Journal of Electronic Imaging

JElectronicImaging.org

Bag of local features for person re-identification on large-scale datasets

Yixiu Liu
Yunzhou Zhang
Jianning Chi

Yixiu Liu, Yunzhou Zhang, Jianning Chi, "Bag of local features for person re-identification on large-scale datasets," *J. Electron. Imaging* **27**(5), 053041 (2018), doi: 10.1117/1.JEI.27.5.053041.

Journal of Electronic Imaging 27(5), 053041 (Sep/Oct 2018)

Bag of local features for person re-identification on large-scale datasets

Yixiu Liu,[a] Yunzhou Zhang,[a,b,*] and Jianning Chi[b]
[a]Northeastern University, College of Information Science and Engineering, Shenyang, China
[b]Northeastern University, Faculty of Robot Science and Engineering, Shenyang, China

Abstract. In recent years, large-scale person re-identification has attracted a lot of attention from video surveillance. Usual approaches addressing this task either learn the effective feature embeddings or design the learning architectures to obtain discriminative metrics. Most of them only focus on improving the accuracy of recognition but neglect retrieval efficiency. To improve the accuracy and efficiency of person re-identification simultaneously, an accurate and fast method is proposed based on the bag of visual words (BoVW) model, which has widely been applied in image retrieval. A bag of local features is developed to simplify feature representation for person re-identification. Cross-view dictionary learning is used to eliminate the redundancy of these local features. These local features consist of scale invariant feature transform and local maximal occurrence representation (LOMO) that are invariant in scale and color, respectively. Finally, integrated BoVW histograms are obtained, which encode the images by *k*-means clustering. Experiments conducted on the CUHK03, Market1501, and MARS datasets show that the proposed method performs favorably against existing approaches. [DOI: 10.1117/1.JEI.27.5.053041]

Keywords: person re-identification; bag of visual words; cross-view dictionary learning; reranking; large-scale datasets.

Paper 180333 received Apr. 15, 2018; accepted for publication Sep. 20, 2018; published online Oct. 17, 2018.

1 Introduction

Identifying whether a person in different images has the same ID under disjoint camera views, referred to as person re-identification, has gained much attention among researchers.[1–6] A person's appearance usually undergoes significant variations due to changes in illumination, camera angle, background clutter, and occlusion over the camera network. Large-scale person re-identification is still a challenging task. As shown in Fig. 1, due to the large variance among images of the same person and the subtle differences among different persons, it is still a hard task to search a most similar image to the query image from the large gallery until now.

In this paper, the bag of visual words (BoVW) model serves to achieve fast and accurate image search. BoVW, as an ensemble visual features, is widely applied in image retrieval. Inspired by natural language processing methods, most state-of-the-art image retrieval techniques rely on the BoVW model.[7] BoVW is derived from the BoW model. It was introduced to computer vision by document analysis. If we view a document as a set of words, the words that are not repeated constitute the dictionary of this document. Then each document can be encoded with the frequency of the elements in this dictionary. Replacing the "words" in the BoW model with visual features, or visual "words," is the so-called BoVW model.

To build upon the bag-of-features representation, local descriptors are extracted to create the visual "words." The field of image search has greatly been advanced since the introduction of the scale invariant feature transform (SIFT) descriptor and the BoW model. Person re-identification exists as a subdomain of image retrieval, in which the local SIFT features still play the important role. For example, to handle viewpoint and illumination change, the SIFT descriptor is used as complementary to color histograms, which effectively improve the accuracy of person re-identification.[8] Based on the two points above, SIFT features are used as the visual "words" for the BoVW model in this paper. Other visual "words" are local maximal occurrence representation (LOMO) features,[9] which perform well for re-id work in recent years. As far as we know, there are only a handful of methods to solve a large gallery problem with the BoVW model, e.g., BoVW + color names (CN).[8] In this paper, CN features are replaced with LOMO features, and lots of experiments[9–11] show that the new BoVW model is more effective than the previous one. In addition, the LOMO features and SIFT features are concatenated to be the enhanced visual "words," which visibly improve the matching rate compared to using them alone.

Dictionary learning is usually adopted to obtain the more discriminative feature representation, such as singular value decomposition by *k*-means (K-SVD),[12] semisupervised coupled dictionary learning,[13] discriminative K-SVD,[14] and projective dictionary pair learning.[15] We adopt cross-view dictionary learning (CDL)[16] due to the view-consistency information. Different from the original, two pairs of dictionaries are learned from cross-view two-level visual data. These visual data are local SIFT and LOMO features in this paper.

