Pattern Recognition 75 (2018) 90–98

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Person re-identification via integrating patch-based metric learning and local salience learning

Zhicheng Zhao[a,b,*], Binlin Zhao[a], Fei Su[a,b]

[a] School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China
[b] Beijing Key Laboratory of Network System and Network Culture, Beijing, China

ARTICLE INFO

Article history:
Received 15 November 2016
Revised 12 March 2017
Accepted 19 March 2017
Available online 24 March 2017

Keywords:
Person re-identification
CNN feature
Patch-based metric learning
Local salience learning
Cross-dataset

ABSTRACT

In this paper, aiming at improving the generalization capability, we propose a cross-dataset person re-identification framework via integrating patch-based metric learning and local salience learning. Firstly, Convolution Neural Network(CNN) features are extracted to represent patches of a person. Secondly, only two positive patch-pairs are chosen and input into a Large Margin Nearest Neighbour(LMNN) network to learn two patch-based metric matrices for feature projection respectively. Thirdly, according to projected new features, a local salience learning algorithm based on Kmeans clustering is proposed to train the weights of patches. Finally, the similarity of image-pair is computed by a weighted summing of all patches. The experimental results indicate that the proposed method outperforms existing conventional approaches based on hand-crafted features and achieves a comparable performance with most recent CNN-based methods, which demonstrates our method's effectiveness and practicality. It does not need a large-scale labeled training dataset, and has a high matching rate with a low computation complexity.

1. Introduction

The goal of person re-identification (re-id) is to match pedestrians across non-overlapping camera views in different scenes. It can be simplified into an image retrieval problem, i.e., for a given person image (query), the re-id system identifies corresponding images from a dataset (gallery). Appearance-based methods are usually adopted to handle this issue. However, due to significant changes from image configurations (i.e., view angle, lighting condition, pose and occlusion etc.), appearances of a individual often show heavy variations, resulting in a complex feature distribution, and accordingly complicates this task.

A number of approaches are proposed to deal with above problems and they mainly focus on two aspects: 1) the feature representation, 2) the similarity metric. The popular features include color, texture-based features such as LBP, Garbor features and local interesting points etc. For similarity measure, the metric learning, which aims at minimizing the intra-class difference and maximizing the inter-class variations, plays an important role.

Recently, many experiments have demonstrated the limitations of hand-crafted features for person representation, thus deep models were introduced. For instance, [1] applied AlexNet to compute the similarity of query and gallery images [2]. used a Siamese network to calculate distances of person-pairs. Compared with conventional methods based on hand-crafted features, deep models achieved better performance on multiple re-id datasets.

However, the training of deep networks exists two difficulties: 1) it needs a large-scale labeled samples, while insufficient dataset is a common problem. 2) the learning of the model is a quite time-consuming task. Moreover, most of methods conduct experiments in a similar way: dividing a dataset into two subsets, the first part is used to train model and the rest one is applied to test. This is an intra-dataset experiment, i.e., the training set (source domain) and test set (target domain) come from the same dataset with the same feature distribution, which actually decreases the difficulty of true re-id.

In the practical application, the re-id system has to face the cross-dataset challenge: target images are collected from different scenes, thus the feature distribution will be very different. Therefore, a high matching rate may be achieved on intra-dataset experiments, while the generalization ability is questionable in a cross-dataset test, which is a vital rule for performance evaluation. In this paper, aiming at improving the generalization capability and practicality of re-id system, we propose a novel re-id framework via integrating patch-based metric learning and local salience learning. Our contributions are summarized as three aspects:

* Corresponding author:
E-mail addresses: zhaozc@bupt.edu.cn, zhao.zc@gmail.com (Z. Zhao), iamthezbl@126.com (B. Zhao), sufei@bupt.edu.cn (F. Su).

http://dx.doi.org/10.1016/j.patcog.2017.03.023
0031-3203/© 2017 Elsevier Ltd. All rights reserved.

1. An extendable re-id framework is proposed. It contains two related parts: patch-based metric learning and local salience learning. First, to handle the problem of pose variant, CNN features are extracted to represent the person, and then a light patch-based metric learning method-pLMNN is leveraged to enhance the discrimination of raw features.
2. A Kmeans-based local salience learning algorithm is presented to train the weights of image patches. Meanwhile, a general similarity computation scheme is designed to relieve an existing training problem, i.e., the parameters need to be re-trained for different datasets.
3. The experimental results on cross-datasets demonstrate our method's generalization ability and the effectiveness.

The rest of paper is organized as follows. Section 2 reviews the related work. Section 3 introduces the proposed algorithm. Section 4 is the experimental results. Finally, Section 5 concludes this paper.

2. The related work

A typical re-id system involves two related issues: feature representation and similarity metric. The former aims to extract discriminative features to represent pedestrian's appearance, and the latter focuses on the learning of optimal similarity measure between the query and gallery images.

Great efforts have been made to the design of invariant features [3–10]. The popular low-level features contain global color features (e.g., HSV and Lab histogram) and texture features such as LBP and Garbor features [9], and SIFT-like features. Additionally, the feature fusion has also been explored to enhance the discrimination [11,12].

On the other hand, the metric learning [13] has been paid growing attention due to its feasibility. Compared with traditional distance metrics, e.g., L_1, L_2 norms and Cosine measure, metric learning was more effective for appearance variations. Mahalanobis metric learning (KISSME)[5], Local Fisher Discriminant Analysis (LFDA)[7], Marginal Fisher Analysis (MFA)[14], large margin nearest neighbour (LMNN)[15], Locally Adaptive Decision Functions (LADF)[16] were typical metric learning algorithms.

However, global metric learning cannot deal with local variations of original data space. Therefore, some local matching schemes were proposed to improve the flexibility of re-id systems. For example, Zheng et.al proposed a partial person re-id method, which used a local patch-level matching model based on a sparse representation [12]. Zhao et.al proposed a salience learning algorithm to assign scores for different image patches [11]. This approach relaxed the spatial constraint and can handle pose variations, thus significantly improved the performance. mFilter [17] also adopted a local patch matching strategy, which learned the mid-level filters to get local discriminative features. Regularized local metric learning (RLML)[18] exploited the discriminative information from local distributions in the original space.

