Person Re-identification by Salience Matching

Rui Zhao Wanli Ouyang Xiaogang Wang
Department of Electronic Engineering, the Chinese University of Hong Kong
{*rzhao, wlouyang, xgwang*}@ee.cuhk.edu.hk

Abstract

Human salience is distinctive and reliable information in matching pedestrians across disjoint camera views. In this paper, we exploit the pairwise salience distribution relationship between pedestrian images, and solve the person re-identification problem by proposing a salience matching strategy. To handle the misalignment problem in pedestrian images, patch matching is adopted and patch salience is estimated. Matching patches with inconsistent salience brings penalty. Images of the same person are recognized by minimizing the salience matching cost. Furthermore, our salience matching is tightly integrated with patch matching in a unified structural RankSVM learning framework. The effectiveness of our approach is validated on the VIPeR dataset and the CUHK Campus dataset. It outperforms the state-of-the-art methods on both datasets.

1. Introduction

Person re-identification is a task of matching persons observed from non-overlapping camera views based on image appearance. It has important applications in video surveillance including threat detection, human retrieval, human tracking, and activity analysis. It saves a lot of human efforts on exhaustively searching for a person from large amounts of video sequences. Nevertheless, person re-identification is a very challenging task. A person observed in different camera views often undergoes significant variations on viewpoints, poses, appearance and illumination, which make intra-personal variations even larger than inter-personal variations. Background clutters and occlusions cause additional difficulties. Our work is mainly motivated by the following several aspects.

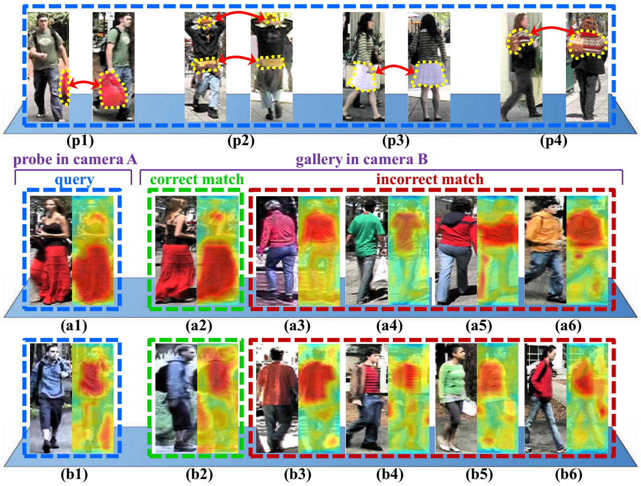


Figure 1. Illustration of human salience and salience matching with examples. In the first row, some salient parts of pedestrians are highlighted with yellow dashed regions. The second row and the third row show examples of salience matching. The salience map of each pedestrian image is shown. **Best viewed in color.**

Misalignments are caused by variations of viewpoints and poses, which are commonly exist in person re-identification. For example in Figure 1, the shoulder of $(b1)$ close to the left boundary becomes a backpack at the same location in $(b2)$. Most existing methods [19, 23, 12, 4, 18] match pedestrian images by directly comparing misaligned features. In our approach, salience matching is integrated with patch matching, and both show robustness to spatial variation and misalignment.

Some local patches are more distinctive and reliable when matching two persons. Some examples are shown in the first row of Figure 1, person $(p1)$ carries a red hand bag, $(p2)$ has an orange cap and a yellow horizontal stripe on his jacket, $(p3)$ wears a white dress, and $(p4)$ is dressed in red sweater with floral texture. Human eyes can easily pick up these persons from other candidates because of these distinctive features. These features can be reliably detected across camera views. If a body part is salient in one camera view, it usually remains salient in another view. However, most existing approaches only consider clothes and trousers as the most important regions for person re-identification. Some distinct features (such as the red bag in $(p1)$) may be considered as outliers to be removed, because they do not belong to body parts. Also, these features may only take up small regions in body parts. If global features are adopted by existing approaches, those small regions have little effect on person matching. In contrast, our approach can well estimate distinctiveness of patches as salience. Patches with high salience values gain large weights in person re-identification, because such patches not only have good discriminative power but also can be reliably detected during patch matching across camera views.

We observe that images of the same person captured from different camera views have some invariance property on their spatial distributions on salience, like pair $(a1, a2)$ in Figure 1. Since the person in image $(a1)$ shows salience in her dress while others in $(a3)$-$(a6)$ have salient blouses. They can be well distinguished simply from the spatial distributions of salience. Therefore, human salience distributions provide useful information in person re-identification. Such information can be encoded during patch matching. If two patches from two images of the same person are matched, they are expected to have the same salience value; otherwise such matching brings salience matching penalty. In the second row in Figure 1, the query image $(b1)$ shows a similar salience distribution as those of gallery images. In this case, visual similarity needs to be considered. This motivates us to relate salience matching penalty to the visual similarity of two matched patches.

Based on above considerations, a new person re-identification approach by salience matching is proposed. This work has three major contributions.

First, a probabilistic distribution of salience is reliably estimated with our approach. Different from general salience detection [6], our salience is especially designed for person re-identification. The estimated human salience is robust across disjoint camera views and is used as a meaningful representation of human appearance in recognition. Second, we formulate person re-identification as a salience matching problem. Dense correspondences between local patches are established based on visual similarity. Matching patches with inconsistent salience brings cost. Images of the same person are recognized by minimizing the salience matching cost, which not only depends on the locations of patches but also the visual similarity of matched patches. Third, salience matching and patch matching are tightly integrated into a unified structural RankSVM learning framework. Structural RankSVM has good training efficiency given a very large number of rank constraints in person re-identification. Moreover, our approach has transformed the original high-dimensional visual feature space to a much lower dimensional salience feature space (80 times lower in this work) to further improve the training efficiency and also avoid overfitting.

The effectiveness of our approach is validated on the VIPeR dataset [7] and the CUHK Campus dataset [12]. It outperforms the state-of-the-art methods on both datasets.

2. Related Works

Existing methods on person re-identification generally fall into two categories: unsupervised and supervised. Our proposed approach is supervised.

