TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 04/09 pp145–156
DOI: 10.26599 / TST.2018.9010060
Volume 23, Number 2, April 2018

Improved Bag-of-Words Model for Person Re-identification

Lu Tian and Shengjin Wang*

Abstract: Person re-identification (person re-id) aims to match observations on pedestrians from different cameras. It is a challenging task in real word surveillance systems and draws extensive attention from the community. Most existing methods are based on supervised learning which requires a large number of labeled data. In this paper, we develop a robust unsupervised learning approach for person re-id. We propose an improved Bag-of-Words (iBoW) model to describe and match pedestrians under different camera views. The proposed descriptor does not require any re-id labels, and is robust against pedestrian variations. Experiments show the proposed iBoW descriptor outperforms other unsupervised methods. By combination with efficient metric learning algorithms, we obtained competitive accuracy compared to existing state-of-the-art methods on person re-identification benchmarks, including VIPeR, PRID450S, and Market1501.

Key words: person re-identification; bag-of-words; unsupervised learning; feature fusion

1 Introduction

Person re-identification[1] is an important task in video surveillance systems. The key challenge is the large intra-class appearance variations, usually caused by various human body poses, illumination, and different camera views. Furthermore, the poor quality of surveillance videos makes it difficult to develop robust and efficient features.

Despite the face that supervised learning methods for person re-identification usually give superior performance and recently works based on Convolutional Neural Networks (CNN) have attracted extensive attention, unsupervised hand-crafted descriptors are still appealing for the following reasons: First, annotating IDs for the pedestrian bounding boxes requires a huge amount of human labor, and it is usually prohibitive to train a good model in a practical environment considering if there is a long recording time and lack of annotated data. Second, re-id models based on supervised learning are often camera specific or dataset specific. A model trained on one dataset is usually not transferable or performs poorly on other datasets. This is because it is challenging for a re-id dataset to cover various camera views, various human clothes, and all illumination situations. Therefore, models pre-trained on some public datasets might not succeed in practical environments. Third, unsupervised methods can be regarded as global re-id models, adaptive to various working conditions, and could be integrated with many supervised methods to improve performance.

Many efforts have been made to design effective and robust feature representation in person re-identification, such as, the Ensemble of Local Features (ELF)[2], Symmetry-Driven Accumulation of Local Features (SDALF)[3], gBiCov[4], Local Descriptors encoded by Fisher Vectors (LDFV)[5], and salience match[6]. It remains an open challenge to design unsupervised descriptors to cope with various environment changes.

Being one of the most widely used unsupervised method in many image retrieval systems, the Bag-of-Words (BoW) model and its variants achieve impressive performance and have recently been adapted to person re-identification with competitive results[7]. The BoW

• Lu Tian and Shengjin Wang are with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China. E-mail: wgsgj@tsinghua.edu.cn.
* To whom correspondence should be addressed.
Manuscript received: 2016-12-04; revised: 2017-01-22; accepted: 2017-01-25

pipeline consists of 4 steps: (1) feature extraction, (2) codebook generation, (3) feature quantization and voting, and (4) score calculation and ranking. For each step, much effort has been made to improve performance[8,9]. In feature extraction, effective hand-crafted features, such as Scale-Invariant Feature Transform (SIFT)[10], Histogram of Oriented Gradient (HOG)[11], and Color Names (CN)[12,13], have been proposed. Despite previous efforts, how to optimize every step and fuse different features for person re-identification remains unknown and requires extensive research[14,15].

In this paper, we propose to use superpixels[16] in a basic pixel segment method to replace traditional patch approaches. By combining the results of superpixel partition and an unsupervised foreground extraction method[17], we extracted perceptually meaningful local regions and reduced the background influence as much as possible. Meanwhile, we carefully investigated three fusion methods: word level fusion[18,19], descriptor level fusion[20,21], and score level fusion[22,23] and examined how they influence the final recognition rate in the BoW model. We formulated feature fusion in the BoW model as a product quantization[24] problem. Our method yields competitive accuracy compared with the state-of-the-art results on existing person re-identification datasets including VIPeR[25], PRID450S[26], and Market1501[27]. In summary, our contributions are two-fold: (1) we improve the conventional BoW model using superpixels as the pixel segment method, and investigate and clarify feature fusion methods in the BoW model; and (2) an unsupervised and robust descriptor is proposed, which achieves state-of-the-art results.

The rest of this paper is organized as follows. In Section 2, a brief discussion of work related to person re-identification is provided. In Section 3 we introduce our method. The experimental results are shown and discussed in Section 4. Finally, we draw our conclusions in Section 5.

2 Related Work

Generally speaking, person re-id includes two basic parts: how to represent pedestrians and how to estimate the similarity between them. The first category focuses on discriminative visual descriptor extraction. Gray and Tao[2] introduced AdaBoost to select good features from 8 color channels (RGB, HS, and YCbCr) and 21 texture features as the ELF. Farenzena et al.[3] proposed the SDALF method, where symmetry and asymmetry are both considered to handle viewpoint variations, and attribute-based features are adopted as mid-level representations. Ma et al.[5] proposed aggregating the local descriptors into an LDFV. Cheng et al.[28] used Pictorial Structures, where part-based color information and color displacement were considered when looking for precise part-to-part correspondence. Recently, saliency information has been investigated for person re-identification[6,29,30]; the 32-dimensional LAB color histogram and the 128-dimensional SIFT descriptor are extracted from each 10×10 patch, which is densely sampled with a step size of 5 pixels. In Ref. [4], gBiCov is proposed as a combination of Biologically Inspired Features (BIF) and covariance descriptors. In Ref. [31], LOMO is proposed to maximize the occurrence of each local pattern among all horizontal sub-windows to tackle viewpoint changes. The Retinex transform and a scale invariant texture operator are applied to handle illumination variations.

The second category learns suitable distance metrics to distinguish true and false match pairs. Specifically, most metric learning methods focus on Mahalanobis-based metrics, which generalizes Euclidean distance using linear scaling and rotations of the feature space, and can be written as

$$d(\boldsymbol{x},\boldsymbol{y})=\sqrt{(\boldsymbol{x}-\boldsymbol{y})^{\mathrm{T}}\boldsymbol{M}(\boldsymbol{x}-\boldsymbol{y})},$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ are feature vectors and $\boldsymbol{M}$ is the positive semi-definite Mahalanobis matrix. Zheng et al.[32] proposed PRDC to optimize relative distance comparisons. KISSME[33] is currently the most popular metric learning method because of its simplicity and efficiency. Hirzer et al.[34] obtained a simplified formula and promising performance by relaxing the PSD constraint required in Mahalanobis matrix. Locally-Adaptive Decision Function (LADF)[35] uses a joint model of a distance metric and a locally adapted thresholding rule for person verification, and extracts local color descriptors from patches. Aside from the Mahalanobis distance, Prosser et al.[36] modeled person re-id as a ranking problem, and applied RankSVM to learn a subspace. In Ref. [37], local experts were considered to learn a common feature space for person re-identification across views. XQDA[31] has been recently proposed as an extension of the Bayesian face[38] and KISSME[33], in which a discriminant subspace is further learned together with a metric. It learns the projection w of a low-dimensional subspace, with the cross-view data solved in a similar manner to Linear Discriminant Analysis (LDA)[39]. Zhang et al.[40] proposed overcoming the small-sample-size problem by matching people in a

discriminative null space, where images of the same person are collapsed to a single point, thus, the intra-class scatter is minimized to zero and the inter-class separation is maximized.

