $\{y^u_{[r-1,r+1],c=1,\ldots,C}\}$, then the corresponding one in the set will be removed in next iteration, as shown above in Figure 3. Finally, each local patch in image $\mathbf{x}_n$ is aligned to a unique one in image $\mathbf{y}_u$. Then, the distance between the two images is the summation of all the pairwise distance of each two local patches from $\mathbf{x}_n, \mathbf{y}_u$, as in Eq. 2. The retrieved image in gallery for the query image is the one gives the smallest distance value $\mathcal{D}(\mathbf{x}_n, \mathbf{y}_u)$.

3.4. Implementation Details

Training of the decision tree plays the main role in learning of the overall inter-camera transformation $\mathcal{F}$. Random forest prevents over-fitting by training multiple decorrelated trees and combining their outputs. To achieve sufficient diversity of trees, we trained 20 trees in the forest. To learn the parameter θ_z at each non-terminal node z in training of each tree, we randomly sub-sample 1024 patch-pairs, 20 pairs of the entry on the feature vector and take 10 random guesses for the threshold τ_z. The decision tree terminates split and creates a leaf once the number of patch pairs is less than 128.



patch	P_Y1	P_Y2	P_Y3	P_Y4	P_Y5	P_Y6
P_X1	0.81	0.28	0.96	0.79	0.68	0.71
P_X2	0.91	0.55	0.49	0.96	0.76	0.03
P_X3	0.13	0.96	0.80	0.66	0.74	0.28
P_X4	0.91	0.97	0.14	0.02	0.39	0.05
P_X5	0.63	0.16	0.42	0.85	0.65	0.09
P_X6	0.08	0.97	0.92	0.03	0.17	0.77

Figure 3. Illustration of the greedy local patches matching via pairwise distance. Suppose 6×2 patches are doing matching from two vertical strips of $\mathbf{x}, \mathbf{y}$, the sequence of the matched patches in this example are denoted as in color yellow, orange and green ... .

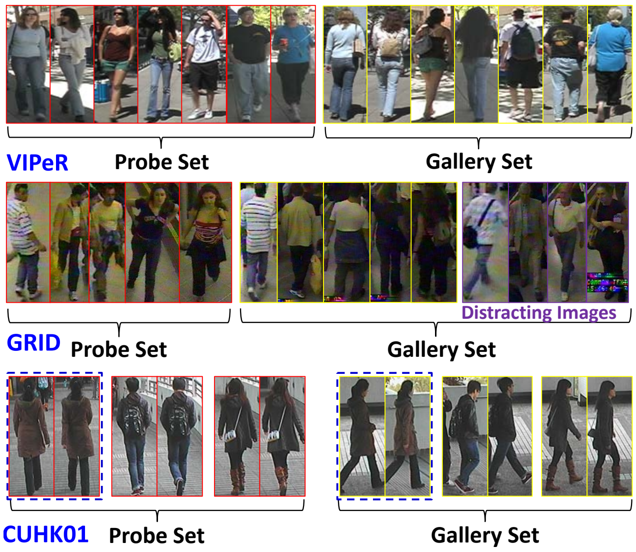


Figure 4. Examplar image pairs in probe set and gallery set from datasets of VIPeR (top), GRID, CUHK01(bottom), respectively.

4. Experimental Results

4.1. Datasets and Protocols

We conduct experiments on three most frequently used datasets: the "viewpoint invariant pedestrian recognition" (VIPeR) [11], "QMUL underround re-identification dataset" (GRID) [28] and the "CUHK person re-identification dataset" (CUHK01) [20]. All three datasets are very challenging for re-id problems due to the significant variations in viewpoints, poses, illuminations, and also their low image resolutions with occlusions and different backgrounds.

VIPeR: it contains 632 pedestrian image pairs that captured by two hand-carried cameras in outdoor environment. All the images are scaled to the same size of 128 × 48 for evaluation. Each pair contains two images of the same person observed from two camera views with pose changes (mostly $> 90^o$ degree) and different lighting conditions.

