Unsupervised Salience Learning for Person Re-identification

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 eyes can recognize person identities based on some small salient regions. However, such valuable salient information is often hidden when computing similarities of images with existing approaches. Moreover, many existing approaches learn discriminative features and handle drastic viewpoint change in a supervised way and require labeling new training data for a different pair of camera views. In this paper, we propose a novel perspective for person re-identification based on unsupervised salience learning. Distinctive features are extracted without requiring identity labels in the training procedure. First, we apply adjacency constrained patch matching to build dense correspondence between image pairs, which shows effectiveness in handling misalignment caused by large viewpoint and pose variations. Second, we learn human salience in an unsupervised manner. To improve the performance of person re-identification, human salience is incorporated in patch matching to find reliable and discriminative matched patches. The effectiveness of our approach is validated on the widely used VIPeR dataset and ETHZ dataset.

1. Introduction

Person re-identification handles pedestrian matching and ranking across non-overlapping camera views. It has many important applications in video surveillance by saving a lot of human efforts on exhaustively searching for a person from large amounts of video sequences. However, this is also a very challenging task. A surveillance camera may observe hundreds of pedestrians in a public area within one day, and some of them have similar appearance. The same person observed in different camera views often undergoes significant variation in viewpoints, poses, camera settings, illumination, occlusions and background, which usually make intra-personal variations even larger than interpersonal variations as shown in Figure 1.

Our work is mainly motivated in three aspects. Most existing works [25, 15, 8, 29, 16, 24] handle the problem of cross-view variations and extract discriminative features by employing supervised models, which require training data with identity labels. Also, most of them require labeling new training data when camera settings change, since the cross-view transforms are different for different pairs of camera views. This is impractical in many applications especially for large-scale camera networks. In this paper, we propose a new approach of learning discriminative and reliable descriptions of pedestrians through unsupervised learning. Therefore, it has much better adaptability to generel camera view settings.

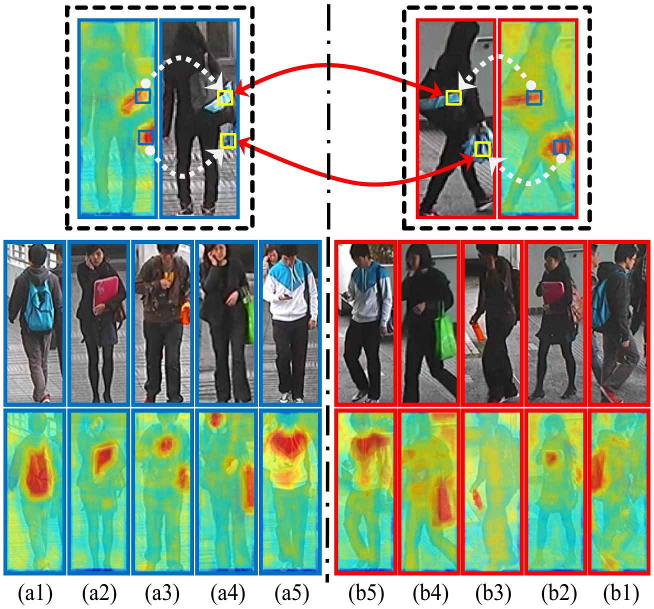


Figure 1. **Examples of human image matching and salience maps.** Images on the left of the vertical dashed black line are from camera view A and those on the right are from camera view B. Upper part of the figure shows an example of matching based on dense correspondence and weighting with salience values, and the lower part shows some pairs of images with their salience maps.

In person re-identification, viewpoint change and pose variation cause uncontrolled misalignment between images. For example in Figure 1, the central region of image $(a1)$ is a backpack in camera view A, while it becomes an arm

1063-6919/13 $26.00 © 2013 IEEE
DOI 10.1109/CVPR.2013.460

in image $(b1)$ in camera view B. Thus spatially misaligned feature vectors cannot be directly compared. In our method, patch matching is applied to tackle the misalignment problem. In addition, based on prior knowledge on pedestrian structures, some constraints are added in patch matching in order to enhance the matching accuracy. With patch matching, we are able to align the blue tilted stripe on the handbag of the lady in the dashed black boxes in Figure 1.