Recently some works based on deep learning have tackled the person re-id problem[41]. The Filter Pairing Neural Network (FPNN)[42] is proposed to jointly handle misalignment, photometric and geometric transforms, and occlusions and background clutter and has the ability to automatically learn optimal features for the re-identification task. Ahmed et al.[43] presented a deep convolutional architecture and proposed a method for simultaneously learning features and a corresponding similarity metric for person re-identification. Compared with hand-crafted features and metric learning methods, Yi et al.[44] proposed a more general way that directly learns a similarity metric from image pixels by using a "Siamese" deep neural network. A scalable distance-driven feature learning framework based on a deep neural network is presented in Ref. [45].

3 Approach

3.1 Review of bag-of-words in person re-id

The BoW model represents an image as a collection of visual words. Previous BoW approaches in person re-id[7,27] employ CNs as low-level features. Pedestrian images are segmented as patches of size $n \times n$. For each patch, CN descriptors of all pixels are calculated and l_1 normalized followed by a $\sqrt{(\cdot)}$ operator. Given the feature descriptors of image patches a codebook is generated by unsupervised clustering, such as standard k-means. Then the image is represented by frequency vectors obtained by quantizing the local descriptors to the visual words in the codebook. Here, Multiple Assignment (MA)[46] is employed to find the near neighbors of the local descriptors. Each visual word histogram is thus weighted using the TF scheme[47,48]. Burstiness[49] is also applied to achieve better performance.

Formally, the BoW method maps a feature vector $\boldsymbol{f} \in \mathbb{R}^d$ to a codeword $\boldsymbol{c}$ in the codebook $\mathcal{C} = \{\boldsymbol{c}(i)\}$ with i as a finite index set. The mapping, termed a quantizer, is denoted by $\boldsymbol{f} \to \boldsymbol{c}(i(\boldsymbol{f}))$. The function $i(\cdot)$ is called an encoder, and function $\boldsymbol{c}(\cdot)$ is called a decoder[50]. The encoder $i(\boldsymbol{f})$ maps any $\boldsymbol{f}$ to the index of its nearest codeword in the codebook $\mathcal{C}$.

An example of the BoW model work-flow is shown in Fig. 1. A pedestrian image is segmented into K horizontal strips. For the k-th strip and l-th patch, a feature vector $\boldsymbol{f}_{k,l}$ is extracted and encoded as $i(\boldsymbol{f}_{k,l})$. Then, a histogram is calculated on the k-th strip of all the visual words $\{i(\boldsymbol{f}_{k,l})\}$, which is denoted as $\boldsymbol{d}_{(k)}$. Encoding and calculating histograms together are called "voting" in this paper. The image BoW descriptor is the concatenation of $\boldsymbol{d} = [\boldsymbol{d}_{(1)}, \cdots, \boldsymbol{d}_{(k)}, \cdots, \boldsymbol{d}_{(K)}]$. The similarity between two images a and b can be calculated as the cosine distance of $\boldsymbol{d}^{(a)}$ and $\boldsymbol{d}^{(b)}$, that is,

$$s(a,b) = \boldsymbol{d}^{(a)} \cdot \boldsymbol{d}^{(b)} = \sum_{k=1}^{K} \boldsymbol{d}_{(k)}^{(a)} \cdot \boldsymbol{d}_{(k)}^{(b)}.$$

Then, the similarity of two images is the summation of the similarities of counterpart image strips. Here, for simplification, we omit (k) and write the similarity score as $s(a,b) = \boldsymbol{d}^{(a)} \cdot \boldsymbol{d}^{(b)}$.

3.2 Superpixels versus patches

Image segmentation using superpixels is an important line of approach. The superpixels algorithm groups pixels into perceptually meaningful atomic regions, which can be used to replace the rigid structure of the pixel grid. Superpixels capture image redundancy and provide a

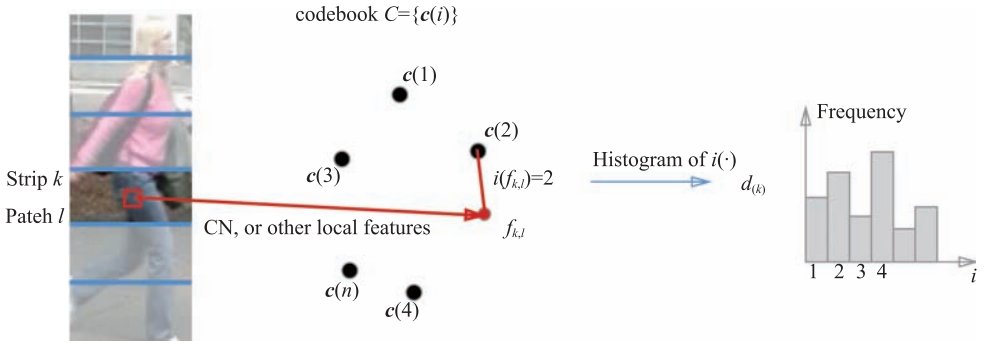

Fig. 1 An example of BoW work-flow. The pedestrian image is partitioned into horizontal strips and square patches. For each patch, a local feature vector $\boldsymbol{f}$ is first extracted then encoded into codewords $i(\cdot)$ according to the codebook $\mathcal{C}$. Then the histogram $\boldsymbol{d}_{(\cdot)}$ of codewords for one strip is calculated. Finally, the BoW descriptor is the concatenation of these histograms.

convenient primitive from which to compute image features, and effectively reduce the complexity of subsequent image processing tasks. They have become a key building block in many computer vision algorithms, such as the top-scoring methods in the multi-task object segmentation challenge in PASCAL VOC[51–53], depth estimation[54], segmentation[55], body model estimation[56], and object localization[51].

Conventional BoW methods segment images into patches ($n \times n$ pixel grids) and extract features from individual patches. Thus, these features are unstable against translation and rotation as image variations may cause shifting, necessitating re-segmentation of the patches, and re-calculation of the features. By comparison, superpixels are clustered according to the similarity of color and texture among pixels, which means they are robust against transformations.