GRID: it contains 250 pedestrian image pairs that captured from 8 disjoint camera views installed in a busy underground station. All the images are scaled to the same size of 300×100 for evaluation. Each pair contains two images of the same individual seen from different camera views. Except for the common challenges (pose changes, *etc*.), the gallery set also contains 775 distracting images which do not match any person in the probe set, bringing much more difficulty in re-identification for a probe (query) image.

CUHK01: this is a multi-shot dataset containing 971 pedestrians captured from two disjoint camera views, with 2 images per person in each view. All the images are scaled to the same size of 160×60. Since it contains much more instances, it has been used for evaluation of deep learning approaches.

Protocols: The pedestrians in each dataset are separated into the training set and the testing set, such that each person appears only once in either the training set or the testing set. The testing set is also partitioned into two sets: the probe set and the gallery set. For the VIPeR dataset, the images in camera A are used as probe images, and the images in camera B are used as gallery images. The GRID dataset already defined the probe set and the gallery set, with 775 distracting images added in the gallery set. For the CUHK01 dataset, the first 2 images of each person are used as probe images and the latter 2 images from another view are stored in the gallery set. According to the existing works in literature, the performances are reported quantitatively as the standard Cumulated Matching Characteristics (CMC) curves, and the performance is the averaged results of 10 trials. In CMC curves, the Rank-κ matching rate is the rate of correct match at rank κ, and the cumulated values of recognition rate at all ranks are recorded as the CMC curve. The parameters in learning of the random kernel forest are illustrated in §3.4. For dense local patches sampling of the images in each dataset, 15×5, 24×8, 19×6 overlapping local patches are extracted in VIPeR, GRID, CUHK01, respectively.

4.2. Empirical Analysis

We investigate how some of the terms in our random kernel forest influence the final re-identification performance. All the analysis and evaluations in this sub-section are based on the VIPeR dataset.

Effect of local kernels: The distance between the query image and each gallery image is the summation of all the pairwise distances of local patches. To tell which local regions contribute the most to discriminate the correct match in the gallery set, we show the similarity distribution of one example image pair in left of Figure 5, from which we can find that these discriminative regions mostly focus on human body parts. In addition, we also show the spatial distributions of the local kernels on one example image in right of Figure 5, which also demonstrates our hypothesis that the inter-camera transforms at different local regions indeed vary accordingly.

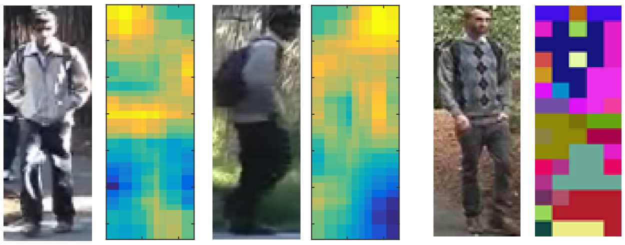


Figure 5. Left-4: similarity distribution of local regions in matching. Right-2: spatial distribution of 127 local kernels in an example image.

Forest diversity: The diversity of trees in the kernel forest is crucial in traditional random forest classifiers. In fact, the accuracy of each single tree is sacrificed in favor of a highly diverse ensemble. Therefore, we vary the number of trees Q in forest and check their influence on the final re-id performance. As shown by the results in Figure 6 (a), a larger number of trees produced higher re-id performance. However, once the number of trees is large enough, the performance becomes stable. Based on the empirical study, we choose the number of trees in our forest as 20, which is relatively small while producing good performance.

Partition of images space and patch-pairs space: As discussed in §2, based on similar motivation that finding a subspace where the cross-view data pair inside have consistent transform, the LAFT [19] partitions the image space while ours partition the more fine local patch space. We thus conduct two tests for evaluation based on the VIPeR dataset, one uses 316 persons in training set (316 gallery images in test) and the other uses only 100 persons in training, resulting in 532 gallery images in test. The performance comparison between the two approaches are shown in Figure 6 (b). We can observe that in the first test, our approach performs better in the range of a small κ (rank 2-15), while in the second more challenging test with much less training samples and a larger gallery set, our performance is obviously much better than the LAFT approach.

4.3. Quantitative Evaluation

In this subsection, we compare our approach to the other existing works on several standard datasets for evaluation.