Unsupervised Methods. This category mainly focuses on feature design. Farenzena *et al.* [5] proposed the Symmetry-Driven Accumulation of Local Features (SDALF) by exploiting the symmetry property in pedestrian images to handle view variation. Ma *et al.* [16] developed the BiCov descriptor based on the Gabor filters and the covariance descriptor to handle illumination change and background variations. Cheng *et al.* [3] utilized the Pictorial Structures to estimate human body configuration and also computed visual features based on different body parts to cope with pose variations. Lu *et al.* [15] employed assembly of part templates to handle the articulation of human body. Zhao *et al.* [22] proposed an unsupervised salience learning method to exploit discriminative features, but they did not consider salience itself as an important feature for patch matching and person re-identification.

Supervised Methods. Distance metric learning has been widely used in person re-identification [23, 4, 12, 13, 18, 24]. They learn metrics by minimizing the intra-class distances while maximizing the inter-class distances. Their performance is limited by the fact that metric is based on the subtraction of misaligned feature vectors, which causes significant information loss and errors. Li and Wang [11] learned a mixture of cross-view transforms and projected features into a common space for alignment. In contrast, our approach handles the problem of feature misalignment through patch matching. Liu *et al.* [14] allowed user feedback in the learning procedure, which achieved significant improvement over metric learning methods. Some other models have also been employed to extract discriminative features. Gray *et al.* [8] used boosting to select a subset of optimal features for matching pedestrian images. Prosser *et al.* [19] formulated person re-identification as a ranking problem, and learned global feature weights based on an ensemble of RankSVM. RankSVM optimizes over the pairwise differences. In this paper, we employ structural RankSVM [10], which considers the ranking difference rather than pairwise difference.

General image salience has been extensively studied [6]. In the context of person re-identification, human salience is different than general image salience in the way of drawing visual attentions.

3. Human Salience

We compute the salience probability map based on dense correspondence with a K-nearest neighbors (KNN) method.

Algorithm 1 Compute human salience.

```
Input: image $\mathbf{x}^{A,u}$ and a reference image set $\mathcal{R} = \{\mathbf{x}^{B,v},\ v = 1, \ldots, N_r\}$
Output: salience probability map $P(l^{A,u}_{m,n} = 1 \mid x^{A,u}_{m,n})$
for each patch $x^{A,u}_{m,n} \in X$ do
    compute $X_{NN}(x^{A,u}_{m,n})$ with Eq. (1)
    compute $score(x^{A,u}_{m,n})$ with Eq. (2)
    compute $P(l^{A,u}_{m,n} = 1 \mid x^{A,u}_{m,n})$ with Eq. (3)
end for
```

3.1. Dense Correspondence

Dense Features. Before building dense correspondence, local patches on a dense grid are extracted. The patch size is 10×10 and the grid step is 5 pixels. 32-bin color histogram in each of LAB channels and 128-dimensional SIFT features are then computed for each patch. To robustly capture the color information, color histograms are also computed on two other downsampled scales for each patch. The color histograms and SIFT features are normalized with $L2$ norm, and are concatenated to form the final dense local features, *i.e.* a 672-dimensional ($32 \times 3 \times 3 + 128 \times 3$) feature vector for each local patch.

Adjacency Constrained Search. Dense local features for an image are denoted by $\mathbf{x}^{A,u} = \{x^{A,u}_{m,n}\}$, and $x^{A,u}_{m,n}$ represents the feature of a local patch at the m-th row and n-th column in the u-th image from camera view A. When patch $x^{A,u}_{m,n}$ searches for its corresponding patch in the v-th image from camera view B, *i.e.* $\mathbf{x}^{B,v} = \{x^{B,v}_{i,j}\}$, the search set of $x^{A,u}_{m,n}$ in $\mathbf{x}^{B,v}$ is $\mathfrak{S}(x^{A,u}_{m,n}, \mathbf{x}^{B,v}) = \{x^{B,v}_{i,j} | j = 1, \ldots, N, i = \max(0, m - l), \ldots, \min(M, m + l)\}$, where l denotes the halft height of adjacency search space, M is the number of rows, and N is the number of columns. If all pedestrian images are well aligned and there is no vertical pose variation, l shall be set zero. However, misalignment, camera view change, and vertical articulation result in vertical movement of the human body in the image. Thus the relaxed adjacency search is necessary to handle spatial variations. Smaller search space cannot tolerate large spatial variation, while larger search space will increases the chance of mismatch. We choose $l = 2$ in our experiment setting.

Patch matching is widely used, and many off-the-shelf methods [1] are available. We simply do a k-nearest-neighbor search for patch $x^{A,u}_{m,n}$ in its search set $\mathfrak{S}(x^{A,u}_{m,n}, \mathbf{x}^{B,v})$. For each patch $x^{A,u}_{m,n}$, a nearest neighbor is sought from its search set in every image within a reference set. The adjacency constrained search is illustrated in Figure 2.

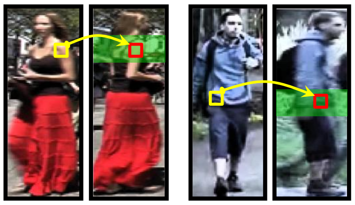


Figure 2. Illustration of adjacency constrained search. Green region represents the adjacency constrained search set of the patch in yellow box. The patch in red box is the target match.

3.2. Unsupervised Salience Learning

Human salience is computed based on previously-built dense correspondence. We utilize the KNN distances to find patch samples in minority, *i.e.* they are unique and special. In the application of person re-identification, we find salient patches that possess property of uniqueness among a reference set $\mathcal{R}$. Denote the number of images in the reference set by N_r. For an image $\mathbf{x}^{A,u} = \{x^{A,u}_{m,n}\}$, a nearest-neighbor (NN) set of size N_r is built for every patch $x^{A,u}_{m,n}$,

$$X_{NN}(x^{A,u}_{m,n}) = \{x | \operatorname*{argmin}_{x^{B,v}_{i,j}} d(x^{A,u}_{m,n}, x^{B,v}_{i,j}), \quad x^{B,v}_{i,j} \in \mathfrak{S}(x^{A,u}_{m,n}, \mathbf{x}^{B,v}), v = 1, \ldots, N_r\}, \tag{1}$$

where $\mathfrak{S}(x^{A,u}_{m,n}, \mathbf{x}^{B,v})$ is the adjacency search set of patch $x^{A,u}_{m,n}$, and function $d(\cdot)$ computes the Euclidean distance between two patch features.