After obtaining the sparse visual "words," the codebook is generated to express the images. We adopt *k*-means clustering to complete this step, as many image retrieval methods based on BoVW model do.[17–19] Then the histograms, which represent the frequency of the feature category, are generated according to the codebook. In this way, all the images can be represented by the statistical histograms, which effectively

*Address all correspondence to: Yunzhou Zhang, E-mail: zhangyunzhou@mail.neu.edu.cn

1017-9909/2018/$25.00

(a) (b)

Fig. 1 Examples of four pairs people from (a) Market1501 and (b) MARS datasets. Large variance among images of the same person (first pair and third pair) and subtle differences among different persons (second pair and fourth pair) make person re-id challenging.

improve the efficiency of retrieval, especially for large datasets image retrieval. Finally, to get more accurate search result, reranking is an essential step because the top results are likely to be very similar in large-scale datasets.

In summary, we make the following contributions for person re-identification:

First, a new strategy of designing the feature representation is proposed based on the BoVW model to achieve fast and robust person research. Then a more discriminative feature representation is obtained by CDL from the extracted local SIFT and local LOMO features. Finally, we quantitatively validate the performance of our algorithm by comparing it to the state-of-the-art methods on three challenging datasets including CUHK03, Market1501,[20] and MARS.[21]

2 Related Works

Most existing approaches for person re-identification consist of two stages: (1) metric learning and (2) feature representation. Generally, metric learning methods derive Mahalanobis distance by projecting the sample $x_i \in R^d$ to $y_i = W^T x_i$, where $W \in R^{d\times m}$ $(m < d)$ is the subspace. The distance between the projected samples is $\|y_i - y_j\|_2^2 = (x_i - x_j)^T WW^T (x_i - x_j)$. Feature is the most important factor to describe the appearance of pedestrians, which is the highlight of this paper.

2.1 *Hand-Craft Features*

Lots of hand-craft features are developed to address the re-id problem. Yang et al.[22] proposed a salient color descriptor based on CN to describe pedestrian appearance, so that color distributions in different spaces were obtained and fused to generate a feature representation. Zhao et al.[23] used dense SIFT descriptors as complementary to color histograms to handle viewpoint and illumination change. Each image was densely divided into overlapping local patches, and each patch was represented by a feature vector concatenating SIFT features and color histograms. LOMO feature was proposed by Liao et al.[9] To make a stable representation against viewpoint changes, LOMO feature analyzes the horizontal occurrence of local features and maximizes the occurrence. Inspired by the above methods, we believe that the combination of several features can play a more powerful role. In this paper, we used both local SIFT and local LOMO features. To offset the reduced retrieval efficiency caused by the sum, we used their BoVW histogram form instead of the original feature, which improves both the accuracy and efficiency of person re-identification.

2.2 *Deep Features*

In recent years, deep learning constantly refreshes the accuracy of recognition tasks, especially for person re-identification. Wu et al.[24] developed an enhanced deep feature representation with feature fusion net, in which convolutional neural network (CNN) features were constrained by the hand-crafted features through back propagation. Ahmed et al.[25] put feature learning and similarity metric in a deep convolutional architecture with layers specially designed to improve person re-identification. Xiao et al.[26] proposed a domain guided dropout algorithm to learning deep feature representations from multiple domains with CNNs. The above methods have one thing in common. They only focus on the recognition accuracy but ignore the retrieval efficiency. Generally, data-driven deep learning methods require a lot of training time and are difficult to identify quickly. Different from them, we take the retrieval efficiency into consideration instead of only focus on recognition accuracy. In this paper, we not only compute the matching rate to get the cumulative curve (CMC), but also compared the time consumption of the proposed method with the state-of-the-art methods.

2.3 *Dictionary Learning*

Dictionary learning is a powerful technique for learning an effective component from sample space. For this reason, it is widely used in person re-identification. More specially, given the feature matrix $X = \{x_1, x_2, \ldots, x_n\} \in R^{d\times n}$, the dictionary $D = \{d_1, d_2, \ldots, d_k\} \in R^{d\times k}$ is learned by solving an optimal objective function:

$$\min_{D,Z} f(X, D, Z) + \lambda g(D), \tag{1}$$

where $Z \in R^{k\times n}$ is the coefficient matrix of X, and λ is a trade-off parameter. There are two terms in this cost function and fall into two categories: the first term $f(\cdot)$ is reconstruction error that makes the learned dictionary D encode the feature matrix X well, and the second term $g(\cdot)$ is the regularization term that enforce the coefficient matrix Z as sparse as possible. Most dictionary learning methods[13–15] are for a single view. Take local coordinate