Recently, inspired by great successes of deep networks in multiple image classification and computer vision tasks, Convolution Neural Networks (CNN) were introduced into the field of person re-id and obtained state-of-the-art performances on several benchmark datasets such as i-LIDS [19], VIPeR [20] and CUHK01[6]. For instance, [21] built a triplet network to learn similarity metrics. DeepReID [22] constructed a pairing-neural-network to resolve problems of occlusion, misalignment and geometric transforms. Ahmed et al.[2] proposed an end-to-end deep learning architecture, which looked pairwise images as inputs, and output a similarity score directly. Using body parts, [23] presented a Siamese network to learn the pairwise similarity. Deep Metric Learning (DML)[24] applied the full-connected layer (fc) features to learn the metric matrix and improved the performance of conventional metric learning.

For cross-dataset person re-id, however, few literatures concerned [25]. proposed a transfer Rank-SVM, which transfered a model from the source domain (i-LIDS or PRID) to the target domain (VIPeR) [24]. adopted a different scheme: it trained the model in the source domain and tested in the target domain [26]. propsoed a deep transfer metric learning algorithm to learn a set of hierarchical nonlinear transformations for cross-domain visual recognition and achieved promising results.

Although impressed results have been obtained, existing studies have several limitations: 1) CNN-based methods show the superiority, while model's training will rely on a large number of positive sample-pairs (labeled sample-pairs are actually deficient). Meanwhile, it is also time-consuming, leading to the flexible updating of the CNN model intractable. 2) most global metric learning algorithms cannot fit local variations of original data. 3) the integration of CNN with metric learning does not been paid abundant attention. 4) cross-dataset experiments are still insufficient, and majority tests are conducted within intra-dataset. Therefore, the generalization capability is still an open question.

To address above problems, we propose an extendable framework via intergrating patch-based metric learning and local salience learning algorithm. The cross-dataset experiments show the effectiveness of the proposed approach.

3. The proposed framework

Fig. 1 shows our re-id framework, which includes four parts: 1) image pre-processing. Image enhancement is firstly applied to restrain the impact of lighting variations, and then the image is normalized into two types of resolutions (224x224 pixels and 128x48 pixels, here, we call them C_image and P_image respectively). The P_image is further segmented into dense patches. 2) feature extraction. CNN features from convolution layers are extracted to represent the image patch. 3) patch-based metric learning. First, two positive patch-pairs are manually chosen from the upper-body and lower-body of a labeled pedestrian-pair. Second, they are input a LMNN network respectively to learn two local metric matrices for feature projection. 4) saliency learning for each patch. According to projected new features, a local salience learning algorithm based on K-means is proposed to train the weights of dense patches. The similarity of image-pair is computed by a weighted summing of all patches.

We observe that the current CNN-based methods mainly focus on the design of network architecture and the training of the model, such as TripleNet and Siamese network etc. In this paper, we pay attention to the generalization ability and extendibility of the system.

3.1. Patch-based metric learning

3.1.1. Image pre-processing and dense segmentation

A big challenge of the re-id is lighting distortion, often resulting in the mismatching among different pedestrians, thus a pre-processing is used to relieve this difficulty. We first transform the image into HSV color space, and then apply the histogram equalization in V-channel to adjust the imbalance of the luminance. Fig. 2 shows an instance. Afterwards, we normalize the image into C_image and P_image to extract features respectively.

The pose variation is another obvious disturbance. For example, in the VIPeR dataset, the pose differences of many sample-pairs exceed 90 °. To weaken this negative impact, insead of global matching, we adopt a local matching method. In our implementation, an image is divided into dense patches by an overlapping slipping-window scheme. The size of the window is 10x10 pixels with a 4-

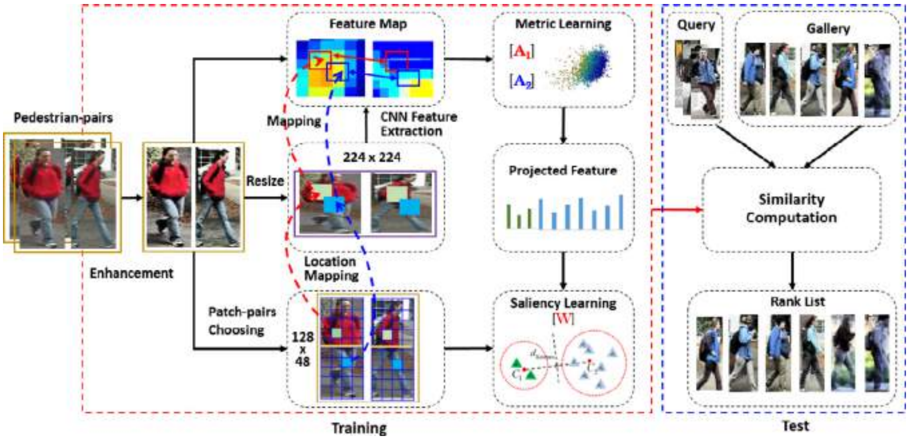


Fig. 1. The overview framework. The training module includes four steps: 1) image enhancement and resizing. 2) CNN feature extraction. 3) positive patch-pair choosing and patch-based metric learning. 4) saliency learning via the Kmeans clustering. Results ranking based on the weighted matching is conducted in the test.

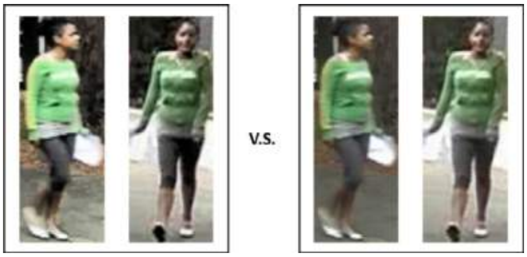


Fig. 2. Image enhancement for V-channel in HSV color space.

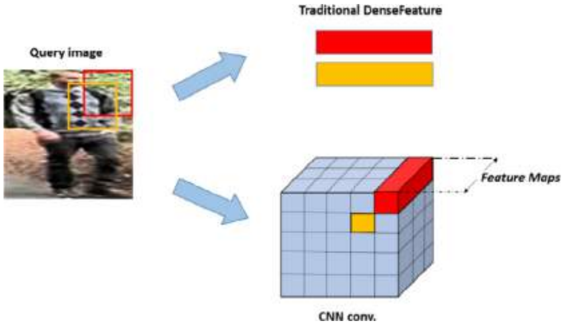


Fig. 3. The corresponding relationship between CNN feature and patches of the image.

pixel stride. As a result, a pedestrian (P_image) is segmented into 300 patches.

