Machine Vision and Applications (2017) 28:725–736
DOI 10.1007/s00138-017-0843-5

SPECIAL ISSUE PAPER

A Siamese inception architecture network for person re-identification

Shuangqun Li[1] · Huadong Ma[1]

Received: 31 October 2016 / Revised: 3 March 2017 / Accepted: 25 April 2017 / Published online: 16 May 2017

Abstract Person re-identification is an extremely challenging problem as person's appearance often undergoes dramatic changes due to the large variations of viewpoints, illuminations, poses, image resolutions, and cluttered backgrounds. How to extract discriminative features is one of the most critical ways to address these challenges. In this paper, we mainly focus on learning high-level features and combine the low-level, mid-level, and high-level features together to re-identify a person across different cameras. Firstly, we design a Siamese inception architecture network to automatically learn effective semantic features for person re-identification in different camera views. Furthermore, we combine multi-level features in null space with the null Foley–Sammon transform metric learning approach. In this null space, images of the same person are projected to a single point, which minimizes the intra-class scatter to the extreme and maximizes the relative inter-class separation simultaneously. Finally, comprehensive evaluations demonstrate that our approach achieves better performance on four person re-identification benchmark datasets, including Market-1501, CUHK03, PRID2011, and VIPeR.

Keywords Person re-identification · Metric learning · Siamese architecture · Null Foley–Sammon transform · Null space

1 Introduction

The purpose of person re-identification is to identify persons across different cameras, or across time in single camera with visual features. Person re-identification has massively important applications in intelligent video surveillance system, such as cross-camera tracking [1] and pedestrian retrieval [2]. Despite the best efforts of researchers, person re-identification is still an extremely challenging problem as a person's appearance often undergoes dramatic changes due to the large variations of viewpoints, illuminations, poses, image resolutions, and cluttered backgrounds. Some examples are shown in Fig. 1. These factors cause pedestrian images in different camera views to have considerable intra-class variations and ambiguous inter-class differences.

As we know, many types of features have been proposed for pedestrian detection, tracking, recognition, and re-identification. Most of them belong to the low-level and mid-level features, such as Local Binary Patterns (LBP) [3], Color Name (CN) [4–6], Histogram of Gradient (HOG) [7], and Local Maximal Occurrence (LOMO) [8]. We discover that these handcrafted features can be extremely hard to represent the mixture of complex transforms. These features are of low discriminative power and not robust to overcome the challenges in person re-identification. Compared with these handcrafted features, the power of deep CNNs in learning discriminative features is very strong in various computer vision tasks [9, 10]. Recently, some deep learning approaches for person re-identification have achieved big progresses [11, 12], which can learn high-level semantic features. In particular, Siamese architecture was firstly introduced for face and signature verification tasks [13]. The network is suited for person re-identification—the number of categories can be very large, with only a few examples per category. Although they have effectively improved the existing re-identification

✉ Huadong Ma
mhd@bupt.edu.cn

Shuangqun Li
shuangqunli@hotmail.com

[1] Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing 100876, China

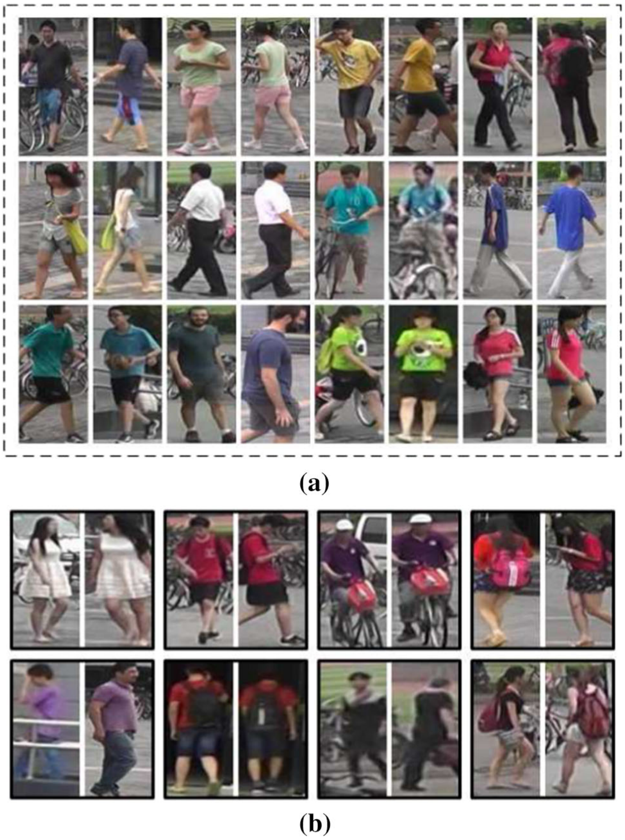


Fig. 1 **a** Samples of pedestrian images in different camera views from the Market-1501 [4] dataset and the two adjacent images are the same identity. **b** Examples of positive samples (*first row*) and negative samples (*second row*) for training our SIAN

performance, optimal feature representation is still the most critical component. We mainly focus on learning high-level features and combine the low-level, mid-level, and high-level features together to re-identify a person across different cameras.

In this paper, we extend the literature [14] to propose a Siamese inception architecture network (SIAN) for person re-identification. First of all, we design a SIAN to learn high-level features for person re-identification, which consists of several BN-Inception [15,16] modules. The network can minimize the distance between the same pedestrian and maximize the distance between different pedestrians simultaneously. Instead of handcrafted features, the deep learning architecture and its fully connected layer can automatically learn semantic features for person re-identification in different camera views. Furthermore, we combine the low-level, mid-level features, and high-level features learned by SIAN in null space by NFST metric learning approach [17] to identify the same pedestrian in different camera views. More importantly, a kernel-based NFST is employed to deal with the non-linearity of the person's appearance, which further boosts the ability of matching in the null space. Therefore, the combination offers a perfect solution for person re-identification. Finally, experimental results on four widely used person re-identification benchmarks, i.e., Market-1501, CUHK03, PRID2011, and VIPeR, demonstrate that the proposed method achieves better performance.

In contrast to the literature [14], we change the SCNN network architecture into SIAN to learn high-level semantic features. The semantic features are of high discriminative power. To reduce overfitting, we mix the Market-1501 and CUHK03 as a joint dataset and train our designed SIAN from scratch on the training set of the joint dataset. Notice that the training identities have no overlap with the test identities. Finally, four widely used benchmark datasets are selected. We further fine-tune the pre-train SIAN on each dataset, respectively, for evaluating the proposed approach.

In summary, the contributions of this paper can be concluded as follows:

- We design a Siamese inception architecture network (SIAN), which can model different camera views and automatically learn high-level semantic features for person re-identification;
- We adopt a kernel-based NFST to combine multi-level features in a new discriminative null space to deal with the non-linearity of the person's appearance;
- Comprehensive evaluations show that our method achieves better performance on four widely used person re-identification benchmarks.