coding dictionary learning[13] as an example, the objective function is

$$\min_{D,Z} \sum_i \left(\frac{1}{2} \|X - DZ\|^2 + \lambda |z_i^j| \|d_j - x_i\|^2 \right) \quad (2)$$

which model data distributed on manifolds. However, the representation power of features might be limited because it mainly focuses on single-view data but cannot directly handle the multiview visual data. We tackle this problem using the CDL algorithm.[16] In this paper, two pairs of dictionaries $D_{i,j}(i \le 2, j \le 2)$ are learned from cross-view two-level visual data. Different from the original, our visual data are from local SIFT and local LOMO features of two camera views.

2.4 *Reranking*

Reranking is a direct method to improve the accuracy of person re-identification. Many researchers have adopted this strategy. An et al.[27] proposed a method that a saliency-based reranking scheme is included to further improve the re-identification accuracy after initial matching with the extracted features. Xie et al.[28] presented a probe-specific reranking framework to refine the initial result measured by the learned metric, and this reranking framework only focused on top-100 images in initial ranking list. In this work, our strategy is similar to the latter. The initial ranking is determined by the cosine distance. Then we adopt cross-view quadratic discriminant analysis (XQDA)[9] to obtain the final result from the top 10% of initial ranking list.

3 BoVW Framework

In this section, we focus on explaining the construction of the feature representation in details. The framework of the feature representation with the BoVW model is shown in Fig. 2. The extracted local features are clustered into the assignment map to describe each image with a BoVW histogram.

3.1 *Feature Extraction*

The purpose of this process is to obtain the visual "words" for the BoVW model. Local SIFT and LOMO features are extracted and merged through concatenation to form the visual "words." The detailed process is shown below.

3.1.1 *Local SIFT features*

SIFT is a local descriptor widely applied in image processing due to its good performance in the scale of invariance. The SIFT feature is based on the interest points of some local appearance of the object and has the properties of scale-invariant and rotation-invariant. At the speed of the computer hardware at present, meanwhile under the small feature database condition, the identification can be achieved in real time, so that it is suitable for fast and accurate matching in mass database. In addition, the SIFT feature has a strong scalability, which allows it to be easily combined with other forms of feature vectors. In this work, the SIFT features are concatenated with the LOMO features, which further improves the matching rate.

Fig 3. The process of SIFT points detection and matching for person re-id.

At first, each image is divided into 10×3 of overlapping local patches. Then the SIFT features are extracted from these patches. Each patch is divided into 4×4 cells, and the orientations of local gradients are quantized into 8 bins. In the end, for each patch, we obtain a $4 \times 4 \times 8 = 128$-dimensional SIFT feature vector, which is ℓ_2 normalized.

From Fig. 3, we can clearly understand the process of SIFT points detection and matching. The picture on the left is the original image from the MARS dataset. The middle picture is the result of SIFT feature point detection. The picture on the right is a schematic of conventional feature matching.

3.1.2 *LOMO features*

LOMO is a traditional hand-crafted descriptor, which behaves with excellent performance in person re-id. Before extracting the features, a multiscale Retinex algorithm[29] is implemented to preprocess person images. The Retinex images have a better consistency in lighting and color, which makes person re-id easier than using original images. Then the scale invariant local ternary pattern (SILTP)[30] descriptor is applied for illumination invariant texture description. It is an improved operator over the well-known local binary pattern,[31] which has good performance under monotonic gray-scale transforms but poor performance for eliminating image noises. Through utilizing a scale invariant local comparison tolerance, SILTP achieves invariance to intensity scale changes and robustness to image noises. In addition, a color feature is extracted and combined with SILTP, making the feature more robust.

The process of LOMO feature extraction is shown in Fig. 4. We analyze the horizontal occurrence of local features and maximize the occurrence to make a stable representation against viewpoint changes. Instead of describing an image

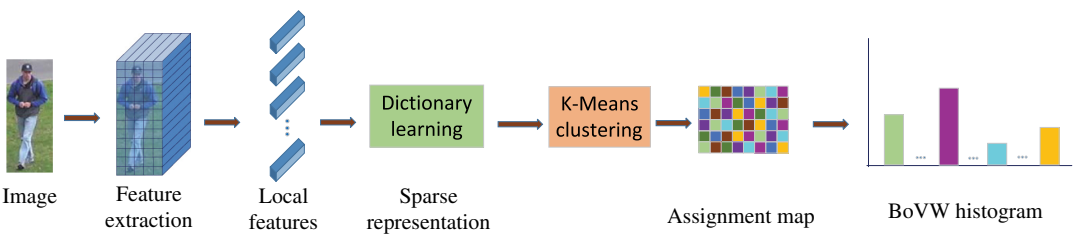

Fig. 2 The architecture of the feature representation with BoVW model.