3.1.2. CNN Feature extraction

In this section, we explore the representation of the person. Here, we extract CNN features from the C_image.

In multiple computer vision tasks, the deep features show better representation than hand-crafted features, thus we extract CNN features to represent the pedestrian [24]. constructed a 5-layer CNN and extracted the fc feature to describe the image. Considering fc features cannot contain clear spatial information, we extract convolution layer features (conv). Specifically, we firstly train our deep model through fine-tuning the VGG-M model [27] on Market1501 pedestrian dataset [28]. Secondly, based on our model, three feature maps from the 3-th, 4-th and 5-th convolution layer are generated. Thirdly, we respectively extract conv features from feature map's local regions, whose locations are corresponding with patches of P_image. Fig. 3 shows the corresponding relationship between the conv feature and patches.

3.1.3. Feature fusion

The feature fusion [29,30] plays an important role for effective re-id. In our implementation, we adopt two strategies to fuse conv features: peer-layer fusion and cross-layer fusion, which are shown in Fig. 4(a) and Fig. 4(b) respectively. For peer-layer fusion, we cascade neighboring 3x3 features to obtain new feature vectors, which contain bigger receptive fields. The cross-layer fusion can be further divided into early fusion an late fusion. In early fusion, 3 conv features is integrated, and results fusion is applied in late fusion phase.

3.2. Patch-based metric learning

We propose a patch-based metric learning method to obtain a stable feature projection. It includes 3 parts: training patch-pair selection, metric matrix learning and the matching algorithm.

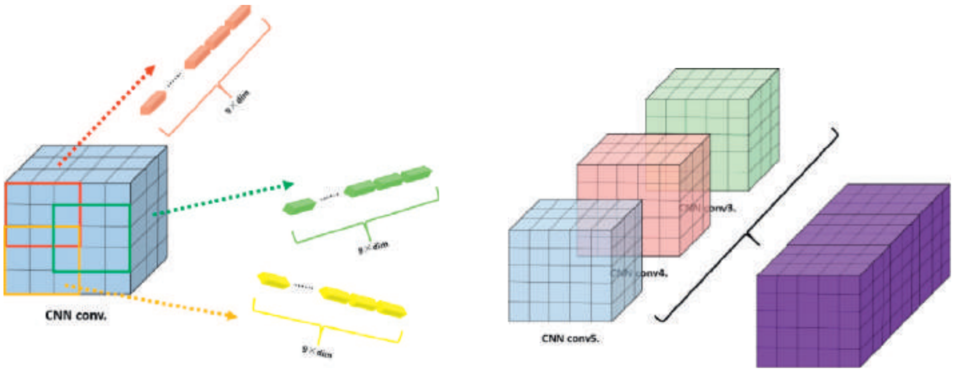
3.2.1. Training patch-pair generation

The proposed patch-base metric learning is different with conventional metric learning. The object labels of the latter are easy to assign, i.e., the same individual certainly is the same label. For the former situation, however, there is not a clear definition whether two patches belong to the same label. Patches which are randomly chosen from the same pedestrian may have a huge visual difference so as to disturb the learning of the metric matrix.

A pedestrian can be naturally divided into two regions: the upper-body and the lower-body. Two parts (clothes and trousers) always show different appearance. Therefore, in order to decrease false patch label, only one positive patch-pair from upper and lower parts are chosen for training respectively. The advantage is that manual labeling is sharply reduced, and meanwhile, a lot of incorrect labels are avoided. Fig. 5 shows the generation of positive patch-pairs. The negative patch-pairs are produced based on the random combination during the metric learning.

3.2.2. Patch-based LMNN:pLMNN

As a representative metric learning method, LMNN [15] has attracted wide attention. The conventional LMNN learns a global projection matrix for person-pairs. However, the global metric cannot adjust to the difference of upper and lower bodies. Therefore, we propose a patch-based LMNN method (pLMNN) to improve the performance. For a pair of pedestrians, two positive patch-pairs are used to learn two local metric matrices respectively to minimize the intra-class difference and maximize the inter-class variation.



(a) Peer-layer feature fusion

(b) Cross-layer fusion

Fig. 4. Two feature fusion schemes.

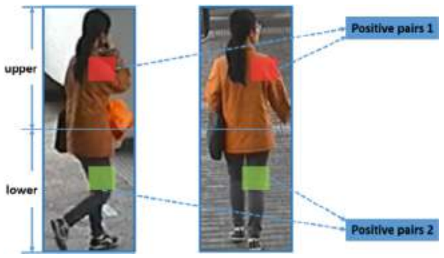


Fig. 5. The positive patch-pair selection.

The core algorithm is written as Eqs. (1) and (2):

$$\min_{A,\xi} \sum_{(x_i,x_j)\in S} d(x_i,x_j) + \gamma \sum_{(x_i,x_j,x_k)\in R} \xi_{ijk}$$
$$\text{s.t. } d(x_i,x_k) - d(x_i,x_j) >= 1 - \xi_{ijk}, \forall (x_i,x_j,x_k) \in R$$
$$A >= 0, \xi_{ijk} >= 0 \quad (1)$$

$$d(x_i,x_j) = (x_i - x_j)^T A (x_i - x_j) \quad (2)$$

where A is the learned projection matrix, which is a semi-definite positive matrix. The triad x_i, x_j, x_k are feature vectors of the patches. x_i, x_j belong to the same class S, and are different from x_k. R denotes the set of all training samples, and ξ_{ijk} is a slack variable. Via matrix A, an Euclidean distance $(x_i - x_j)$ is transformed into a new metric $d(x_i, x_j)$. In experiments, the weighting parameter was set to $\gamma = 0.5$. Finally, we apply a gradient descent solver [15] to solve this optimization.