2 Related works

Existing approaches on person re-identification can be roughly divided into three types: feature representation [4–8,18–22], deep learning [11,12,18], and metric learning [3,8,17].

The *feature representation*-based methods try to design robust and discriminative features in different camera views. The general trend is that the dimensions of the proposed features are getting higher. For instance, Dalal et al. [7] designed HOG feature descriptors, which would be discriminated cleanly for human form, even in cluttered backgrounds and difficult illuminations. Yang et al. [5] exploited salient color names to generate a feature representation. Liao et al. [8] proposed LOMO feature representation, which provided a stable representation against viewpoint changes. Peng et al. [21] developed a novel cross-dataset transfer learning approach to learn a discriminative feature representation. Specifically, it was a multitask dictionary learning method which was able to learn a dataset-shared feature representation. However, no matter how robust the obtained features are, they are unlikely to be completely invariant to the large variations of viewpoints, illuminations, body poses, image resolutions, and cluttered backgrounds.

The *deep learning*-based methods can automatically learn high-level semantic features [9–12,18,23]. Gan et al. [9] designed a deep neural network, namely DevNet, that extracted the location of temporal key evidences and provided discriminative spatial regions simultaneously. By using these features, their approach achieved a promising performance in event detection. Yao et al. [10] put forward to a multi-level coarse-to-fine object description which could automatically generate discriminative visual features for visual classification. Li et al. [12] proposed a novel filter pairing neural network (FPNN) to jointly handle misalignment, photometric and geometric transforms, occlusions, and background clutter by maximizing the strength of each component when cooperating with others. Wang et al. [23] designed a joint learning framework to unify the matching of single-image representation (SIR) and the classification of cross-image representation (CIR) using convolutional neural network. Bromley et al. [24] proposed the two-stream Siamese architecture for signature verification. Later, the two-stream network architecture had been applied to face verification [13]. In this work, we design a SIAN to automatically learn effective semantic features for person re-identification, aiming to be invariant to different camera views.

The *metric learning*-based methods mainly focus on learning discriminative distance metrics or subspaces for matching pedestrian in different camera views. For example, Kostinger et al. [3] introduced a simple but effective strategy to learn a distance metric from equivalence constraints, based on a statistical inference perspective. Liao et al. [8] designed the Cross-view Quadratic Discriminant Analysis (XQDA) method, which learned a discriminant low-dimensional subspace. Zhang et al. [17] proposed the NFST approach by matching pedestrian in a discriminative null space of the training data. In this paper, we exploit the NFST metric learning approach to combine HOG, CN, LOMO, and deep semantic features for addressing these challenges in person re-identification. In summary, different from the above methods, our aim is to model different camera views and automatically learn effective semantic features for person re-identification. Furthermore, we fuse multi-level features and then feed them into the most advanced distance metric method [17] to identify the same pedestrian. As a result, our approach achieves better performance on four person re-identification benchmarks.

3 The proposed method

3.1 Overview

Figure 2 shows the architecture of the proposed person re-identification approach. Firstly, we extract the low-level, mid-level, and high-level semantic features from person images. For the high-level semantic features, we design a SIAN, which can model different camera views and automatically learn effective semantic features. Finally, we combine multi-level features and then feed them into the NFST metric learning approach [17] to identify the same pedestrian. Next, we will describe each component in detail.

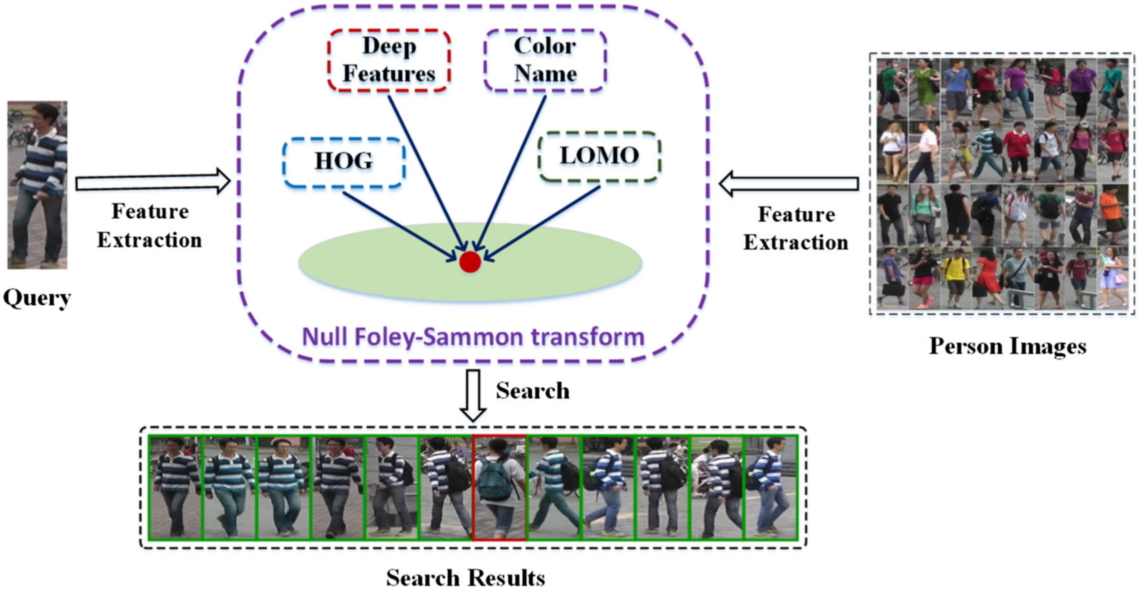


Fig. 2 Overview of our method for person re-identification. The top 12 results are listed. The true positive results are in *green box*; otherwise, *red* (color figure online)

