Person Re-identification by saliency Learning

Rui Zhao, *Student Member, IEEE,* Wanli Oyang, *Member, IEEE,* and Xiaogang Wang, *Member, IEEE*

Abstract—Human eyes can recognize person identities based on small salient regions, i.e. person saliency is distinctive and reliable in pedestrian matching across disjoint camera views. However, such valuable information is often hidden when computing similarities of pedestrian images with existing approaches. Inspired by our user study result of human perception on person saliency, we propose a novel perspective for person re-identification based on learning person saliency and matching saliency distribution. The proposed saliency learning and matching framework consists of four steps: (1) To handle misalignment caused by drastic viewpoint change and pose variations, we apply adjacency constrained patch matching to build dense correspondence between image pairs. (2) We propose two alternative methods, i.e. K-Nearest Neighbors and One-class SVM, to estimate a saliency score for each image patch, through which distinctive features stand out without using identity labels in the training procedure. (3) saliency matching is proposed based on patch matching. Matching patches with inconsistent saliency brings penalty, and images of the same identity are recognized by minimizing the saliency matching cost. (4) Furthermore, saliency matching is tightly integrated with patch matching in a unified structural RankSVM learning framework. The effectiveness of our approach is validated on the four public datasets. Our approach outperforms the state-of-the-art person re-identification methods on all these datasets.

Index Terms—Person re-identification, person saliency, patch matching, video surveillance.

1 Introduction

Person re-identification [7], [17], [58] is to match pedestrians observed from non-overlapping camera views based on image appearance. It has important applications in video surveillance such as human retrieval, human tracking, and activity analysis. It saves a lot of human efforts on exhaustively searching for a person from large amounts of images and videos. Nevertheless, person re-identification is a very challenging task. A person observed in different camera views undergoes significant variations on viewpoints, poses, and illumination, which make intra-personal variations even larger than inter-personal variations. Image blurring, background clutters and occlusions also cause additional difficulties.

Variations of viewpoints and poses commonly exist in person re-identification, and cause misalignment between images. In Figure 1, the lower right region of $(p1a)$ is a red bag, while a leg appears in this region in $(p1b)$; the central region of $(p3a)$ is an arm, while it becomes a backpack in $(p3b)$. Most existing methods [14], [36], [50], [53], [69] match pedestrian images by first computing the difference of feature vectors and then the similarities based on such difference vectors, which is problematic due to the spatial misalignment. In our work, patch matching is employed to handle misalignment, and it is integrated with saliency matching to improve the discriminative power and robustness to spatial variation.

Salient regions in pedestrian images provide valuable information in identification. However, if they are small in size, saliency information is often overwhelmed by other features when computing similarities of images. In this paper, *saliency* means regions with attributes that 1) make a person *distinctive* against potential distractors, and 2) are *reliable* in finding the same person across camera views. In many cases, humans can easily recognize matched pedestrian pairs because they have distinct features. For example, in Figure 1, person $p1$ takes a red bag, $p2$ dresses bright white skirt, $p3$ takes a blue bag, and $p4$ carries a red folder in arm. These features are discriminative in distinguishing one person from others. Intuitively, if a body part is salient in one camera view, it usually remains salient in another camera view. Therefore, saliency also has view invariance.

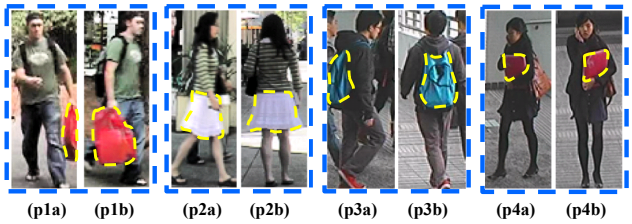


Fig. 1. Salient region could be a body part or a carrying accessory. Some salient regions of pedestrians are highlighted with yellow dashed boundaries.

Salient regions are not limited to body parts (such as clothes and trousers), but also include accessories (such as baggage, folders and umbrellas as shown in Figure 1), which are often considered as outliers and removed in existing approaches. Our computation of saliency is based on the comparison with images from a large scale reference dataset rather than a small group of persons. Therefore, it is quite stable in most circumstances.

We observe that images of the same person captured from different camera views have some invariance property in vertical direction on their spatial distributions of

• *R. Zhao, W. Ouyang, and X. Wang are with the Department of Electronic Engineering, the Chinese University of Hong Kong, Hong Kong. W. Ouyang is the corresponding author.*
E-mail: {rzhao, wlouyang, xgwang}@ee.cuhk.edu.hk

This work is supported by the General Research Fund sponsored by the Research Grants Council of Hong Kong (Project Nos. CUHK14206114, CUHK14205615, CUHK417011, CUHK419412, CUHK 14203015, and CUHK14207814) and National Natural Science Foundation of China(NSFCNO.61371192).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2544310, IEEE Transactions on Pattern Analysis and Machine Intelligence

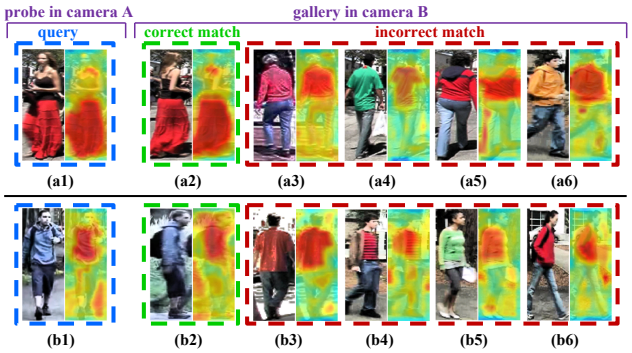


Fig. 2. Illustration of saliency matching with examples. Saliency map of each pedestrian image is shown. **Best viewed in color**.

saliency, like pair $(a1, a2)$ in Figure 2. Since the person in image $(a1)$ shows saliency in her dress while others $(a3)$-$(a6)$ are salient in blouses, they can be well distinguished simply from the spatial distributions of saliency. Therefore, not only the visual features from salient regions are discriminative, the spatial distributions of person saliency also provide useful information in person re-identification. Such information can be encoded into patch matching. If two patches from two images of the same person are matched, they are expected to have similar saliency values; otherwise such matching brings penalty on saliency matching. In the second row in Figure 2, the query image $(b1)$ shows a similar saliency distribution as those of gallery images. In this case, visual similarity needs to be considered. This motivates us to relate saliency matching penalty to the visual similarity of two matched patches.

2 Our Approach

Although saliency plays an important role in person re-identification, it has not been well explored in literature. In this paper, a novel framework of person saliency learning and matching is proposed for person re-identification. Our major contributions can be summarized from the following aspects.

We propose a way of estimating what is salient to humans. It is estimated from the number of trials that a human subject recognizes a query image from a candidate pool only based on a local region. It shows that most pedestrian images can be matched by humans from local salient regions without looking at whole images. The saliency estimated from user study is compared with the result of our saliency computation model. Compared with general image saliency detection methods [8], [16], our proposed saliency computation has much stronger correlation with human perception in person re-identification.

A computation model is proposed to estimate the probabilistic saliency map. Different from general image saliency detection methods, it is specially designed for person re-identification, and has the following properties. 1) It is robust to changes of viewpoints, poses and articulation. 2) Distinct patches are considered as salient only when they are matched and distinct in both camera views. 3) person saliency itself is a useful descriptor for pedestrian matching. For example, a person only with salient upper body and a person only with salient lower body must be different identities.

We formulate person re-identification as a saliency matching problem. Dense correspondences between patches are established by patch matching based on visual similarity, and matching patches with inconsistent saliency brings cost. Images of the same person are recognized by minimizing the saliency matching cost, which depends on both locations and visual similarity of matched patches.

Saliency matching and patch matching are tightly integrated into a unified structural RankSVM framework. Structural RankSVM has good training efficiency given a large number of rank constraints in person re-identification. Our approach transforms the original high-dimensional visual feature space to a 80 times lower dimensional saliency feature space to further improve training efficiency and also avoid overfitting.

3 Related Works

Existing works on person re-identification mainly focus on two aspects: 1) *features and representations*, and 2) *distance metric*. A review can be found in [17].

3.1 Features and Representations

A lot of research efforts [5], [12], [13], [15], [45]–[47], [59], [62], [65], [66], [68], [70] have been devoted to exploiting discriminative features in person re-identification. Wang *et al.* [59] proposed shape and appearance context to model the spatial distributions of appearance relative to body parts in order to extract discriminative features robust to misalignment. Farenzena *et al.* [15] proposed the Symmetry-Driven Accumulation of Local Features (SDALF) by exploiting the symmetry property in pedestrian images to handle view variation. Bak *et al.* [5], Xu *et al.* [62] and Cheng *et al.* [12], [13] applied human part models and pictorial structures to cope with pose variations by establishing the spatial correspondence. Wei *et al.* [60] proposed a cascade ranking model to utilize human gait information. Ma *et al.* [45]–[47] developed the BiCov descriptor based on the Gabor filters and the covariance descriptor to handle illumination change and background variation. Zheng *et al.* [68], [70] used the contextual visual cues from surrounding people to enrich human signatures. Information on salient regions exploited in our work can be integrated with many of these feature designs by putting more weights on features from salient regions.

Features vary in their usefulness in person matching, and some works have been done on feature selection and importance learning. Gray *et al.* [19] used AdaBoost to select features. Schwartz [54] assigned weights to features with Partial Least Squares (PLS). Liu *et al.* [41] developed an unsupervised approach to learn bottom-up feature importance, and adaptively weight features. Instead of globally weighting features across all the pedestrian

images, our approach adaptively weights features based on individual person pairs to be matched, since different persons have different salient regions.

Visual features suffer from a range of variations across camera views. Feature transforms are learned to improve the invariance to cross-view transforms. Prosser *et al.* [52] learned the Cumulative Brightness Transfer Function to handle color transforms. Avraham *et al.* [3], [4] learned both implicit and explicit transforms of visual features. Martinel *et al.* [48] modeled the feature transforms by classifying feasible and infeasible warp functions. Li *et al.* [35] proposed a cross-view projective dictionary learning approach to learn view-invariant features. Rather than learning feature transforms for specific camera view settings, our approach flexibly handles the cross-view variations by performing a constrained patch matching technique, which can be generalize to any disjoint camera-view transition. Recently, a similar work [72] employed patch matching to handle proposed partial person re-identification problem.

Some works explored higher level features [30]–[32], [43], [56], [67] to assist person re-identification. Vaquero *et al.* [56] first introduced mid-level facial attributes in human recognition. Layne *et al.* [30], [31] proposed 15 human attributes for person re-identification. Song *et al.* [43] used human attributes to match persons with Bayesian decision. Shi *et al.* [55] learned semantic representation by transferring attribute information from fashion photography datasets. Li *et al.* [37] and Ahmed *et al.* [2] designed deep convolutional neural networks to learn deep features. Saliency distribution can also be considered as one kind of high-level features.

3.2 Rank and Metric Learning

Given a query image, an image of the same person is expected to have a high rank on the candidate list after matching. Prosser *et al.* [53] formulated person re-identification as a ranking problem, and learned global feature weights with RankSVM. Wu *et al.* [61] introduced rank-loss optimization to improve accuracy in re-identification. Loy *et al.* [44] exploited unlabeled gallery data to propagate labels to query instances with a manifold ranking model. Liu *et al.* [42] presented a man-in-loop method to allow users to quickly refine ranking result. In this paper, we employ structural RankSVM [26], which considers ranking difference.