Afterwards, we use L_2-norm to compute the distance of local patch-pairs. Finally, this distance is converted into a score by Eq. (3). The bigger the score is, the higher the similarity will be. The scale parameter σ is set to balance the difference from upper-body and lower-body metrics. We apply two patch-based LMNN matrixs to project features, thus a weighted summing for all patches will leverage at different scales to obtain the optimal effect. As a result, σ plays a key role to adjust the metric of each patch. In experiments, we will discuss the impact of σ in detail.

$$s(x,y) = \mathbf{exp}\left(\frac{d(x,y)^2}{2\sigma^2}\right) \quad (3)$$

3.2.3. Horizontal matching scheme

Although pose differences are common, large variations cannot regularly appear in vertical direction. Therefore, patches matching are only applied in horizontal direction is reasonable. As shown in Fig. 6, we adopt a horizontal constraint scheme: retrieve a query patch with neighbor patches of gallery images in the same row, and then regard the nearest patch as the matching-pair. Moreover, as we use conv3, conv4 and conv5 features, thus each pixel in feature maps will correspond to a bigger receptive field in the original image. As a result, vertical retrieval in row neighbor is omit. Finally, only one matching-pair is determined for each row.

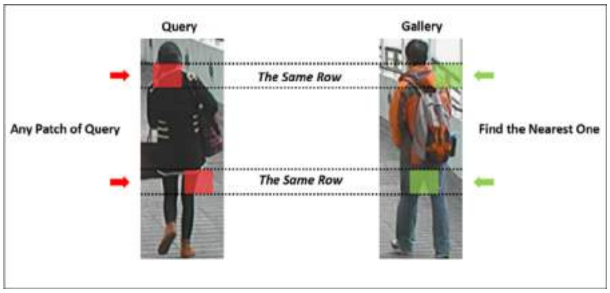


Fig. 6. An instance for horizontal constraint.

3.3. Kmeans-based local salience learning

The goal of the salience learning is to learn a weight for each patch to embody its importance: the reliable and discriminative patch should have a high value. The final similarity of two images is obtained by a weighted summing over all patches [11]. adopted KNN-based and OCSVM-based methods to train the weights. In KNN-based approach, the parameter K not only is hard to determine, but also not fixed in different datasets. OCSVM-based method has to face a problem of the time-consuming training.

In this paper, we propose a Kmeans-based local salience learning algorithm, which is original from the hypothesis: if the majority training patches are similar with a given patch, it should be assigned a small weight (the higher the similarity is, the smaller the weight will be), and vice versa. Hence, an unsupervised Kmeans clustering (K=2) can adaptively get the hyperplane of binary classification.

Take two cameras A, B as an example. Let TR_A, TR_B and T_A, T_B represent the training dataset and test dataset from A and B respectively. Their sizes are NR_A, NR_B, N_A and N_B. In our cross-dataset experiments, the saliences of T_A are learned from TR_B: for a patch p_i in T_A, compute its NR_B neighbors q_{j*} and corresponding distances $d_{ij}*$ from TR_B. After clustering $d_{ij}*$ into 2 classes, we look the mean of two centers C_1 and C_2 as the salience value of p_i. Similarly, the weights of T_B are trained from TR_A. The algorithm is described in Algorithm 1.

Algorithm 1 Kmeans-based salience learning for P_image.

```
Require: $x_i$ and $x_{jk}$: the feature vectors, which extracted from
    test patch $p_i$ and training patch $q_{jk}$ (in the same row), $j = 1, 2, \ldots, NR_B$, $k = 1, 2, \ldots, 10$;
Ensure: Salience score of patch $p_i$: $S(x_i)$.
For an arbitrary $x_i$ in $T_A$;
for $j = 1$ to $NR_B$ from $TR_B$ do
    for $k = 1$ to 10 do
        Compute the distances: $d(x_i, x_{jk})$;
    end for
    Find the nearest neighbors: $d_{ij}* \leftarrow \arg \min_j d(x_i, x_{jk})$;
    Cluster $d_{ij}*$ into 2 classes with centers $C_1, C_2$;
    Compute the salience value: $C(x_i) = (C1 + C2)/2$;
    Max-min normalization. The final salience score of $p_i$ can be
    written as:
        $$S(x_i) = \frac{C(x_i) - min}{max - min} \quad (4)$$
end for
```

The salience score is determined by Kmeans clustering, which is adaptive and no parameter need to be tuned. Hence, it is not only more tractable than KNN-based method, but also has a lower computation complexity than OCSVM-based approach.

3.4. The weighted matching scheme

Based on salience scores of patches, the final matching value of image-pair (i.e., query and gallery images, and are denoted by Q, G subsequently) can be naturally obtained via a linearly weighted summing. Zhao et.al [11] adopted Eq. (5) to compute this value.

$$V_{qg} = \sum_{x \in Q, y^* \in G} \frac{S(x)d(x, y^*)S(y^*)}{\alpha + |S(x) - S(y^*)|} \tag{5}$$

where x comes from Q, y^* belongs to G and $y^* \leftarrow \arg \min_{y \in L} d(x, y)$. The denominator is use to adjust the scale differences of salience scores, and α is a balance parameter. In practice, α is hard to empirically get. Additionally, experimental decision for α in different datasets is also unfeasible. Considering the cross-dataset test, we revise the equation into Eq. (6). The new function speedups the computation and is not disturbed by α.

$$V_{qg} = \sum_{x \in Q, y^* \in G} S(x)d(x, y^*)S(y^*) \tag{6}$$

4. Experiments

In this section, we evaluate the performance of our approach. Two types of experiments are set up: the intra-dataset experiments are conducted to illustrate the basic performance of the proposed method and the cross-dataset experiments demostrate the generalization capability.

In order to compare the performance of CNN features and conventional features, like [11], two hand-crafted features, i.e., multiscale Lab color histograms and dense-SIFT features are extracted to represent each patch.

4.1. Datasets and setting

We use ETHZ dataset [31] to learn the weights of patches, and then evaluate the algorithm on two challenging datasets: VIPeR and CUHK01.

ETHZ dataset. It contains 3 video sequences(SEQ_1, SEQ_2 and SEQ_3) of street scenes, which are captured by 2 moving cameras. SEQ_1 includes 83 persons with 4857 images. SEQ_2 contains other 35 pedestrains and 1936 images, and SEQ_3 has 28 individuals and 1762 images.

As the performance on ETHZ dataset tends to be saturated. Therefore, we only use it to learn the salience scores of patches. In our implementation, SEQ_1, SEQ_2 and SEQ_3 are used individually, and meanwhile, for each person in three sequences, only one image is randomly chosen.