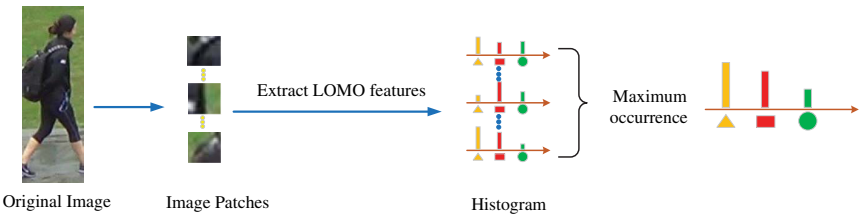

Fig. 4 The process of LOMO feature extraction.

with a feature vector as LOMO + XQDA[9] do, we view the horizontal maximized occurrence of patches features in the same row as local features. In this way, we can obtain 10 feature vectors from each image, and 10 denote the numbers of patches in vertical direction.

3.2 *Cross-View Dictionary Learning*

Dictionary learning aims to learn expressive feature representations and has widely been applied in many learning tasks. Typical dictionary learning is as follows:

$$X \approx DZ, \tag{3}$$

where $X \in R^{d\times n}$ denotes a set of samples, $D \in R^{d\times m}$ is the dictionary we learned, and $Z \in R^{m\times n}$ is the sparse coefficients.

However, existing dictionary learning methods based on Eq. (3) do not solve any real visual recognition applications, especially when samples are captured across different camera views. We tackle this problem by utilizing an efficient CDL model. To improve the efficiency of the calculation, Eq. (3) can be reformulated as $X \approx DPX$, and $P \in R^{m\times d}$ ($m \ll d$) denotes the low-dimensional projection matrix.

We consider the dictionary learning in a two-view setting. Let $X_1 \in R^{d\times n}$ and $X_2 \in R^{d\times n}$ denote two training sets of two camera views. Then the process of dictionary learning can be expressed as

$$X_1 = D_1P_1X_1, \quad X_2 = D_2P_2X_2, \tag{4}$$

here D_1 and D_2 represent the dictionaries of two different camera views, whereas P_1 and P_2 denote the dimensional projection matrices.

After feature extraction, we obtain two kinds of feature representations: local SIFT and local LOMO. We define the local SIFT feature representation in one camera view as $X_{1,1}$, then the corresponding dictionary and projection matrix are $D_{1,1}$ and $P_{1,1}$. The local LOMO feature representation is defined as $X_{1,2}$, and the corresponding dictionary and projection matrix are $D_{1,2}$ and $P_{1,2}$. Similarly, $X_{2,1}$, $D_{2,1}$, $P_{2,1}$, $X_{2,2}$, $D_{2,2}$, and $P_{2,2}$ are the variables defined in another camera view. The schematic of CDL is shown in Fig. 5.

Then we can present the objective function of CDL model as

$$\begin{aligned}\min_{\substack{D_{1,v},D_{2,v},\\ P_{1,v},P_{1,v}}} & \|X_{1,l} - D_{1,l}P_{1,l}X_{1,l}\|_F^2 \\ & + \|X_{2,l} - D_{2,l}P_{2,l}X_{2,l}\|_F^2 + \lambda_l f_l(D_{1,l}, D_{2,l}, P_{1,l}, P_{2,l}) \\ \text{s.t. } & \|d_{1l(:,i)}\| \le 1, \|d_{2l(:,i)}\| \le 1, \quad i = 1,\ldots,m, l = 1,2,\end{aligned} \tag{5}$$

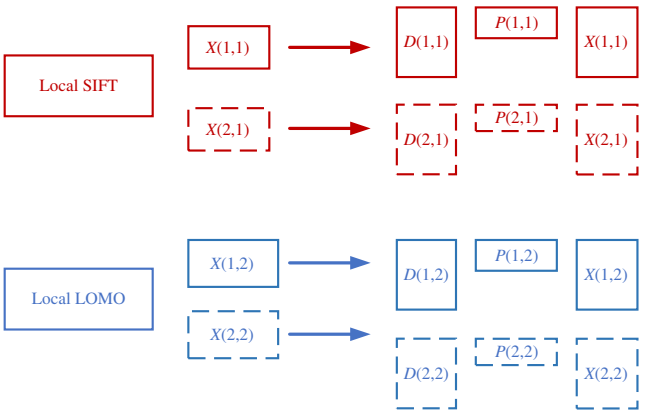

Fig. 5 The pipeline of CDL.