Many research works [14], [21], [22], [28], [36], [38], [41], [50], [51], [69] focused on optimizing distance metrics for matching persons. Zheng *et al.* [69] learned the metric by maximizing the likelihood of true matches to have a smaller distance than that of a wrongly matched pair. Dikmen *et al.* [14] proposed to learn a Mahalanobis distance that is optimal for k-nearest neighbor classification by using a maximum margin formulation. Mignon and Jurie [50] learned a joint projection for dimension reduction, satisfying distance constraints added by image pairs. Li *et al.* [38] proposed to learn a decision function for matching, which jointly models a distance metric and a locally adaptive thresholding rule. Pedagadi *et al.* [51] employed Local Fisher Discriminant Analysis to learn a distance metric. Zheng *et al.* [71] proposed a transfered local relative distance comparison model to mine and transfer information from the open-world non-target pedestrian images. Liao *et al.* [39] learned a discriminative subspace and a distance metric by cross-view quadratic discriminant analysis. The above learned metrics are based on subtraction of misaligned feature vectors, which causes significant information loss and errors. Our approach handles feature misalignment through patch matching.

3.3 Person saliency vs. General Image saliency

General image saliency has been well studied [16], [23], [24], [27], [34], [64]. In the context of person re-identification, person saliency is different from general image saliency in the way of drawing visual attention. With the aim to improve the performance of re-identification, person saliency is considered as visual patterns that distinguish a person from others, while general saliency draws visual attention within a single image to capture salient foreground objects from background.

4 Method Overview

The diagram of the proposed saliency learning and matching framework is shown in Figure 3. Section 5 conducts a user study to estimate person saliency based on human perception in the person re-identification task. We investigate the discriminative power of different body regions in identifying a target person from a gallery set. The saliency of each local region of a query image is quantitatively estimated by measuring the averaged number of trials that human labelers find the target person only based on that region of the query image. An illustration is shown in Figure 3 (a). The red and green bounding boxes indicate incorrect and correct targets chosen by the labeler from the gallery. The red skirt has higher saliency and causes fewer failure trials compared with the arm. Our result shows that subjects can recognize a query person only based on a small salient part without looking at the whole image. Salient regions vary on different persons.

An unsupervised approach for saliency learning is proposed in Section 6 and illustrated in Figure 3 (b). With constrained patch matching, each patch finds its matched neighbors from a reference set of training images. K-Nearest Neighbor and One-Class SVM models are employed to learn a saliency measure suitable for person re-identification. Our experimental results show both qualitative and quantitative evaluation of the correlation between the learned saliency and human perception. With obtained person saliency, matching image pairs can be performed in unsupervised and supervised ways as described in Section 6. For the unsupervised manner, saliency is used to weight patch matching similarity and penalize inconsistence of saliency distribution across camera views, as shown by the blue lines in Figure 3 (c). For the supervised manner, person matching is formulated as a saliency matching problem, which considers four types of saliency matching cases, as shown

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2544310, IEEE Transactions on Pattern Analysis and Machine Intelligence

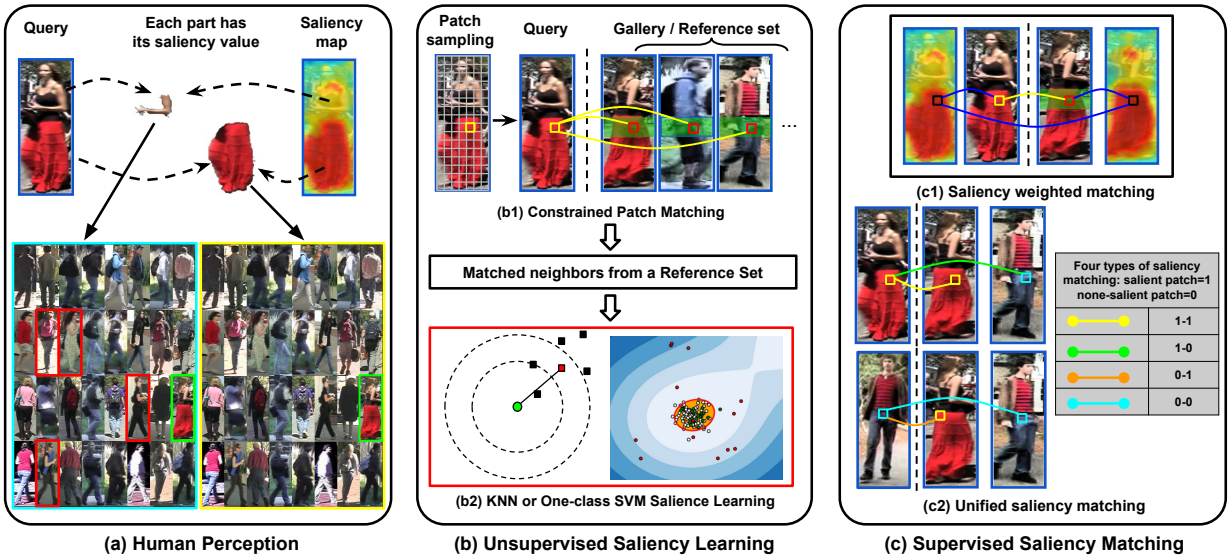


Fig. 3. Diagram of our novel framework of person saliency learning and matching for person re-identification.

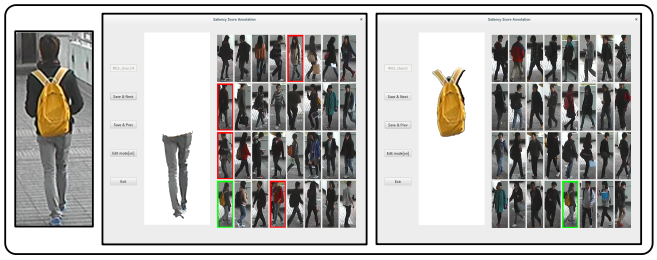


Fig. 4. Interface of user study to obtain person saliency.

in the table in Figure 3 (c). The matching cost is a linear function of patch matching similarities, which is learned with Structural RankSVM. The learned saliency matching function is used to measure similarities between images.

5 Saliency from Human Perception

We define person saliency in the context of person re-identification and estimate it by user study.

Given an image, we apply superpixel segmentation [1], and then manually merge superpixels that are coherent in appearance. Superpixels with different semantic meanings are not merged. For example, hair and jacket may share similar appearance, but they are treated as two parts. Only foreground superpixels are considered. Note that applying superpixel segmentation and manual merging are only for user study. Later in our proposed saliency learning approach, the saliency region is automatically estimated.

A segmented body part is randomly selected and presented to a labeler. Labelers are asked to perform part-based re-identification task. Each part is shown multiple times to different labelers. The user study results are combined into a saliency value. In Figure 4, a body part from a query image is revealed (on the left) at its original spatial location in the image while other parts are masked, and a list of 32 images randomly sampled from the gallery set are also shown (on the right) to the labeler. The true target (observed in a different camera view from the query image) is among the sampled images, but the order is randomly shuffled. In each trial, the labeler is asked to select the most likely image from the list based on visual perception. The labeler is allowed to select for multiple times until the correct match is found. In Figure 4, the red bounding boxes indicate wrong selection and the green one indicates the correct match found in the end. A part is considered as salient if labelers try fewer times to found the target.

Denote the i-th revealed part by p_i. Then the saliency value of the revealed part is estimated as

$$\textbf{\textit{score}}(p_i) = \exp(-\frac{m_{p_i}^2}{\sigma_{avg}^2}) \exp(-\frac{s_{p_i}^2}{\sigma_{std}^2}). \quad (1)$$

m_{p_i} and s_{p_i} are the average and standard deviation of number of trials over all the labelers. σ_{avg} and σ_{std} are bandwidth parameters. For the first term, smaller mean number of trials indicate the revealed body part is helpful to find the target person, and should have a higher saliency score. Larger mean number of trials lead to lower saliency score. In the second term, a large standard deviation of number of trials means the revealed body part cannot consistently help users to identify target person. The larger the standard deviation, smaller the second term is. Thus, they are combined.

The user study is conducted on 524 body parts of 100 images from camera view A of the VIPeR dataset [18]. Some examples of the saliency maps obtained by user study are shown in Figure 5. In order to investigate whether salient regions exist in pedestrian images, Figure 6 (left) shows the histogram on the numbers of trials used to find the targets only based on the most salient parts on query images. It shows that more than half of the pedestrians can be recognized, if the labelers only observe the most salient part of a query image. As comparison, Figure 6 (right) plots the histogram on the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2544310, IEEE Transactions on Pattern Analysis and Machine Intelligence

Fig. 5. Examples of saliency obtained from user study. Each body part obtains a saliency value. Saliency map is overlaid on the gray-level image. The original color image is on the left.

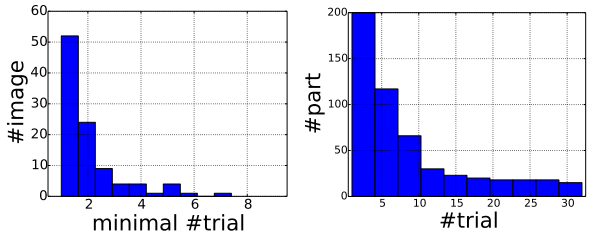


Fig. 6. Statistics on saliency user study. Left: Histogram on the numbers of trials used to find the targets only based on the most salient parts on query images. Right: Histogram on the numbers of trials for all the parts.

numbers of trials for all the parts. It shows that most other body parts are not salient enough. The correlation between the user study saliency and that obtained with the proposed computation model will be validated in experiments in Section 8.

6 Person Saliency Learning

Groundtruth person saliency costs large amount of human labors, and it usually becomes different in changed camera settings. Thus, we propose to automatically learn person saliency in an unsupervised manner. Dense correspondence between images is first built with patch matching, and two alternative approaches (K-nearest neighbor and One-Class SVM) are proposed to estimate person saliency without using identity labels or user study saliency.

6.1 Feature Extraction

Each image is densely divided into a grid $M \times N$ of overlapping local patches, and each patch is represented by a feature vector concatenating color histograms and SIFT features computed around its local region.

Dense Color Histogram. A color histogram in LAB color space is extracted from each patch. LAB color histograms are computed on multiple downsampled scales and L2 normalized.

Dense SIFT. To handle viewpoint and illumination change, SIFT descriptor is used as complementary to color histograms. We divide each patch into 4×4 cells, quantize the orientations of local gradients into 8 bins, and obtain a $4 \times 4 \times 8 = 128$ dimentional SIFT feature vector, which is also L2 normalized.

In our experiment, scales of pedestrian images range from 128×48 to 160×60. Patches of size 10×10 pixels are sampled on a dense grid with a step size 4. 32-bin color histograms are computed in each LAB channels, and in each channel, 3 levels of downsampling are used with scaling factors 0.5, 0.75 and 1. SIFT features are also extracted in 3 color channels and thus produces a 128×3 feature vector for each patch. In a summary, each patch is finally represented with a discriminative descriptor vector of length $32 \times 3 \times 3 + 128 \times 3 = 672$. We denote the combined Color-SIFT feature vector as $DenseFeats$.

6.2 Dense Correspondence

To deal with misalignment, we build dense correspondence between images by adjacency constrained search. $DenseFeats$ features of a pedestrian image is represented as $X^{A,u} = \{\mathbf{x}_{p_i}^{A,u} \mid p_i = 1 \ldots, MN\}$, where (A, u) denotes the u-th image in camera A, p_i denotes the position of the patch in this image, and $\mathbf{x}_{p_i}^{A,u}$ is the dense Color-SIFT feature vector of the patch. A natural baseline is to compute image similarity with concatenated patch features,

$$\boldsymbol{sim}_{DenseFeats}(X^{A,u}, X^{B,v}) = \sum_{p_i} s(\mathbf{x}_{p_i}^{A,u}, \mathbf{x}_{p_i}^{B,v}), \quad (2)$$

where

$$s(\mathbf{x}_{p_i}^{A,u}, \mathbf{x}_{p_i}^{B,v}) = exp(-\frac{d(\mathbf{x}_{p_i}^{A,u}, \mathbf{x}_{p_i}^{B,v})^2}{2\sigma^2}), \quad (3)$$

is the similarity between two patch features, $d(\cdot)$ is the Euclidean distance, and σ is a bandwidth parameter.