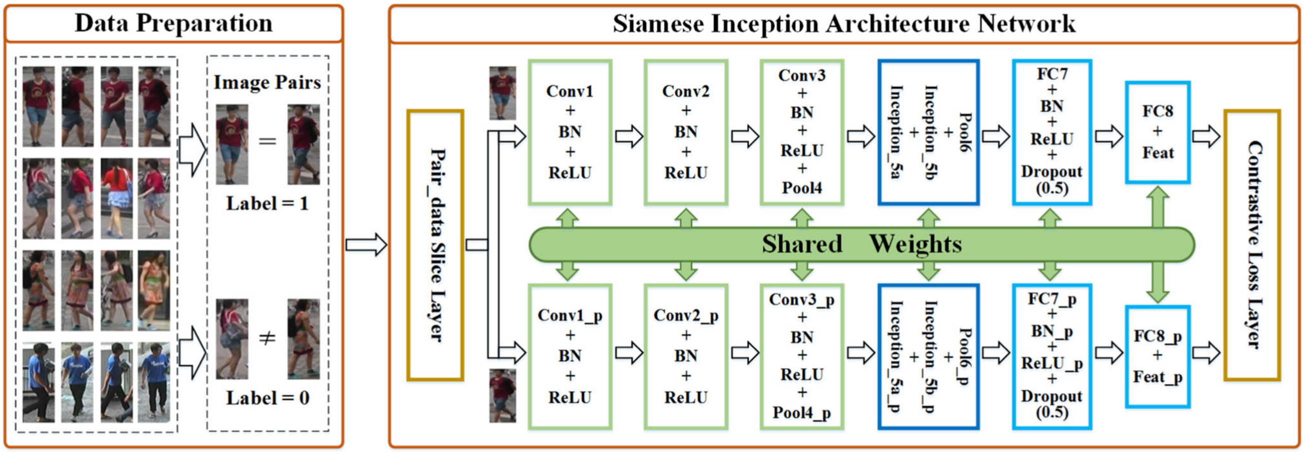

Data Preparation
Image Pairs
=
Label = 1
≠
Label = 0
Siamese Inception Architecture Network
Pair_data Slice Layer
Conv1 + BN + ReLU
Conv2 + BN + ReLU
Conv3 + BN + ReLU + Pool4
Inception_5a + Inception_5b + Pool6
FC7 + BN + ReLU + Dropout (0.5)
FC8 + Feat
Shared Weights
Conv1_p + BN + ReLU
Conv2_p + BN + ReLU
Conv3_p + BN + ReLU + Pool4_p
Inception_5a_p + Inception_5b_p + Pool6_p
FC7_p + BN_p + ReLU_p + Dropout (0.5)
FC8_p + Feat_p
Contrastive Loss Layer

Fig. 3 Structure of our designed SIAN

3.2 Siamese inception architecture network

To learn high-level semantic features robust to all sorts of challenges as described earlier, we need a model that operates on pairs of data. A network architecture that has been successfully applied on pairs of data is the two-stream Siamese architecture [13] which consists of two identical convolutional networks sharing the same set of weights. So, we design a two-stream Siamese network which is called Siamese inception architecture network (SIAN) to learn high-level semantic features. Next, we analyze the different components of the proposed SIAN.

Network architecture The SIAN includes two sub-networks working in "image pairs → label" mode, as shown in Fig. 3. The two sub-networks are parallel in structure and share the same parameters. The SIAN processes two input images separately through individual sub-networks. For individual sub-networks, we design a deeper and wider network by utilizing the inception architecture [16]. In the following, we mainly describe the designed network.

Because inception architecture is designed to perform well even under high computational efficiency and low parameter count, we increase the depth and width of the sub-network by employing the inception architecture to achieve good performance. Batch normalization [15] layers are employed before each ReLU layer, which can further accelerate the convergence process.

Max pooling only selects the strongest activations from local neighborhood as input for the subsequent layers, so it can maintain invariance to local deformations. There are many deformations owing to the variations of camera viewpoints and poses in person images. So, we apply one max pooling at the very early stage, which is important to maintain robustness to local noise. Inspired by [25], we introduce two inception architectures to expand the network's depth and width for capturing progressively high-level semantic features. Detailed structures are listed in Table 1.

Table 1 Details of our designed SIAN for person re-identification

Name	Patch size	Stride	Output size
Pair_data			6 × 128 × 64
Slice_pair			3 × 128 × 64
			3 × 128 × 64
Conv1–Conv3	3 × 3	1	32 × 128 × 64
Pool4(MAX)	2 × 2	2	32 × 64 × 32
Inception_5a			256 × 64 × 32
Inception_5b		2	384 × 32 × 16
Pool6(AVG)	8 × 4	1	384 × 25 × 13
FC7			2048
FC8			M
Feat			2
Conv1_p–Conv3_p	3 × 3	1	32 × 128 × 64
Pool4_p(MAX)	2 × 2	2	32 × 64 × 32
Inception_5a_p			256 × 64 × 32
Inception_5b_p		2	384 × 32 × 16
Pool6_p(AVG)	8 × 4	1	384 × 25 × 13
FC7_p			2048
FC8_p			M
Feat_p			2

Network input Due to the challenges as described earlier in person re-identification, we need plentiful, various and balancing data for training the SIAN. To form the balancing training sample pairs, we randomly select two images belonging to the same identity as positive pairs, whereas we randomly select two images belonging to the different identities as negative pairs. The training sample pairs can be formulated as follows. Suppose we have two images x_1 and x_2, let y be a binary label of the pair, $y = 1$ if the images x_1 and x_2 belong to the same identity and $y = 0$ otherwise. We use all the positive sample pairs and randomly select the same number of negative sample pairs for training the SIAN.

Contrastive loss layer Finally, the two branches in the SIAN are connected with a loss layer. For person re-identification, we want to learn a non-linearly function which maps person images to points in a low-dimensional space. Moreover, it makes positive pairs close enough, whereas negative pairs are far away at least by a margin.

Inspired by [13], we apply the contrastive loss. The outputs $G_W(x_1)$ and $G_W(x_2)$ of the two IANs are the two points in the low-dimensional space generated by mapping x_1 and x_2. W is the shared parameter vector throughout the Siamese architecture which needs to be learned. Then, the parameterized distance function $E_W(x_1, x_2)$ between x_1 and x_2 can be defined as:

$$E_W(x_1, x_2) = ||G_W(x_1) - G_W(x_2)||_2. \tag{1}$$

We can define the contrastive loss function as follows:

$$\mathcal{L}(W) = \frac{1}{2P}\sum_{i=1}^{P} L\left(W, (y, x_1, x_2)^i\right), \tag{2}$$

$$L(W, (y, x_1, x_2)^i) = y \cdot \{E_W(x_1, x_2)^i\}^2 + (1 - y) \cdot \{\max(m - E_W(x_1, x_2)^i, 0)\}^2, \tag{3}$$

where $(y, x_1, x_2)^i$ is the ith pair, which is composed of a pair of images with corresponding label y, P is the number of the training pairs. The positive number m is the minimum distance margin of different identities' images.

Training the SIAN We consider the person re-identification as binary classification. Training data include image pairs and label. In the training stage, the two sub-networks will be optimized simultaneously with the weight sharing mechanism. Pairwise images with similar or dissimilar labels separately entrance the two IANs. Then, the outputs of two IANs are combined by the contrastive loss layer to compute the contrastive loss. After that, the back-propagating with contrastive loss is used to train the model.

Our training algorithm adopts the mini-batch stochastic gradient descent for optimizing the objective function. The training data are divided into mini-batches. Training errors are calculated upon each mini-batch in the contrastive loss layer and backward propagated to the lower layers, and network weights are updated simultaneously.

As positive pairs and negative pairs have different data distribution, they can bring about data imbalance and overfitting. To avoid these problems, we randomly dropout 50% neurons of the FC7 layer in the training process. With more rounds over the training data, the model is trained until it converges.