Salient regions in pedestrian images provide valuable information in identification. However, if they are small in size, salience information is often hidden when computing similarities of images. In this paper, *salience* means distinct features that 1) are *discriminative* in making a person standing out from their companions, and 2) are *reliable* in finding the same person across different views. For example, in Figure 1, if most persons in the dataset wear similar clothes and trousers, it is hard to identify them. However, human eyes are easy to identify the matching pairs because they have distinct features, *e.g.* person $(a1 - b1)$ has a backpack with tilted blue stripes, person $(a2 - b2)$ has a red folder under her arms, and person $(a3 - b3)$ has a red bottle in his hand. These distinct features are discriminative in distinguishing one from others and robust in matching themselves across different camera views. Intuitively, if a body part is salient in one camera view, it is usually also salient in another camera view. Moreover, our computation of salience 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. However, these distinct features may be considered by existing approaches as outliers to be removed, since some of they (such as baggages or folders) do not belong to body parts. Clothes and trousers are generally considered as the most important regions for person re-identification. Aided by patch matching, these discriminative and reliable features are employed in this paper for person re-identification.

The contributions of this paper can be summarized in three-folds. First, an unsupervised framework is proposed to extract distinctive features for person re-identification without requiring manually labeled person identities in the training procedure. Second, patch matching is utilized with adjacency constraint for handling the misalignment problem caused by viewpoint change, pose variation and articulation. We show that the constrained patch matching greatly improves person re-identification accuracy because of its flexibility in handling large viewpoint change. Third, human salience is learned in an unsupervised way. Different from general image salience detection methods [4], our salience is especially designed for human matching, and has the following properties. 1) It is robust to viewpoint change, pose variation and articulation. 2) Distinct patches are considered as salient only when they are matched and distinct in both camera views. 3) Human salience 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 have different identities.

2. Related Work

Discriminative models like SVM and boosting [25, 13, 15] are widely used for feature learning. Prosser *et al.* [25] formulated person re-identification as a ranking problem, and used ensembled RankSVMs to learn pairwise similarity. Gray *et al.* [13] combined spatial and color information in an ensmeble of local features by boosting. Schwartz *et al.* [26] extracted high-dimensional features including color, gradient, and texture, and then utilized the partial least square (PLS) for dimension reduction. Another direction is to learn task-specific distance functions with metric learning algorithms [29, 8, 24, 16]. Li and Wang [17] partitioned the image spaces of two camera views into different configurations and learned different metrics for different locally aligned common feature spaces. Li *et al.* [18] proposed a transferred metric learning framework for learning specific metric for individual query-candidate settings. In all these supervised methods, training samples with identity labels are required.

Some unsupervised methods have also been developed for person re-identification [10, 21, 22, 19]. Farenzena *et al.* [10] proposed the Symmetry-Driven Accumulation of Local Features (SDALF). They exploited the property of symmetry in pedestrian images and obtained good view invariance. Ma *et al.* [21] developed the BiCov descriptor, which combined the Gabor filters and the covariance descriptor to handle illumination change and background variations. Malocal *et al.* [22] employed Fisher Vector to encode higher order statistics of local features. All these methods focused on feature design, but rich information from the distribution of samples in the dataset has not been fully exploited. Our approach exploit the salience information among person images, and it can be generalized to take use of these features.

Several appoaches were developed to handle pose variations [27, 11, 1, 7]. Wang *et al.* [27] 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. Gheissari *et al.* [11] fit a triangluar graph model. Bak *et al.* [1] and Cheng *et al.* [7] adopted part-based models to handle pose variation. However, these appoaches are not flexible enough and only applicable when the pose estimators work accurately. Our approach differs from them in that patch matching is employed to handle spatial misalignment.

Contextual visual knowledge coming from surrounding people was used to enrich human signature [28]. Liu *et al.* [19] used an attribute-based weighting scheme, which shared similar spirit with our salience in finding the unique and inherent appearance property. They clustered prototypes in an unsupervised manner, and learned attribute-based feature importance for feature weighting. Their approach was based on global features. They weighted different types of features instead of local patches. Therefore they could not pick up salient regions as shown in Figure 1. Experimental results show that our defined salience is much more effective.

3. Dense Correspondence

Dense correpondence has been applied to face and scene alignment [23, 20]. Inheriting the characteristics of part-based and region-based approaches, fine-grained methods including optical flow in pixel-level, keypoint feature matching and local patch matching are often better choices for more robust alignment. In our approach, considering moderate resolution of human images captured by far-field surveillance cameras, we adopt the mid-level local patches for matching persons. To ensure the robustness in matching, local patches are densely sampled in each image. Different than general patch matching approaches, a simple but effective horizontal constraint is imposed on searching matched patches, which makes patch matching more adaptive in person re-identification.

3.1. Feature Extraction

Dense Color Histogram. Each human image is densely segmented into a grid of local patches. A LAB color histogram is extracted from each patch. To robustly capture color information, LAB color histograms are also computed on downsampled scales. For the purpose of combination with other features, all the histograms are L2 normalized.

Dense SIFT. To handle viewpoint and illumination change, SIFT descriptor is used as a complementary feature to color histograms. The same as the setting of extracting dense color histograms, a dense grid of patches are sampled on each human image. 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. SIFT features are also L2 normalized.