Adjacency Searching. $\boldsymbol{sim}_{DenseFeats}$ does not consider misalignment. We propose adjacency constrained searching to allow flexible matching among patches in image pairs. When the patches are matched with those from another image, patches in the same row have the same search set, denoted as

$$\mathcal{S}(\mathbf{x}_{p_i}^{A,u}, X^{B,v}) = \{\mathbf{x}_{\hat{p}_i}^{B,v} \mid I_y(\hat{p}_i) = I_y(p_i)\}, \quad (4)$$

where $I_y(p_i)$ indicate the row index of position p_i. $\mathcal{S}(\mathbf{x}_{m,n}^{A,u}, X^{B,v})$ restricts the search set in $X^{B,v}$ within the $I_y(p_i)$-th row. However, bounding boxes produced by a human detector are not always well aligned, and also uncontrolled human pose variations exist. We relax the horizontal constraint to have a larger search range:

$$\hat{\mathcal{S}}(\mathbf{x}_{p_i}^{A,u}, X^{B,v}) = \{\mathbf{x}_{\hat{p}_i}^{B,v} \mid I_y(\hat{p}_i) \in \mathcal{N}(I_y(p_i))\}, \quad (5)$$

where

$$\begin{aligned}\mathcal{N}(I_y(p_i)) = \{ &\max\big(0, I_y(p_i) - l\big), \ldots, \\ &I_y(p_i), \ldots, \min\big(I_y(p_i) + l, M\big)\}, \end{aligned} \quad (6)$$

and l defines the size of the relaxed adjacent vertical space. Less relaxed search space cannot well tolerate the spatial variation while more relaxed search space increases the chance of matching different body parts. $l = 2$ is chosen in our setting.

We perform the nearest neighbor search for each $\mathbf{x}_{p_i}^{A,u}$ in its search set $\hat{\mathcal{S}}(\mathbf{x}_{p_i}^{A,u}, X^{B,v})$ in $X^{B,v}$,

$$\begin{aligned} p_i' = \quad &\underset{\hat{p}_i}{\operatorname{argmin}}\, d(\mathbf{x}_{p_i}^{A,u}, \mathbf{x}_{\hat{p}_i}^{B,v}), \\ s.t.\ &\mathbf{x}_{\hat{p}_i}^{B,v} \in \hat{\mathcal{S}}(\mathbf{x}_{p_i}^{A,u}, X^{B,v}), \end{aligned} \quad (7)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2544310, IEEE Transactions on Pattern Analysis and Machine Intelligence

Thus, dense correspondences $P^{u,v} = \{(p_i, p_i')\}$ between patches in image $X^{A,u}$ and $X^{B,v}$ are obtained by the adjacency searching. We denote the method of only using patch matching without saliency information as $PatMatch$, and the image similarity is expressed as

$$\boldsymbol{sim}_{PatMatch}(X^{A,u}, X^{B,v}) = \sum_{(p_i,p_i')\in P^{u,v}} s(\mathbf{x}_{p_i}^{A,u}, \mathbf{x}_{p_i'}^{B,v}), \quad (8)$$

where $s(\mathbf{x}_{p_i}^{A,u}, \mathbf{x}_{p_i'}^{B,v})$ is defined in Eq. (3).

To estimate person saliency, we randomly sample N_r images from the training set as a reference set $\mathcal{R}$ without using the identity labels. For each patch $\mathbf{x}_{p_i}^{A,u}$, a nearest neighbor is found in every reference image, and these nearest neighbors are collected to build a reference patch set $X_{ref}(\mathbf{x}_{p_i}^{A,u})$ for each patch,

$$X_{ref}(\mathbf{x}_{p_i}^{A,u}) = \{\mathbf{x}_{p_i'}^{B,v} \mid X^{B,v} \in \mathcal{R},\ (p_i, p_i') \in P^{u,v}\}. \quad (9)$$

The reference set uses training images in different camera because the learned saliency serves for person re-identification, which is to match pedestrians across different camera views. Using reference images in different camera fits such cross-view setting. We use the reference patch set as opposed to all patches because saliency measures the ability of a patch to distinguish identities rather than different patches from the same person. A salient query patch could have many similar patches in one reference image if the corresponding salient region is large, and if all these similar patches are used in computing saliency in Eq. (10), then this salient query patch will have a low saliency score. So we constrain that one reference image can only contribute one patch.

6.3 Unsupervised saliency Learning

6.3.1 K-Nearest Neighbor (KNN) saliency

Byers *et al.* [9] found the KNN distances can be used for clutter removal. Since person saliency detection shares a similar goal as abnormality detection, which also measures how unusual a data sample is. KNN should also be viable in finding person saliency. By searching for the K-nearest neighbors of a test patch in the set of matched patches obtained with dense correspondence, KNN is adapted to the re-identification problem. The saliency score of the test patch is computed with the KNN distance.

The distance between $\mathbf{x}_{p_i}^{A,u}$ and its k-th nearest neighbor in $X_{ref}(\mathbf{x}_{p_i}^{A,u})$ is used as the saliency score:

$$\boldsymbol{score}_{knn}(\mathbf{x}_{p_i}^{A,u}) = d_k(X_{ref}(\mathbf{x}_{p_i}^{A,u})), \quad (10)$$

where d_k denotes the distance of the k-th nearest neighbor. Salient patches only find a limited number ($k = \alpha_k N_r$) of visually similar neighbors, and then $\boldsymbol{score}_{knn}(x_{p_i}^{A,p})$ is expected to be large. $0 < \alpha_k < 1$ is a proportion parameter reflecting our expectation on the statistical distribution of salient patches.

Choosing k. The goal of saliency detection for person re-identification is to identify parts with unique appearance. We set $\alpha_k = 0.5$ with an empirical assumption that a patch is considered to have unique appearance such that more than half of the people in the reference set do not share similar patches with it. N_r reference images are randomly sampled from training set in our experiments. Enlarging the reference dataset will not deteriorate saliency detection, because saliency is defined in the statistical sense. It is robust as long as the distribution of the reference dataset well reflects the test scenario.

6.3.2 One-Class SVM saliency

One-class SVM [20] has been widely used for outlier detection. The basic idea is to use a hypersphere to describe data in the feature space and put most of the data into the hypersphere. It is formulated as an objective function:

$$\min_{R\in\mathbb{R},\xi\in\mathbb{R}^l,c\in F} R^2 + \frac{1}{vl}\sum_i \xi_i, \quad (11)$$
$$s.t. \|\Phi(\mathbf{x}_i) - c\|^2 \le R^2 + \xi_i, \quad \forall i \in \{1, ...l\} : \xi_i \ge 0,$$

where $\Phi(\mathbf{x}_i)$ is the multi-dimensional feature vector of i-th training sample, l is the number of training samples, R and c are the radius and center of the hypersphere learned by One-Class SVM, and $v \in [0, 1]$ is a trade-off hyperparameter. The goal is to keep the hypersphere as small as possible and include most of the training data. It can be solved in a dual form by QP optimization [11]. The decision function is:

$$f(\mathbf{x}) = R^2 - \|\Phi(\mathbf{x}) - c\|^2, \quad (12)$$
$$\|\Phi(\mathbf{x}) - c\|^2 = k(\mathbf{x}, \mathbf{x}) - 2\sum_i \alpha_i k(\mathbf{x}_i, \mathbf{x}) + \sum_{i,j} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j),$$

where α_i and α_j are the parameters for each constraint in the dual problem. We use the radius basis function (RBF) $K(\mathbf{x}, \mathbf{y}) = \exp\{-\|\mathbf{x} - \mathbf{y}\|^2/2\sigma^2\}$ as kernel to deal with high-dimensional, non-linear, and multi-mode distributions. As shown in [11], the decision function $f(x)$ of kernel One-class SVM can well capture the density and modality of the feature distribution. Saliency score is defined in terms of kernel One-class SVM decision function:

$$\boldsymbol{score}_{ocsvm}(\mathbf{x}_{p_i}^{A,u}) = d(\mathbf{x}_{p_i}^{A,u}, \mathbf{x}^*), \quad (13)$$
$$\mathbf{x}^* = \operatorname*{argmax}_{\mathbf{x}\in X_{ref}(\mathbf{x}_{p_i}^{A,u})} f(\mathbf{x}),$$

where x^* is the patch with the highest density (we say density center). Then the ocsvm score is the distance between the current patch and the density center, and this is reasonable because it describe how far away from the majority (density center). Our experiments show very similar results in person re-identification with both saliency detection methods. $\boldsymbol{score}_{ocsvm}$ performs slightly better than $\boldsymbol{score}_{knn}$ in some circumstances. The probability of $\mathbf{x}_{m,n}^{A,u}$ being a salient patch is

$$p(l_{p_i}^{A,u} = 1 \mid \mathbf{x}_{p_i}^{A,u}) = 1 - \exp(-\boldsymbol{score}_{opt}(\mathbf{x}_{p_i}^{A,u})^2/\sigma_0^2), \quad (14)$$

where $opt \in \{knn,\ ocsvm\}$. The person saliency learning is summarized in Algorithm 1.

7 SALIENCY MATCHING

One of our main contributions is to match human images based on their saliency probability maps. It is based on our observation that people in different camera views show consistency in saliency probability maps, as shown

Algorithm 1 Person saliency learning.

Input: image $X^{A,u}$ and a reference image set $\mathcal{R} = \{X^{B,v},\ v = 1, \ldots, N_r\}$

Output: saliency probability map $p(l_{p_i}^{A,u} = 1 \mid \mathbf{x}_{p_i}^{A,u})$

1: **for** each patch $\mathbf{x}_{p_i}^{A,u} \in X$ **do**
2: compute $X_{ref}(\mathbf{x}_{p_i}^{A,u})$ with Eq. (9)
3: compute $\boldsymbol{score}_{opt}(\mathbf{x}_{p_i}^{A,u}),\ opt \in \{knn, ocsvm\}$ with Eq. (10) or Eq. (13)
4: compute $p(l_{p_i}^{A,u} = 1 \mid \mathbf{x}_{p_i}^{A,u})$ with Eq. (14)
5: **end for**

in Figure 2. Since matching is applied to arbitrary image pairs, we omit the image index in notation for concise clarity, *i.e.* change $X^{A,u}$ to X^A, $X^{B,v}$ to X^B, $\mathbf{x}_{m,n}^{A,u}$ to $\mathbf{x}_{p_i}^A$ and $\mathbf{x}_{i,j}^{B,v}$ to $\mathbf{x}_{p_i'}^B$. p_i is the patch index in image X^A and p_i' is the corresponding matched patch index in image X^B produced by dense correspondence. We denote the dense correspondence between X^A and X^B as $P = \{(p_i, p_i')\}_{i=1,\ldots,MN}$.

7.1 Saliency Weighted Matching

A saliency weighted matching scheme is designed to incorporate saliency information. We denote this method as saliency guided dense correspondence (SDC), as illustrated in Figure 3(c1), and the similarity between two images is computed as

$$\boldsymbol{sim}_{SDC_{opt}} = \sum_{(p_i,p_i')\in P} \frac{\boldsymbol{score}_{opt}(\mathbf{x}_{p_i}^A) \cdot s(\mathbf{x}_{p_i}^A, \mathbf{x}_{p_i'}^B) \cdot \boldsymbol{score}_{opt}(\mathbf{x}_{p_i'}^B)}{\alpha_{sdc} + |\boldsymbol{score}_{opt}(\mathbf{x}_{p_i}^A) - \boldsymbol{score}_{opt}(\mathbf{x}_{p_i'}^B)|}, \tag{15}$$

where α_{sdc} is a parameter representing a base penalty. Intuitively, large saliency scores in both matched patches are expected to enhance the similarity score of matched patches. In another aspect, images of the same person would be more likely to have similar saliency distributions than those of different persons, so the difference in saliency score can be used as a penalty to the similarity score. We set $\alpha_{sdc} = 1$ in experiments. The matching weights are inversely proportional to $\alpha_{sdc} + |\boldsymbol{score}_{opt}(\mathbf{x}_{p_i}^A) - \boldsymbol{score}_{opt}(\mathbf{x}_{p_i'}^B)|$. If saliency scores of a pair of patches are not consistent, the matching weights will be low. The weights are manually designed without using identity information. In next section, we address how the weights of saliency matching can be learned in a supervised approach with identity labels.