3.3 Metric learning-based multi-level features combination

In this section, we firstly explore multi-level features for person re-identification. Subsequently, we adopt a kernel-based NFST to fuse multi-level features for better identification performance. The details can be found as follows.

Low-level and mid-level features In real-world practice, low-level and mid-level features such as HOG, CN, LOMO are usually designed to capture the invariant visual information of different views. For HOG feature [7], it is very effective to characterize person's variable appearance and wide range of poses for human detection, even in cluttered backgrounds and difficult illumination. We extract 1680-D HOG descriptors as low-level feature for identifying person.

For BOW-CN feature [4], this is a benchmark method in person re-identification, which applies BOW with CN feature. Person image is normalized to 128 × 64 and is densely sampled using 4 × 4 patch with the 4-pixel sampling step. The codebook of BOW is trained on the training set of each dataset using standard k-means, and codebook size is 350. Moreover, we also adopt the TF-IDF, weak geometric constrains, and background suppression techniques. We obtain 5600-D BOW-CN descriptors as mid-level feature.

For LOMO feature [8], it is an effective feature representation for person re-identification. The Retinex algorithm is firstly applied to preprocess person images for handling illumination variations. By analyzing and maximizing the horizontal occurrence of local features, we can get a stable representation against viewpoint changes, namely LOMO feature. Finally, we extract 26960-D LOMO descriptors as mid-level feature for identifying person.

These features are effective to filter out the dissimilar person according to visual appearance. However, it is difficult to tell the difference between persons with similar visual appearance. These features are of low discriminative power and not enough to tackle the challenges in person re-identification. So, we utilize high-level features which represent the rich semantic information of person for accurate person re-identification.

High-level features For the high-level features learned by our designed SIAN, we send the probe images and gallery images into one of the IANs and then compute the feedforward network to extract effective deep semantic features (high-level features). According to our comprehensive experiments, we find that the FC7 layer gives the best performance, so we choose the features extracted from the FC7 layer to report results for person re-identification.

Table 2 Statistics comparison of the datasets

Datasets	Market-1501 [4]	CUHK03 [12]	PRID2011 [26]	VIPeR [27]
# Identities	1501	1467	385	632
# Train images	10,350	21,012	2997	506
# Val images	2586	5252	749	126
# ID Gallery	751	100	649	316
# ID Probe	751	100	100	316

In real-world practice, low-level and mid-level features such as HOG, CN, LOMO are effective to filter out the dissimilar person according to visual appearance. However, it is difficult to tell the difference between persons with similar visual appearance. High-level semantic features such as highlighted regions, shoulder bag, and backpack can be used to identify the same person from others. Therefore, we complementarily fuse these features by an early fusion scheme.

Distance metric method In order to better combine multi-level features, we adopt the state-of-the-art method NFST [17]. So, let us first briefly revisit it.

The objective of NFST is to learn a discriminative subspace where the training data of each of the C classes are projected to a single point, resulting in C points in the space. Formally, we aim to learn the optimal projection matrix W so that each column, denoted as w, is an optimal discriminant direction that maximizes the Fisher discriminant criterion:

$$\mathcal{J}(w) = \frac{w^T S_b w}{w^T S_w w}, \tag{4}$$

where S_b is the inter-class scatter matrix and S_w is the intra-class scatter matrix. It satisfies the following two conditions, i.e., zero intra-class scatter and positive inter-class scatter:

$$w^T S_w w = 0, \tag{5}$$

$$w^T S_b w > 0, \tag{6}$$

This guarantees the best separability of the training data in the sense of Fisher discriminant criterion.

In this null space, images of the same identity are projected to a single point, thus minimizing the intra-class scatter to the extreme and maximizing the relative inter-class separation simultaneously. This method has a strong discriminative power and has been a most advanced technique for person re-identification until now. Therefore, we exploit the NFST approach to combine multi-level features together to re-identify pedestrian across different cameras for further improvements. More specifically, we concatenate HOG, CN, LOMO, and semantic features in turn and then feed them into the NFST method to identify the same pedestrian in different camera views. In the fusion scheme, we take into account the complementary property between multi-level features such as color, local shape, local maximal occurrence, and deep semantic features. The experimental results show that the overall performance can be improved by using our proposed approach.

4 Experiments

4.1 Datasets and settings

Datasets In our experiments, four widely used datasets are selected, including Market-1501, CUHK03, PRID2011, and VIPeR. Market-1501 and CUHK03 are the largest person re-identification benchmark datasets. For each dataset, we randomly draw roughly 20% of all training images for validation. Notice that both the training and validation identities have no overlap with the test identities. A statistics comparison is shown in Table 2.

Market-1501 [4] contains 32,668 detected person bounding boxes of 1501 identities. Each identity is captured by six cameras at most and two cameras at least. The dataset is randomly divided into training and testing sets, containing 750 and 751 identities, respectively. During testing, for each identity, one query image in each camera is selected. Each identity has multiple images under each camera. We employ the single query and multi-query to evaluate our method on the provided fixed training and test set as in [4].

CUHK03 [12] includes 28,193 images of 1467 identities, which is captured with six surveillance cameras. Each identity is captured in two views and has an average of 4.8 images in each view. Pedestrian images collected from multiple pairs of camera views are all mixed, and they form complex cross-view modality. It provides both manually cropped person images and automatically detected person images by the Deformable Part Model (DPM) detector [28]. We use the manually cropped images in the experiment. The 20 training/test splits provided in [12] are used. Two images are randomly chosen for testing: one is for probe and the other for gallery.

PRID2011 [26] consists of person images recorded from two camera views. Camera view A shows 385 persons, while camera view B shows 749 persons. The first 200 persons

Table 3 Results by combining multi-level features on Market-1501

Features	mAP	HIT@1	HIT@10	HIT@20	HIT@30
HOG	0.84	3.62	10.42	14.49	17.34
HOG + CN	21.31	47.71	77.52	83.97	87.20
HOG + CN + LOMO	33.37	58.61	86.07	90.56	92.90
HOG + CN + LOMO + DF	34.63	60.01	86.79	91.63	93.68
HOG + CN + LOMO + DF + MultiQ	45.20	70.61	92.41	95.29	96.65

appear in both camera views. The single-shot version is used in our experiments. We randomly select 100 ones in the first 200 persons for testing, view A is used as probe, and view B is used as gallery. The remaining persons are used as the training set.

VIPeR [27] contains 632 identities. Each identity has two images captured from different viewpoints. Each image pair was taken from an arbitrary viewpoint under varying illumination conditions. All images are scaled to 128 × 48 pixels. We randomly divide the 632 identities into two equal halves: one for training and the other for testing.

Implementation details Our experiments are based on Caffe [29]. Our training process is divided into two stages: (1) We mix the Market-1501 and CUHK03 as a joint dataset and train our designed SIAN from scratch on the joint dataset. We use the "step" learning rate policy. We initialize the learning rate to 0.05 and gamma to 0.04. Momentum and weight decay are set to 0.9 and 0.0005, respectively. With more rounds over the training data, the model is trained until it converges. (2) We further fine-tune all fully connected layers of the pre-trained SIAN on each dataset, respectively. A small learning rate of 0.0001 is employed in the process of fine-tuning.