Dense color histograms and dense SIFT features are concatenated as the final multi-dimensional descriptor vector for each patch. In our experiment, the parameters of feature extraction are as follows: patches of size 10×10 pixels are sampled on a dense grid with a grid step size 4; 32-bin color histograms are computed in L, A, B channels respectively, 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 by a discriminative descriptor vector with length $32 \times 3 \times 3 + 128 \times 3 = 672$. We denote the combined feature vector as dColorSIFT.

3.2. Adjacency Constrained Search

In order to deal with misalignment, we conduct adjacency constrained search. dColorSIFT features in human image are represented as $x_{m,n}^{A,p}$, where (A, p) denotes the p-th image in camera A, and (m, n) denotes the patch centered at the m-th row and the n-th column of image p. The m-th row $\mathcal{T}$ of image p from camera A are represented as:

$$\mathcal{T}^{A,p}(m) = \{x_{m,n}^{A,p} | n = 1, 2, ..., N\}. \tag{1}$$

All patches in $\mathcal{T}^{A,p}(m)$ have the same search set $\mathcal{S}$ for patch matching in image q from camera B:

$$\mathcal{S}(x_{m,n}^{A,p}, \mathbf{x}^{B,q}) = \mathcal{T}^{B,q}(m), \ \forall x_{m,n}^{A,p} \in \mathcal{T}^{A,p}(m), \tag{2}$$

where $\mathbf{x}^{B,q}$ represent the collection of all patch features in image q from camera B. The $\mathcal{S}$ restricts the search set in image q within the m-th row. However, bounding boxes produced by a human detector are not always well aligned, and also uncontrolled human pose variations exist in some conditions. To cope with the spatial variations, we relax the strict horizontal constraint to have a larger search range.

$$\begin{aligned} \hat{\mathcal{S}}(x_{m,n}^{A,p}, \mathbf{x}^{B,q}) = \ & \{\mathcal{T}^{B,q}(b) | b \in \mathcal{N}(m)\}, \\ & \forall x_{m,n}^{A,p} \in \mathcal{T}^{A,p}(m), \end{aligned} \tag{3}$$

where $\mathcal{N}(m) = \{m - l, ..., m, ...m + l\}$, $m - l \geq 0$ and $m + l \leq M$. l defines the size of the relaxed adjacent vertical space. If l is very small, a patch may not find correct match due to vertical misalignment. When l is set to be very large, a patch in the upper body would find a matched patch on the legs. Thus 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.

Adjacency Searching. Generalized patch matching is a very mature technique in computer vision. Many off-the-shelf methods [2, 3] are available to boost the performance and efficiency. In this work, we simply do a k-nearest neighbor search for each $x_{m,n}^{A,p}$ in search set $\hat{\mathcal{S}}(x_{m,n}^{A,p}, \mathbf{x}^{B,q})$ of every image in the reference set. The search returns the nearest neighbor for each image according to the Euclidean distance. As suggested in [23], aggregaing similarity scores is much more effective than minimizing accumulated distances, especially for those misaligned or background patches which could generate very large distances during matching. By converting to similarity, their effect could be reduced. We convert distance value to similarity score with the Gaussian function:

$$s(x, y) = exp(-\frac{d(x, y)^2}{2\sigma^2}), \tag{4}$$

where $d(x, y) = \|x - y\|_2$ is the Euclidean distance between patch features x and y, and σ is the bandwidth of

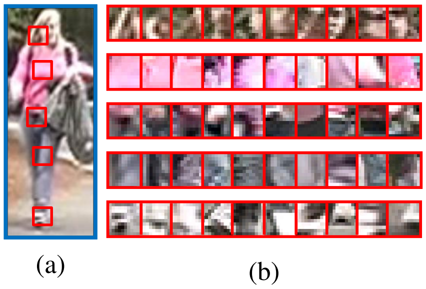


Figure 2. **Examples of adjacency search.** (a) A test image from the VIPeR dataset. Local patches are densely sampled, and five exemplar patches on different body parts are shown in red boxes. (b) One nearest neighbor from each reference image is returned by adjacency search for each patch on the left, and then N nearest neighbors from N reference images are sorted. The top ten nearest neighbor patches are shown. Note that the ten nearest neighbors are from ten different images.

the Gaussian function. Figure 2 shows some visually similar patches returned by the discriminative adjacency constrained search.

4. Unsupervised Salience Learning

With dense correspondence, we learn human salience with unsupervised methods. In this paper, we propose two methods for learning human salience: the K-Nearest Neighbor (KNN) and One-Class SVM (OCSVM).