7.2 Unified saliency Matching

Previous saliency scores have continuous values, which can be understood as the probability of a patch being salient or non-salient. From this point of view, we can regard patch saliency as binary hidden variables. To formulate the person re-identification as a saliency matching problem in a probabilistic way, we introduce hidden variables $L^A = \{l_{p_i}^A \mid l_{p_i}^A \in \{0,1\}\}_{p_i}, L^B = \{l_{p_i}^B \mid l_{p_i}^B \in \{0,1\}\}_{p_i}$ to consider four different saliency matching cases separately, *i.e.* salient/salient ($l_{p_i}^A = 1, l_{p_i'}^B = 1$), salient/non-salient ($l_{p_i}^A = 1, l_{p_i'}^B = 0$), non-salient/salient ($l_{p_i}^A = 0, l_{p_i'}^B = 1$), and non-sailient/non-salient ($l_{p_i}^A = 0, l_{p_i'}^B = 0$). L^A, L^B do not need to be inferred, and they are marginalized later in Eq. (18). The saliency matching score in Eq. (19) can be computed from continuous saliency probabilities, estimated in Algorithm 1.

If all the saliency labels are known, we can perform person matching by computing the saliency matching score, and each matching case should contribute to the matching score f_z differently,

$$\begin{aligned} f_z(X^A, X^B, L^A, L^B; P, Z) = \\ \sum_{(p_i,p_i')\in P} \Big\{ & z_{p_i,1} l_{p_i}^A l_{p_i'}^B + z_{p_i,2} l_{p_i}^A (1 - l_{p_i'}^B) \\ & + z_{p_i,3} (1 - l_{p_i}^A) l_{p_i'}^B + z_{p_i,4} (1 - l_{p_i}^A)(1 - l_{p_i'}^B) \Big\}, \end{aligned} \tag{16}$$

where $Z = \{z_{p_i,k}\}_{i=1,\ldots,MN,\ k=1,2,3,4}$ are the matching scores for four different saliency matching results. For example, if a salient patch is matched with a non-salient patch, its contribution could be negative. $z_{p_i,k}$ is not a constant for all the patch pairs. Instead, it is modeled as a linear function of visual similarity of patch pairs. 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 patches $\mathbf{x}_{p_i}^A$ and patch $\mathbf{x}_{p_i'}^B$. Instead of directly using the Euclidean distance $d(\mathbf{x}_{p_i}^A, \mathbf{x}_{p_i'}^B)$, we convert it to similarity as in Eq. (3) to reduce the side effect in summation of very large distances in incorrect matching, 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,

$$z_{p_i,k} = \alpha_{p_i,k} \cdot s(\mathbf{x}_{p_i}^A, \mathbf{x}_{p_i'}^B) + \beta_{p_i,k}. \tag{17}$$

Thus Eq. (16) considers both saliency matching and visual similarity. Note that $z_{p_i,k}$ are not parameters. $\alpha_{p_i,k}$ and $\beta_{p_i,k}$ are weighting parameters, which are independent on image pairs. Once learned, $\alpha_{p_i,k}$ and $\beta_{p_i,k}$ are used in testing for any pairs without re-learning.

Since the saliency labels $l_{p_i}^A$ and $l_{p_i'}^B$ in Eq. (16) are hidden variables, they can be marginalized by computing the expectation of the saliency matching score as

$$\begin{aligned} & f^*(X^A, X^B; P, Z) \\ & = \sum_{L^A, L^B} f_z(X^A, X^B, L^A, L^B; P, Z) p(L^A, L^B | X^A, X^B) \\ & = \sum_{(p_i,p_i')\in P} \sum_{k=1}^{4} \Big[\alpha_{p_i,k} \cdot s(\mathbf{x}_{p_i}^A, \mathbf{x}_{p_i'}^B) + \beta_{p_i,k} \Big] c_{p_i,k}(\mathbf{x}_{p_i}^A, \mathbf{x}_{p_i'}^B), \end{aligned} \tag{18}$$

where parameters $\alpha_{p_i,k}$ weight both visual similarities $s(\mathbf{x}_{p_i}^A, \mathbf{x}_{p_i'}^B)$ and saliency similarities $c_{p_i,k}(\mathbf{x}_{p_i}^A, \mathbf{x}_{p_i'}^B)$, parameters $\beta_{p_i,k}$ weight only saliency similarities. Besides visual similarity, saliency similarity itself is also useful in re-identification. For example, even if visual similarity is low (*i.e.* $s(\mathbf{x}_{p_i}^A, \mathbf{x}_{p_i'}^B) \approx 0$), but matched patches are salient. Then $\beta_{p_i,k} \cdot c_{p_i,k}(\mathbf{x}_{p_i}^A, \mathbf{x}_{p_i'}^B)$ is large and provides evidence of the same identitiy. That is why Figure 2 and the fifth paragraph of Section 1 show that even without considering visual similarity, the spatial distribution of saliency itself has some power on matching

identity. $c_{p_i,k}(\mathbf{x}^A_{p_i}, \mathbf{x}^B_{p'_i})$ depends on saliency probabilities $P(l^A_{p_i} = 1 \mid \mathbf{x}^A_{p_i})$ and $P(l^B_{p'_i} = 1 \mid \mathbf{x}^B_{p'_i})$ given in Eq. (14),

$$c_{p_i,k}(\mathbf{x}^A_{p_i}, \mathbf{x}^B_{p'_i}) = \begin{cases} p(l^A_{p_i} = 1 \mid \mathbf{x}^A_{p_i}) \cdot p(l^B_{p'_i} = 1 \mid \mathbf{x}^B_{p'_i}), & k = 1, \\ p(l^A_{p_i} = 1 \mid \mathbf{x}^A_{p_i}) \cdot p(l^B_{p'_i} = 0 \mid \mathbf{x}^B_{p'_i}), & k = 2, \\ p(l^A_{p_i} = 0 \mid \mathbf{x}^A_{p_i}) \cdot p(l^B_{p'_i} = 1 \mid \mathbf{x}^B_{p'_i}), & k = 3, \\ p(l^A_{p_i} = 0 \mid \mathbf{x}^A_{p_i}) \cdot p(l^B_{p'_i} = 0 \mid \mathbf{x}^B_{p'_i}), & k = 4. \end{cases} \tag{19}$$

To better formulate this learning problem, we extract out all the weighting parameters in Eq. (18) as $\mathbf{w}$, and have

$$\begin{aligned} f^*(X^A, X^B; P, Z) &= \mathbf{w}^{\mathrm{T}} \Phi(X^A, X^B; P) \\ &= \sum_{(p_i, p'_i) \in P} \mathbf{w}^{\mathrm{T}}_{p_i} \phi(\mathbf{x}^A_{p_i}, \mathbf{x}^B_{p'_i}), \end{aligned} \tag{20}$$

where

$$\begin{aligned} \Phi(X^A, X^B; P) &= [\phi(\mathbf{x}^A_{p_1}, \mathbf{x}^B_{p'_1})^{\mathrm{T}}, \ldots, \phi(\mathbf{x}^A_{p_{MN}}, \mathbf{x}^B_{p'_{MN}})^{\mathrm{T}}]^{\mathrm{T}}, \\ \mathbf{w} &= [\mathbf{w}_{p_1}, \ldots, \mathbf{w}_{p_{MN}}]^{\mathrm{T}}, \\ \mathbf{w}_{p_i} &= [\{\alpha_{p_i,k}\}_{k=1,2,3,4}, \ \{\beta_{p_i,k}\}_{k=1,2,3,4}]. \end{aligned} \tag{21}$$

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

$$\phi(\mathbf{x}^A_{p_i}, \mathbf{x}^B_{p'_i}) = \begin{bmatrix} s(\mathbf{x}^A_{p_i}, \mathbf{x}^B_{p'_i}) \cdot p(l^A_{p_i} = 1 \mid \mathbf{x}^A_{p_i}) \cdot p(l^B_{p'_i} = 1 \mid \mathbf{x}^B_{p'_i}) \\ s(\mathbf{x}^A_{p_i}, \mathbf{x}^B_{p'_i}) \cdot p(l^A_{p_i} = 1 \mid \mathbf{x}^A_{p_i}) \cdot p(l^B_{p'_i} = 0 \mid \mathbf{x}^B_{p'_i}) \\ s(\mathbf{x}^A_{p_i}, \mathbf{x}^B_{p'_i}) \cdot p(l^A_{p_i} = 0 \mid \mathbf{x}^A_{p_i}) \cdot p(l^B_{p'_i} = 1 \mid \mathbf{x}^B_{p'_i}) \\ s(\mathbf{x}^A_{p_i}, \mathbf{x}^B_{p'_i}) \cdot p(l^A_{p_i} = 0 \mid \mathbf{x}^A_{p_i}) \cdot p(l^B_{p'_i} = 0 \mid \mathbf{x}^B_{p'_i}) \\ p(l^A_{p_i} = 1 \mid \mathbf{x}^A_{p_i}) \cdot p(l^B_{p'_i} = 1 \mid \mathbf{x}^B_{p'_i}) \\ p(l^A_{p_i} = 1 \mid \mathbf{x}^A_{p_i}) \cdot p(l^B_{p'_i} = 0 \mid \mathbf{x}^B_{p'_i}) \\ p(l^A_{p_i} = 0 \mid \mathbf{x}^A_{p_i}) \cdot p(l^B_{p'_i} = 1 \mid \mathbf{x}^B_{p'_i}) \\ p(l^A_{p_i} = 0 \mid \mathbf{x}^A_{p_i}) \cdot p(l^B_{p'_i} = 0 \mid \mathbf{x}^B_{p'_i}) \end{bmatrix}. \tag{22}$$

As shown in Eq. (22), the pairwise feature map $\Phi(X^A, X^B; P)$ combines the saliency probability map with appearance matching similarities. There are three advantages of matching with person saliency : (1) the person saliency probability distribution is more invariant than other features in different camera views; (2) because the saliency probability map is built based on dense correspondence, 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 following sections by formulating the person re-identification problem with $\Phi(X^A, X^B; P)$ in the structural RankSVM framework.