Evaluation protocol For the Market-1501 dataset, there are multiple cross-camera ground truths for each query. Therefore, we adopt mean average precision (mAP) to evaluate the overall performance for all datasets in this paper as in [30,31]. The mAP is computed over all queries as

$$mAP = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j} Precision(R_{jk}), \tag{7}$$

where Q is the query set, m_j is the number of ground truths images in each query, $Precision(R_{jk})$ is the average precision (AP) at the position of returned kth ground truths images.

Besides, we also adopt the Cumulative Matching Characteristic (CMC) curve, HIT@1, HIT@10, HIT@20, and HIT@30 which are widely used to evaluate the performance of person re-identification methods.

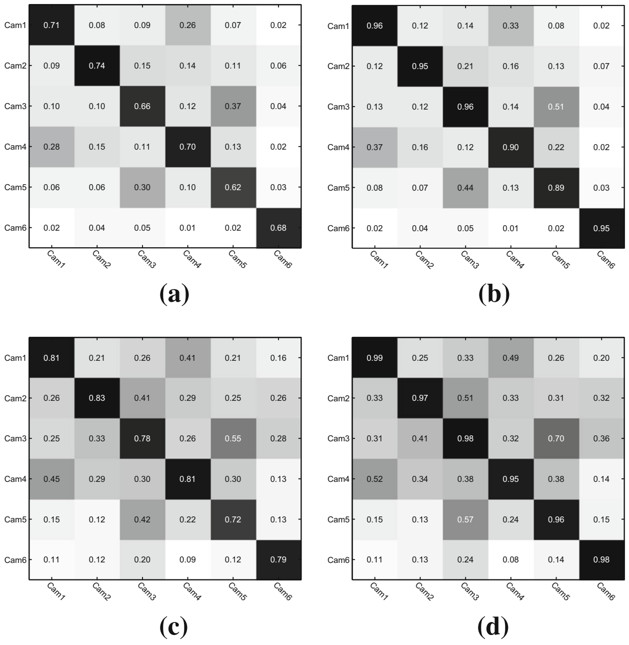
4.2 Evaluations on Market-1501

Combination of multi-level features Market-1501 is one of the largest person re-identification benchmark datasets. We show experimental results by using HOG, CN, LOMO, our learned SIAN Deep Features (DF), and multiple queries (MultiQ) on Market-1501 in Table 3.

Firstly, HOG feature produces a relatively low accuracy: HIT@1 accuracy = 3.62% on Market-1501. Secondly, when integrating CN, LOMO, and DF, we observe consistent improvement in accuracy. For example, mAP increases from 0.84 to 34.63% (+33.79%), and a larger improvement can be seen from HIT@1 accuracy, from 3.62 to 60.01% (+56.39%). It is clear that multi-level features combination is a very efficient way to improve the identification accuracy. Finally, we test multiple queries, where each query identity has multiple query images in a single camera. In the baseline scenario [4], multi-query by max pooling is slightly superior to average pooling because max pooling gives more weights to the rare but salient features. However, from our experimental results, multi-query by average pooling is greatly superior to max pooling in our approach, because average pooling takes into account the complementary property of multi-level features and further improves recognition accuracy.

Results between camera pairs To comprehensively evaluate our method, we provide comparisons between our approach and baseline in terms of the re-identification results between all camera pairs, as shown in Fig. 4. We can see that our approach is considerably superior to baseline in mAP and HIT@1 accuracy. For different camera pairs, the cross-camera average mAP and HIT@1 accuracy of baseline are calculated to be 10.68 and 13.72%, while the cross-camera average mAP and HIT@1 accuracy of our method are 25.06 and 30.42%. It is clear that our method outperforms the baseline by a large margin between different camera pairs. This demonstrates that the combination of multi-level features can overcome the challenges of different camera views in a certain degree.

Comparison with state-of-the-art methods We compare our approach with baseline [4], NFST [17], XQDA [8], KISSME [3], ITML [32], and LMNN [33]. When using baseline and metric learning methods, CN features and LOMO



	Cam1	Cam2	Cam3	Cam4	Cam5	Cam6
Cam1	0.71	0.08	0.09	0.26	0.07	0.02
Cam2	0.09	0.74	0.15	0.14	0.11	0.06
Cam3	0.10	0.10	0.66	0.12	0.37	0.04
Cam4	0.28	0.15	0.11	0.70	0.13	0.02
Cam5	0.06	0.06	0.30	0.10	0.62	0.03
Cam6	0.02	0.04	0.05	0.01	0.02	0.68

(a)

	Cam1	Cam2	Cam3	Cam4	Cam5	Cam6
Cam1	0.96	0.12	0.14	0.33	0.08	0.02
Cam2	0.12	0.95	0.21	0.16	0.13	0.07
Cam3	0.13	0.12	0.96	0.14	0.51	0.04
Cam4	0.37	0.16	0.12	0.90	0.22	0.02
Cam5	0.08	0.07	0.44	0.13	0.89	0.03
Cam6	0.02	0.04	0.05	0.01	0.02	0.95

(b)

	Cam1	Cam2	Cam3	Cam4	Cam5	Cam6
Cam1	0.81	0.21	0.26	0.41	0.21	0.16
Cam2	0.26	0.83	0.41	0.29	0.25	0.26
Cam3	0.25	0.33	0.78	0.26	0.55	0.28
Cam4	0.45	0.29	0.30	0.81	0.30	0.13
Cam5	0.15	0.12	0.42	0.22	0.72	0.13
Cam6	0.11	0.12	0.20	0.09	0.12	0.79

(c)

	Cam1	Cam2	Cam3	Cam4	Cam5	Cam6
Cam1	0.99	0.25	0.33	0.49	0.26	0.20
Cam2	0.33	0.97	0.51	0.33	0.31	0.32
Cam3	0.31	0.41	0.98	0.32	0.70	0.36
Cam4	0.52	0.34	0.38	0.95	0.38	0.14
Cam5	0.15	0.13	0.57	0.24	0.96	0.15
Cam6	0.11	0.13	0.24	0.08	0.14	0.98

(d)

Fig. 4 Re-identification performance between camera pairs on Market-1501: **a** mAP and **b** HIT@1 accuracy of baseline [4], and **c** mAP and **d** HIT@1 accuracy of our method. Cameras on the *vertical* and *horizontal axis* are probe and gallery, respectively