4.1. K-Nearest Neighbor Salience

Byers *et al*. [5] found the KNN distances can be used for clutter removal. To apply the KNN distance to person re-identification, we search for the K-nearest neighbors of a test patch in the output set of the dense correspondence. With this strategy, salience is better adapted to re-identification problem. Following the shared goal of abnormality detection and salience detection, we redefine the salient patch in our task as follows:

Salience for person re-identification: *salient patches are those possess uniqueness property among a specific set.*

Denote the number of images in the reference set by N_r. After building the dense correspondeces between a test image and images in reference set, the most similar patch in every image of the reference set is returned for each test patch, *i.e.*, each test patch $x^{A,p}_{m,n}$ have N_r neighbors in set $\boldsymbol{X}_{nn}(x^{A,p}_{m,n})$,

$$\boldsymbol{X}_{nn}(x^{A,p}_{m,n}) = \{x| \operatorname*{argmax}_{\hat{x}\in\hat{\mathcal{S}}_{p,q}} s(x^{A,p}_{m,n},\hat{x}), q=1,2,...,N_r\},$$

where $\hat{\mathcal{S}}_{p,q} = \hat{\mathcal{S}}(x^{A,p}_{m,n}, \mathbf{x}^{B,q})$ is the search set in Eq. (3), and s is the similarity score function in Eq. (4).

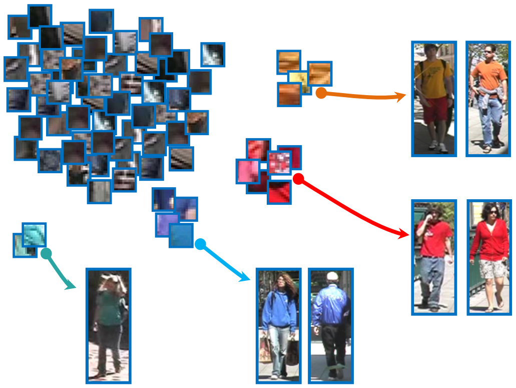


Figure 3. **Illustration of salient patch distribution.** Salient patches are distributed far way from other pathes.

We apply a similar scheme in [5] to $\boldsymbol{X}_{nn}(x^{A,p}_{m,n})$ of each test patch, and the KNN distance is utilized to define the salience score:

$$\boldsymbol{score}_{knn}(x^{A,p}_{m,n}) = D_k(\boldsymbol{X}_{nn}(x^{A,p}_{m,n})), \tag{5}$$

where D_k denotes the distance of the k-th nearest neighbor. If the distribution of the reference set well relects the test scenario, the salient patches can only find limited number ($k = \alpha N_r$) of visually similar neighbors, as shown in Figure 3(a), and then $\boldsymbol{score}_{knn}(x^{A,p}_{m,n})$ is expected to be large. $0 < \alpha < 1$ is a proportion parameter relecting our expectation on the statistical distribution of salient patches. Since k depends on the size of the reference set, the defined salience score works well even if the reference size is very large.

Choosing the Value of k. The goal of salience detection for person re-identificatioin is to identify persons with unique appearance. We assume that if a person has such unique appearance, more than half of the people in the reference set are dissimilar with him/her. With this assumption, $k = N_r/2$ is used in our experiment. For seeking a more principled method to compute human salience, one-class SVM salience is discussion in Section 4.2.

To qualitatively compare with sophiscated supervised learning methods, Figure 4(a) shows the feature weighting map estimated by partial least square (PLS) [26]. PLS is used to reduce the dimensionality and the weights of the first projection vector are shown as the average of the feature weights in each block. Our results of unsupervised KNN salience are show in Figure 4(b) on the ETHZ dataset and 4(c) on the VIPeR dataset. Salience scores are assigned to the center of patches, and the salience map is upsampled for better visualization. Our unsupervised learning method better captures the salient regions.

4.2. One-class SVM Salience

One-class SVM [14] has been widely used for outlier detection. Only positive samples are used in training. The basic idea of one-class SVM is to use a hypersphere to describe data in the feature space and put most of the data into the hypersphere. The problem is formulated into an objective function as follows:

$$\min_{R\in\mathbb{R},\xi\in\mathbb{R}^l,c\in F} R^2 + \frac{1}{vl}\sum_i \xi_i, \tag{6}$$
$$s.t. \|\Phi(X_i) - c\|^2 \le R^2 + \xi_i, \ \ \forall i \in \{1, ...l\} : \xi_i \ge 0,$$

where $\Phi(X_i)$ is the multi-dimensional feature vector of training sample X_i, l is the number of training samples, R and c are the radius and center of the hypersphere, and $v \in [0, 1]$ is a trade-off parameter. The goal of optimizing the objective function is to keep the hypersphere as small as possible and include most of the training data. The optimization problem can be solved in a dual form by QP optimization methods [6], and the decision function is:

$$f(X) = R^2 - \|\Phi(X) - c\|^2, \tag{7}$$
$$\|\Phi(X) - c\|^2 = k(X, X) - 2\sum_i \alpha_i k(X_i, X) + \sum_{i,j} \alpha_i \alpha_j k(X_i, X_j),$$

where α_i and α_j are the parameters for each constraint in the dual problem. In our task, we use the radius basis function (RBF) $K(X, Y) = \exp\{-\|X-Y\|^2/2\sigma^2\}$ as kernel in one-class SVM to deal with high-dimensional, non-linear, multi-mode distributions. As shown in [6], the decision function of kernel one-class SVM can well capture the density and modality of feature distribution. To approximate the KNN salience algorithm (Section 4.1) in a nonparametric form, the sailence score is re-defined in terms of kernel one-class SVM decision function:

$$\boldsymbol{score}_{ocsvm}(x^{A,p}_{m,n}) = d(x^{A,p}_{m,n}, x^*), \tag{8}$$
$$x^* = \operatorname*{argmax}_{x\in \boldsymbol{X}_{nn}(x^{A,p}_{m,n})} f(x),$$

where d is the Euclidean distance between patch features.

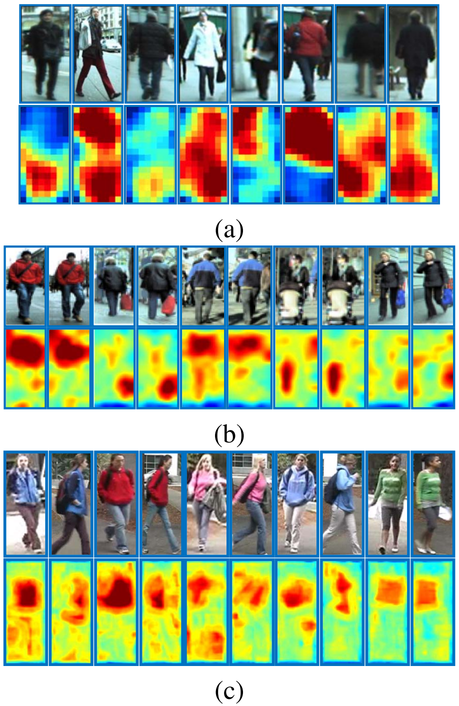
Our experiments show very similar results in person re-identification with the two salience detection methods. $\boldsymbol{score}_{ocsvm}$ performs slightly better than $\boldsymbol{score}_{knn}$ in some circumstances.

5. Matching for re-identification

Dense correspondence and salience described in Section 3 and 4 are used for person re-identification.

5.1. Bi-directional Weighted Matching

A bi-directional weighted matching mechanism is designed to incorporate salience information into dense correspondence matching. First, we consider matching between a pair of images. As mentioned in Section 4.1, patch $x^{A,p}_{m,n}$ is matched to $x^{B,q}$ within search range $\hat{\mathcal{S}}_{p,q} = \hat{\mathcal{S}}(x^{A,p}_{m,n}, x^{B,q})$. Denote the nearest neighbor produced by dense correspondence algorithm as



(a)

(b)

(c)

Figure 4. **Qualitative comparison on salience.** (a) shows the feature weighting maps estimated by partial least square [26]. (b) shows our KNN salience estimation. Red indicates large weights.

$$x^{B,q}_{i,j} = \operatorname*{argmax}_{\hat{x}\in\hat{\mathcal{S}}_{p,q}} s(x^{A,p}_{m,n}, \hat{x}). \tag{9}$$

Then searching for the best matched image in the gallery can be formulated as finding the maximal similarity score.

$$q^* = \operatorname*{argmax}_q \boldsymbol{Sim}(\mathbf{x}^{A,p}, \mathbf{x}^{B,q}), \tag{10}$$

where $\mathbf{x}^{A,p}$ and $\mathbf{x}^{B,q}$ are collection of patch features in two images, *i.e.* $\mathbf{x}^{A,p} = \{x^{A,p}_{m,n}\}_{m\in\mathcal{M},n\in\mathcal{N}}$, and $\mathbf{x}^{B,q} = \{x^{B,q}_{i,j}\}_{m\in\mathcal{M},n\in\mathcal{N}}$, and the similarity between two image is computed with a bi-directional weighting mechanism illustrated in Figure 5. Intuitively, images of the same person would be more likely to have similar salience distributions than those of different persons. Thus, the difference in salience score can be used as a penalty to the similarity score. In another aspect, large salience scores are used to enhance the similarity score of matched patches. Finally, we formulate the bi-directional weighting mechanism as follows:

$$\boldsymbol{Sim}(\mathbf{x}^{A,p}, \mathbf{x}^{B,q}) = \sum_{m,n} \frac{\boldsymbol{score}_{knn}(x^{A,p}_{m,n}) \cdot s(x^{A,p}_{m,n}, x^{B,q}_{i,j}) \cdot \boldsymbol{score}_{knn}(x^{B,q}_{i,j})}{\alpha + |\boldsymbol{score}_{knn}(x^{A,p}_{m,n}) - \boldsymbol{score}_{knn}(x^{B,q}_{i,j})|}, \tag{11}$$