7.3 Ranking by Partial Order

We cast person re-identification as a ranking problem for supervised training. The ranking problem will be solved by finding an optimal partial order, mathematically defined in Eq. (23)(24)(27). Given a dataset of pedestrian images, $\mathcal{D}^A = \{X^{A,u}, id^{A,u}\}_{u=1}^U$ from camera view A and $\mathcal{D}^B = \{X^{B,v}, id^{B,v}\}_{v=1}^V$ from camera view B, where $X^{A,u}$ is the u-th image, $id^{A,u}$ 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 $X^{A,u}$ 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 $X^{A,u}$, we rank the relevant images before irrelevant ones, but no information of the orders within relevant images or irrelevant ones is provided. The partial order $\mathbf{y}^{A,u}$ is denoted as,

$$\mathbf{y}^{A,u} = \{y^{A,u}_{v,v'}\}, \quad y^{A,u}_{v,v'} = \begin{cases} +1 & X^{B,v} \prec X^{B,v'} \\ -1 & X^{B,v} \succ X^{B,v'} \end{cases}, \tag{23}$$

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

The partial order feature [25], [49] is appropriate for our goal and can encode the difference between relevant pairs and irrelevant pairs with only partial orders. The partial order feature for image $X^{A,u}$ is formulated as,

$$\Psi_{po}(X^{A,u}, \mathbf{y}^{A,u}; \{X^{B,v}\}_{v=1}^V, \{P^{u,v}\}_{v=1}^V) = \sum_{X^{B,v} \in S^+_{X^{A,u}}} \sum_{X^{B,v'} \in S^-_{X^{A,u}}} y^{A,u}_{v,v'} \frac{\Phi(X^{A,u}, X^{B,v}; P^{u,v}) - \Phi(X^{A,u}, X^{B,v'}; P^{u,v'})}{|S^+_{X^{A,u}}| \cdot |S^-_{X^{A,u}}|}, \tag{24}$$

$$S^+_{X^{A,u}} = \{X^{B,v} \mid id^{B,v} = id^{A,u}\}, \tag{25}$$

$$S^-_{X^{A,u}} = \{X^{B,v} \mid id^{B,v} \neq id^{A,u}\}, \tag{26}$$

where $\{P^{u,v}\}_{v=1}^V$ are the dense correspondences between image $X^{A,u}$ and every gallery image $X^{B,v}$, $S^+_{X^{A,u}}$ is relevant image set of $X^{A,u}$, $S^-_{X^{A,u}}$ is irrelevant image set, $\Phi(X^{A,u}, X^{B,v}; P^{u,v})$ is the feature map defined in Eq. (21), and the difference vector of two feature maps $\Phi(X^{A,u}, X^{B,v}; P^{u,v}) - \Phi(X^{A,u}, X^{B,v'}; P^{u,v'})$ is added if $X^{B,v} \prec X^{B,v'}$ or 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 the following score function,

$$\mathbf{y}^{A,u}_* = \underset{\mathbf{y}^{A,u} \in \mathcal{Y}^{A,u}}{\operatorname{argmax}} \ \mathbf{w}^{\mathrm{T}} \Psi_{po}(X^{A,u}, \mathbf{y}^{A,u}; \{X^{B,v}\}_{v=1}^V, \{P^{u,v}\}_{v=1}^V), \tag{27}$$

where $\mathcal{Y}^{A,u}$ is the space consisting of all the possible partial orders. As discussed in [25], [63], good ranking can be obtained by sorting gallery images by $\{\mathbf{w}^{\mathrm{T}} \Phi(X^{A,u}, X^{B,v}; P^{u,v})\}_v$ in a descending order. The remaining problem is how to learn $\mathbf{w}$. With an optimized $\mathbf{w}_*$, we denote the unified saliency matching similarity as

$$\boldsymbol{sim}_{SalMatch_{opt}}(X^A, X^B) = \mathbf{w}^{\mathrm{T}}_* \boldsymbol{\Phi}(X^A, X^B; P), \tag{28}$$

where $opt \in \{knn, \ ocsvm\}$.

7.4 Structural RankSVM Training

We employ structural SVM to learn the weighting parameters $\mathbf{w}$. Different than many previous SVM-based approaches [10], [53] doing optimization over pairwise differences, structural SVM optimizes over ranking differences and can incorporate non-linear multivariate loss functions into global optimization in SVM training.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2544310, IEEE Transactions on Pattern Analysis and Machine Intelligence

Objective function. Our goal is to learn a linear model and the training is based on n-slack structural SVM [26]. The objective function is as follows,

$$\begin{aligned} \min_{\mathbf{w},\xi} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{u=1}^{U}\xi_u, \\ s.t. \quad & \mathbf{w}^{\mathrm{T}}\delta\Psi_{po}(X^{A,u},\mathbf{y}^{A,u},\hat{\mathbf{y}}^{A,u};\{X^{B,v}\}_{v=1}^{V},\{P^{u,v}\}_{v=1}^{V}) \\ & \geq \Delta(\mathbf{y}^{A,u},\hat{\mathbf{y}}^{A,u}) - \xi_u, \\ & \forall \hat{\mathbf{y}}^{A,u} \in \mathcal{Y}^{A,u} \backslash \mathbf{y}^{A,u},\ \xi_u \geq 0,\ for\ u = 1,\ldots,U, \end{aligned} \tag{29}$$

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

$$\begin{aligned} &\delta\Psi_{po}(X^{A,u},\mathbf{y}^{A,u},\hat{\mathbf{y}}^{A,u};\{X^{B,v}\}_{v=1}^{V},\{P^{u,v}\}_{v=1}^{V}) \\ &= \Psi_{po}(X^{A,u},\mathbf{y}^{A,u};\{X^{B,v}\}_{v=1}^{V},\{P^{u,v}\}_{v=1}^{V}) \\ &\quad -\Psi_{po}(X^{A,u},\hat{\mathbf{y}}^{A,u};\{X^{B,v}\}_{v=1}^{V},\{P^{u,v}\}_{v=1}^{V}), \end{aligned} \tag{30}$$

$\mathbf{w}$ is the weight vector, C is a parameter to balance between margin and 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. (29) 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 person re-identification, we choose the ROC Area loss, which is also known as Area Under Curve (AUC) loss. It is computed from the number of swapped pairs,

$$\begin{aligned} &N_{swap} = \{(v,v') : X^{B,v} \prec X^{B,v'}\ and \\ &\mathbf{w}^{\mathrm{T}}\Phi(X^{A,u},X^{B,v};P^{u,v}) < \mathbf{w}^{\mathrm{T}}\Phi(X^{A,u},X^{B,v'};P^{u,v'})\}, \end{aligned} \tag{31}$$

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

$$\begin{aligned} \Delta(\mathbf{y}^{A,u},\hat{\mathbf{y}}^{A,u}) &= |N_{swap}|/|S^{+}_{X^{A,u}}| \cdot |S^{-}_{X^{A,u}}|, \\ &= \sum_{v,v'}(1-\hat{y}^{A,u}_{v,v'})/(2\cdot|S^{+}_{X^{A,u}}|\cdot|S^{-}_{X^{A,u}}|), \end{aligned} \tag{32}$$

which is a non-linear, and multivariate function. We note that there are an exponential number of constraints in Eq. (29) due to the huge dimensionality of $\mathcal{Y}^{A,u}$. Joachims *et al.* [26] showed that the problem could be efficiently solved by a cutting plane algorithm. In our problem, the discriminative model is learned by the structural RankSVM algorithm, and the weight vector $\mathbf{w}$ in our model means how important it is for each term in Eq. (22). In Eq. (22), $\{\alpha_{p_i,k}\}_{k=1,2,3,4}$ correspond to the first four terms based on saliency matching with visual similarity, and $\{\beta_{p_i,k}\}_{k=1,2,3,4}$ correspond to the last four terms only depending on saliency matching.

We visualize the learning result of $\mathbf{w}$ in Figure 7, and find that the first four terms in Eq. (22) are heavily weighted in the central part of human body which implies the importance of saliency 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.

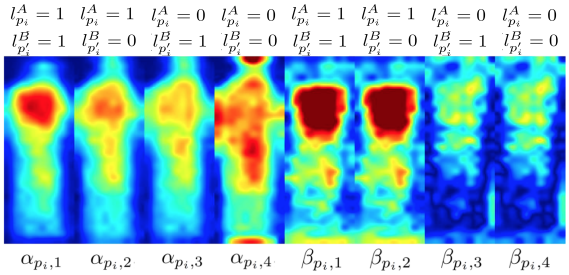


Fig. 7. We take the absolute value of the learned weight vector $\mathbf{w}$, and reshape it to a 2-dimensional importance map for different spatial locations. Eight importance maps correspond to $\{\alpha_{p_i,k}\}_{k=1,2,3,4}$ and $\{\beta_{p_i,k}\}_{k=1,2,3,4}$ in Eq. (18).

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.

7.5 Combination with existing approaches

Our approach is complementary to existing approaches. In order to combine existing approaches with the matching score in Eq. (20), the distance between two images can be computed as follows:

$$\begin{aligned} \boldsymbol{dist}_{eSalMatch_{opt}}(X^A,X^B) &= \sum_i \mu_i \cdot \boldsymbol{dist}_i(X^A,X^B) \\ &\quad -\mu_{Sal}\cdot \boldsymbol{sim}_{SalMatch_{opt}}(X^A,X^B) \end{aligned} \tag{33}$$

where $\mu_i(>0)$ is the weight for the ith similarity measure, $\mu_{Sal}(>0)$ the weight for unified saliency matching similarity. $\boldsymbol{dist}_i$ corresponds to the dissimilarity measures using wHSV and MSCR in [15] or LADF [38]. In the experiment, $\{\mu_i\}$ are chosen the same as in [15], [38]. μ_{Sal} is fixed as 1. The testing procedures are summarized in Algorithm 2.

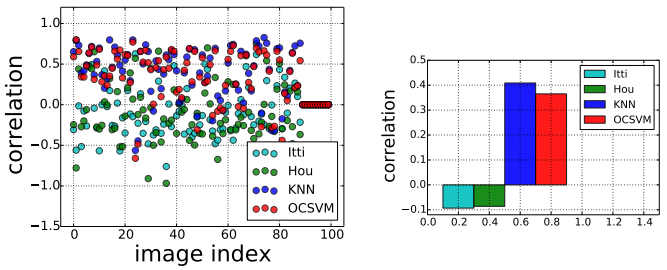
Algorithm 2 Testing procedures of our approach.

Input: probe images $\{X^{A,u}\}_u$, gallery images $\{X^{B,v}\}_v$, and learned structural SVM weights $\mathbf{w}_*$.

Output: matching similarities $\boldsymbol{sim}(X^{A,u},X^{B,v})$ or distances $\boldsymbol{dist}(X^{A,u},X^{B,v})$

1: extract feature for each local patch in an image, as described in Section 6.1.
2: build dense correspondences by adjacency search with Eq. (7).
3: compute saliency probability for each patch $p(l^{A,u}_{p_i}=1 \mid \mathbf{x}^{A,u}_{p_i})$ and $p(l^{B,v}_{p_i}=1 \mid \mathbf{x}^{B,v}_{p_i})$ following the Algorithm 1.
4: compute matching similarities / distances with one of components in our approach, including Eq. (2), Eq. (8), Eq. (15), Eq. (28), or Eq. (33).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2544310, IEEE Transactions on Pattern Analysis and Machine Intelligence



(a) (b)

Fig. 8. Correlation between automatically estimated saliency by different approaches (Itti [24], Hou [23], our KNN model and our One-Class SVM (OCSVM) model) and estimation from human perception. (a) Scatter plot of correlations over 100 images. (b) Average correlations.

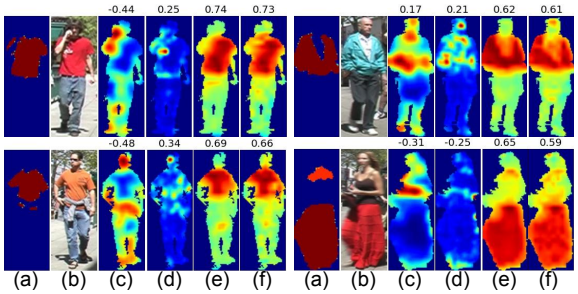
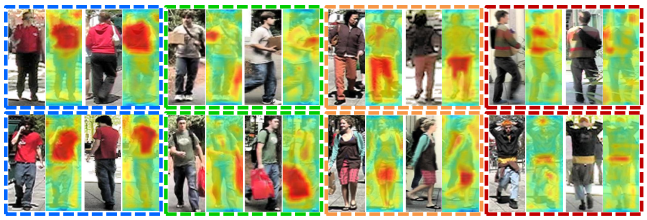
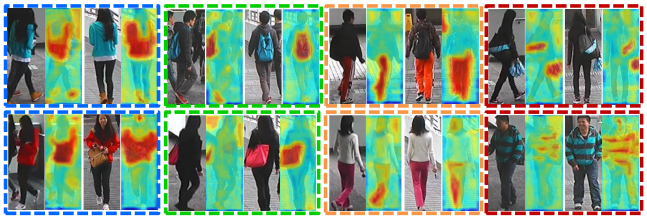


Fig. 9. Examples of estimated saliency map (only body parts are shown). (a) Groundtruth saliency for all parts are shown in column (a). Many parts have low saliency scores (in blue color), but a person may have multiple salient parts, *e.g.* in column (a) at bottom right. (b) Pedestrian images. (c) and (d) are general image saliency estimated by Itti [24] and Hou [23]. (e) and (f) are person saliency estimated by KNN and OCSVM. Number on top of each saliency map indicates the correlation with person saliency estimated from user study.