VIPeR: two protocals were defined for evaluation on this dataset: the first one randomly selects 316 persons to form the training set and results in 316 persons in testing set; the other one randomly selects 100 persons to form the training set and results in 532 persons in test. Our approach is com-

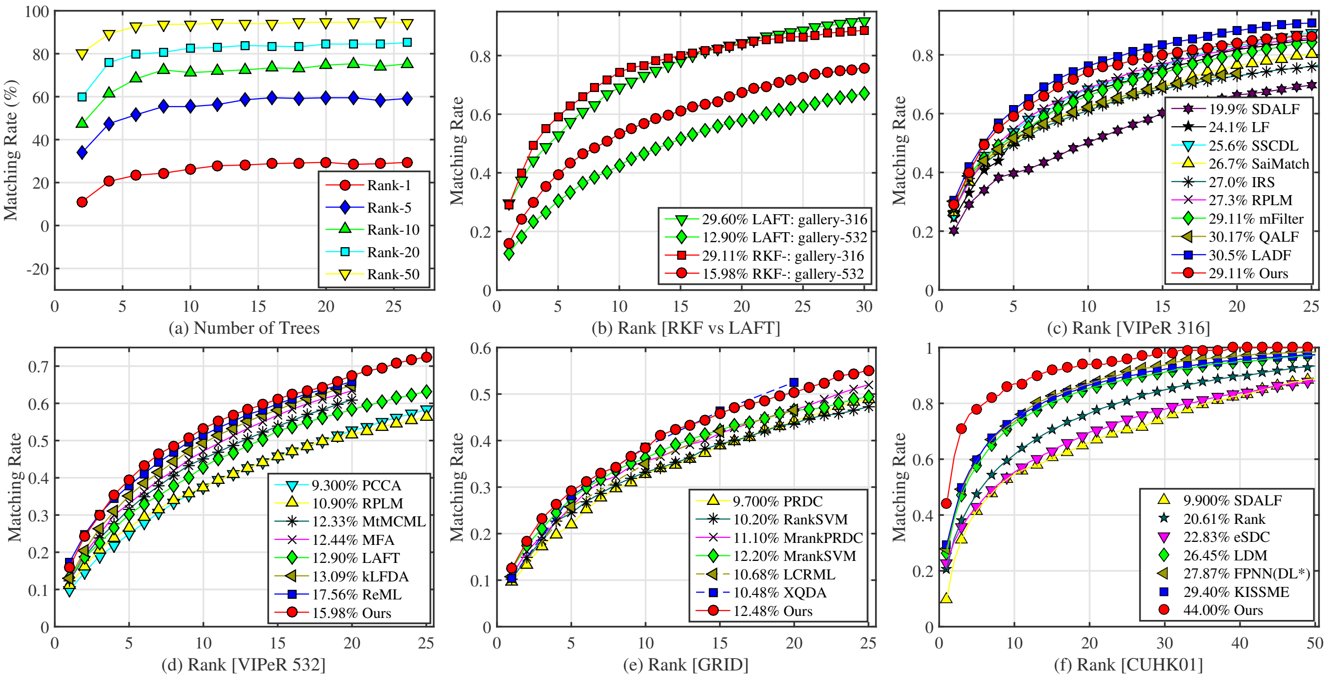


Figure 6. Evaluations: (a) Performance comparison of different numbers of trees in random kernel forest. (b) Comparison between RKF and LAFT via CMC curves. (c) CMC curves on VIPeR dataset with 316 gallery images. (d) CMC curves on VIPeR dataset with 532 gallery images. (e) CMC curves on GRID dataset with 900 gallery images. (f) CMC curves on CUHK01 dataset with 100 gallery images.

Method	$\kappa = 1$	$\kappa = 5$	$\kappa = 10$	$\kappa = 20$	ref
SSCDL	25.6	53.7	68.1	83.6	CVPR14 [26]
SalMat	26.7	50.7	62.4	76.4	CVPR13 [47]
IRS	27.0	49.4	61.1	72.8	PAMI15 [24]
RPLM	27.3	54.5	68.8	82.4	ECCV12 [14]
mFilter	29.1	52.3	66.0	79.9	CVPR14 [48]
QALF	30.2	51.6	62.4	73.8	CVPR15 [49]
LADF	30.5	61.2	76.2	88.2	CVPR13 [22]
Ours	29.1	59.2	74.4	83.8	

Table 1. Top ranked matching rates (%) on VIPeR dataset with 316 gallery images (highest 3 are colored as red, blue and magenta).