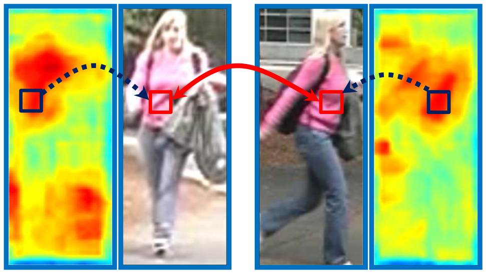


Figure 5. **Illustration of bi-directional weighting for patch matching.** Patches in red boxes are matched in dense correspondence with the guidence of corresponding salience scores in dark blue boxes.

where α is a parameter controlling the penalty of salience difference. One can also change the salience score to $\textbf{\textit{score}}_{ocsvm}$ in a more principled framework without choosing the parameter k in Eq. (5).

5.2. Combination with existing approaches

Our approach is complementary to existing approaches. In order to combine the similarity scores of existing approaches with the similarity score in Eq. (11), the distance between two images can be computed as follows:

$$d_{eSDC}(I_p^A, I_q^B) = \sum_i \beta_i \cdot \boldsymbol{d}_i(f_i(I_p^A), f_i(I_q^B)) \\ -\beta_{SDC} \cdot \textbf{\textit{Sim}}(\mathbf{x}^{A,p}, \mathbf{x}^{B,q}), \qquad (12)$$

where $\beta_i(> 0)$ is the weight for the ith distance measure and $\beta_{SDC}(> 0)$ the weight for our approach. $\boldsymbol{d}_i$ and f_i correspond to the distance measures and features (wHSV and MSCR) in [10]. In the experiment, $\{\beta_i\}$ are chosen the same as in [10]. β_{SDC} is fixed as 1.

6. Experiments

We evaluated our approach on two publicly available datasets, the VIPeR dataset [12], and the ETHZ dataset [26]. These two datasets are the most widely used for evaluation and reflect most of the challenges in real-world person re-identification applications, e.g., viewpoint, pose, and illumination variation, low resolution, background clutter, and occlusions. The results are show in standard Cumulated Matching Characteristics (CMC) curve [27]. Comparisons to the state-of-the-art feature based methods are provided, and we also show the comparison with some classical metric learning algorithms.

VIPeR Dataset [12]. The VIPeR dataset[1] is captured by two cameras in outdoor academic environment with two images for each persons seen from different viewpoints. It is one of the most challenging person re-identification datasets, which suffers from significant viewpoint change, pose variation, and illumination difference between two camera views. It contains 632 pedestrian pairs, each pair contains two images of the same individual seen from different viewpoints, one from CAM A and another from CAM B. All images are normalized to 128 × 48 for experiments. CAM A captured images mainly from 0 degree to 90 degree while CAM B mostly from 90 degree to 180 degree, and most of the image pairs show viewpoint change larger than 90 degree.

Following the evaluation protocol in [13], we randomly sample half of the dataset, *i.e.*, 316 image pairs, for training (however, the identity information is not used), and the remaining for test. In the first round, images from CAM A are used as probe and those from CAM B as gallery. Each probe image is matched with every gallery image, and the correctly matched rank is obtained. Rank-k recognition rate is the expectation of the matches at rank k, and the CMC curve is the cumulated values of recognition rate at all ranks. After this round, the probe and gallery are switched. We take the average of the two rounds of CMC curves as the result of one trial. 10 trials of evaluation are repeated to achieve stable statistics, and the average result is reported.

Since ELF[13], SDALF[10], and LDFV[22] have published their results on the VIPeR dataset, they are used for comparison. The splitting assignments [2] in these approaches are used in our experiments. Figure 6 report the comparison results. It is observed that our two salience detection based methods (SDC_knn and SDC_ocsvm) outperform all the three benchmarking approaches. In particular, rank 1 matching rate is around 24% for SDC_knn and 25% for SDC_ocsvm, versus 20% for SDALF, 15% for LDFV, and 12% for ELF. The matching rate at rank 10 is around 52% for SDC_knn, and 56% for SDC_ocsvm, versus 49% for SDALF, 48% for LDFV, and 44% for ELF. The improvement is due to two aspects of our approach. First, the dense correspondece matching can tolerate larger extent of pose and appearance variations. Second, we incorporate human salience information to guide dense correspondence. By combining with other descriptors, the rank 1 matching rate of eSDC_knn goes to 26.31% and eSDC_ocsvm goes to 26.74%. This shows the complementarity of our SDC approach to other features. More comparison results are show in Table 1. The compared methods includes the classical metric learning approaches, such as LMNN [29], and ITML [29], and their variants modified for person re-identification, such as PRDC [29], attribute PRDC (denoted as aPRDC) [19], and PCCA [24].