Our goal of computing human salience is to identify patches with special appearance. We use the KNN distances to define the salience score:

$$score(x^{A,u}_{m,n}) = d_k(X_{NN}(x^{A,u}_{m,n})), \tag{2}$$

and the probability of $x^{A,u}_{m,n}$ being a salient patch is

$$P(l^{A,u}_{m,n} = 1 \mid x^{A,u}_{m,n}) = 1 - exp(-score(x^{A,u}_{m,n})^2/\sigma_0^2), \tag{3}$$

where d_k denotes the distance of the k-th nearest neighbor, $l^{A,u}_{m,n}$ is a binary salience label and σ_0 is a bandwidth parameter. We set $k = N_r/2$ in the salience learning scheme with an empirical assumption that a patch is considered to have special appearance such that more than half of the people in the reference set do not share similar patch with it. N_r reference images are randomly sampled from training set, and we set $N_r = 100$ in our experiments. Enlarging the reference dataset will not deteriorate salience detection, because the *salience* is defined in the statistical sense. It is robust as long as the distribution of the reference dataset well reflects the test scenario. Our human salience learning method is summarized in algorithm 1.

4. Supervised Salience Matching

One of the main contributions of this work is to match pedestrian images based on the salience probability map. In contrast with most of the works on person re-identification, which focus on feature selection, feature weighting, or distance metric learning, we instead exploit the consistence property of human salience and incorporate it in person matching. This is based on our observation that person in different camera views shows consistence in the salience probability map, as shown in Figure 1.

4.1. Matching based on Salience

Since matching is applied for arbitrary image pairs, we omit the image index in notation for clarity, *i.e.* change $\mathbf{x}^{A,u}$ to $\mathbf{x}^A$ and $\mathbf{x}^{B,v}$ to $\mathbf{x}^B$. Also, the patch notations are changed accordingly, *i.e.* $x^{A,u}_{m,n}$ to $x^A_{p_i}$ and $x^{B,v}_{i,j}$ to $x^B_{p'_i}$, where p_i is the patch index in image $\mathbf{x}^A$ and p'_i is the corresponding matched patch index in image $\mathbf{x}^B$ produced by previously built dense correspondence. To incorporate the salience into matching, we introduce $\mathbf{l}^A = \{l^A_{p_i} \mid l^A_{p_i} \in \{0,1\}\}$ and $\mathbf{l}^B = \{l^B_{p'_i} \mid l^B_{p'_i} \in \{0,1\}\}$ as salience labels for all the patches in image $\mathbf{x}^A$ and $\mathbf{x}^B$ respectively. If all the salience labels are known, we can perform person matching by computing salience matching score as follows:

$$\begin{aligned} f_z(\mathbf{x}^A, \mathbf{x}^B, \mathbf{l}^A, \mathbf{l}^B; \mathbf{p}, \mathbf{z}) = \\ \sum_{p_i} \Big\{ z_{p_i,1} l^A_{p_i} l^B_{p'_i} + z_{p_i,2} l^A_{p_i}(1 - l^B_{p'_i}) \\ + z_{p_i,3}(1 - l^A_{p_i}) l^B_{p'_i} + z_{p_i,4}(1 - l^A_{p_i})(1 - l^B_{p'_i}) \Big\}, \end{aligned} \tag{4}$$

where $\mathbf{p} = \{(p_i, p'_i)\}$ are dense correspondence patch index pairs, and $\mathbf{z} = \{z_{p_i,k}\}_{k=1,2,3,4}$ are the matching scores for four different salience matching results at one local patch. $z_{p_i,k}$ is not a constant for all the patches. Instead, it depends on the spatial location p_i. For example, the score of matching patches on the background should be different than those on legs. $z_{p_i,k}$ also depends on the visual similarity between patch $x^A_{p_i}$ and patch $x^B_{p'_i}$,

$$s(x^A_{p_i}, x^B_{p'_i}) = \exp\Big(- \frac{d(x^A_{p_i}, x^B_{p'_i})^2}{2\sigma_0^2} \Big), \tag{5}$$

where σ_0 is bandwidth of the Gaussian function. Instead of directly using the Euclidean distance $d(x^A_{p_i}, x^B_{p'_i})$, we convert it to similarity to reduce the side effect in summation of very large distances in incorrect matching, which may be caused by misalignment, occlusion, or background clutters.

Therefore, we define the matching score $z_{p_i,k}$ as a linear function of the similarity as follows,

$$z_{p_i,k} = \alpha_{p_i,k} \cdot s(x^A_{p_i}, x^B_{p'_i}) + \beta_{p_i,k}, \tag{6}$$

where $\alpha_{p_i,k}$ and $\beta_{p_i,k}$ are weighting parameters. Thus Eq.(4) jointly considers salience matching and visual similarity.

Since the salience label $l^A_{p_i}$ and $l^B_{p'_i}$ in Eq.(4) are hidden variables, they can be marginalized by computing the expectation of the salience matching score as

$$\begin{aligned} & f^*(\mathbf{x}^A, \mathbf{x}^B; \mathbf{p}, \mathbf{z}) \\ & = \sum_{\mathbf{l}^A, \mathbf{l}^B} f_z(\mathbf{x}^A, \mathbf{x}^B, \mathbf{l}^A, \mathbf{l}^B; \mathbf{p}, \mathbf{z}) p(\mathbf{l}^A, \mathbf{l}^B | \mathbf{x}^A, \mathbf{x}^B) \\ & = \sum_{p_i} \sum_{k=1}^{4} \Big[\alpha_{p_i,k} \cdot s(x^A_{p_i}, x^B_{p'_i}) + \beta_{p_i,k} \Big] c_{p_i,k}(x^A_{p_i}, x^B_{p'_i}), \end{aligned} \tag{7}$$