(a) VIPeR dataset



(b) CUHK01 dataset

Fig. 10. Examples of saliency matching in our experiments. It shows four types of saliency distributions: saliency in upper body (in blue dashed box), saliency of taking bags (in green dashed box), saliency of lower body (in orange dashed box), and saliency of stripes on human body (in red dashed box). **Best viewed in color**.

8 Experimental Results

We evaluated our approach on four public datasets, *i.e.* VIPeR [18], CUHK01 [36], i-LIDS [68], and 3DPeS [6]. All these public datasets are very challenging datasets for person re-identification because they contain significant variations on viewpoints, poses, and illuminations, and their images are with occlusions and background clutters. Qualitative results of saliency learning were shown, and quantitative results were reported in standard Cumulated Matching Characteristics (CMC) curves [59].

8.1 Evaluation Protocol

Our experiments followed the evaluation protocol in [19] for the VIPeR and CUHK01 datasets, and the protocol in [69] for the i-LIDS and 3DPeS datasets. We randomly partitioned the dataset into two even parts, 50% for training (donoted by $\mathcal{D}_{trn}$) and 50% for testing (denoted by $\mathcal{D}_{tst}$). Images from one view were used as probe and those from another view as gallery. Each probe image was matched with every image in gallery, and the rank of correct match was obtained. We computed the expectation of correct match at rank k as rank-k matching rate, and the cumulated values of matching rate at all ranks was recorded as one-trial CMC result. 10 trials of evaluation were conducted to achieve stable statistics, and the expectation was reported.

For training the structural SVM, all images with identity labels in $\mathcal{D}_{trn}$ were used. For person saliency learning, 100 images in both camera views were randomly sampled from $\mathcal{D}_{trn}$ as our reference set. In fact, there was overlap in image data for training the structural SVM and person saliency learning, because both tasks aimed to learn statistics of the testing camera view setting, and $\mathcal{D}_{trn}$ was a good training set to approximate the testing data. Only 100 images from $\mathcal{D}_{trn}$ were used in person saliency learning to reduce the computational cost in saliency estimation, and our experimental results showed the reference set is large enough to obtain good estimation.

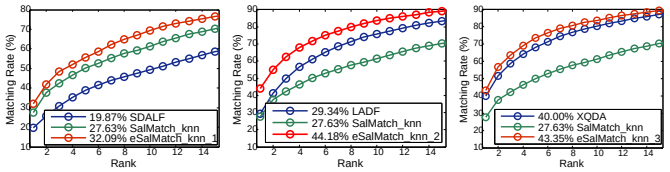
8.2 Evaluation on saliency Learning

We investigated the correlation between the person saliency estimated from human perception through user study and that automatically estimated by computation models. The computation models included those design for general image saliency (such as Itti [24] and Hou [23]) and our KNN and One-Class SVM (OCSVM) models specially desgined for person saliency. We computed the mean saliency score of each body part, and the Pearson correlation between the automatically estimated saliency and estimation from human perception. Results were shown , the scatter map in Figure 8(a) showed our learned saliency (KNN and OCSVM) had high positive correlations with human perception over the 100 images in user study, while general image saliency (Itti and Hou) exhibited slight negative correlations. Figure 8(b) showed averaged correlations. Some compared examples were shown in Figure 9. The approaches for general image saliency detection could separate body parts from background. However, the identified body parts might not be effective on recognizing identities.

More interesting results of saliency estimation were shown in Figure 10(a)(b) both on the VIPeR dataset and the CUHK01 dataset. Qualitative results showed our saliency learning approach could well approximate human perception and captured important salient regions on human body.

We also quantitatively compared the effectiveness of the saliency estimated from user study and our computation models in person re-identification. We regarded the 100 images (of 100 different persons) with saliency estimated from user study as the probe set for evaluation, and images of the corresponding identities in another camera view were included as the gallery set. Saliency weighted matching was adopted in testing competing saliency estimation methods, including general image saliency (Itti and Hou), our learned person saliency (SDC_knn and SDC_ocsvm), and saliency estimated

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2544310, IEEE Transactions on Pattern Analysis and Machine Intelligence



(a) +SDALF [15] (b) +LADF [38] (c) +XQDA [39]

Fig. 11. CMC curves of different ensemble approaches on the VIPeR dataset combining our approach $SalMatch\_knn$ with (a) SDALF, (b) LADF, and (c) XQDA.

Denotation	Description of component combination in test
$DenseFeats$	Matching with concatenated patch features
$PatMatch$	Use patch matching to handle misalignment
SDC_knn	Saliency weighted matching (KNN saliency)
$SalMatch\_fix$	Unified saliency matching (averaged silhouette as saliency)
$SalMatch\_knn$	Unified saliency matching (KNN saliency)
$eSalMatch\_knn\_1$	Combine SalMatch_knn with SDALF [15]
$eSalMatch\_knn\_2$	Combine SalMatch_knn with LADF [38]
$eSalMatch\_knn\_3$	Combine SalMatch_knn with XQDA [39]
SDC_ocsvm	Saliency weighted matching (OCSVM saliency)
$SalMatch\_ocsvm$	Unified saliency matching (OCSVM saliency)
$eSalMatch\_ocsvm\_1$	Combine SalMatch_ocsvm with SDALF [15]
$eSalMatch\_ocsvm\_2$	Combine SalMatch_ocsvm with LADF [38]

TABLE 1
Description of all the test settings in components evaluation. Refer to evaluation results in Figure 13(a) and Figure 14(a).

from user study (SDC_gt). CMCs were reported in Figure 12(a). Results showed that the our learned person saliency could well approximate the saliency estimated from user study in person re-identification, while general image saliency significantly degraded the re-identification performance.

8.3 Component-wise Evaluation

The effectiveness of different components in our framework was evaluated. Different settings of component combination were described in Table 1 and their results were shown in Figure 13(a) and Figure 14(a). DenseFeats in Eq. (2) performed the worst since it directly matched misaligned patches. PatMatch in Eq. (8) performed better by handling misalignment. SDC_knn (SDC_ocsvm) in Eq. (15) improved the performance by incorporating the estimated KNN (One-class SVM) saliency in patch matching. SalMatch_knn (SalMatch_ocsvm) in Eq. (28) formulated person re-identification as saliency matching, and learned matching weights in a supervised way. If we replaced the KNN saliency in SalMatch_knn by a fixed saliency map obtained by averaging all the pedestrian silhouettes, denoted by SalMatch_fix in Table 1 of revised version, we found it had worse matching rates than PatMatch, which did not utilize saliency information at all, *i.e.* 19.21% vs. 20.76% at rank 1, and 37.5% vs. 41.77% at rank 5. The reason could be that due to pose variation, some patches from human body might be outside of the averaged silhouettes template and their matching scores were weighted improperly. We combined our approach with other methods, including SDALF, LADF, and XQDA, and found it was complementary to each of these methods, as shown in Figure 11. eSalMatch_knn_1 (eSalMatch_ocsvm_1) in Eq. (33) ensembled SDALF feature matching scores in SalMatch_knn (SalMatch_ocsvm) matching scores, and eSalMatch_knn_2 (eSalMatch_ocsvm_2) ensembled LADF similarity measures. By combining with either of the two methods, the fusion methods outperformed each component, showing that our approach was complementary to other methods. One-class SVM saliency achieved slightly better than its counterpart settings using KNN saliency.

8.4 Comparison with the state-of-the-art

Figure 13(b) showed significant improvement of SDC (unsupervised) comparing with existing unsupervised methods, *i.e.* SDALF [15], CPS [13], eBiCov [45], eLDFV [46], and Comb [29] in the VIPeR dataset. For the CUHK01 dataset, we included the $DenseFeats$, SDALF, and Comb in comparison, as shown in Figure 14(b). In the evaluation of Comb, we used automatically extracted silhouettes. Specifically, we applied human pose estimator to find human skeleton, and used a Gaussian kernel to depict the silhouette. Among the methods in comparison, the methods denoted with "*method*-DF" were implemented with the source code provided by authors and using our features. Histogram equalization was applied to these methods. For other methods, their published results on public datasets were directly used for comparison.

Figure 13(c) compared our supervised saliency matching ($SalMatch$ and $eSalMatch$) with several alternative supervised methods, including seven benchmarking distance metric learning methods, *i.e.* PRDC [69], LMNN-R [14], KISSME [28], LADF [38], PCCA [50], WFS [48], XQDA [39], attribute-based PRDC (aPRDC) [41] and LF [51], a boosting approach (ELF) [19], an ensemble of RankSVM (PRSVM) [53], and a sparse ranking method (ISR) [40]. Also we compared with KISSME and LFDA using our $DenseFeats$ as baselines, which were denoted by KISSME-DF and LFDA-DF. Our approach outperformed all these methods. They ignored the domain knowledge on spatial variation caused by misalignment and poses as mentioned in Section 3. Although aPRDC shared a similar spirit as ours in finding unique and inherent appearance, it weighted different types global features instead of local patches. Its Rank-1 accuracy was only half of ours. ELF had a low performance since it selected features in the original feature space in which features of different classes were highly correlated. RankSVM was similar to our method in formulating person re-identification as ranking problem. Combined approach $eSalMatch$ was not evaluated in CUHK01 dataset because the weights μ_i in Eq. (33) were not carefully tuned for this dataset in SDALF method, and features of this dataset were not available in combining method LADF [38]. Compared with classical metric learning methods (CCA, LMNN, ITML, KISSME, and LFDA) based on our $DenseFeats$ features in CUHK01 dataset, our approach also had generally superior performance, as shown in Figure 14(c).

Also, as shown in Figure 12(b) our approach outperformed other approaches at rank-1 matching rate on the i-LIDS dataset, but did not obtain best performance after rank 5. This was mainly because that images in the i-LIDS dataset present frequent occlusion in lower body (people taking suitcases), and there was no module in our approach handling heavy occlusions. On the 3DPeS dataset, all images had very clean background, and the main problem was the lighting variations. Histogram equalization could mostly handle the main problem, and our approach outperformed other methods by a large margin on this dataset in Figure 12(c).

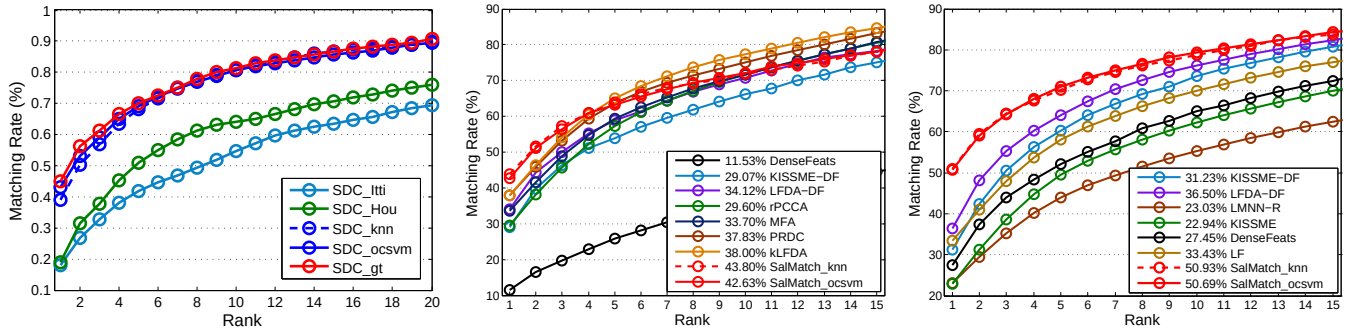
In general, our approach had much better performance because we adopted the discriminative saliency matching strategy for pairwise matching, and the structural RankSVM incorporated ranking loss in global optimization. This implied the importance of exploiting person saliency matching and its effectiveness in training structural RankSVM.