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

[2] The splitting assignment of SDALF can be found in their code at `http://www.lorisbazzani.info/code-datasets/sdalf-descriptor/`

Method	r=1	r=5	r=10	r=20
LMNN[29]	6.23	19.65	32.63	52.25
ITML [29]	11.61	31.39	45.76	63.86
PRDC[29]	15.66	38.42	53.86	70.09
aPRDC[19]	16.14	37.72	50.98	65.95
PCCA [24]	19.27	48.89	64.91	80.28
ELF[13]	12.00	31.00	41.00	58.00
SDALF [10]	19.87	38.89	49.37	65.73
CPS [7]	21.84	44.00	57.21	71.00
eBiCov [21]	20.66	42.00	56.18	68.00
eLDFV [22]	22.34	47.00	60.04	71.00
eSDC_knn	**26.31**	**46.61**	**58.86**	**72.77**
eSDC_ocsvm	**26.74**	**50.70**	**62.37**	**76.36**

Table 1. VIPeR dataset: top ranked matching rates in [%] with 316 persons.

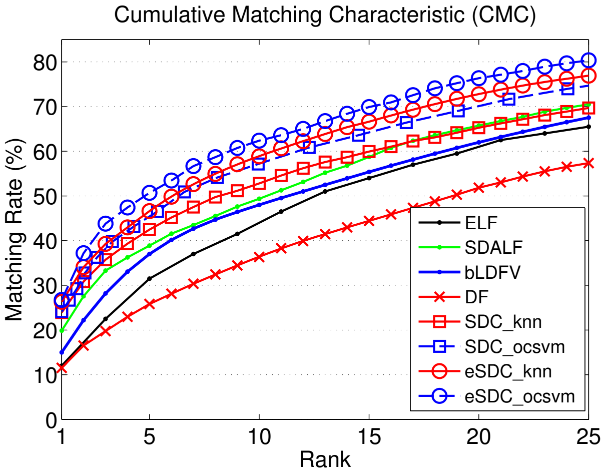


Figure 6. Performance on the VIPeR dataset. Our approach: SDC_knn and SDC_ocsvm. Our approach combined with wHSV and MSCR [10]: eSDC_knn and eSDC_ocsvm.

ETHZ Dataset [9]. This dataset[3] contains three video sequences captured from moving cameras. It contains a large number of different people in uncontrolled conditions. With these videos sequences, Schwartz, et al. [26] extracted a set of images for each people to test their Partial Least Square method. Since the original video sequences are captured from moving cameras, images have a range of variations in human appearance and illumination, and some even suffer from heavy occlusions. Following the settings in [26], all image samples are normalized to 64 × 32 pixels, and the dataset is structured as follows: SEQ.#1 contains 83 persons (4,857 images); SEQ.#2 contains 35 persons (1,936 images); SEQ.#3 contains 28 persons (1,762 images).

The same settings of experiments in [10, 26] are reproduced to make fair comparisons. Similar to them, we use a single-shot evaluation strategy. For each person, one image is randomly selected to build gallery set while the rest images form the probe set. Each image in probe is matched to every gallery image and the correct matched rank is obtained. The whole procedure is repeated for 10 times, and the average CMC curves are plotted in Figure 7.

As shown in Figure 7, our approach outperforms the three benchmarking methods, PLS, SDALF and eBiCov[21] on all three sequences. Comparisons with supervised learning methods PLS and RPLM are reported in Table 2. On SEQ.#2 and SEQ.#3, our eSDC_knn and eSDC_ocsvm outperforms all other methods. On SEQ.#1, our SDC approach has better results than supervised methods, PLS and RPLM, and has comparable performance with the recently proposed eLDFV[22].

7. Conclusion

In this work, we propose an unsupervised framework with salience detection for person re-identification. Patch matching is utilized with adjacency constraint for handling the viewpoint and pose variation. It shows great flexibility in matching across large viewpoint change. Human salience is unsupervisedly learned to seek for discriminative and reliable patch matching. Experiments show that our unsupervised salience learning approach greatly improve the performance of person re-identification.

8. Acknowledgement

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

References

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

[2] C. Barnes, E. Shechtman, A. Finkelstein, and D. Goldman. Patchmatch: a randomized correspondence algorithm for structural image editing. *TOG*, 2009.