Method	$\kappa = 1$	$\kappa = 5$	$\kappa = 10$	$\kappa = 20$	ref
PCCA	9.3	24.9	37.4	52.9	CVPR12 [34]
RPLM	10.9	26.7	37.7	51.6	ECCV12 [14]
MtMCML	12.3	31.6	45.1	61.1	TIP14 [32]
MFA	12.4	33.3	47.2	63.5	ECCV14 [44]
LAFT	12.9	30.3	42.7	58.0	CVPR13 [19]
kLFDA	13.1	35.2	49.4	65.0	ECCV14 [44]
ReML	17.5	37.9	51.8	66.0	TIP15 [5]
Ours	16.0	39.5	53.3	67.4	

Table 2. Top ranked matching rates (%) on VIPeR dataset with 532 gallery images (highest 3 are colored as red, blue and magenta).

pared to the other existing works including: SDALF [10], LF [36], SSCDL [26], SalMat [47], IRS [24], RPLM [14], mFilter [48], QALF [49], LADF [22], PCCA [34], MtMCML [32], MFA [44], LAFT [19], kLFDA [44], ReML [5]. The performance comparison is shown in Figure 6 (c) and (d) by CMC curves. From these results, we can find that our approach gives the second best performance in the first test and the best performance in the second test. We also summarize the performance comparison in Tables 1&2 to show the matching rate values more straightforwardly. It is clear that our approach achieves 29.1% and 16.0% rank-1 matching rate in the two tests, which is very competitive compared to the other results in literature. The rank-20 matching rate for our approach is 83.8% and 67.4% in the two tests, which also outperform most of the other methods.

GRID: experiments on this dataset were conducted according to the 10 data partitions provided along with the dataset. In each partition, the image pairs from 125 randomly selected individuals are used for training, and the rest 125 persons together with the 775 irrelevant distracting images form the gallery set in test. Our approach is compared to some recently published results by: PRDC [50], RankSVM [37], MrankPRDC [27], MrankSVM [27], LCRML [4], XQDA [23] in Figure 6 (e) and Table 3. The CMC curves and top rank matching rates show our approach achieves very competitive results on this benchmark.

CUHK01: this multi-shot dataset contains 971 persons. 100 persons are randomly selected in test, and the rest 871

Method	$\kappa = 1$	$\kappa = 5$	$\kappa = 10$	$\kappa = 20$	ref
PRDC	9.7	22.0	33.0	44.3	CVPR11 [50]
RankSVM	10.2	24.6	33.3	43.7	BMVC10 [37]
MrankPRDC	11.1	26.1	35.8	46.6	ICIP13 [27]
MrankSVM	12.2	27.8	36.3	46.6	ICIP13 [27]
LCRML	10.7	25.8	35.0	46.5	ICPR14 [4]
XQDA	10.5	28.1	38.6	52.6	CVPR15 [23]
Ours	12.5	29.2	38.3	50.3	

Table 3. Top ranked matching rates (%) on GRID dataset with 900 gallery images (highest 3 are colored as red, blue and magenta).

Method	$\kappa = 1$	$\kappa = 5$	$\kappa = 10$	$\kappa = 20$	ref
SDALF	9.9	41.5	54.7	66.0	CVPR10 [10]
Rank	20.6	47.6	61.6	76.5	ICML10 [33]
eSDC	22.8	43.0	55.3	69.7	CVPR13 [47]
LDM	26.5	57.6	72.6	85.5	ICCV09 [12]
FPNN	27.9	59.7	73.4	87.3	CVPR14 [20]
KISSME	29.4	59.8	74.5	86.6	CVPR12 [16]
Ours	44.0	78.5	86.7	94.0	

Table 4. Top ranked matching rates (%) on CUHK01 with 100 gallery images (highest 3 are colored as red, blue and magenta).

persons are used for training. This protocol was designed for deep learning in FPNN [20]. Figure 6 (f) and Table 4 compares the performance of our approach to the other existing works including FPNN, eSDC [47], KISSME [16], LDM [12] *etc*. The results show that our approach outperforms the other existing works by a large margin ($> 15\%$ than FPNN), with the rank-1 matching rate being 44%. In summary, all the above results also show that our approach is able to achieve competitive performance without the strict requirements on training data as in deep learning.