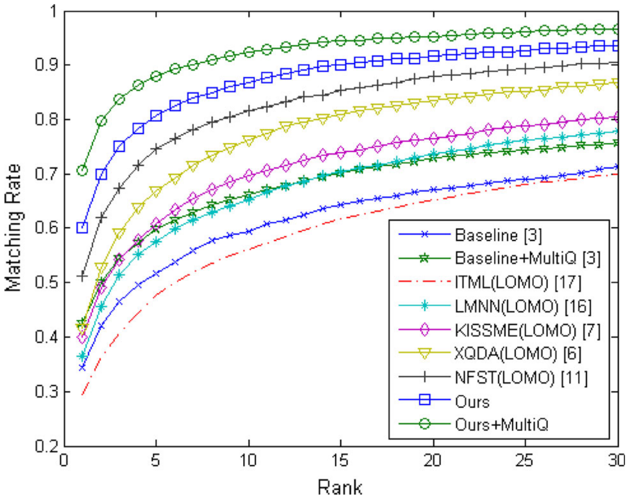


Fig. 5 Comparison with the state-of-the-art methods on Market-1501

features are adopted, respectively. Figure 5 plots the CMC curves of all these methods on the Market-1501. It can be seen that our approach outperforms all these methods with large margins. For our method with single query and multiple query, the HIT@1 accuracy is 60.01% and 70.61%, respectively. In this sense, it is practicable to adopt multi-level features combination for improving person re-identification accuracy.

Figure 6 shows several sample results of our method on Market-1501 dataset. As you can see clearly, there are some challenging problems in these samples such as the large variations of viewpoints, illuminations, poses, and cluttered backgrounds, but our method can still give a much higher identification accuracy, which are shown in the first six rows of Fig. 6. In particular, even if there are dramatic color variances caused by the changes in illuminations in the second row and the fifth row of Fig. 6, our method can still work well. Besides, if the appearances of samples are very similar as shown in the last two rows of Fig. 6, our method does not work well. In view of these special cases, our approach still has room for future improvements.

4.3 Evaluations on CUHK03

Combination of multi-level features CUHK03 is another of the largest person re-identification benchmark dataset, and the labeled version with manually cropped person images is used for experiments. We conduct experiments to validate the proposed approach by using HOG, CN, LOMO, and DF. Experimental results show that a larger improvement can be found in mAP and accuracy by gradually integrating CN, LOMO, and DF. For example, mAP increases from 4.83% to 55.26% (+50.43%), and HIT@1 accuracy increases from 4.85% to 62.03% (+57.18%), as shown in Table 4.

Comparison with state-of-the-art methods We compare the proposed method with BoW + Geo + Gauss [4], NFST [17], XQDA [8], KISSME [3], ITML [32], and LMNN [33]. When using BoW + Geo + Gauss and metric learning methods, CN features and LOMO features are used, respectively. Figure 7 plots the CMC curves of all these methods on CUHK03. It can be seen that our approach outperforms all these methods. It is also worth mentioning that BoW+Geo+Gauss by employing the CN feature does not work well owing to the poor color information of person images on CUHK03. It only achieves 14.35% in HIT@1 accuracy.

4.4 Evaluations on PRID2011 & VIPeR

Combination of multi-level features PRID2011 and VIPeR are all small-scale person re-identification benchmark datasets. We conduct experiments for evaluating the proposed approach by using HOG, CN, LOMO, and DF. As shown in Table 5, when CN, LOMO, and DF are integrated, a consistent improvement in accuracy is observed on PRID2011. For example, mAP increases from 5.34 to 32.40% (+27.06%), and a larger improvement can be seen from HIT@1 accuracy, from 3.00 to 28.00% (+25.00%). As for VIPeR, Experimental results show that a larger improvement can be found in mAP and accuracy by gradually integrating CN, LOMO, and DF. For example, mAP increases from 6.51 to 47.76% (+41.25%), and HIT@1 accu-

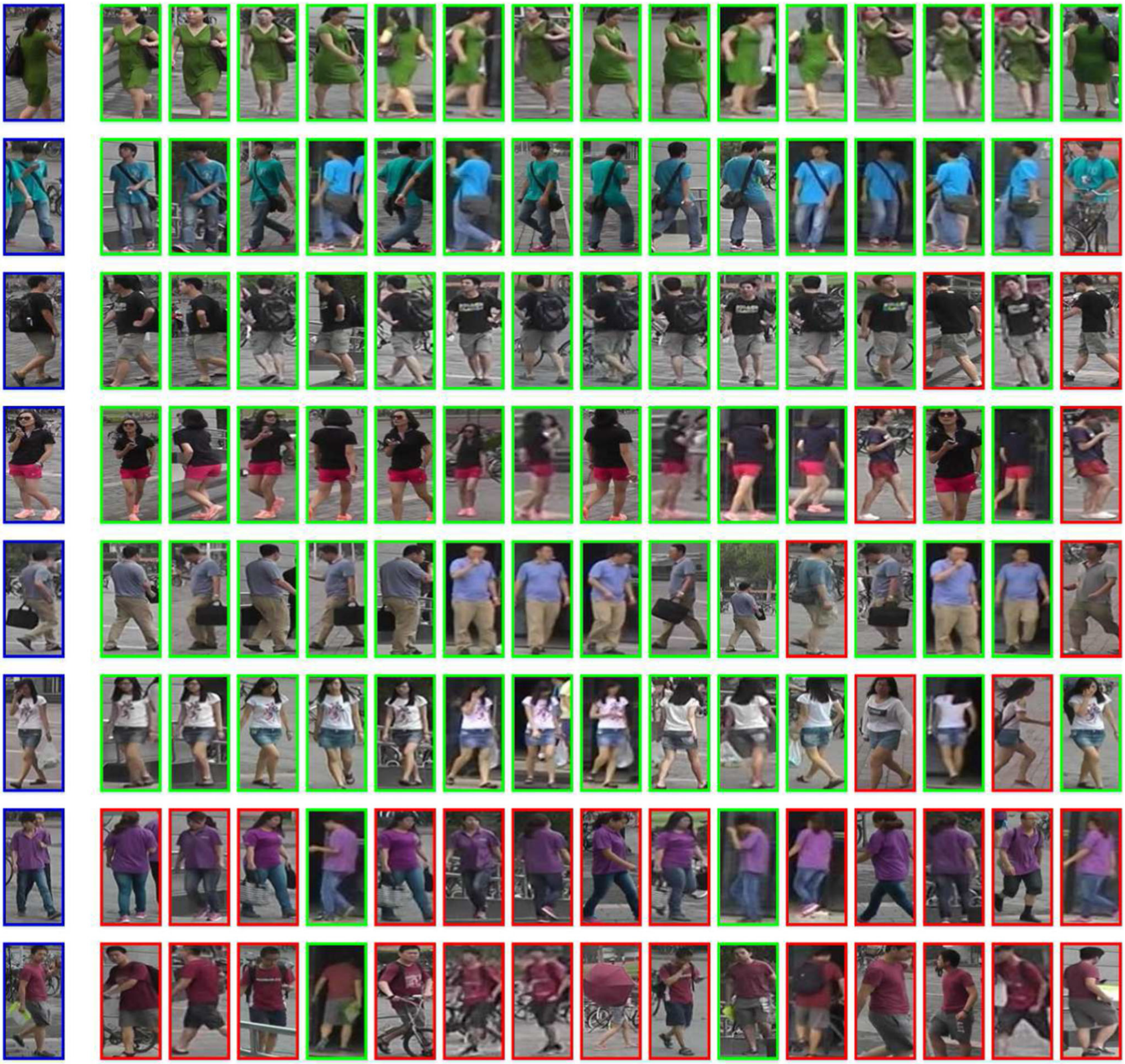


Fig. 6 Sample results of our method on Market-1501 dataset. The query is in *blue box*, and the top 15 results are listed. The true positive results are in *green box*; otherwise, *red* (color figure online)