In this paper, we employ SLIC[16] to generate superpixels and use an unsupervised pedestrian parsing model[17] to obtain a human body mask. Only superpixels whose intersection with the body mask is larger than 50% are considered as foreground. Then low-level features such as HOG[11] and SILTP[57] are extracted, respectively, from these foreground superpixels. We examine the performance gain in Section 4 and discuss the parameter tuning later. We carefully examine two parameters: the average area of a superpixel and its compactness, as illustrated in Fig. 2.

3.3 Feature fusion

Fusing different low-level features may provide richer information. We tested and compared feature fusion at different stages in the BoW model. We considered four different appearance-based features: color histograms (HSV)[7], CN[12,13], HOG[11], and SILTP[57] to cover both color and texture characteristics.

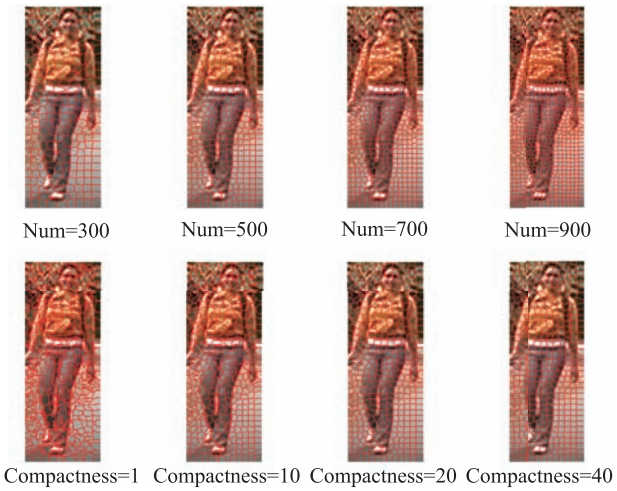

Fig. 2 Pedestrian images segmented using SLIC into superpixels of superpixel number 300, 500, 700, and 900 (approximately), superpixel compactness 1, 10, 20, and 40.

3.3.1 Feature extraction

A Color Histogram (CH) is widely used to describe color characteristics within one region. First, the original image is transferred to the HSV color space, then the statistical distribution of the hue (H) and saturation (S) channels is calculated separately. Each channel is quantized to 10 bins. The luminance (V) channel is excluded because of illumination changes.

CNs are semantic attributes obtained by assigning linguistic color labels to image pixels. Here, we use off-the-shelf descriptors, learned from real-world images such as Google Images, to map the RGB values of a pixel to 11 color terms[13]. The CN descriptor assigns each pixel an 11-dimensional vector, each dimension corresponding to one of the 11 basic colors. Then, the CN descriptor of a superpixel region is computed as the average value of each pixel.

HOG is a classical texture descriptor that counts the occurrences of gradient orientation in localized portions of an image. We partitioned gradient orientation into 9 bins and computed the descriptor using gray-level images. Then, the HOG descriptor was extracted from each superpixel region.

The SILTP[57] descriptor is an improved operator over the well-known Local Binary Pattern (LBP)[58]. LBP has a nice invariant property under monotonic gray-scale transforms, but it is not robust to image noise. SILTP improves LBP by introducing a scale invariant local comparison tolerance, achieving invariance to intensity scale changes and robustness to image noise. Within each superpixel, we extract 2 scales of SILTP histograms ($\mathrm{SILTP}^{0.3}_{4,3}$ and $\mathrm{SILTP}^{0.3}_{4,5}$) as suggested in Ref. [31].

Root descriptors have proven effective[59]. We apply root descriptors to these four features. Euclidean distance is the most general but probably sub-optimal choice considering histogram similarity. The Hellinger kernel performs better[59]. The root transformation can be regarded as an explicit feature map from the original space to the root space. Then, the Euclidean distance in the root space is equivalent to the Hellinger distance in the original space.

3.3.2 Fusion strategies

In the BoW model we can apply feature fusion in three stages and in the following, we will show how it can be

formulated as product quantization[24] problems. Here, we denote the feature vector generated by each feature method in one superpixel as $\boldsymbol{f}_1,\cdots,\boldsymbol{f}_n,\cdots,\boldsymbol{f}_N$ for a total of N feature extraction methods. We denote the overall feature space as $\boldsymbol{f}=[\boldsymbol{f}_1,\cdots,\boldsymbol{f}_n,\cdots,\boldsymbol{f}_N]$.

Product Quantization. Given a feature vector $\boldsymbol{f}\in\mathbb{R}^d$, Product Quantization (PQ) aims to decompose the original high-dimensional vector space into the Cartesian product of subspaces, then quantize these subspaces separately.

We denote the Cartesian product $\mathcal{C}=\mathcal{C}_1\times\cdots\times\mathcal{C}_N$ for $\boldsymbol{f}\in\mathbb{R}^d$ as the set in which a codeword $\boldsymbol{c}\in\mathcal{C}$ is formed by concatenating N sub-codewords: $\boldsymbol{c}=[\boldsymbol{c}_1,\cdots,\boldsymbol{c}_n,\cdots,\boldsymbol{c}_N]$, with each $\boldsymbol{c}_n\in\mathcal{C}_n$. It is easy to show that the nearest codeword $\boldsymbol{c}$ of $\boldsymbol{f}$ in $\mathcal{C}$ is the concatenation of the N nearest sub-codewords $\boldsymbol{c}=[\boldsymbol{c}_1,\cdots,\boldsymbol{c}_n,\cdots,\boldsymbol{c}_N]$ where $\boldsymbol{c}_n$ is the nearest sub-codeword of the subvector $\boldsymbol{f}_n$. That is,

$$\boldsymbol{c}(i(\boldsymbol{f}))=\boldsymbol{c}(i([\boldsymbol{f}_1,\cdots,\boldsymbol{f}_n,\cdots,\boldsymbol{f}_N]))=$$
$$[\boldsymbol{c}_1(i(\boldsymbol{f}_1)),\cdots,\boldsymbol{c}_n(i(\boldsymbol{f}_n)),\cdots,\boldsymbol{c}_N(i(\boldsymbol{f}_N))].$$

The benefit of product quantization is that it can easily generate a codebook $\mathcal{C}$ with an exponentially large number of codewords. If each sub-codebook has k sub-codewords, then their Cartesian product $\mathcal{C}$ has k^N codewords.