As for the efficiency of the proposed approach, the time cost to discover local transforms in training of the random forest with 20 trees is $\approx$ 18-min on the VIPER dataset with 316 training persons. The testing time for a query image is $<$ 3-sec with 316 gallery persons. Time costs are measured in Matlab on a laptop with i7 2.7G CPU.

5. Conclusion

This paper presented a novel approach based on the random kernel forest for person re-identification across disjoint camera views with complicated appearance variations. The complex inter-camera transformation is modelled by a combination of many local functions, which formulate each local transform in a much simpler but effective manner. Both the decomposition of the overall inter-camera transformation and the local metric kernels for re-identification are discovered automatically by the aligned local training patch-pairs using the random forest framework. Any local patch in a query image is assigned a specific kernel in the tree structure, then the local metric is able to generate a minimized distance between the true patch-pairs. Extensive experimental results showed that the proposed random kernel forest achieved very competitive re-identification performance as compared to the other existing works.

References

[1] E. Ahmed, M. Jones, and T. Marks. An improved deep learning architecture for person re-identification. In *CVPR*, 2015. 2, 3

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

[3] D. Chen, Z. Yuan, G. Hua, N. Zheng, and J. Wang. Similarity learning on an explicit polynomial kernel feature map for person re-identification. In *CVPR*, 2015. 2

[4] J. Chen, Z. Zhang, and Y. Wang. Relevance metric learning for person re-identification by exploiting global similarities. In *ICPR*, 2014. 7, 8

[5] J. Chen, Z. Zhang, and Y. Wang. Relevance metric learning for person re-identification by exploiting listwise similarities. *IEEE Trans. on Image Processing*, 24(12):4741–4756, 2015. 7

[6] A. Criminisi and J. Shotton. Decision forests for computer vision and medical image analysis. *Springer*, 2013. 3, 4

[7] P. Dollar and C. L. Zitnick. Structured forests for fast edge detection. In *ICCV*, 2013. 3, 4

[8] K. Duan, L. Marchesotti, and D. Crandall. Attribute-based vehicle recognition using viewpoint-aware multiple instance svms. In *WACV*, 2014. 3

[9] S. Fanello, C. Keskin, P. Kohli, S. Izadi, J. Shotton, A. Criminisi, U. Pattacini, and T. Paek. Filter forests for learning data-dependent convolution kernels. In *CVPR*, 2014. 3

[10] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010. 2, 7, 8

[11] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an emsemble of local features. In *ECCV*, 2008. 1, 2, 5

[12] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? metric learning approaches for face identification. In *ICCV*, 2009. 8

[13] R. Guo and H. Qi. Partially-sparse restricted boltzmann machine for background modeling and subtraction. In *ICMLA*, 2013. 5

[14] M. Hirzer, P. Roth, M. Kostinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *ECCV*, 2012. 2, 7

[15] X. Jing, X. Zhu, F. Wu, X. You, Q. Liu, D. Yue, R. Hu, and B. Xu. Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning. In *CVPR*, 2015. 2

[16] M. Kostinger, M. Hirzer, P. Wohlhart, P. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012. 2, 8

[17] E. Krupka, A. Vinnikov, B. Klein, A. B. Hillel, and D. Freedman. Discriminative ferns ensemble for hand pose recognition. In *CVPR*, 2014. 3

[18] R. Layne, T. Hospedales, S. Gong, and Q. Mary. Person reidentification by attributes. In *BMVC*, 2012. 3

[19] W. Li and X. Wang. Locally aligned feature transforms across views. In *CVPR*, 2013. 3, 6, 7

[20] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-idnetification. In *CVPR*, 2014. 2, 3, 5, 8