Table 4 Results by combining multi-level features on CUHK03

Features	mAP	HIT@1	HIT@10	HIT@20	HIT@30
HOG	4.83	4.85	27.64	36.08	46.41
HOG + CN	32.52	43.04	78.69	86.71	89.45
HOG + CN + LOMO	53.51	59.70	88.61	94.51	96.84
HOG + CN + LOMO + DF	55.26	62.03	91.35	95.78	96.84

racy increases from 3.80 to 41.14% (+37.34%), as shown in Table 6.

Comparison with state-of-the-art methods We compare our method with BoW + Geo + Gauss [4], NFST [17], XQDA [8], KISSME [3], ITML [32], and LMNN [33]. When using BoW + Geo + Gauss and metric learning methods, CN features and LOMO features are adopted, respectively. Figures 8 and 9 plot the CMC curves of all these methods on

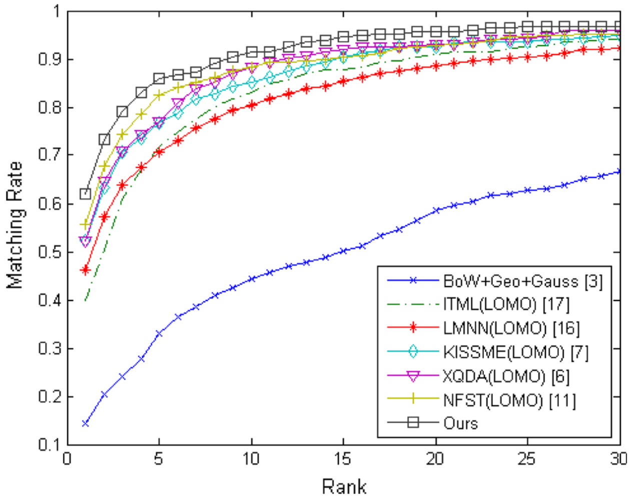


Fig. 7 Comparison with the state-of-the-art methods on CUHK03

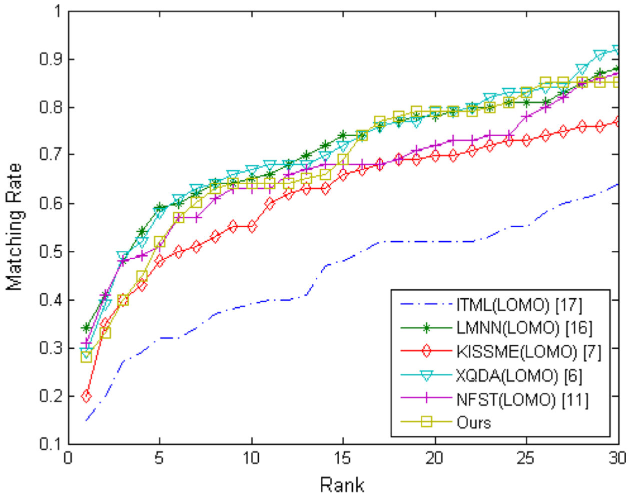


Fig. 8 Comparison with the state-of-the-art methods on PRID2011

PRID2011 and VIPeR. Because the scales of PRID2011 and VIPeR are all very small, our approach only outperforms all these methods by a small margin on VIPeR, and XQDA by using the LOMO feature is slightly superior to our method on PRID2011. As before, BoW + Geo + Gauss by employing the CN feature gives a bad result owing to the poor color information of person images on VIPeR. It only achieves 16.14% in HIT@1 accuracy.

5 Conclusion

In this paper, we have comprehensively analyzed the properties of multi-level features. We mainly focus on learning high-level features and combine multi-level features together to re-identify a person across different cameras. Firstly,

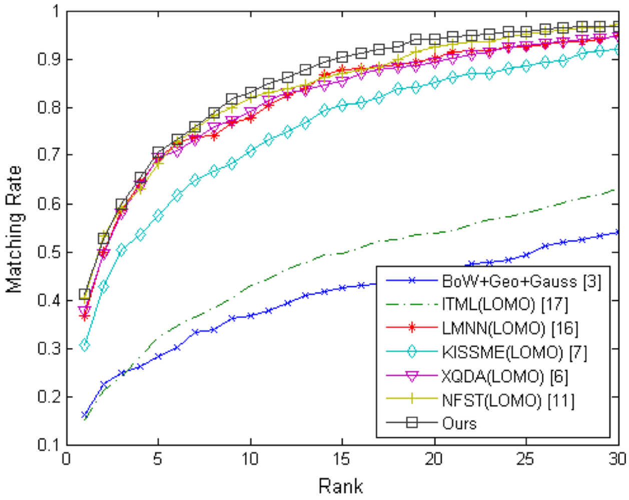


Fig. 9 Comparison with the state-of-the-art methods on VIPeR

Table 5 Results by combining multi-level features on PRID2011

Features	mAP	HIT@1	HIT@10	HIT@20	HIT@30
HOG	5.34	3.00	11.00	27.00	37.00
HOG + CN	10.27	7.00	29.00	39.00	51.00
HOG + CN + LOMO	25.84	20.00	59.00	74.00	80.00
HOG + CN + LOMO + DF	32.40	28.00	64.00	79.00	85.00

Table 6 Results by combining multi-level features on VIPeR

Features	mAP	HIT@1	HIT@10	HIT@20	HIT@30
HOG	6.51	3.80	17.41	27.22	37.34
HOG + CN	27.62	21.20	65.19	81.96	88.61
HOG + CN + LOMO	48.01	41.46	83.23	94.30	97.15
HOG + CN + LOMO + DF	47.76	41.14	82.91	93.99	96.84

we design a SIAN to automatically learn semantic features for person re-identification in different camera views. Furthermore, multi-level features are combined into the discriminative null space with the kernel-based NFST method. Finally, experimental results on four widely used benchmarks demonstrate that the proposed method achieves the best performance.

Although our approach obtains a promising performance in person re-identification, it does not work well for person with very similar appearance. Nevertheless, as biometric information, gait has the unique capability to identify a person from others. In the future work, we will further combine gait feature with appearance for better recognition performance.