As discussed above, $\boldsymbol{d}$ is the histogram of $i(\boldsymbol{f})$ in the BoW model. A joint histogram of two independent variables is equivalent to the product of histograms of these two variables separately, because for two independent random variables X and Y, we always have $p(X,Y)=p(X)\times p(Y)$. On the assumption that $\boldsymbol{f}_n$ is independent, the BoW descriptor can be written as $\boldsymbol{d}=\boldsymbol{d}_1\otimes\cdots\otimes\boldsymbol{d}_n\otimes\cdots\otimes\boldsymbol{d}_N$. Here we should use the outer product because $\boldsymbol{d}_n$ are on different axes, i.e., $\boldsymbol{d}_m\otimes\boldsymbol{d}_n$ should be an $L\times L$ vector flattened by the outer product L-by-L matrix with L as the vector length of $\boldsymbol{d}_m$ and $\boldsymbol{d}_n$. Thus, the similarity score of image a and b could be written as

$$\begin{aligned}s(a,b)&=\boldsymbol{d}^{(a)}\cdot\boldsymbol{d}^{(b)}=\\&(\boldsymbol{d}_1^{(a)}\otimes\cdots\otimes\boldsymbol{d}_N^{(a)})\cdot(\boldsymbol{d}_1^{(b)}\otimes\cdots\otimes\boldsymbol{d}_N^{(b)})=\\&(\boldsymbol{d}_1^{(a)}\cdot\boldsymbol{d}_1^{(b)})\times\cdots\times(\boldsymbol{d}_N^{(a)}\cdot\boldsymbol{d}_N^{(b)})=\\&s_1(a,b)\times\cdots\times s_N(a,b),\end{aligned}$$

under the assumption that each feature subspace is independent. The proof of the commutative law of the outer product and dot product is shown in Lemma 1. The feature space independence assumption does not always hold strictly true, but could help solve the dimension explosion problem.

Lemma 1 Given four vectors each with length L, $\boldsymbol{X}=[x_1,\cdots,x_l,\cdots,x_L]$, $\boldsymbol{Y}=[y_1,\cdots,y_l,\cdots,y_L]$, $\boldsymbol{Z}=[z_1,\cdots,z_l,\cdots,z_L]$, $\boldsymbol{W}=[w_1,\cdots,w_l,\cdots,w_L]$. There should be the commutative law as $(\boldsymbol{X}\otimes\boldsymbol{Y})\cdot(\boldsymbol{Z}\otimes\boldsymbol{W})=(\boldsymbol{X}\cdot\boldsymbol{Z})\times(\boldsymbol{Y}\cdot\boldsymbol{W})$.

$$\begin{aligned}&(\boldsymbol{X}\otimes\boldsymbol{Y})\cdot(\boldsymbol{Z}\otimes\boldsymbol{W})=[\cdots,x_iy_j,\cdots]\cdot[\cdots,z_iw_j,\cdots]=\\&\sum_{i,j=1}^{L}x_iy_jz_iw_j=\left(\sum_{i=1}^{L}x_iz_i\right)\times\left(\sum_{j=1}^{L}y_jw_j\right)=\\&(\boldsymbol{X}\cdot\boldsymbol{Z})\times(\boldsymbol{Y}\cdot\boldsymbol{W}).\end{aligned}$$

Score Level Fusion. Score level fusion generates different codebooks $\mathcal{C}_1,\cdots,\mathcal{C}_n,\cdots,\mathcal{C}_N$ for each feature separately. Then in the query stage, the BoW descriptors are calculated separately as $\boldsymbol{d}_1,\cdots,\boldsymbol{d}_n,\cdots,\boldsymbol{d}_N$, where $\boldsymbol{d}_n$ is the n-th feature vector $\boldsymbol{f}_n$ voting on the n-th codebook $\mathcal{C}_n$. Then the similarity score s between two images i and j is calculated as $s=\sqrt[N]{s_1\times\cdots\times s_N}$[22,23], whereas s_n is the similarity score of the n-th descriptor. As discussed above, score level fusion is equivalent to jointly voting on the overall $\mathcal{C}$ codebook, under the assumption that each feature subspace is independent. It can be regarded as one kind of product quantization where each feature space is equivalent to each PQ subspace. The overall feature space is naturally decomposed by these different feature subspaces, which could be a good choice because these features are not related and unlikely to be correlated.

In this paper, four codebooks are computed for four features (HSV, CN, HOG, and SILTP), each with a size of 350. Thus, the effective number of codewords in the overall feature space is 350^4. In the query phase, four feature vectors of one superpixel are quantized using the four generated codebooks, respectively, then four similarity scores are calculated by the four resulting BoW descriptor vectors.

Word Level Fusion. During codebook generation, different feature vectors of one superpixel can be concatenated as a uniform fusion feature vector $\boldsymbol{f}=[\boldsymbol{f}_1,\cdots,\boldsymbol{f}_n,\cdots,\boldsymbol{f}_N]$. Then a BoW codebook $\mathcal{C}$ is learned from these concatenated feature vectors. In the query stage, the BoW descriptor of a superpixel is calculated by the codebook $\mathcal{C}$ and the concatenated feature vector $\boldsymbol{f}=[\boldsymbol{f}_1,\cdots,\boldsymbol{f}_n,\cdots,\boldsymbol{f}_N]$. Thus, the word-level fusion is equivalent to no product quantization at all in the overall feature space. In this paper, we generated a codebook of size 1400, thus keeping the final BoW vector length equal to that of other fusion methods with codebook sizes far less effective than the fusion score level.

Descriptor Level Fusion. Descriptor level fusion merges different features in the query stage. It also has N codebooks and N BoW descriptors $\boldsymbol{d}_1,\cdots,\boldsymbol{d}_n,\cdots,\boldsymbol{d}_N$,

which is the same as score level fusion. Compared with score level fusion that computes the N scores $s_1, \cdots, s_n, \cdots, s_N$ separately as $s_n = \boldsymbol{d}_n^{(a)} \cdot \boldsymbol{d}_n^{(b)}$, descriptor level fusion concatenates different BoW descriptors together as $\boldsymbol{d} = [\boldsymbol{d}_1, \cdots, \boldsymbol{d}_n, \cdots, \boldsymbol{d}_N]$, and the similarity score between two images a and b is calculated by the concatenated descriptor as $s = \boldsymbol{d}^{(a)} \cdot \boldsymbol{d}^{(b)} = \sum_{n=1}^{N} \boldsymbol{d}_n^{(a)} \cdot \boldsymbol{d}_n^{(b)} = \sum_{n=1}^{N} s_n$.

In summary, score level fusion is equivalent to product quantization under the assumption that each feature space is independent and has a codebook of k^N codewords. Word level fusion is equivalent to no product quantization in the overall feature space, whose codebook has $k \times N$ codewords. Descriptor level fusion differs from score level fusion where the final score is calculated using the arithmetic mean rather than the geometric mean of the sub-scores. We researched these three fusion strategies in detail and discuss their performance in Section 4.