[21] Y. Li, Z. Wu, and R. Radke. Multi-shot re-identification with random projection based random forests. In *WACV*, 2015. 3

[22] Z. Li, S. Chang, F. L. T. Huang, L. Cao, and J. Smith. Learning locally adaptive decision functions for person verification. In *CVPR*, 2013. 1, 2, 7

[23] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015. 2, 7, 8

[24] G. Lisanti, I. Masi, A. D. Bagdanov, and A. D. Bimbo. Person re-identification by iterative re-weighted sparse ranking. *IEEE Trans. on PAMI*, 37(8):1629–1643, 2015. 7

[25] C. Liu, S. Gong, C. Loy, and X. Lin. Person re-identification: what features are important? In *ECCV*, 2012. 2

[26] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu. Semi-supervised coupled dictionary learning for person re-identification. In *CVPR*, 2014. 2, 7

[27] C. Loy, C. Liu, and S. Gong. Person re-identification by manifold ranking. In *ICIP*, 2013. 7, 8

[28] C. Loy and X. Tang. Multi-camera activity correlation analysis. In *CVPR*, 2009. 1, 5

[29] J. Luo and H. Qi. Distributed object recognition via feature unmixing. In *ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC)*, 2010. 2

[30] J. Luo, W. Wang, and H. Qi. Feature extraction and representation for distributed multi-view human action recognition. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 3(2):145–154, 2013. 2

[31] B. Ma, Y. Su, and F. Jurie. Local descriptors encoded by fisher vectors for person re-identification. In *ECCV*, 2012. 2

[32] L. Ma, X. Yang, and D. Tao. Person re-identification over camera networks using multi-task distance metric learning. *IEEE Trans. on Image Processing*, 23(8):3656–3670, 2014. 2, 7

[33] B. Mcfee and G. Lanckriet. Metric learning to rank. In *ICML*, 2010. 8

[34] A. Mignon and F. Jurie. Pcca: A new approach for distance learning from sparse pairwise constraints. In *CVPR*, 2012. 2, 7

[35] S. Paisitkriangkrai, C. Shen, and A. Hengel. Learning to rank in person re-identification with metric ensemble. In *CVPR*, 2015. 2

[36] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian. Local fisher discriminant analysisi for pedestrian re-identification. In *CVPR*, 2013. 1, 2, 5, 7

[37] B. Prosser, W. Zheng, S. Gong, T. Xiang, and Q. Mary. Person re-identification by support vector ranking. In *BMVC*, 2010. 2, 7, 8

[38] M. Ristin, J. Gall, M. Guillaumin, and L. V. Gool. From categories to subcategories: Large-scale image classification with partial class label refinement. In *CVPR*, 2015. 3

[39] Z. Shi, T. Hospedales, and T. Xiang. Transferring a semantic representation for person re-identification and search. In *CVPR*, 2015. 3

[40] W. Wang, J. Luo, and H. Qi. Action recogntion across cameras by exploring reconstructable paths. In *ACM/IEEE International Conference on Distributed Smart Cameras*, 2013. 2

[41] X. Wang. Intelligent multi-camera video surveillance: A review. *Pattern Recognition Letter*, 34:3–19, 2013. 1

[42] X. Wang, R. Guo, and C. Kambhamettu. Deeply-learned feature for age estimation. In *WACV*, 2015. 2

[43] Z. Wu, Y. Li, and R. Radke. Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features. *IEEE Trans. on PAMI*, 37(5):1095–1108, 2015. 2

[44] F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel based metric learning methods. In *ECCV*, 2014. 1, 2, 7

[45] D. Yi, Z. Lei, and S. Li. Deep metric learning for practical person re-identification. In *ICPR*, 2014. 3

[46] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In *ICCV*, 2013. 2

[47] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013. 1, 2, 3, 5, 7, 8

[48] R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identification. In *CVPR*, 2014. 3, 7

[49] L. Zheng, S. Wang, L. Tian, F. He, Z. Liu, and Q. Tian. Query adaptive late fusion for image search and person re-identification. In *CVPR*, 2015. 7

[50] W. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. In *CVPR*, 2011. 2, 7, 8