where $c_{p_i,k}(x^A_{p_i}, x^B_{p'_i})$ is the probabilistic salience matching cost depending on salience probabilities $P(l^A_{p_i} = 1 \mid x^A_{p_i})$ and $P(l^B_{p'_i} = 1 \mid x^B_{p'_i})$ given in Eq.(3),

$$\begin{aligned} & c_{p_i,k}(x^A_{p_i}, x^B_{p'_i}) \\ & = \begin{cases} P(l^A_{p_i} = 1 \mid x^A_{p_i}) P(l^B_{p'_i} = 1 \mid x^B_{p'_i}), & k = 1, \\ P(l^A_{p_i} = 1 \mid x^A_{p_i}) P(l^B_{p'_i} = 0 \mid x^B_{p'_i}), & k = 2, \\ P(l^A_{p_i} = 0 \mid x^A_{p_i}) P(l^B_{p'_i} = 1 \mid x^B_{p'_i}), & k = 3, \\ P(l^A_{p_i} = 0 \mid x^A_{p_i}) P(l^B_{p'_i} = 0 \mid x^B_{p'_i}), & k = 4. \end{cases} \end{aligned} \tag{8}$$

To better formulate this learning problem, we extract out all the weighting parameters in Eq.(7) as $\mathbf{w}$, and have

$$\begin{aligned} f^*(\mathbf{x}^A, \mathbf{x}^B; \mathbf{p}, \mathbf{z}) &= \mathbf{w}^{\mathrm{T}} \Phi(\mathbf{x}^A, \mathbf{x}^B; \mathbf{p}) \\ &= \sum_{p_i} w^{\mathrm{T}}_{p_i} \phi(x^A_{p_i}, x^B_{p'_i}), \end{aligned} \tag{9}$$

where

$$\begin{aligned} \Phi(\mathbf{x}^A, \mathbf{x}^B; \mathbf{p}) &= [\phi(x^A_{p_1}, x^B_{p'_1})^{\mathrm{T}}, \ldots, \phi(x^A_{p_N}, x^B_{p'_N})^{\mathrm{T}}]^{\mathrm{T}}, \\ \mathbf{w} &= [w_{p_1}, \ldots, w_{p_N}]^{\mathrm{T}}, \\ w_{p_i} &= [\{\alpha_{p_i,k}\}_{k=1,2,3,4}, \ \{\beta_{p_i,k}\}_{k=1,2,3,4}]. \end{aligned} \tag{10}$$

$\Phi(\mathbf{x}^A, \mathbf{x}^B; \mathbf{p})$ is the feature map describing the matching between $\mathbf{x}^A$ and $\mathbf{x}^B$. For each patch p_i, the matching feature $\phi(x^A_{p_i}, x^B_{p'_i})$ is an eight dimensional vector:

$$\phi(x^A_{p_i}, x^B_{p'_i}) = \begin{bmatrix} s(x^A_{p_i}, x^B_{p'_i}) P(l^A_{p_i} = 1 \mid x^A_{p_i}) P(l^B_{p'_i} = 1 \mid x^B_{p'_i}) \\ s(x^A_{p_i}, x^B_{p'_i}) P(l^A_{p_i} = 1 \mid x^A_{p_i}) P(l^B_{p'_i} = 0 \mid x^B_{p'_i}) \\ s(x^A_{p_i}, x^B_{p'_i}) P(l^A_{p_i} = 0 \mid x^A_{p_i}) P(l^B_{p'_i} = 1 \mid x^B_{p'_i}) \\ s(x^A_{p_i}, x^B_{p'_i}) P(l^A_{p_i} = 0 \mid x^A_{p_i}) P(l^B_{p'_i} = 0 \mid x^B_{p'_i}) \\ P(l^A_{p_i} = 1 \mid x^A_{p_i}) P(l^B_{p'_i} = 1 \mid x^B_{p'_i}) \\ P(l^A_{p_i} = 1 \mid x^A_{p_i}) P(l^B_{p'_i} = 0 \mid x^B_{p'_i}) \\ P(l^A_{p_i} = 0 \mid x^A_{p_i}) P(l^B_{p'_i} = 1 \mid x^B_{p'_i}) \\ P(l^A_{p_i} = 0 \mid x^A_{p_i}) P(l^B_{p'_i} = 0 \mid x^B_{p'_i}) \end{bmatrix}. \tag{11}$$

As shown in Eq.(11), the pairwise feature map $\Phi(\mathbf{x}^A, \mathbf{x}^B; \mathbf{p})$ combines the salience probability map with appearance matching similarities. For each query image $\mathbf{x}^A$, the images in the gallery are ranked according to the expectations of salience matching scores in Eq.(7). There are three advantages of matching with human salience : (1) the human salience probability distribution is more invariant than other features in different camera views; (2) because the salience probability map is built based on dense correspondence, so it inherits the property of tolerating spatial variation; and (3) it can be weighted by visual similarity to improve the performance of person re-identification. We will present the details in next section by formulating the person re-identification problem with $\Phi(\mathbf{x}^A, \mathbf{x}^B; \mathbf{p})$ in structural RankSVM framework, and the effectiveness of salience matching will be shown in experimental results.

4.2. Ranking by Partial Order

We cast person re-identification as a ranking problem for training. The ranking problem will be solved by finding an optimal partial order, which will be mathematically defined in Eq.(12)(13)(16). Given a dataset of pedestrian images, $\mathcal{D}^A = \{\mathbf{x}^{A,u}, id^{A,u}\}_{u=1}^{U}$ from camera view A and $\mathcal{D}^B = \{\mathbf{x}^{B,v}, id^{B,v}\}_{v=1}^{V}$ from camera view B, where $\mathbf{x}^{A,u}$ is the u-th image, y_u is its identity label, and U is the total number of images in $\mathcal{D}^A$. Similar notations apply for variables of camera view B. Each image $\mathbf{x}^{A,u}$ has its relevant images (same identity) and irrelevant images (different identities) in dataset $\mathcal{D}^B$. Our goal is to learn the weight parameters $\mathbf{w}$ that order relevant gallery images before irrelevant ones. For the image $\mathbf{x}^{A,u}$, the orders in its groundtruth ranking are not all known, *i.e.*, we rank the relevant images before irrelevant ones, but no information of the orders within relevant images or irrelevant ones is provided in groundtruth. The partial order $\mathbf{y}^{A,u}$ is denoted as,