VIPeR dataset. This dataset contains 632 persons (2 images per subject) and are collected from two cameras (camera A and camera B). It is one of the most challenging dataset due to huge intra-variations. We randomly split VIPeR into 2 groups: 316 persons for training and the resting 316 subjects for testing.

CUHK01 dataset. It is a large-scale dataset, containing 971 persons who are captured from two camera views in campus. Camera A captures frontal or back views of a person while camera B takes the person's profile. Each person has 4 images(each camera captures 2 ones). 485 persons are randomly chosen to train and the remaining 486 ones are used to test.

4.1.1. Evaluation protocol

We adopt CMC metric for quantitative evaluation. In the phase of test, the algorithm will return the top-n nearest images from the gallery set. If the list contains the target image at the k-th position, then this query is considered as the success of Rank-k. We repeat the procedure 10 times and use the average as the final result.

4.1.2. Parameter setting

In experiments, except the scale σ, other parameters are fixed. Moreover, the VIPeR and CUHK01 datasets keep the same setting so as to accord with the practical application, i.e., it is inconvenient to re-tune parameters when the system is used to different scenes. Additionally, for the hand-crafted feature and CNN feature, we have two different configurations.

- Hand-crafted feature: the P_image is divided into 30 × 10=300 patches (the size of a single patch is 10 × 10 × 00A0;pixels and the stride is 4; The upper body and lower one includes 15 × 10 patches respectively). Then, Lab histogram and dense-SIFT are extracted: 1) Lab histogram: from 3 color channels and at 3 scales (1, 0.75 and 0.5), Lab histograms are first extracted and L_2-normalized, and then, each color channel is quantized into a 32-bin histogram at each scale. Hence, a 32 × 3 × 3-dimensional Lab feature vector is produced. 2) dense-SIFT: each patch is divided into 4 × 4 cells, and then SIFT feature is extracted at 3 scales and 8 orientations. After L_2-normalization, a 4 × 4 × 8 × 3-dimensional feature vector is generated. Finally, Lab histogram and dense-SIFT feature are cascaded into a 672-dimensional feature vector to represent a patch.
- CNN feature: based on VGG-M model and Caffe toolbox [32], the C_image is input to extract conv3, conv4 and conv5 features. At each layer, we get 13 × 13=169 512-dimension feature vectors.
- In KNN and OCSVM methods, the parameter α is set to 2.5 for Lab and SIFT features and is set to 1.5 for CNN feature.
- The number of iteration of pLMNN is set to 10000, and the input feature vector is reduced to 100 dimensions via PCA.

4.2. Intra-dataset comparison on the VIPeR dataset

We evaluate the performance in different configurations on intra-dataset experiments, where one image of a person from camera A is looked as a query image and one image of the same person from camera B is regarded as a gallery image.

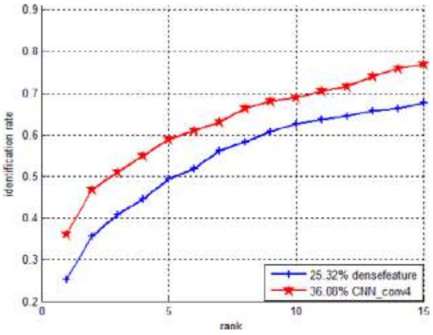


Fig. 7. Comparison for features.

Table 1
The impact of the scale parameter σ.

σ (%)	Rank1	Rank5	Rank10	Rank15	Rank20
0.5	38.0	63.0	74.1	82.0	86.1
0.6	39.9	66.5	76.0	84.8	88.9
0.9	40.8	67.1	76.0	84.5	89.3
1.1	41.5	67.4	78.2	84.8	88.9
1.3	**41.8**	**68.7**	**78.8**	**85.4**	**90.5**
1.5	40.8	67.4	78.5	85.4	90.2
1.6	40.2	67.1	79.1	86.4	90.2
2.0	40.2	66.8	79.4	86.1	89.9
2.2	38.9	66.1	78.8	85.4	89.6

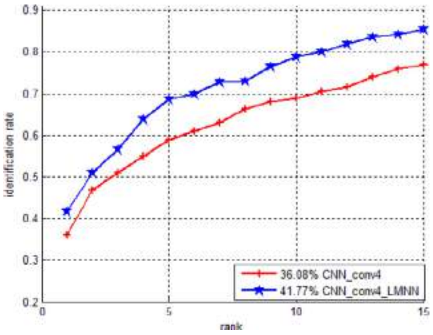


Fig. 8. Comparison for metrics.

4.2.1. The comparison of features

At first, we compare the performance of two types of features in the same setting, and L_2 distance is used to compute the similarity of the image-pair. In Fig. 7, the CMC shows that CNN feature (conv4) outperforms Lab+SIFT feature about 11% and indicates a better representation.

4.2.2. Comparison for scale parameter σ and metrics

In Eq. (3), the scale parameter σ is set to balance the difference of upper-body and lower-body metrics. Table 1 shows the impacts of different σ. We can see when σ varies from 0.9 to 1.6, the CMC has not an obvious variation. In the following test, we choose σ=1.3.

Fig. 8 gives the results of proposed pLMNN-based method and L_2 distance. It is clear that pLMNN is much better than L_2 metric and can discriminate features more effective. In the course of pLMNN learning, only two positive patch-pairs need to be labeled, thus it is a worthy labour to exchange a significant improvement of the performance.

4.2.3. Comparison for salience learning methods

In this subsection, we compare our Kmeans-based salience learning with existing KNN and OCSVM methods [11] on the VIPeR dataset.

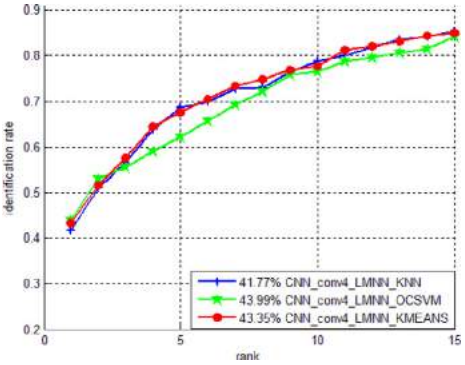


Fig. 9. The comparison of three salience learning methods.