4 Experiments

To evaluate the effectiveness of our method, we conducted experiments on three public benchmark datasets: VIPeR, PRID450S, and Market1501. The conventional evaluation protocol splits the dataset into training and test parts. For unsupervised method evaluation, only test samples are used. Considering re-identification as a ranking problem, the performance is measured in Cumulative Matching Characteristics (CMC). Hereby, we denote R1 as the rank 1 recognition rate and R20 as the rank 20 recognition rate.

4.1 Datasets

4.1.1 VIPeR

The 1264 images, which were normalized to 128 × 48 pixels in the VIPeR dataset, were captured from two different cameras in an outdoor environment, and included 632 individuals and 2 images of each person. It is the large variations in viewpoint, pose, resolution, and illumination that makes VIPeR very challenging. In conventional evaluation, the dataset is randomly divided into two equal parts, one for training, and the other for testing. In one trial, images were taken sequentially and matched against the opposite camera. Ten trials were repeated and the average result is given.

4.1.2 PRID450S

450 single-shot image pairs depicting walking humans were captured from two surveillance cameras. Pedestrian bounding boxes were manually labeled with a vertical resolution of 100−150 pixels, while the resolution of the original images was 720×576 pixels. In addition, part-level segmentation was provided describing the following regions: head, torso, legs, carried object at torso level (if any) and carried object below torso (if any). As with VIPeR, we randomly partitioned the dataset into two equal parts, one for training and the other for testing. 10 trials were repeated.

4.1.3 Market1501

Market1501 consists of 32 668 detected person bounding boxes of 1501 individuals captured by six cameras (five high-resolution and one low-resolution) with overlaps. Each identity was captured by at least two cameras but multiple images may occur in one camera. For each identity on test, one query image in each camera was selected, so that multiple queries were used for each identity. Note that, the selected 3 368 queries were hand-drawn, instead of DPM-detected as in the gallery. The provided fixed training and test sets were used under both single-query and multi-query evaluation settings.

4.2 Superpixel evaluation

We first studied the performance of superpixels against patches using four different descriptors (HSV, CN, HOG, and SILTP) separately on the VIPeR dataset. Then, fusion was performed at descriptor level to merge these four features. As shown in Fig. 3 and Table 1, the superpixels approach performs better than the patch approach on every feature; about 0.5% R1 gain on HSV and CN, 1.8% on SILTP, and 3.4% on HOG. The overall performance gain

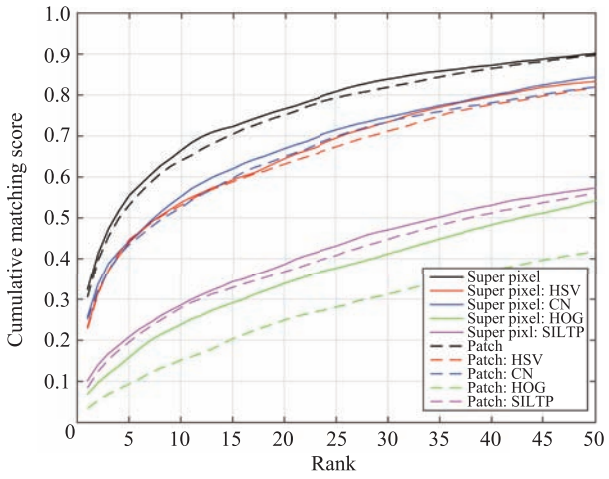

Fig. 3 CMC curves on the VIPeR dataset, by comparing the proposed superpixel approach to conventional patch methods. HSV, CN, HOG, SILTP, and descriptor level fusion are employed respectively.

Table 1 Superpixel performance on VIPeR.

	R1/R20				
	HSV	CN	HOG	SILTP	Fusion
Patches	22.8/63.2	25.1/65.0	3.45/24.9	8.43/36.8	30.6/75.2
Superpixels	23.3/64.5	25.7/67.0	6.87/33.9	10.2/38.7	32.4/76.7

reached 1.8% for R1 when these four features were fused at the descriptor level.

The improvement in HOG is the most notable and this is because HOG and SILTP are texture features, which are more sensitive to image boundaries, while HSV and CN are color features.

Next, two parameters, the average area of a superpixel and its compactness, were examined, as illustrated in Fig. 2. In the conventional patch approach, patch size is important but usually chosen heuristically. A small patch size degrades the performance of a single pixel and a micro disturbance in several pixels can cause big changes in local features and image descriptors. Whereas, a large patch size can cause descriptors to be less discriminating as a large patch carries too many pixels could lead to local information loss. This also applies to superpixels, but the influence is insignificant. In our study, the performance of the superpixel approach is robust against superpixel size, and poorly chosen superpixel size degrade performance by at most 0.8% R1 in total. As shown in Fig. 4, the best performance is achieved by segmenting images to 400−500 superpixels, i.e., an average of about 12−16 pixels in one superpixel.

Compactness also plays an important role. Highly compact superpixels can be regarded as a retrogradation to patches. Conversely, compact and regular superpixels are often desirable because their bounded size and few neighbors form a more interpretable graph, and more locally relevant features can be extracted. In our experiment, we tested re-id accuracy against a superpixel compactness of 1, 10, 20, and 30. As shown in Fig. 5, high compactness is harmful to all features and a value of 30 can cause performance degradation up to 1% in R1. Lower compactness usually results in better performance, which demonstrates that regular boundaries contribute little to the performance of the iBoW descriptor.

In summary, our method is robust to the two superpixel parameter settings of average size and compactness. This is a result of the essential characteristics of superpixels, as explained in Section 3.2. The SLIC[16] superpixel algorithm itself groups similar pixels into atomic regions and captures image redundancy with really stable performance under different parameter settings, as shown in Fig. 2. Contrarily, local features extracted from conventional patches can be significantly changed under different patch size settings.

4.3 Exploration of feature fusion methods

Local features usually have multiple descriptors, such as HOG, HSV, CN, and SILTP, each of which corresponds to a specific view of the image data. For the empirical study in the previous section, we chose descriptor level fusion. In this section, we analyze the influence of different fusion

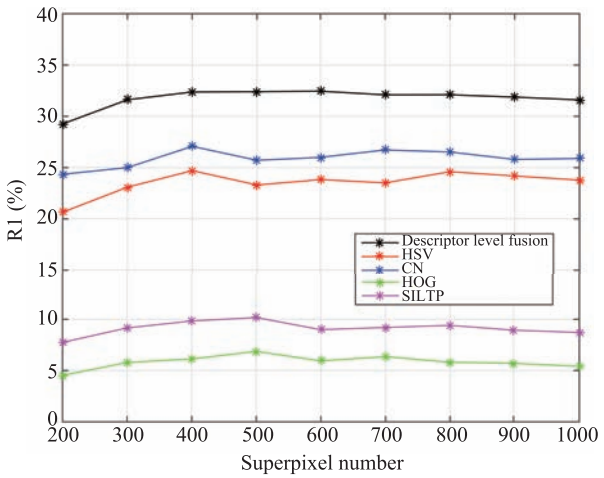

Fig. 4 R1 on the VIPeR dataset, by comparing different superpixel numbers of an image. HSV, CN, HOG, SILTP, and descriptor level fusion are employed respectively.