$$\mathbf{y}^{A,u} = \{y^{A,u}_{v,v'}\}, \quad y^{A,u}_{v,v'} = \begin{cases} +1 & \mathbf{x}^{B,v} \prec \mathbf{x}^{B,v'}, \\ -1 & \mathbf{x}^{B,v} \succ \mathbf{x}^{B,v'}, \end{cases} \tag{12}$$

where $\mathbf{x}^{B,v} \prec \mathbf{x}^{B,v'}$ ($\mathbf{x}^{B,v} \succ \mathbf{x}^{B,v'}$) represents that $\mathbf{x}^{B,v}$ is ranked before (after) $\mathbf{x}^{B,v'}$ in partial order $\mathbf{y}^{A,u}$.

The partial order feature [9, 17] is appropriate for our goal and can well encode the difference between relevant pairs and irrelevant pairs with only partial orders. The partial order feature for image $\mathbf{x}^{A,u}$ is formulated as,

$$\Psi_{po}(\mathbf{x}^{A,u}, \mathbf{y}^{A,u}; \{\mathbf{x}^{B,v}\}_{v=1}^{V}, \{\mathbf{p}^{u,v}\}_{v=1}^{V}) = \sum_{\substack{\mathbf{x}^{B,v} \in S^{+}_{\mathbf{x}^{A,u}} \\ \mathbf{x}^{B,v'} \in S^{-}_{\mathbf{x}^{A,u}}}} \sum y^{A,u}_{v,v'} \frac{\Phi(\mathbf{x}^{A,u}, \mathbf{x}^{B,v}; \mathbf{p}^{u,v}) - \Phi(\mathbf{x}^{A,u}, \mathbf{x}^{B,v'}; \mathbf{p}^{u,v'})}{|S^{+}_{\mathbf{x}^{A,u}}| \cdot |S^{-}_{\mathbf{x}^{A,u}}|}, \tag{13}$$

$$S^{+}_{\mathbf{x}^{A,u}} = \{\mathbf{x}^{B,v} \mid id^{B,v} = id^{A,u}\}, \tag{14}$$

$$S^{-}_{\mathbf{x}^{A,u}} = \{\mathbf{x}^{B,v} \mid id^{B,v} \neq id^{A,u}\}, \tag{15}$$

where $\{\mathbf{p}^{u,v}\}_{v=1}^{V}$ are the dense correspondences between image $\mathbf{x}^{A,u}$ and every gallery image $\mathbf{x}^{B,v}$, $S^{+}_{\mathbf{x}^{A,u}}$ is relevant image set of $\mathbf{x}^{A,u}$, $S^{-}_{\mathbf{x}^{A,u}}$ is irrelevant image set, $\Phi(\mathbf{x}^{A,u}, \mathbf{x}^{B,v}; \mathbf{p}^{u,v})$ is the feature map defined in Eq.(10), and the difference vector of two feature maps $\Phi(\mathbf{x}^{A,u}, \mathbf{x}^{B,v}; \mathbf{p}^{u,v}) - \Phi(\mathbf{x}^{A,u}, \mathbf{x}^{B,v'}; \mathbf{p}^{u,v'})$ is added if $\mathbf{x}^{B,v} \prec \mathbf{x}^{B,v'}$ and subtracted otherwise.

A partial order may correspond to multiple rankings. Our task is to find a good ranking satisfying the optimal partial order $\mathbf{y}^{A,u}_{*}$ that maximizes following score function,

$$\mathbf{y}^{A,u}_{*} = \underset{\mathbf{y}^{A,u} \in \mathcal{Y}^{A,u}}{\operatorname{argmax}} \ \mathbf{w}^{\mathrm{T}} \Psi_{po}(\mathbf{x}^{A,u}, \mathbf{y}^{A,u}; \{\mathbf{x}^{B,v}\}_{v=1}^{V}, \{\mathbf{p}^{u,v}\}_{v=1}^{V}), \tag{16}$$

where $\mathcal{Y}^{A,u}$ is space consisting of all possible partial orders. As discussed in [9, 21], the good ranking can be obtained simply by sorting gallery images by $\{\mathbf{w}^{\mathrm{T}}\Phi(\mathbf{x}^{A,u}, \mathbf{x}^{B,v}; \mathbf{p}^{u,v})\}_v$ in descending order. The remaining problem is how to learn $\mathbf{w}$.

4.3. Structural RankSVM

In this work, we employ structural SVM to learn the weighting parameters $\mathbf{w}$. Different than many previous SVM-based approaches optimizing over the pairwise differences (*e.g.*, [2, 19]), structural SVM optimizes over ranking differences and it can incorporate non-linear multivariate loss functions directly into global optimization in SVM training.

Objective function. Our goal is to learn a linear model and the training is based on n-slack structural SVM [10]. The objective function is as follows,

$$\min_{\mathbf{w},\xi} \ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{u=1}^{U} \xi_u, \tag{17}$$

$$\begin{aligned} s.t. \ \ & \mathbf{w}^{\mathrm{T}} \delta\Psi_{po}(\mathbf{x}^{A,u}, \mathbf{y}^{A,u}, \hat{\mathbf{y}}^{A,u}; \{\mathbf{x}^{B,v}\}_{v=1}^{V}, \{\mathbf{p}^{u,v}\}_{v=1}^{V}) \\ & \qquad \geq \Delta(\mathbf{y}^{A,u}, \hat{\mathbf{y}}^{A,u}) - \xi_u, \\ & \forall \hat{\mathbf{y}}^{A,u} \in \mathcal{Y}^{A,u} \setminus \mathbf{y}^{A,u}, \ \xi_u \geq 0, \ for \ u = 1, \ldots, U, \end{aligned}$$

where $\delta\Psi_{po}$ is defined as

$$\begin{aligned} & \delta\Psi_{po}(\mathbf{x}^{A,u}, \mathbf{y}^{A,u}, \hat{\mathbf{y}}^{A,u}; \{\mathbf{x}^{B,v}\}_{v=1}^{V}, \{\mathbf{p}^{u,v}\}_{v=1}^{V}) \\ & \quad = \Psi_{po}(\mathbf{x}^{A,u}, \mathbf{y}^{A,u}; \{\mathbf{x}^{B,v}\}_{v=1}^{V}, \{\mathbf{p}^{u,v}\}_{v=1}^{V}) \\ & \quad\quad - \Psi_{po}(\mathbf{x}^{A,u}, \hat{\mathbf{y}}^{A,u}; \{\mathbf{x}^{B,v}\}_{v=1}^{V}, \{\mathbf{p}^{u,v}\}_{v=1}^{V}), \end{aligned} \tag{18}$$