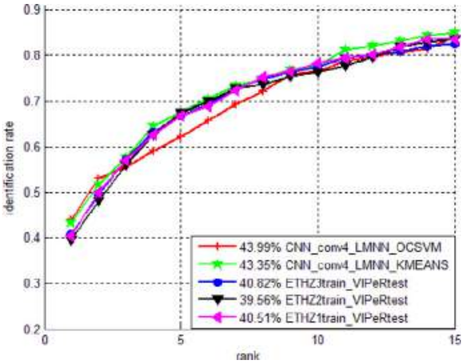


Fig. 10. Intra-dataset and cross-dataset comparison.

Fig. 9 shows the comparison. We can see that OCSVM obtains the top results at Rank1 and Rank2, while starting from Rank3, OCSVM loses the accuracy and is worse than Kmeans and KNN, and Kmeans becomes the best. In practice, Rank5~ Rank15 are also important indicators. Therefore, among above three algorithms, Kmeans method holds the highest application value, due to simple learning and low computation complexity. Note that k is set to half size of the training set when KNN is applied.

4.3. Cross-dataset experiments

4.3.1. Intra and cross dataset comparison

Most of re-id methods can achieve promising performance on intra-dataset experiments, while they have to face a difficulty that the performance falls off obviously on the cross-dataset test. Moreover, a large-scale training set is often required to guarantee a relatively stable result. However, in practice, the pedestrians are captured from different scenes with varied data distribution, thus it is impractical to train a model for each camera. Hence, how to decrease the adjustment of parameters and the dependency for data scale while retain the performance is a crucial problem. In the following experiments, ETHZ dataset is used to train the salience scores, and the VIPeR dataset is applied to test.

In Fig. 10, the first two curves (CNN_conv4_LMNN_OSCVM and CNN_conv_ 4_ LMNN_Kmeans) denote two results on an intra-dataset (VIPeR). ETHZ*train_ VIPeRtest means training on three subsets of ETHZ dataset and testing on the VIPeR dataset. According to this figure, we can observe that cross-dataset results are slightly lower than the intra-dataset test, illustrating a good generalization ability of our method. Note that three subsets of ETHZ dataset are trained individually, and for each person only one image is used, thus the scale of cross-dataset training is very small (only 83, 35 and 28 images respectively), which indicates a advantage of our method: a small-scale training can meet the requirement of the test.

Table 2
The results of three peer-layer fusion schemes on the VIPeR dataset.

Rank	1	2	3	4	5
Cascade	0.22	0.30	0.35	0.39	0.41
Average	0.26	0.35	0.39	0.43	0.48
Sum	0.13	0.18	0.23	0.29	0.32
Rank	10	15	20	30	40
Cascade	0.55	0.62	0.66	0.74	0.78
Average	0.59	0.66	0.70	0.77	0.81
Sum	0.43	0.51	0.56	0.62	0.67

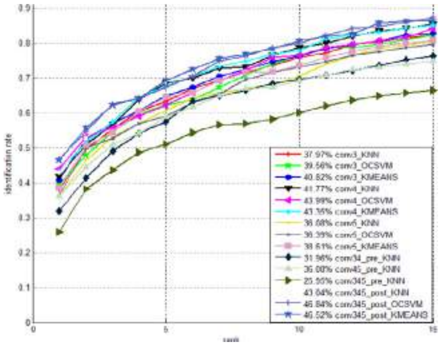


Fig. 11. Cross-layer fusion on VIPeR.

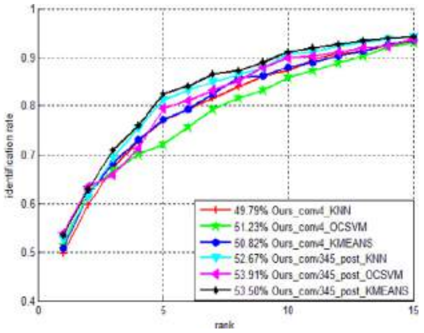


Fig. 12. Cross-layer fusion on CUHK01.

4.3.2. *Peer-layer fusion of CNN feature*

Inspired by [29], we cascade neighboring 3 × 3 conv feature vectors at the same conv layer to form a new local feature vector. Afterwards, the feature vector is reduced to 100 dimensions to test. Table 2 gives the results. We find that three peer-layer fusion schemes (cascade, average and sum) cannot play positive roles. Because conv layer has a big receptive field, and above fusion methods will further extend the range of receptive field, which destroys the context of local information, thus cannot handle problems such as the pose and viewpoint variations.

4.4. *Cross-layer fusion of CNN feature*

Cross-layer fusion can be classified into the early fusion and late fusion. In early fusion, we cascade three feature vectors from different conv layers and then apply L_2-normalization. The late fusion is to combine three single-layer results by a linear weighting. Figs. 11 and 12 show the results of single-layer feature, early fusion and late fusion on the VIPeR and CUHK01 datasets respectively. The results indicate: 1) conv4 layer is better than conv3 and conv5 layers. 2) late fusion obtains the best performance, where the empirical weights corresponding to conv3, conv4 and conv5 for KNN is (0.08, 0.6, 0.32), OCSVM is (0.45,0.46, 0.09) and Kmeans is (0.16,0.74,0.1). 3) Kmeans-based salience learning method still leads in the overall advantage.

Furthermore, we compare our method with 18 state-of-the-art approaches: 1) 12 conventional and distance learning methods such as MtMCML [33], SDALF [34], eBiCov [4], eSDC [11], PRDC [35], aPRDC [36], PCCA [37], KISSME [5], SalMatch [6], LMLF [17], mFilter+LADF [17] and Sakrapee [38]. 2) 6 deep learning based approaches like DML [24], CDML [23], Improved DML [39], DeepFeature [40], MTCP [41] and LSSL [42].

Table 3
The comparison with the state-of-the-art methods on the VIPeR dataset.

Method	Rank1	Rank5	Rank10	Rank15
MtMCML [33]	28.8	59.3	75.8	83.4
SDALF [34]	19.9	38.4	49.4	58.5
eBiCov [4]	20.7	42.0	56.2	63.3
eSDC [11]	26.3	46.4	58.6	66.6
PRDC [35]	15.7	38.4	53.9	63.3
aPRDC [36]	16.1	37.7	51.0	59.5
PCCA [37]	19.3	48.9	64.9	73.9
KISSME [5]	19.6	48.0	64.9	70.9
SalMatch [6]	30.2	52.3	66.0	73.4
LMLF [17]	29.1	52.3	66.0	73.9
mFilter+LADF [17]	43.4	–	–	–
Sakrapee [38]	45.9	–	–	–
DML [24]	28.2	59.3	73.5	81.2
CDML [23]	40.9	–	–	–
Improved DML [39]	34.4	62.2	75.9	82.6
Deepfeature [40]	40.5	60.8	70.4	78.3
MTCP [41]	47.8	–	–	–
LSSL [42]	47.8	–	–	–
Ours_Kmeans	46.5	69.3	80.7	86.5

Table 4
The comparison with the state-of-the-art methods on the CUHK01 dataset.