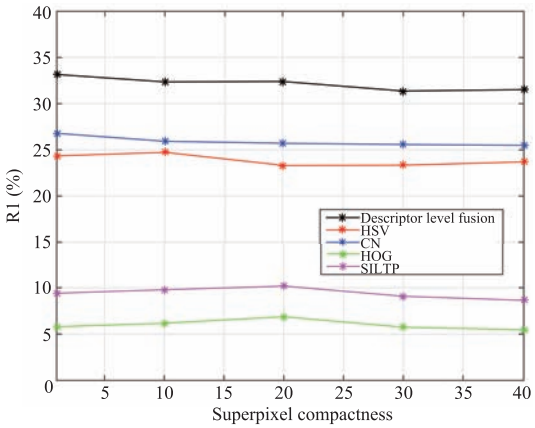

Fig. 5 R1 on the VIPeR dataset, by comparing different superpixel compactness settings. HSV, CN, HOG, SILTP, and descriptor level fusion are employed respectively.

methods on the final re-identification performance.

We evaluated three types of method, word level fusion, descriptor level fusion, and score level fusion, as described in Section 3. For score level fusion, we used the geometric mean to combine scores. The experimental results on VIPeR are shown in Table 2.

As shown in Table 2, score level fusion outperforms word level fusion by a 2.4% R1 recognition rate. As explained in Section 3, a codebook size of k_n in a $\boldsymbol{f}_n$ feature space owns k_n codewords. In word level fusion, the fused codebook, size $k = k_1 + \cdots + k_n + \cdots + k_N$, is a very sparse representation in this huge overall Cartesian product feature space. While in score level fusion, the N feature codebooks together could be regarded equivalent to a codebook size of $k = k_1 \times \cdots \times k_n \times \cdots \times k_N$ in the overall Cartesian product feature space, which is much more dense and discriminative. As for descriptor level fusion, the fusion operator is replaced from "$\times$" in score level fusion by "$+$" and outperforms it by a 1.8% R1 recognition rate. The accuracy of the different features varies hugely. Thus the "$\times$" operator in score level fusion can propagate substantial errors, which could cause performance degradation, while the "+" in descriptor level fusion is much more robust.

Based on the observations and analysis above, we conclude that the fusion method is a very important component for the combination of multiple features. Descriptor level fusion performed better in our experiments.

4.4 Comparison with state-of-the-art results

In this section, we compare our proposed method with state-of-the-art approaches. Specifically, we chose HSV, CN, HOG, and SILTP features. As for feature fusion, we adopted the descriptor level fusion method.

We first compared our approach with state-of-the-art results from VIPeR and PRID450S in Table 3. Within all unsupervised approaches, we obtained a Rank 1 re-identification rate of 32.41% with VIPeR and 30.16% with PRID450S, which are superior to the best result obtained from VIPeR and PRID450S, by 5.7% and 5.6%, respectively. We also integrated the proposed unsupervised method with three supervised metric learning methods, KISSME[33], XQDA[31], and Null Space[40]. The best result was achieved by integrating our descriptor with Null Space[40] metric learning, which reached a Rank 1 precision of 50.0% from VIPeR and 68.04% from PRID450S, and outperformed these state-of-the-art methods by 2.2% and 7.5%, respectively.

Table 2 Comparison of different feature fusion methods.

	R1 (%)	R20 (%)
Word level fusion	28.28	71.95
Score level fusion	30.66	76.11
Descriptor level fusion	32.41	76.66

Table 3 Comparison to the state-of-the-art results on VIPeR and PRID450S.

	Method	R1 (%) VIPeR	R1 (%) PRID450S
Unsupervised	SDALF[3]	19.9	-
	eSDC[29]	26.7	-
	CPS[28]	22.0	-
	ELF6[60]	8.73	-
	HSV+Lab+LBP[33]	12.47	13.0
	gBiCov[4]	9.87	-
	LOMO[31]	19.91	24.6
	BoW[27]	21.74	-
	Proposed	**32.41**	**30.16**
Supervised	WARCA[61]	40.2	24.6
	Cheng et al.[62]	47.8	-
	LSSCDL[63]	42.7	60.5
	X-KPLS[64]	33.1	52.8
	Kernel HPCA[65]	39.4	52.8
	ECM[66]	38.9	41.9
	SCNCD[67]	37.8	37.8
	CBRA[68]	31.2	26.4
	LOMO+KISSME[31]	34.05	48.8
	LOMO+XQDA[31]	40.00	59.64
	LOMO+NullSpace[40]	42.3	-
	Proposed+KISSME	37.18	52.47
	Proposed+XQDA	43.23	63.07
	Proposed+NullSpace	**50.00**	**68.04**

As for large-scale datasets, such as Market1501, our method yielded a Rank 1 recognition of 48.37% and mAP of 19.98% under the single query mode. This was the best of all the unsupervised approaches, as shown in Table 4. We roughly classified supervised learning methods into two categories, a conventional metric learning based approach and a deep learning based approaches. Our method gave a Rank 1 recognition of 64.13% and mAP of 36.21% with Null Space[40] metric learning, which outperforms the best metric learning approaches by 8.7% and 6.3%, respectively. Our result even outperformed many other deep learning based approaches and is comparable to the recent state-of-the-art method Gated Siamese CNN[77]. This result is quite outstanding as Market1501 is generally considered more suitable for deep learning based methods due to its large image volume.

Table 4 Comparison to the state-of-the-art results on Market1501.

	Method	R1 (%)	mAP (%)
Unsupervised	gBiCov[4]	8.28	2.23
	HistLBP[69]	9.62	2.72
	LOMO[31]	26.07	7.75
	BoW[27]	34.38	14.10
	Proposed	**48.37**	**19.98**
Metric learning	WARCA[61]	45.16	-
	TMA[70]	47.92	22.31
	SCSP[71]	51.90	26.35
	LOMO+KISSME[31]	40.50	19.02
	LOMO+XQDA[31]	43.79	22.22
	LOMO+Null Space[40]	55.43	29.87
	Proposed+KISSME	54.81	27.65
	Proposed+Null Space	**64.13**	**36.21**
Deep-learning	PersonNet[72]	37.21	18.57
	CAN[73]	48.24	24.43
	SSDAL[74]	39.4	19.6
	Triplet CNN[75]	45.1	-
	Histogram Loss[76]	59.47	-
	Gated Siamese CNN[77]	**65.88**	**39.55**

Figure 6 shows the example retrieval results from Market1501.