$\mathbf{w}$ is the weight vector, C is a parameter to balance between the margin and the training error, $\mathbf{y}^{A,u}$ is a correct partial order that ranks all correct matches before incorrect matches, and $\hat{\mathbf{y}}^{A,u}$ is an incorrect partial order that violates some of the pairwise relations, *e.g.* a correct match is ranked after an incorrect match in $\hat{\mathbf{y}}^{A,u}$. The constraints in Eq. (17) force the discriminant score of correct partial order $\mathbf{y}^{A,u}$ to be larger than that of incorrect one $\hat{\mathbf{y}}^{A,u}$ by a margin, which is determined by a loss function $\Delta(\mathbf{y}^{A,u}, \hat{\mathbf{y}}^{A,u})$ and a slack variable ξ_u.

AUC loss function. Many loss functions can be applied in structural SVM. In the application of person re-identification, we choose the ROC Area loss, which is also

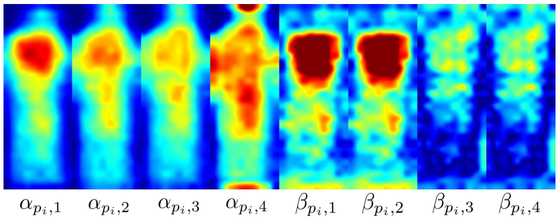


Figure 3. We normalize the learnt weight vector $\mathbf{w}$ to a 2-dimensional importance map for different spatial location. Eight importance maps correspond to $\{\alpha_{p_i,k}\}_{k=1,2,3,4}$ and $\{\beta_{p_i,k}\}_{k=1,2,3,4}$ in Eq. (7).

known as Area Under Curve (AUC) loss. It is computed from the number of swapped pairs,

$$N_{swap} = \{(v, v') : \mathbf{x}^{B,v} \succ \mathbf{x}^{B,v'} \; and \\ \mathbf{w}^{\mathrm{T}}\Phi(\mathbf{x}^{A,u}, \mathbf{x}^{B,v}; \mathbf{p}^{u,v}) < \mathbf{w}^{\mathrm{T}}\Phi(\mathbf{x}^{A,u}, \mathbf{x}^{B,v'}; \mathbf{p}^{u,v'})\}, \tag{19}$$

i.e. the number of pairs of samples that are not ranked in correct order. In the case of partial order ranking, the loss function is

$$\begin{aligned}\Delta(\mathbf{y}^{A,u}, \hat{\mathbf{y}}^{A,u}) &= |N_{swap}|/|S^{+}_{\mathbf{x}^{A,u}}| \cdot |S^{-}_{\mathbf{x}^{A,u}}|, \\ &= \sum_{v,v'} (1 - \hat{y}^{A,u}_{v,v'})/(2 \cdot |S^{+}_{\mathbf{x}^{A,u}}| \cdot |S^{-}_{\mathbf{x}^{A,u}}|).\end{aligned} \tag{20}$$

We note that there are an exponential number of constraints in Eq.(17) due to the huge dimensionality of $\mathcal{Y}^{A,u}$. [10] shows that the problem can be efficiently solved by a cutting plane algorithm. In our problem, the discriminative model is learned by structural RankSVM algorithm, and the weight vector $\mathbf{w}$ in our model means how important it is for each term in Eq.(11). In Eq.(11), $\{\alpha_{p_i,k}\}_{k=1,2,3,4}$ correspond to the first four terms based on salience matching with visual similarity, and $\{\beta_{p_i,k}\}_{k=3,4}$ correspond to the last four terms only depending on salience matching.

We visualize the learning result of $\mathbf{w}$ in Figure 3, and find that the first four terms in Eq.(11) are heavily weighted in the central part of human body which implies the importance of salience matching based on visual similarity. $\{\beta_{p_i,k}\}_{k=1,2}$ are not relevant to visual similarity and they correspond to the two cases when $l^{A}_{p_i} = 1$, *i.e.* the patches on the query images are salient. It is observed that their weighting maps are highlighted on the upper body, which matches to our observation that salient patches usually appear on the upper body. $\{\beta_{p_i,k}\}_{k=3,4}$ are not relevant to visual similarity either, but they correspond to the cases when $l^{A}_{p_i} = 0$, *i.e.* the patches on the query images are not salient. We find that their weights are very low on the whole maps. It means that non-salient patches on query images have little effect on person re-identification if the contribution of visual similarity is not considered.

5. Experimental Results

We evaluate our approach on two public datasets, *i.e.* the VIPeR dataset [7], and the CUHK Campus dataset [12]. The VIPeR dataset is the mostly used person re-identification dataset for evaluation, and the recently published CUHK Campus dataset contains more images than VIPeR (3884 $vs.$ 1264 specifically). Both are very challenging datasets for person re-identification because they contain significant variations on viewpoints, poses, and illuminations, and their images are in low resolutions, with occlusions and background clutters. All the quantitative results are reported in standard Cumulated Matching Characteristics (CMC) curves [20].