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

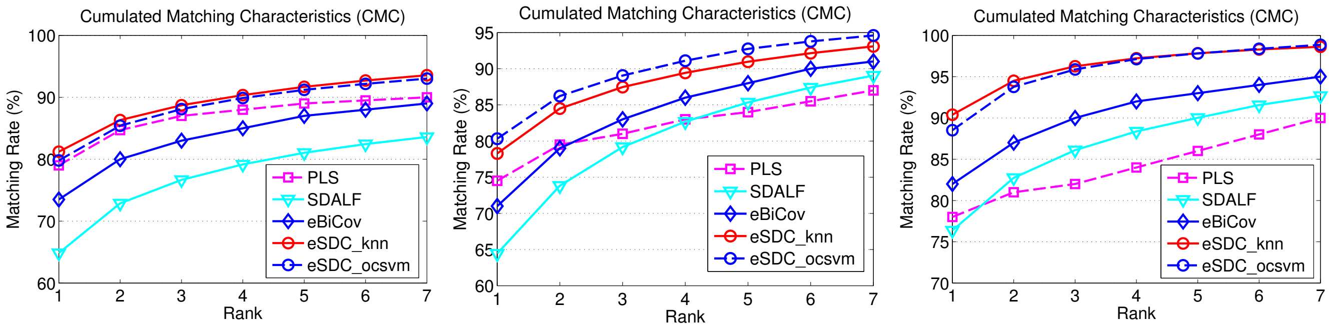
[4] A. Borji and L. Itti. Exploiting local and global patch rarities for saliency detection. In *CVPR*, 2012.

[5] S. Byers and A. Raftery. Nearest-neighbor clutter removal for estimating features in spatial point processes. *Journal of the American Statistical Association*, 1998.

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

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

[3] The ETHZ dataset is available to download at the website `http://homepages.dcc.ufmg.br/~william/datasets.html`



SEQ.#1 SEQ.#2 SEQ.#3

Figure 7. Performances comparison using CMC curves on SEQ.#1, SEQ.#2, and SEQ.#3 of the ETHZ dataset. According to [10], only the first 7 ranks are shown. All the compared methods are reported under single-shot setting.

Method	SEQ.#1							SEQ.#2							SEQ.#3						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7
PLS [26]	79	85	86	87	88	89	90	74	79	81	83	84	85	87	77	81	82	84	85	87	89
RPLM [16]	77	83	87	90	91	92	92	65	77	81	82	86	89	90	83	90	92	94	96	96	97
SDALF [10]	65	73	77	79	81	82	84	64	74	79	83	85	87	89	76	83	86	88	90	92	93
eBiCov [21]	74	80	83	85	87	88	89	71	79	83	86	88	90	91	82	87	90	92	93	94	95
eLDFV [22]	**83**	**87**	**90**	**91**	92	93	94	79	85	88	90	**92**	**93**	**94**	**91**	94	**96**	**97**	97	97	97
eSDC_knn	**81**	**86**	**89**	**90**	**92**	**93**	**94**	**79**	**84**	**87**	**90**	**91**	**92**	**93**	**90**	**95**	**96**	**97**	**98**	**98**	**99**
eSDC_ocsvm	**80**	**85**	**88**	**90**	**91**	**92**	**93**	**80**	**86**	**89**	**91**	**93**	**94**	**95**	**89**	**94**	**96**	**97**	**98**	**98**	**99**

Table 2. Matching rates in [%] on the *ETHZ* dataset. Our approach (eSDC_knn and eSDC_ocsvm) is compared with supervised learning methods PLS and RPLM, and unsupervised methods SDALF, eBiCov, and eLDFV. In accordance with what reported in other methods, only the matching rates at the first 7 ranks are shown.

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

[9] A. Ess, B. Leibe, and L. Van Gool. Depth and appearance for mobile scene analysis. In *ICCV*, 2007.

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

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

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

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

[14] 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 (DMSEC)*, 2003.

[15] M. Hirzer, C. Beleznai, P. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. *Image Analysis*, 2011.

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

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

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

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

[20] C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *TPAMI*, 2011.

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

[22] B. Ma, Y. Su, and F. Jurie. Local descriptors encoded by fisher vectors for person re-identification. 2012.

[23] K. Ma and J. Ben-Arie. Vector array based multi-view face detection with compound exemplars. In *CVPR*, 2012.

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

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

[26] W. Schwartz and L. Davis. Learning discriminative appearance-based models using partial least squares. In *XXII Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI)*, 2009.

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

[28] W. Zheng, S. Gong, and T. Xiang. Associating groups of people. In *BMVC*, 2009.

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