5 Conclusion

In this paper, we proposed an unsupervised descriptor for person re-identification. The approach uses an improved BoW model based on superpixels and descriptor level fusion, combining both color and texture features. We carefully examined the parameter settings for superpixel generation, and different fusion methods were compared theoretically and practically. Experiments demonstrated the effectiveness and robustness of our method. The proposed descriptor outperforms other unsupervised methods in VIPeR, PRID450S, and Market1501. Meanwhile, our descriptor can be effectively integrated with efficient supervised metric learning algorithms and outperforms current state-of-the-art results.

In our work, there is still much room for improvement and expansion. Deep neural networks attract a lot of attention nowadays, but the connection between our unsupervised descriptor and deep learning has not been explored. Generally speaking, as our unsupervised descriptor can be regarded as a global model, it would be interesting to combine it with deep learning models by connecting it in the first convolutional layer.

Fig. 6 Example results from Market1501 using our proposed method with Null Space metric learning[40]. The images in the first column are the query images. The gallery images are sorted according to their similarity scores from left to right, with a highlighted green box for correct matches.

Acknowledgment

The work was supported by the National Natural Science Foundation of China (No. 61071135) and the National Science and Technology Support Program (No. 2013BAK02B04).

References

[1] S. Gong, M. Cristani, S. Yan, and C. C. Loy, *Person Re-identification*. Springer, 2014.

[2] D. Gray and H. Tao, Viewpoint invariant pedestrian recognition with an ensemble of localized features, in *European Conference on Computer Vision*. Springer, 2008, pp. 262–275.

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

[4] B. Ma, Y. Su, and F. Jurie, Covariance descriptor based on bio-inspired features for person re-identification and face verification, *Image and Vision Computing*, vol. 32, no. 6, pp. 379–390, 2014.

[5] B. Ma, Y. Su, and F. Jurie, Local descriptors encoded by fisher vectors for person re-identification, in *European Conference on Computer Vision*, 2012, pp. 413–422.

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

[7] L. Tian and S. Wang, Person re-identification as image retrieval using bag of ensemble colors, *IEICE TRANSACTIONS on Information and Systems*, vol. 98,

no. 1, pp. 180–188, 2015.

[8] L. Zheng, S. Wang, Z. Liu, and Q. Tian, Packing and padding: Coupled multi-index for accurate image retrieval, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1939–1946.

[9] L. Zheng, S. Wang, J. Wang, and Q. Tian, Accurate image search with multi-scale contextual evidences, *International Journal of Computer Vision*, pp. 1–13, 2016.

[10] D. G. Lowe, Distinctive image features from scale-invariant keypoints, *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[11] N. Dalal and B. Triggs, Histograms of oriented gradients for human detection, in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005, pp. 886–893.

[12] B. Berlin and P. Kay, *Basic Color Terms: Their Universality and Evolution*. Oakland, CA, USA: Univ. of California Press, 1991.

[13] J. Van de Weijer, C. Schmid, and J. Verbeek, Learning color names from real-world images, in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[14] L. Zheng, Y. Yang, and A. G. Hauptmann, Person re-identification: Past, present and future, arXiv preprint arXiv:1610.02984, 2016.

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

[16] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, Slic superpixels compared to state-of-the-art superpixel methods, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[17] P. Luo, X. Wang, and X. Tang, Pedestrian parsing via deep decompositional network, in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2648–2655.

[18] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, Learning realistic human actions from movies, in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[19] X. Wang, L. Wang, and Y. Qiao, A comparative study of encoding, pooling and normalization methods for action recognition, in *Asian Conference on Computer Vision*, 2012, pp. 572–585.

[20] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, Dense trajectories and motion boundary descriptors for action recognition, *International Journal of Computer Vision*, vol. 103, no. 1, pp. 60–79, 2013.

[21] H. Wang and C. Schmid, Lear-inria submission for the thumos workshop, in *ICCV Workshop on Action Recognition with a Large Number of Classes*, 2013, p. 8.

[22] K. Tang, B. Yao, L. Fei-Fei, and D. Koller, Combining the right features for complex event recognition, in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2696–2703.

[23] G. K. Myers, C. G. Snoek, R. Nevatia, R. Nallapati, J. van Hout, S. Pancoast, C. Sun, A. Habibian, D. C. Koelma, K. E. van de Sande, et al., Evaluating multimedia features and fusion for example-based event detection, in *Fusion in Computer Vision*. Springer, 2014, pp. 109–133.

[24] T. Ge, K. He, Q. Ke, and J. Sun, Optimized product quantization, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 4, pp. 744–755, 2014.

[25] D. Gray, S. Brennan, and H. Tao, Evaluating appearance models for recognition, reacquisition, and tracking, in *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, Citeseer, 2007.

[26] P. M. Roth, M. Hirzer, M. Koestinger, C. Beleznai, and H. Bischof, Mahalanobis distance learning for person re-identification, in *Person Re-Identification*, S. Gong, M. Cristani, S. Yan, and C. C. Loy, eds. Springer, 2014, pp. 247–267.

[27] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, Scalable person re-identification: A benchmark, in *Computer Vision, IEEE International Conference on*, 2015.

[28] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, Custom pictorial structures for re-identification, presented at the 22nd British Machine Vision Conference, 2011.

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

[30] Y. Liu, Y. Shao, and F. Sun, Person re-identification based on visual saliency, in *2012 12th International Conference on Intelligent Systems Design and Applications (ISDA)*, 2012, pp. 884–889.

[31] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, Person re-identification by local maximal occurrence representation and metric learning, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2197–2206.

[32] W.-S. Zheng, S. Gong, and T. Xiang, Person re-identification by probabilistic relative distance comparison, in *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 649–656.

[33] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, Large scale metric learning from equivalence constraints, in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012, pp. 2288–2295.

[34] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof, Relaxed pairwise learned metric for person re-

identification, in *European Conference on Computer Vision*, 2012, pp. 780–793.

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

[36] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary, Person re-identification by support vector ranking, presented at the 21st British Machine Vision Conference, 2010.

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

[38] B. Moghaddam, T. Jebara, and A. Pentland, Bayesian face recognition, *Pattern Recognition*, vol. 33, no. 11, pp. 1771–1782, 2000.

[39] B. Scholkopft and K.-R. Mullert, Fisher discriminant analysis with kernels, *Neural Networks for Signal Processing IX*, vol. 1, no. 1, p. 1, 1999.

[40] L. Zhang, T. Xiang, and S. Gong, Learning a discriminative null space for person re-identification, arXiv preprint arXiv:1603.02139, 2016.

[41] L. Zheng, Y. Yang, and Q. Tian, Sift meets cnn: A decade survey of instance retrieval, arXiv preprint arXiv:1608.01807, 2016.

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

[43] E. Ahmed, M. Jones, and T. K. Marks, An improved deep learning architecture for person re-identification, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3908–3916.