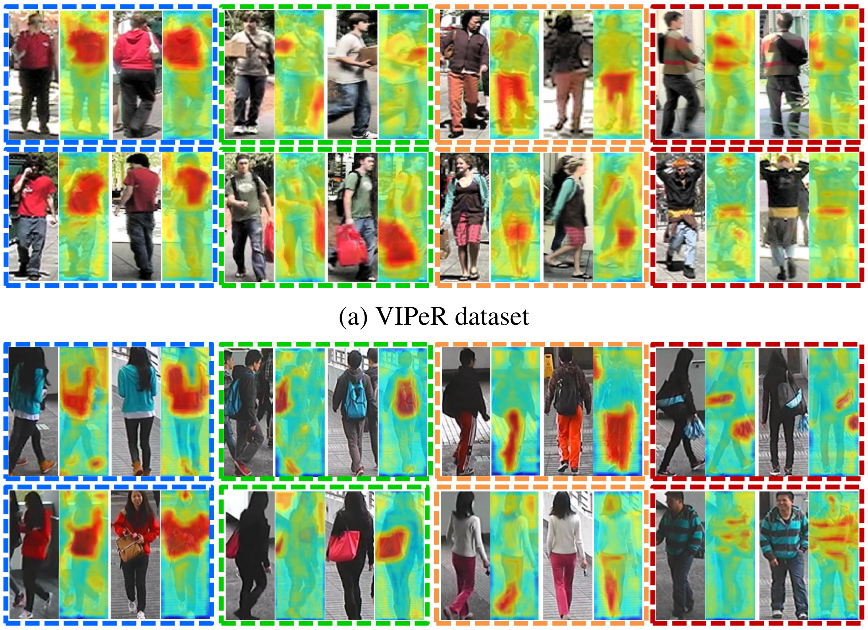
Evaluation Protocol. Our experiments on both datasets follow the evaluation protocol in [8], *i.e.* we randomly partition the dataset into two even parts, 50% for training and 50% for testing, without overlap on person identities. Images from camera A are used as probe and those from camera B as gallery. Each probe image is matched with every image in gallery, and the rank of correct match is obtained. Rank-k recognition rate is the expectation of correct match at rank k, and the cumulated values of recognition rate at all ranks is recorded as one-trial CMC result. 10 trials of evaluation are conducted to achieve stable statistics, and the expectation is reported. We denote our salience matching approach by $SalMatch$. To validate the usefulness of salience matching, we repeat all the training and testing evaluation on our approach, but without using salience. This control experiment is denoted by $PatMatch$.

VIPeR Dataset [7]. The VIPeR dataset [1] contains images from two cameras in outdoor academic environment. It contains 632 pedestrian pairs, and each pair contains two images of the same person observed from different camera views. Most of the image pairs show viewpoint change larger than 90 degrees. All images are normalized to 128 × 48 for experiments.

On VIPeR dataset, comparing $PatMatch$ and $SalMatch$ with several existing unsupervised methods, *i.e.* SDALF [5], CPS [3], eBiCov [16] and eSDC [22], experimental results show significant improvements in Figure 5 (a).

We also compare our approaches with six alternative supervised learning methods, including four benchmarking distance metric learning methods, *i.e.* PRDC [23], LMNN-R [4], PCCA [18], and attribute-based PRDC (aPRDC) [13], a boosting approach (ELF) [8] and Rank SVM (RankSVM) [19]. As seen from the the comparison results in Figure 5 (a), our approach $SalMatch$ achieves 30.16% at rank one with standard deviation 1.23%, and outperforms all these methods. The control experiment $PatMatch$ achieves 26.90%, which shows the effectiveness of integrating salience matching into patch matching. For distance metric learning methods, they ignore the domain knowledge of person re-identification that pedestrian images suf-

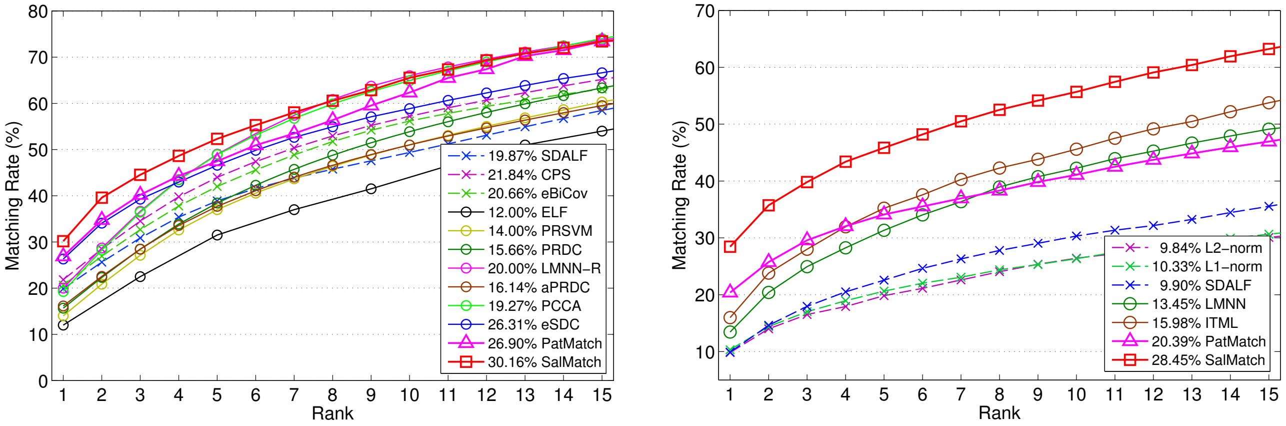
[1] The VIPeR dataset is available to download at: http://vision.soe.ucsc.edu/?q=node/178



(a) VIPeR dataset

(b) CUHK Campus dataset

Figure 4. Some interesting examples of salience matching in our experiments. This figure shows four categories of salience probability types: salience in upper body (in blue dashed box), salience of taking bags (in green dashed box), salience of lower body (in orange dashed box), and salience of stripes on human body (in red dashed box). **Best viewed in color**.