9 Discussion

When salient region does not exist in image. If pedestrian images have no salient regions (*e.g.* many pedestrians wear similar uniforms), our saliency matching approach degenerates to be patch matching in Eq. (8), which only depends on visual similarity. However, it will not hurt the performance. Our approach may also encounter difficulty when saliency regions are occluded by other pedestrians or self-occluded due to viewpoint change.

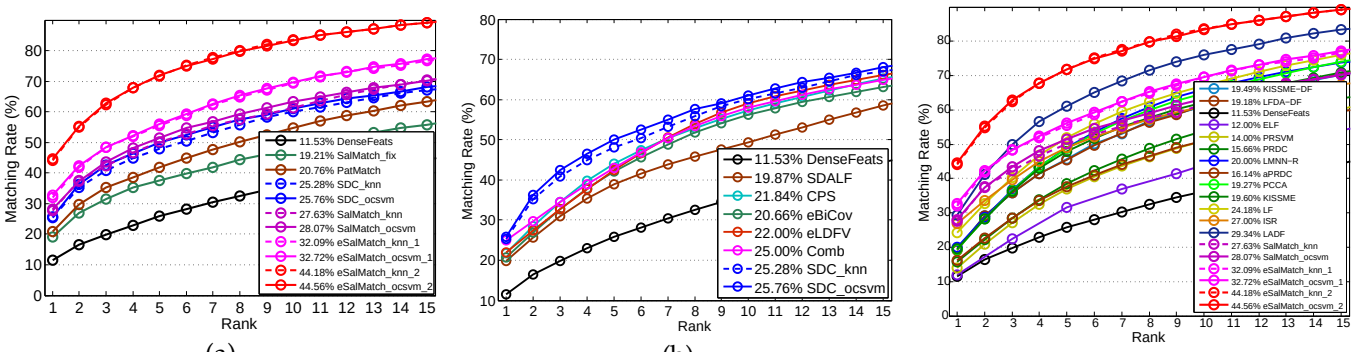
Salient / Non-salient Matching. The saliency label indicates whether a patch is salient or not. Although, saliency is expected to be invariant across camera views, such invariance is not absolute. A salient patch may become non-salient in the other camera view because of the change of lighting, viewpoint and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2544310, IEEE Transactions on Pattern Analysis and Machine Intelligence



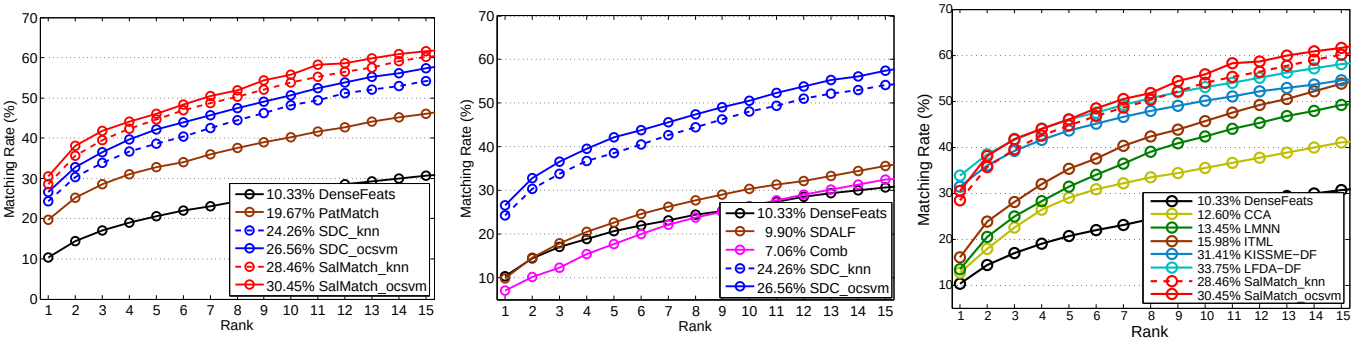
(a) (b) (c)

Fig. 12. (a) CMC curves of saliency weighted matching (denoted by SDC) using different saliency on the VIPeR dataset; (b) CMC curves of compared methods on the i-LIDS dataset; (c) CMC curves of compared methods on the 3DPeS dataset.



(a) (b) (c)

Fig. 13. CMC curves on the VIPeR dataset. (a) Component-wise evaluation; (b) Comparison of unsupervised approaches; (c) Comparison of supervised approaches.



(a) (b) (c)

Fig. 14. CMC curves on the CUHK01 dataset. (a) Component-wise evaluation; (b) Comparison of unsupervised approaches; (c) Comparison of supervised approaches.

pose change. There is no clear boundary between salient and non-salient patches. In our approach, patch matching always finds a nearest neighbor for each query patch based on visual similarity from another image, even if they have different saliency or these two images belong to different persons. Therefore, a salient/non-salient match could provide evidence that two images belonging to different identities. $z_{p_i,2}$ and $z_{p_i,3}$ could be negative values. Figure 7 visualizes the absolute values of $\alpha_{p_i,k}, \beta_{p_i,k}$. The Appendix shows that $\beta_{p_i,2} = -\beta_{p_i,1}$, and $\beta_{p_i,3} = -\beta_{p_i,4}$. So Eq. (16) allows a salient patch to be matched with a non-salient patch. Since the reliability of such match is lower, or it indicates different persons, $z_{p_i,2}$ and $z_{p_i,3}$ are expected to have lower values or even being negative.

Extension to Multi-shot Setting. In testing, our method can be applied to multi-shot setting, since the match score computed with our method can be easily applied to any multi-shot setting. Our training stage can also be naturally applied to multi-shot setting.

Evaluation on Auto-detected Pedestrian Images. In [37], when our approach is evaluated on the CUHK03 dataset with pedestrian images automatically detected by DPM, and the performance only drops 1.08% at rank-1 matching rate compared to the result evaluated on manually cropped pedestrian images.

10 Conclusion and Future Work

We propose a novel person saliency learning and matching framework for person re-identification. Adjacency constrained patch matching is applied to build dense correspondence between image pairs to handle misalignment caused by drastic viewpoint change and pose variations. Then K-Nearest Neighbor and One-class SVM approaches are proposed to estimate saliency score for each image patch without using identity labels. User study shows that the automatically estimated

person saliency has good correlation with human perception. It is more effective than general image saliency in person re-identification. The estimated saliency can be incorporated into patch matching in both the saliency weighted matching scheme and the unified saliency matching framework, and images of the same identity can be recognized by maximizing the saliency matching score. Learning the weights in unified saliency matching framework is formulated as solving a structural RankSVM problem. Experimental results valid the effectiveness of our approach and show superior performances on both the VIPeR and CUHK01 datasets.

The proposed framework can be extended by being integrated with other person re-identification approaches. For example, $DenseFeats$ used in this work can be replaced by other more advanced descriptors of characterizing local patches. Patch matching in our framework can be replaced by more sophisticated feature matching techniques [33]. Since saliency information is complementary to appearance, our saliency matching result can be combined with the matching results of existing approaches to boost their performance as shown in Section 7.5.

11 Appendix

The learned weights by Structural RankSVM. Here we explain why the learned weights by structural RankSVM are identical in $\beta_{p_i,1}$ and $\beta_{p_i,2}$, as shown in in Figure 7. The structural SVM learning is based on the partial order feature in Eq. (24), We find the numerator in Eq. (24) further depends on the subtraction between saliency matching features $\Delta\phi(\mathbf{x}_{p_i}^{A,u}, \mathbf{x}_{p_i'}^{B,v}, \mathbf{x}_{p_i''}^{B,v'}) = \phi(\mathbf{x}_{p_i}^{A,u}, \mathbf{x}_{p_i'}^{B,v}) - \phi(\mathbf{x}_{p_i}^{A,u}, \mathbf{x}_{p_i''}^{B,v'})$

$$\Phi(X^{A,u}, X^{B,v}; P^{u,v}) - \Phi(X^{A,u}, X^{B,v'}; P^{u,v'}) = \Big[\dots, \big[\phi(\mathbf{x}_{p_i}^{A,u}, \mathbf{x}_{p_i'}^{B,v}) - \phi(\mathbf{x}_{p_i}^{A,u}, \mathbf{x}_{p_i''}^{B,v'})\big]^{\mathrm{T}}, \dots\Big]^{\mathrm{T}},$$

where $P^{u,v} = \{(p_i, p_i')\}_{i=1,\dots,MN}$ are dense correspondence between patches in matching $X^{A,u}$ and $X^{B,v}$, $P^{u,v'} = \{(p_i, p_i'')\}_{i=1,\dots,MN}$ are dense correspondence between patches in matching $X^{A,u}$ and $X^{B,v'}$, and the saliency matching feature is defined in Eq. (22). Thus, we have the fifth term in subtraction $\Delta\phi(\mathbf{x}_{p_i}^{A,u}, \mathbf{x}_{p_i'}^{B,v}, \mathbf{x}_{p_i''}^{B,v'})$

$$\begin{aligned}\Delta\phi_5(\mathbf{x}_{p_i}^{A,u}, \mathbf{x}_{p_i'}^{B,v}, \mathbf{x}_{p_i''}^{B,v'}) =& p(l_{p_i}^{A,u} = 1|x_{p_i}^{A,u}) \cdot p(l_{p_i'}^{B,v} = 1|x_{p_i'}^{B,v}) \\ & - p(l_{p_i}^{A,u} = 1|x_{p_i}^{A,u}) \cdot p(l_{p_i''}^{B,v'} = 1|x_{p_i''}^{B,v'}),\end{aligned}$$

and the sixth term

$$\begin{aligned}\Delta\phi_6(\mathbf{x}_{p_i}^{A,u}, \mathbf{x}_{p_i'}^{B,v}, \mathbf{x}_{p_i''}^{B,v'}) &= p(l_{p_i}^{A,u} = 1|x_{p_i}^{A,u}) \cdot p(l_{p_i'}^{B,v} = 0|x_{p_i'}^{B,v}) \\ &\quad - p(l_{p_i}^{A,u} = 1|x_{p_i}^{A,u}) \cdot (l_{p_i''}^{B,v'} = 0|x_{p_i''}^{B,v'}) \\ &= p(l_{p_i}^{A,u} = 1|x_{p_i}^{A,u}) \cdot \Big[1 - p(l_{p_i'}^{B,v} = 1|x_{p_i'}^{B,v})\Big] \\ &\quad - p(l_{p_i}^{A,u} = 1|x_{p_i}^{A,u}) \cdot \Big[1 - p(l_{p_i''}^{B,v'} = 1|x_{p_i''}^{B,v'})\Big] \\ &= -p(l_{p_i}^{A,u} = 1|x_{p_i}^{A,u}) \cdot p(l_{p_i'}^{B,v} = 1|x_{p_i'}^{B,v}) \\ &\quad + p(l_{p_i}^{A,u} = 1|x_{p_i}^{A,u}) \cdot p(l_{p_i''}^{B,v'} = 1|x_{p_i''}^{B,v'}) \\ &= -\Delta\phi_5(\mathbf{x}_{p_i}^{A,u}, \mathbf{x}_{p_i'}^{B,v}, \mathbf{x}_{p_i''}^{B,v'}).\end{aligned}$$

We find the $\Delta\phi_5$ and $\Delta\phi_6$ are of the same value but opposite signs, and they are the actual features used in Structural SVM training. Weights $\beta_{p_i,1}$ and $\beta_{p_i,2}$ correspond to the $\Delta\phi_5$ and $\Delta\phi_6$ in Ψ_{po}. That is the reason that the normalized weights in Figure 11 are identical in the fifth and sixth terms, but please note that they are of opposite signs.