[44] D. Yi, Z. Lei, and S. Z. Li, Deep metric learning for practical person re-identification, arXiv preprint arXiv:1407.4979, 2014.

[45] S. Ding, L. Lin, G. Wang, and H. Chao, Deep feature learning with relative distance comparison for person re-identification, *Pattern Recognition*, vol. 48, no. 10, pp. 2993–3003, 2015.

[46] H. Jegou, M. Douze, and C. Schmid, Hamming embedding and weak geometric consistency for large scale image search, in *European Conference on Computer vision*, Springer, 2008, pp. 304–317.

[47] J. Sivic and A. Zisserman, Video google: A text retrieval approach to object matching in videos, in *Proceedings of Ninth IEEE International Conference on Computer Vision*, 2003, pp. 1470–1477.

[48] L. Zheng, S. Wang, and Q. Tian, Lp-norm idf for scalable image retrieval, *IEEE Transactions on Image Processing*, vol. 23, no. 8, pp. 3604–3617, 2014.

[49] H. Jégou, M. Douze, and C. Schmid, On the burstiness of visual elements, in *Computer Vision and Pattern Recognition, IEEE Conference on CVPR 2009*, 2009, pp. 1169–1176.

[50] R. Gray, Vector quantization, *IEEE Assp Magazine*, vol. 1, no. 2, pp. 4–29, 1984.

[51] B. Fulkerson, A. Vedaldi, and S. Soatto, Class segmentation and object localization with superpixel neighborhoods, in *ICCV*, 2009, vol. 9, pp. 670–677.

[52] Y. Yang, S. Hallman, D. Ramanan, and C. Fowlkes, Layered object detection for multi-class segmentation, in *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 3113–3120.

[53] S. Gould, J. Rodgers, D. Cohen, G. Elidan, and D. Koller, Multi-class segmentation with relative location prior, *International Journal of Computer Vision*, vol. 80, no. 3, pp. 300–316, 2008.

[54] C. L. Zitnick and S. B. Kang, Stereo for image-based rendering using image over-segmentation, *International Journal of Computer Vision*, vol. 75, no. 1, pp. 49–65, 2007.

[55] Y. Li, J. Sun, C.-K. Tang, and H.-Y. Shum, Lazy snapping, in *ACM Transactions on Graphics (ToG)*, vol. 23, no. 3, pp. 303–308, 2004.

[56] G. Mori, Guiding model search using segmentation, in *Tenth IEEE International Conference on Computer Vision (ICCV'05)*, 2005, pp. 1417–1423.

[57] S. Liao, G. Zhao, V. Kellokumpu, M. Pietikäinen, and S. Z. Li, Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes, in *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 1301–1306.

[58] T. Ojala, M. Pietikäinen, and D. Harwood, A comparative study of texture measures with classification based on featured distributions, *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[59] R. Arandjelović and A. Zisserman, Three things everyone should know to improve object retrieval, in *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 2911–2918.

[60] W.-S. Zheng, S. Gong, and T. Xiang, Reidentification by relative distance comparison, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 3, pp. 653–668, 2013.

[61] C. Jose and F. Fleuret, Scalable metric learning via weighted approximate rank component analysis, arXiv preprint arXiv:1603.00370, 2016.

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

[63] Y. Zhang, B. Li, H. Lu, A. Irie, and X. Ruan, Sample-

specific svm learning for person re-identification, in *IEEE Conference on Computer Vision and Pattem Recognition*, 2016, pp. 1278–1287.

[64] R. F. Prates and W. R. Schwartz, Kernel hierarchical pca for person re-identification, in *2016 23rd Internationa Conference on Pattem Recognition (ICPR)*, 2016.

[65] R. Prates, M. Oliveira, and W. R. Schwartz, Kernel partial least squares for person re-identification, in *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2016, pp. 249–255.

[66] X. Liu, H. Wang, Y. Wu, J. Yang, and M.-H. Yang, An ensemble color model for human re-identification, in *2015 IEEE Winter Conference on Applications of Computer Vision*, 2015, pp. 868–875.

[67] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, Salient color names for person re-identification, in *European Conference on Computer Vision*, 2014, pp. 536–551.

[68] R. F. de Carvalho Prates and W. R. Schwartz, Cbra: Color-based ranking aggregation for person re-identification, in *2015 IEEE International Conference on Image Processing (ICIP)*, 2015, pp. 1975–1979.

[69] F. Xiong, M. Gou, O. Camps, and M. Sznaier, Person re-identification using kernel-based metric learning methods, in *European Conference on Computer Vision*, 2014, pp. 1–16.

[70] N. Martinel, A. Das, C. Micheloni, and A. K. Roy-Chowdhury, Temporal model adaptation for person re-identification, in *European Conference on Computer Vision*, 2016, pp. 858–877.

[71] D. Chen, Z. Yuan, B. Chen, and N. Zheng, Similarity learning with spatial constraints for person re-identification, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1268–1277.

[72] L. Wu, C. Shen, and A. van den Hengel, Personnet: Person re-identification with deep convolutional neural networks, arXiv preprint arXiv:1601.07255, 2016.

[73] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, End-to-end comparative attention networks for person re-identification, arXiv preprint arXiv:1606.04404, 2016.

[74] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, Deep attributes driven multi-camera person re-identification, arXiv preprint arXiv:1605.03259, 2016.

[75] J. Liu, Z.-J. Zha, Q. Tian, D. Liu, T. Yao, Q. Ling, and T. Mei, Multi-scale triplet cnn for person re-identification, in *Proceedings of the 2016 ACM on Multimedia Conference*, 2016, pp. 192–196.

[76] E. Ustinova and V. Lempitsky, Learning deep embeddings with histogram loss, in *Advances in Neural Information Processing Systems*, 2016, pp. 4170–4178.

[77] R. R. Varior, M. Haloi, and G. Wang, Gated siamese convolutional neural network architecture for human re-identification, in *European Conference on Computer Vision*, 2016, pp. 791–808.

Shengjin Wang received the BEng degree from Tsinghua University, China, and the PhD degree from the Tokyo Institute of Technology, Tokyo, Japan, in 1985 and 1997, respectively. From 1997 to 2003, he was a member of the researcher with the Internet System Research Laboratories, NEC Corporation, Japan. Since 2003, he has been a professor with the Department of Electronic Engineering, Tsinghua University, where he is currently the director of the Research Institute of Image and Graphics. His current research interests include image processing, computer vision, video surveillance, and pattern recognition. He is a member of IEEE and IEICE.

Lu Tian received the BEng degree from Tsinghua University, China, in 2011. She has been studying for the PhD degree in electronic engineering at Tsinghua University from 2011. Her current research interests include pattern recognition and human feature extraction, in particular person re-identification.