(a) VIPeR dataset

(b) CUHK Campus dataset

Figure 5. CMC statistics on the VIPeR dataset and the CUHK Campus dataset. (a) On VIPeR dataset, our approach ($PatMatch$ and $SalMatch$) is compared with benchmarking methods including SDALF [5], CPS [3], eBiCov [16], eSDC [22], ELF [8], PRSVM [19], PRDC [23], LMNN-R [4], aPRDC [13], and PCCA [18]; (b) On CUHK Campus dataset, our approach is compared with $L1$-norm distance, $L2$-norm distance, SDALF, LMNN [12], and ITML [12]. All the rank-1 performances are marked in the front of method names. Unsupervised methods are drawn in dashed lines while supervised method in solid lines.

fer spatial variation caused by misalignment and pose variation, as discussed in Section 2. Among these metric learning approaches, although the aPRDC also tries to find the unique and inherent appearance property of pedestrian images, our approach is almost more than doubled on rank-1 accuracy against aPRDC. aPRDC weights global features instead local patches based on their distinctiveness. It did not consider the consistency of salience distribution as a cue or matching pedestrian images. ELF gains a lower performance since it selects features in original feature space in which features of different identities are highly correlated. The RankSVM also formulate person re-identification as a ranking problem, but ours shows much better performance because it adopts discriminative salience matching strategy for pairwise matching, and the structural SVM incorporates ranking loss in global optimization. This implies the importance of exploiting human salience matching and the effectiveness of structural SVM training.

CUHK Campus Dataset [12]. The CUHK Campus dataset [2] is also captured with two camera views in a campus environment. Different than the VIPeR dataset, images in this dataset are of higher resolution and are more suitable to show the effectiveness of salience matching. The CUHK Campus dataset contains 971 persons, and each person has two images in each camera view. Camera A captures the frontal view or back view of pedestrians, while camera B captures the side views. All the images are normalized to 160×60 for evaluations.

Since no unsupervised methods are published on the CUHK Campus dataset, so we compare with L_1-norm and L_2-norm distances of our dense features introduced in Section 3.1. Features of all local patches are directly concatenated regardless of spatial misalignment problem (therefore, patch matching is not used), and the pairwise distance is simply computed by L_1-norm and L_2-norm. Also, we compare with the result of a benchmarking unsupervised method, *i.e*. SDALF [5], which is obtained by running the original implementation[3] on the CUHK Campus dataset. As shown in Figure 5 (b), our approach greatly outperforms these unsupervised methods.

Our approach is also compared with available results of distance learning methods including LMNN [12], and ITML [12]. On the CUHK Campus dataset, $SalMatch$ obtains a matching rate of 28.45% at rank one with standard deviation 1.02% while $PatMatch$ achieves 20.39%. Apparently, our salience matching approach outperforms the others methods, and similar conclusions as in the VIPeR dataset can be drawn from the comparisons.

6. Conclusion

In this paper, we formulate person re-identification as a salience matching problem. The dense correspondences of local patches are established by patch matching. Salience probability maps of pedestrian images are reliably estimated to find the distinctive local patches. Matching patches with inconsistent salience brings penalty. Images of the same person are recognized by minimizing the salience matching cost. We tightly integrate patch matching and salience matching in the partial order feature and feed them into a unified structural RankSVM learning framework. Experimental results show our salience matching approach greatly improved the performance of person re-identification.

7. Acknowledgement

This work is supported by the General Research Fund sponsored by the Research Grants Council of Hong Kong (Project No. CUHK 417110, CUHK 417011, CUHK 429412) and National Natural Science Foundation of China (Project No. 61005057).

References

[1] C. Barnes, E. Shechtman, D. B. Goldman, and A. Finkelstein. The generalized patchmatch correspondence algorithm. In *ECCV*. 2010. 3

[2] B. Carterette and D. Petkova. Learning a ranking from pairwise preferences. In *ACM SIGIR*, 2006. 5

[3] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In *BMVC*, 2011. 2, 6, 7

[4] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja. Pedestrian recognition with a learned metric. In *ACCV*. 2011. 1, 2, 6, 7

[5] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010. 2, 6, 7, 8

[6] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. *PAMI*, 2012. 2

[7] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *PETS*, 2007. 2, 6

[8] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*. 2008. 2, 6, 7

[9] T. Joachims. A support vector method for multivariate performance measures. In *ICML*, 2005. 5

[10] T. Joachims, T. Finley, and C.-N. J. Yu. Cutting-plane training of structural svms. *Machine Learning*, 2009. 2, 5, 6

[11] W. Li and X. Wang. Locally aligned feature transforms across views. In *CVPR*, 2013. 2

[12] W. Li, R. Zhao, and X. Wang. Human reidentification with transferred metric learning. In *ACCV*, 2012. 1, 2, 6, 7, 8

[13] C. Liu, S. Gong, C. C. Loy, and X. Lin. Person re-identification: what features are important? In *ECCV*, 2012. 2, 6, 7

[14] C. Liu, C. C. Loy, S. Gong, and G. Wang. Pop: Person re-identification post-rank optimisation. In *ICCV*, 2013. 2

[15] Y. Lu, L. Lin, and W.-S. Zheng. Human re-identification by matching compositional template with cluster sampling. In *ICCV*, 2013. 2

[16] B. Ma, Y. Su, and F. Jurie. Bicov: a novel image representation for person re-identification and face verification. 2012. 2, 6, 7

[17] B. McFee and G. Lanckriet. Metric learning to rank. 2010. 5

[18] A. Mignon and F. Jurie. Pcca: A new approach for distance learning from sparse pairwise constraints. In *CVPR*, 2012. 1, 2, 6, 7

[19] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary. Person re-identification by support vector ranking. In *BMVC*, 2010. 1, 2, 5, 6, 7

[20] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu. Shape and appearance context modeling. In *ICCV*, 2007. 6

[21] Y. Yue, T. Finley, F. Radlinski, and T. Joachims. A support vector method for optimizing average precision. In *ACM SIGIR*, 2007. 5

[22] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013. 2, 6, 7

[23] W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. In *CVPR*, 2011. 1, 2, 6, 7

[24] W.-S. Zheng, S. Gong, and T. Xiang. Re-identification by relative distance comparison. In *PAMI*, 2013. 2

[2] The CUHK Campus is available to download at: http://www.ee.cuhk.edu.hk/~xgwang/CUHK_identification.html

[3] The implementation of SDALF method is provided by authors at the website: http://www.lorisbazzani.info/code-datasets/sdalf-descriptor/