Method	Rank1	Rank5	Rank10	Rank15	Rank20
mFilter [17]	34.3	55.0	65.3	70.5	–
SalMatch [11]	28.5	46.3	57.2	64.1	–
PalMatch [11]	20.4	34.1	41.0	47.3	–
TransferM [43]	20.0	44.1	57.1	64.3	–
ITML [44]	16.0	28.5	45.3	53.5	–
LMNN [45]	13.5	31.2	41.8	48.5	–
eSDC [11]	19.7	33.1	40.5	46.8	–
Sakrapee [38]	53.4	76.4	84.4	–	90.5
Deepreid [22]	27.9	–	–	–	–
Deepmodel [2]	47.5	–	–	–	–
M3T [41]	46.0	67.7	78.7	85.3	88.7
M3TC [41]]	49.3	76.5	86.6	93.7	94.7
M3TP [41]	52.3	82.1	90.3	94.0	95.6
M3TCP [41]	53.7	84.3	91.0	93.3	96.3
Ours_Kmeans	53.5	82.5	91.2	94.3	96.1

Table 3 show the comparison on the VIPeR dataset. We can see that our method outperforms all methods based on handcrafed features and most of deep learning based ones, and achieves comparable results with two most recent CNN-based approaches [41] and [42].

We also compare existing approaches on the CUHK01 dataset, and we use the same parameter setting on the VIPeR dataset so as to demostrate the adaptability. The representative methods include mFilter [17], SalMatch [11] and PatMatch [11], TransferM [43], ITML [44], LMNN [45], eSDC [11], Sakrapee [38], and deep learning based methods: Deepreid [22], Deepmodel [2], M3T [41], M3TC [41], M3TP [41] and M3TCP [41]. Table 4 gives the comparison.

The experimental results show that under the condition of not adjusting the parameters (directly use the configuration on the VIPeR dataset), our algorithm still maintain a higher accuracy on the CUHK01 dataset, which indicates a good generalization capability.

4.5. *Discussion*

From CMC, Rank1 of our algorithm on VIPeR and CUHK01 datasets gets 46.5% and 53.5% respectively, and outperforms typically conventional methods and achieves comparable results with recent CNN-based models. We have several observations: 1) Compared with hand-crafted features, CNN feature can obtain a better image repersentation. 2) Among three conv layer features from VGG-M model, conv4 can get the best result, since conv3 feature is not abstract while conv5 is a little too sparse. Moreover, cross-layer late fusion is the optimal fusion scheme. 3) Compared with KNN and OCSVM, the proposed Kmeans-based salience learning can hold an overall advantage, meanwhile, it has a lower computation cost. 4) Cross-dataset experiments show under the condition of not adjusting the parameters, our algorithm still maintains a high matching rate, which demonstrates the effectiveness.

It is noted that in this paper, we do not focus on the design of CNN architecture and the training of the deep model, and we propose an open framework, in which the performance could be further improved if more effective CNN model is introduced. Furthermore, our method does not need a large-scale samples and dozens of patch-pairs can guarantee a high matching rate.

5. Conclusion

In this paper, we propose a cross-dataset person re-id framework via integrating patch-based metric learning and local salience learning. Firstly, CNN features are extracted to represent patches of a person. Secondly, only two positive patch-pairs are chosen and input a LMNN network to learn two patch-based metric matrices for feature projection respectively. Afterwards, a local salience learning algorithm based on Kmeans clustering is proposed to train the weights of patches. Finally, the similarity of image-pair is computed by a weighted summing of patches. Experimental results indicate that the proposed method outperforms conventional approaches and achieves a comparable performance with recent CNN-based methods, which demonstrates the effectiveness of our method.

Acknowledgments

This work is supported by Chinese National Natural Science Foundation(61471049, 61372169, 61532018).

References

[1] S. Chen, C. Guo, J. Lai, Deep ranking for person re-identification via joint representation learning, IEEE Trans. Image Process. 25 (5) (2016) 2353–2367.

[2] E. Ahmed, M.J. Jones, T.K. Marks, An improved deep learning architecture for person re-identification, in: CVPR 2015, 2015, pp. 3908–3916.

[3] R.R. Varior, G. Wang, J. Lu, T. Liu, Learning invariant color features for person reidentification, IEEE Trans. Image Process. 25 (7) (2016) 3395–3410.

[4] B. Ma, Y. Su, F. Jurie, Bicov: a novel image representation for person re-identification and face verification, in: British Machive Vision Conference, 2012. 11—pages

[5] M. Köstinger, M. Hirzer, P. Wohlhart, P.M. Roth, H. Bischof, Large scale metric learning from equivalence constraints, in: Computer Vision and Pattern Recognition, 2012, pp. 2288–2295.

[6] R. Zhao, W. Ouyang, X. Wang, Person re-identification by salience matching, in: IEEE International Conference on Computer Vision, 2013, pp. 2528–2535.

[7] S. Pedagadi, J. Orwell, S. Velastin, B. Boghossian, Local fisher discriminant analysis for pedestrian re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3318–3325.

[8] S. Khamis, C. Kuo, V.K. Singh, V.D. Shet, L.S. Davis, Joint learning for attribute-consistent person re-identification, in: Computer Vision - ECCV 2014 Workshops, 2014, pp. 134–146.

[9] W. Li, X. Wang, Locally aligned feature transforms across views, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3594–3601.

[10] J. Lu, V.E. Liong, X. Zhou, J. Zhou, Learning compact binary face descriptor for face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 37 (2015) 2041–2056.

[11] R. Zhao, W. Ouyang, X. Wang, Unsupervised salience learning for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3586–3593.

[12] L. Zheng, S. Wang, L. Tian, F. He, Z. Liu, Q. Tian, Query-adaptive late fusion for image search and person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1741–1750.