Implementation settings. We use the Matlab interface of SVM^{struct} [26], [57] to implement the structural RankSVM. All experiments are performed in Matlab 2012b on Windows x64 with 3.33 GHz Intel Xeon CPU, and 48 GB RAM. We show the value settings for all the parameters in our approach in Table 2. These parameters are chosen empirically. Most of them were chosen with reasonable values without being carefully tuned. For example, σ_0 and σ_{sdc} are set as 1. $\alpha_k = 0.5$ and k is decided by $\alpha_k * N_r$, where N_r is the size of the reference set in Eq. (9). So k is simply chosen as half of the reference set size. σ is set to the average of patch disances in the training set. In One-class SVM (Eq. (11)), c is automatically learned. These parameters are kept the same across datasets.

Param.	M	N	l	σ_{avg}	σ_{std}	σ	α_k	v	σ_0	α_{sdc}	k	N_r
Eq.	2	2	6	1	1	3	10	11	14	15	10	9
Val.	38	13	1	3	3	-	0.5	2	1	1	50	100

TABLE 2
Parameter settings. All parameters in our approach, associated equations, and values in experiment are listed.

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. on PAMI*, 34:2274–2282, 2012.

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

[3] T. Avraham, I. Gurvich, M. Lindenbaum, and S. Markovitch. Learning implicit transfer for person re-identification. In *Workshops ECCV*, pages 381–390. Springer, 2012.

[4] T. Avraham and M. Lindenbaum. Learning appearance transfer for person re-identification. In *Person Re-Identification*, pages 231–246. Springer, 2014.

[5] S. Bak, E. Corvee, F. Brémond, and M. Thonnat. Person re-identification using spatial covariance regions of human body parts. In *AVSS*, 2010.

[6] D. Baltieri, R. Vezzani, and R. Cucchiara. 3dpes: 3d people dataset for surveillance and forensics. In *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*, pages 59–64. ACM, 2011.

[7] A. Bedagkar-Gala and S. K. Shah. A survey of approaches and trends in person re-identification. *Image and Vision Computing*, 32(4):270–286, 2014.

[8] A. Borji and L. Itti. Exploiting local and global patch rarities for saliency detection. In *CVPR*, 2012.

[9] S. Byers and A. E. Raftery. Nearest-neighbor clutter removal for estimating features in spatial point processes. *Journal of the American Statistical Association*, 93:577–584, 1998.

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

[11] Y. Chen, X. Zhou, and T. Huang. One-class svm for learning in image retrieval. In *ICIP*, 2001.

[12] D. S. Cheng and M. Cristani. Person re-identification by articulated appearance matching. In *Person Re-Identification*, pages 139–160. Springer, 2014.

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

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

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

[16] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. *IEEE Trans. on PAMI*, 34:1915–1926, 2012.

[17] S. Gong, M. Cristani, S. Yan, and C. C. Loy. Person re-identification. Springer, 2013.

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

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

[20] K. Heller, K. Svore, A. Keromytis, and S. Stolfo. One class support vector machines for detecting anomalous windows registry accesses. In *Workshop on Data Mining for Computer Security*, 2003.

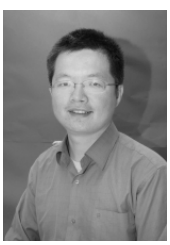
[21] M. Hirzer, C. Beleznai, M. Kostinger, P. M. Roth, and H. Bischof. Dense appearance modeling and efficient learning of camera transitions for person re-identification. In *ICIP*, 2012.

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

[23] X. Hou, J. Harel, and C. Koch. Image signature: Highlighting sparse salient regions. *IEEE Trans. on PAMI*, 34(1):194–201, 2012.
[24] L. Itti, C. Koch, E. Niebur, et al. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. on PAMI*, 20:1254–1259, 1998.
[25] T. Joachims. A support vector method for multivariate performance measures. In *ICML*, 2005.
[26] T. Joachims, T. Finley, and C.-N. J. Yu. Cutting-plane training of structural svms. *Machine Learning*, 77:27–59, 2009.
[27] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *ICCV*, 2009.
[28] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012.
[29] I. Kviatkovsky, A. Adam, and E. Rivlin. Color invariants for person reidentification. *IEEE Trans. on PAMI*, 35(7):1622–1634, 2013.
[30] R. Layne, T. M. Hospedales, and S. Gong. Towards person identification and re-identification with attributes. In *Workshops ECCV*, pages 402–412. Springer, 2012.
[31] R. Layne, T. M. Hospedales, S. Gong, et al. Person re-identification by attributes. In *BMVC*, volume 2, page 3, 2012.
[32] A. Li, L. Liu, and S. Yan. Person re-identification by attribute-assisted clothes appearance. In *Person Re-Identification*, pages 119–138. Springer, 2014.
[33] H. Li, X. Huang, J. Huang, and S. Zhang. Feature matching with affine-function transformation models. *IEEE Trans. on PAMI*, 36(12):2407–2422, Dec 2014.
[34] H. Li and K. N. Ngan. A co-saliency model of image pairs. *IEEE Trans. on Image Processing*, 20:3365–3375, 2011.
[35] S. Li, M. Shao, and Y. Fu. Cross-view projective dictionary learning for person re-identification. In *IJCAI*, 2015.
[36] W. Li, R. Zhao, and X. Wang. Human reidentification with transferred metric learning. In *ACCV*, 2012.
[37] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014.
[38] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith. Learning locally-adaptive decision functions for person verification. In *CVPR*, 2013.
[39] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015.
[40] G. Lisanti, I. Masi, A. Bagdanov, and A. Del Bimbo. Person re-identification by iterative re-weighted sparse ranking. *IEEE Trans. on PAMI*, 2014.
[41] C. Liu, S. Gong, C. C. Loy, and X. Lin. Person re-identification: what features are important? In *ECCV*, 2012.
[42] C. Liu, C. C. Loy, S. Gong, and G. Wang. Pop: Person re-identification post-rank optimisation. In *International Conference on Computer Vision*, 2013.
[43] X. Liu, M. Song, Q. Zhao, D. Tao, C. Chen, and J. Bu. Attribute-restricted latent topic model for person re-identification. *Pattern Recognition*, 45(12):4204–4213, 2012.
[44] C. C. Loy, C. Liu, and S. Gong. Person re-identification by manifold ranking. In *ICIP*, volume 20, 2013.
[45] B. Ma, Y. Su, and F. Jurie. Bicov: a novel image representation for person re-identification and face verification. In *BMVC*, 2012.
[46] B. Ma, Y. Su, and F. Jurie. Local descriptors encoded by fisher vectors for person re-identification. 2012.
[47] B. Ma, Y. Su, and F. Jurie. Discriminative image descriptors for person re-identification. In *Person Re-Identification*, pages 23–42. Springer, 2014.
[48] N. Martinel, A. Das, C. Micheloni, and A. K. Roy-Chowdhury. Re-identification in the function space of feature warps. *IEEE Trans. on PAMI*, 24(12):5645–5658, 2015.
[49] B. McFee and G. Lanckriet. Metric learning to rank. In *ICML*, 2010.
[50] A. Mignon and F. Jurie. Pcca: A new approach for distance learning from sparse pairwise constraints. In *CVPR*, 2012.
[51] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *CVPR*, 2013.
[52] B. Prosser, S. Gong, and T. Xiang. Multi-camera matching using bi-directional cumulative brightness transfer functions. In *BMVC*, volume 8, pages 164–1. Citeseer, 2008.
[53] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary. Person re-identification by support vector ranking. In *BMVC*, 2010.
[54] W. Schwartz and L. Davis. Learning discriminative appearance-based models using partial least squares. In *XXII Brazilian Symposium on Computer Graphics and Image Processing*, 2009.
[55] Z. Shi, T. M. Hospedales, and T. Xiang. Transferring a semantic representation for person re-identification and search. In *CVPR*, 2015.
[56] D. A. Vaquero, R. S. Feris, D. Tran, L. Brown, A. Hampapur, and M. Turk. Attribute-based people search in surveillance environments. In *WACV*, pages 1–8. IEEE, 2009.
[57] A. Vedaldi. A matlab wrapper of svmstruct. http://www.vlfeat.org/vedaldi/code/svm-struct-matlab.html, 2008.
[58] R. Vezzani, D. Baltieri, and R. Cucchiara. People reidentification in surveillance and forensics: A survey. *ACM Computing Surveys (CSUR)*, 46(2):29, 2013.
[59] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu. Shape and appearance context modeling. In *ICCV*, 2007.
[60] L. Wei, Y. Tian, Y. Wang, and T. Huang. Swiss-system based cascade ranking for gait-based person re-identification. In *AAAI*, 2015.
[61] Y. Wu, M. Mukunoki, T. Funatomi, M. Minoh, and S. Lao. Optimizing mean reciprocal rank for person re-identification. In *AVSS*, 2011.
[62] Y. Xu, L. Lin, W.-S. Zheng, and X. Liu. Human re-identification by matching compositional template with cluster sampling. In *ICCV*, 2013.
[63] Y. Yue, T. Finley, F. Radlinski, and T. Joachims. A support vector method for optimizing average precision. In *ACM SIGIR*, 2007.
[64] R. Zhao, W. Ouyang, H. Li, and X. Wang. Saliency detection by multi-context deep learning. In *CVPR*, 2015.
[65] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In *ICCV*, 2013.
[66] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013.
[67] R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identification. In *CVPR*, 2014.
[68] W.-S. Zheng, S. Gong, and T. Xiang. Associating groups of people. In *BMVC*, 2009.
[69] W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. In *CVPR*, 2011.
[70] W.-S. Zheng, S. Gong, and T. Xiang. Group association: Assisting re-identification by visual context. In *Person Re-Identification*, pages 183–201. Springer, 2014.
[71] W.-S. Zheng, S. Gong, and T. Xiang. Towards open-world person re-identification by one-shot group-based verification. *IEEE Trans. on PAMI*, 2015.
[72] W.-S. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, and S. Gong. Partial person re-identification. In *ICCV*, 2015.

Rui Zhao (S'12) received the B.Eng. degree in Electronic Engineering and Information Science from University of Science and Technology of China in 2010. He is currently a PhD student in the Department of Electronic Engineering at the Chinese University of Hong Kong. His research interests include computer vision, pattern recognition and machine learning.

Wanli Ouyang (S'08-M'11) received the B.S. degree in computer science from Xiangtan University, Hunan, China, in 2003. He received the M.S. degree in computer science from the College of Computer Science and Technology, Beijing University of Technology, Beijing, China. He received the PhD degree in the Department of Electronic Engineering, The Chinese University of Hong Kong, where he is now a Research Assistant Professor. His research interests include image processing, and computer vision.



Xiaogang Wang (S'03-M'10) received the B.S. degree from University of Science and Technology of China in 2001, the M.S. degree from Chinese University of Hong Kong in 2004, and the PhD degree in Computer Science from Massachusetts Institute of Technology. He is currently an assistant professor in the Department of Electronic Engineering at the Chinese University of Hong Kong. He received the Outstanding Young Researcher Award in Automatic Human Behaviour Analysis in 2011, and the Hong Kong Early Career Award in 2012. His research interests include computer vision and machine learning.