Acknowledgements This work is supported by the National Key Research and Development Plan (2016YFC0801005), the Funds for Creative Research Groups of China (61421061), the NSFC-Guangdong Joint Fund (U1501254), the Beijing Training Project for the Leading Talents in S&T (ljrc 201502), the Cosponsored Project of Beijing Committee of Education.

References

1. Wang, X.: Intelligent multi-camera video surveillance: a review. Pattern Recognit. Lett. **34**(1), 3–19 (2013)
2. Loy, C.C., Xiang, T., Gong, S.: Multi-camera activity correlation analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1988–1995 (2009)
3. Köstinger, M., Hirzer, M., Wohlhart, P., Roth, P.M., Bischof, H.: Large scale metric learning from equivalence constraints. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2288–2295 (2012)
4. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: a benchmark. In: Proceedings of the International Conference on Computer Vision, pp. 1116–1124 (2015)
5. Yang, Y., Yang, J., Yan, J., Liao, S., Yi, D., Li, S.Z.: Salient color names for person re-identification. In: Proceedings of the European Conference on Computer Vision, pp. 536–551 (2014)
6. Liu, X.C., Liu, W., Ma, H.D., Fu, H.Y.: Large-scale vehicle re-identification in urban surveillance videos. In: Proceedings of the International Conference on Multimedia and Expo, pp. 1–6 (2016)
7. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 886–893 (2005)
8. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2197–2206 (2015)
9. Gan, C., Wang, N., Yang, Y., Yeung, D.Y., Hauptmann, A.G.: DevNet: A deep event network for multimedia event detection and evidence recounting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2568–2577 (2015)
10. Yao, H., Zhang, S., Zhang, Y., Li, J., Tian, Q.: Coarse-to-fine description for fine-grained visual categorization. IEEE Trans. Image Process. **25**(10), 4858–4872 (2016)
11. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3908–3916 (2015)
12. Li, W., Zhao, R., Xiao, T., Wang, X.: DeepReID: deep filter pairing neural network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 152–159 (2014)
13. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 539–546 (2005)
14. Li, S.Q., Liu, X.C., Liu, W., Ma, H.D., Zhang, H.T.: A discriminative null space based deep learning approach for person re-identification. In: Proceedings of the IEEE Conference on Cloud Computing and Intelligent Systems, pp. 480–484 (2016)
15. Ioffe, S., Szegedy, C.: Batch Normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of the International Conference on Machine Learning, pp. 448–456 (2015)
16. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)
17. Zhang, L., Xiang, T., Gong, S.: Learning a discriminative null space for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1239–1248 (2016)
18. Zhang, C., Liu, W., Ma, H.D., Fu, H.Y.: Siamese neural network based gait recognition for human identification. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing, pp. 2832–2836 (2016)
19. Liu, W., Zhang, Y., Tang, S., Tang, J., Hong, R., Li, J.: Accurate estimation of human body orientation from RGB-D sensors. IEEE Trans. Cybern. **43**(5), 1442–1452 (2013)
20. Wang, B., Tang, S., Zhao, R., Liu, W., Cen, Y.: Pedestrian detection based on region proposal fusion. In: Proceedings of the International Workshop on Multimedia Signal Processing, pp. 1–6 (2015)
21. Peng, P., Xiang, T., Wang, Y., Pontil, M., Gong, S., Huang, T., Tian, Y.: Unsupervised cross-dataset transfer learning for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1306–1315 (2016)
22. Liu, W., Mei, T., Zhang, Y.: Instant mobile video search with layered audio-video indexing and progressive transmission. IEEE Trans. Multimed. **16**(8), 2242–2255 (2014)
23. Wang, F., Zuo, W., Lin, L., Zhang, D., Zhang, L.: Joint learning of single-image and cross-image representations for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1288–1296 (2016)
24. Bromley, J., Bentz, J.W., Bottou, L., Guyon, I., LeCun, Y., Moore, C., Säckinger, E., Shah, R.: Signature verification using a "Siamese" time delay neural network. IJPRAI **7**(4), 669–688 (1993)
25. Xiao, T., Li, H., Ouyang, W., Wang, X.: Learning deep feature representations with domain guided dropout for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1249–1258 (2016)
26. Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H.: Person re-identification by descriptive and discriminative classification. In: Proceedings of the Scandinavian Conference on Image Analysis, pp. 91–102 (2011)
27. Doug, G., Shane, B., Hai, T.: Evaluating appearance models for recognition, reacquisition, and tracking. In: Proceedings of the IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (2007)
28. Felzenszwalb, P.F., Girshick, R.B., McAllester, D.A., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1627–1645 (2010)
29. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R.B., Guadarrama, S., Darrell, T.: Caffe: convolutional architecture

for fast feature embedding. In: Proceedings of the ACM International Conference on Multimedia, pp. 675–678 (2014)

30. Liu, W., Mei, T., Zhang, Y., Che, C., Luo, J.: Multi-task deep visual-semantic embedding for video thumbnail selection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3707–3715 (2015)
31. Chu, L., Wang, S., Zhang, Y., Huang, Q.: Robust spatial consistency graph model for partial duplicate image retrieval. IEEE Trans. Multimed. **15**(8), 1982–1996 (2013)
32. Davis, J.V., Kulis, B., Jain, P., Sra, S., Dhillon, I.S.: Information-theoretic metric learning. In: Proceedings of the International Conference on Machine Learning, pp. 209–216 (2007)
33. Weinberger, K.Q., Blitzer, J., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. In: Advances in Neural Information Processing Systems, pp. 1473–1480 (2005)

Shuangqun Li is a PhD candidate at School of Computer Science, Beijing University of Posts and Telecommunications, China. He received his BS and MS degree from the College of Computer and Information Engineering, Henan Normal University, China, in 2001 and 2011, respectively. His research interests include gait recognition, person re-identification, computer vision, and deep learning.

Huadong Ma is a Chang Jiang Scholar professor and director of Beijing Key Lab of Intelligent Telecommunications Software and Multimedia, executive dean of School of Computer Science, Beijing University of Posts and Telecommunications, China. He received his PhD degree in Computer Science from the Institute of Computing Technology, Chinese Academy of Science in 1995, his MS degree in Computer Science from Shenyang Institute of Computing Technology, Chinese Academy of Science in 1990, and his BS degree in Mathematics from Henan Normal University in 1984. He visited UNU/IIST as a research fellow in 1998 and 1999, respectively. From 1999 to 2000, he held a visiting position in the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan. He was a visiting professor at the University of Texas at Arlington from July to September 2004, and a visiting professor at Hong Kong University of Science and Technology from December 2006 to February 2007. His current research focuses on multimedia system and networking, sensor networks and internet of things, and he has published over 200 papers and four books on these fields. He is a senior member of IEEE and serves for Chair of ACM SIGMOBILE CHINA.