[13] J. Lu, G. Wang, P. Moulin, Localized multifeature metric learning for image-set-based face recognition, IEEE Trans. Circuits Syst. Video Techn. 26 (2016) 529–540.

[14] F. Xiong, M. Gou, O.I. Camps, M. Sznaier, Person re-identification using kernel-based metric learning methods, in: Computer Vision - ECCV 2014, 2014, pp. 1–16.

[15] K.Q. Weinberger, L.K. Saul, Distance metric learning for large margin nearest neighbor classification, J. Mach. Learn. Res. 10 (2009) 207–244.

[16] Z. Li, S. Chang, F. Liang, T.S. Huang, L. Cao, J.R. Smith, Learning locally-adaptive decision functions for person verification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3610–3617.

[17] R. Zhao, W. Ouyang, X. Wang, Learning mid-level filters for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 144–151.

[18] V.E. Liong, J. Lu, Y. Ge, Regularized local metric learning for person re-identification, Pattern Recognit. Lett. 68 (2015) 288–296.

[19] T. Wang, S. Gong, X. Zhu, S. Wang, Person re-identification by video ranking, in: Computer Vision - ECCV 2014, 2014, pp. 688–703.

[20] D. Gray, S. Brennan, H. Tao, Evaluating appearance models for recognition, reacquisition, and tracking, IEEE International Workshop on Performance Evaluation for Tracking and Surveillance, 3, 2007.

[21] T. Xiao, H. Li, W. Ouyang, X. Wang, Learning deep feature representations with domain guided dropout for person re-identification, CoRR (2016).

[22] W. Li, R. Zhao, T. Xiao, X. Wang, Deepreid: deep filter pairing neural network for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 152–159.

[23] H. Shi, X. Zhu, S. Liao, Z. Lei, Y. Yang, S.Z. Li, Constrained deep metric learning for person re-identification, CoRR (2015).

[24] D. Yi, Z. Lei, S. Liao, S.Z. Li, Deep metric learning for person re-identification, in: IEEE International Conference on Pattern Recognition, 2014, pp. 34–39.

[25] A.J. Ma, P.C. Yuen, J. Li, Domain transfer support vector ranking for person re-identification without target camera label information, in: IEEE International Conference on Computer Vision, 2013, pp. 3567–3574.

[26] J. Hu, J. Lu, Y. Tan, J. Zhou, Deep transfer metric learning, IEEE Trans. Image Process. (2016) 5576–5588.

[27] K. Chatfield, K. Simonyan, A. Vedaldi, A. Zisserman, Return of the devil in the details: delving deep into convolutional nets, arXiv preprint arXiv:1405.3531 (2014).

[28] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: a benchmark, in: IEEE International Conference on Computer Vision, 2015, pp. 1116–1124.

[29] L. Liu, C. Shen, A. van den Hengel, The treasure beneath convolutional layers: cross-convolutional-layer pooling for image classification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4749–4757.

[30] B. Hariharan, P. Arbeláez, R. Girshick, J. Malik, Hypercolumns for object segmentation and fine-grained localization, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 447–456.

[31] A. Ess, B. Leibe, L.J.V. Gool, Depth and appearance for mobile scene analysis, in: International Conference on Computer Vision, 2007, pp. 1–8.

[32] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, Caffe: convolutional architecture for fast feature embedding, in: Proceedings of Multimedia, 2014, pp. 675–678.

[33] L. Ma, X. Yang, D. Tao, Person re-identification over camera networks using multi-task distance metric learning, IEEE Trans. Image Process. (2014) 3656–3670.

[34] M. Farenzena, L. Bazzani, A. Perina, V. Murino, M. Cristani, Person re-identification by symmetry-driven accumulation of local features, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 2360–2367.

[35] W.-S. Zheng, S. Gong, T. Xiang, Person re-identification by probabilistic relative distance comparison, in: Computer vision and pattern recognition, 2011, pp. 649–656.

[36] C. Liu, S. Gong, C.C. Loy, X. Lin, Person re-identification: what features are important? in: European Conference on Computer Vision, 2012, pp. 391–401.

[37] A. Mignon, F. Jurie, Pcca: a new approach for distance learning from sparse pairwise constraints, in: Computer Vision and Pattern Recognition, 2012, pp. 2666–2672.

[38] S. Paisitkriangkrai, C. Shen, A. van den Hengel, Learning to rank in person re-identification with metric ensembles, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1846–1855.

[39] D. Yi, Z. Lei, S.Z. Li, Deep metric learning for practical person re-identification, CoRR (2014).

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

[41] D. Cheng, Y. Gong, S. Zhou, J. Wang, N. Zheng, Person re-identification by multi-channel parts-based cnn with improved triplet loss function, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1335–1344.

[42] Y. Yang, S. Liao, Z. Lei, S.Z. Li, Large scale similarity learning using similar pairs for person verification, in: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, 2016, pp. 3655–3661.

[43] W. Li, R. Zhao, X. Wang, Human reidentification with transferred metric learning, in: 11th Asian Conference on ComputerVision, 2012, pp. 31–44.

[44] J.V. Davis, B. Kulis, P. Jain, S. Sra, I.S. Dhillon, Information-theoretic metric learning, in: IEEE proceedings of the Machine Learning, 2007, pp. 209–216.

[45] A.J. Ma, P. Li, Query based adaptive re-ranking for person re-identification, in: Asian Conference on Computer Vision, 2014, pp. 397–412.

Zhicheng Zhao is an associate professor of Beijing University of Posts and Telecommunications. He was a visiting scholar at School of Computer Science, Carnegie Mellon University from 2015 to 2016. His research interests are computer vision, image and video semantic understanding and retrieval. He has authored and coauthored more than 50 journal and conference papers.

Binlin Zhao is a master student of School of Information and Telecommunication, Beijing University of Posts and Telecommunications. Her research interests include computer vision, and image processing.

Fei Su is a female professor in School of Information and Telecommunication, Beijing University of Posts and Telecommunications. She received the Ph.D. degree majoring in Communication and Electrical Systems from BUPT in 2000. She was a visiting scholar at electrical computer engineering department, Carnegie Mellon University from 2008 to 2009. Her current interests include pattern recognition, image and video processing and biometrics. She has authored and coauthored more than 70 journal and conference papers and some textbooks.