where $f_l(D_{1,l}, D_{2,l}, P_{1,l}, P_{2,l})$ is the regularization function, and λ_l is the corresponding trade-off parameter. $\|\cdot\|_F$ means Frobenius norm. After training, the obtained optimal dictionary pairs $\{D_{1,1}, D_{2,1}\}$ and $\{D_{1,2}, D_{2,2}\}$ can be utilized to generate new representation for test samples.

3.3 *k-Means for Multiview Clustering*

To get the codebook that represents the entire database, k-means clustering is utilized to categorize the features in the database, and the process is unsupervised.

After feature extraction and CDL, we gain the sparse feature database:

$$D = \{x_i | i = 1,\ldots,n\} \quad x_i \in R^d, \tag{6}$$

where n denotes the number of the features and d denotes the length of a feature vector.

The k-means algorithm will divide the set D into k clusters, so that each feature vector is assigned in one of the k clusters by minimizing the nonnegative cost function:

$$\text{cost} = \sum_{i=1}^{n}(\arg\min_j \|x_i - \mu_j\|_2^2 \quad j = 1,\ldots,k). \tag{7}$$

In Eq. (7), μ_j denotes the center of clustering. The object of k-means to minimize is the sum of squares of Euclidean distance between each feature x_i and its nearest cluster μ_j.

Integrating Sec. 3.1 and 3.2 with k-means clustering method, our algorithm can be expressed as Algorithm 1, the local features are specified as SIFT/LOMO, or SIFT + LOMO.

Using the linear combination of the elements in this codebook, all the images can be described as the statistical histograms of the frequency of the elements.

Algorithm 1 *K*-means clustering for local features

Input: Feature database $D = \{x_i | i = 1, \ldots, n\}$; Number of clusters k; Iterations N

Output: The set of clusters $C = \{c_j | j = 1, \ldots, k\}$; Cluster member vectors L

1: Initialize the clusters C, $t = 1$

2: **for** $t <= N$ **do**

3: Redistribute the elements in D to nearest cluster

4: Update L (l_i denotes the label of the i'th element in D)

5: Update C (c_j denotes the label of the j'th cluster)

6: **while** Eq. (7) converges **do**

7: Return L

8: **end while**

9: **end for**

4 Experimental Result and Discussions

In this section, lots of experiments are done on three large-scale datasets: CUHK03, Market1501, and MARS, which indicate that the proposed algorithm is more accurate and faster than many state-of-the-art algorithms.[9,10,14,20,21,32–36] Some samples of CUHK03, Market1501, and MARS datasets are shown in Fig. 6. In addition, we also discuss the influence of reranking on re-id work in this paper.

4.1 *Implementation Details*

The following is some details of the proposed algorithm. LOMO features are composed of HSV and SILTP histograms. We utilize sliding windows with a size of 10×10 pixels and an overlapping step of 5 pixels to locate local patches in 128×48 pixel images. HSV histogram is extracted which has $8 \times 8 \times 8$ bins $= 512$ dimensions per patch. Two scales of SILTP histograms ($\text{SILTP}^{0.3}_{4,3}$ and $\text{SILTP}^{0.3}_{4,5}$) are extracted for each patch, and the dimension is $3^4 \times 2 = 81$. In addition, a three-scale pyramid representation is built for utilizing the multiscale information, which down-samples the original 128×48 image by two 2×2 local average pooling operations. So the LOMO feature has ($8 \times 8 \times 8$ color bins $+3^4 \times 2$ SILTP bins) $\times$ ($24 + 11 + 5$ horizontal groups) = 26,960 dimensions. After cascading with SIFT features, each image can be represented by a 27,088-dimensional vector ($n = 27{,}088$). By solving Eq. (5), m is set to 2488 ($\lambda_1 = 0.2$ and $\lambda_2 = 0.3$). It can be seen that the feature vector is reduced by an order of magnitude after CDL. As for k-means, we present $k = 0.25d$, where 0.25 is the empirical value.

4.2 *Comparison with State-of-the-Art Methods*

In this part, three datasets are introduced, and then we show that our method is superior to many methods in accuracy and search efficiency through lots of comparison experiments. Our method is the trade-off between accuracy and efficiency, so we only compare our method with the suboptimal method in search efficiency.

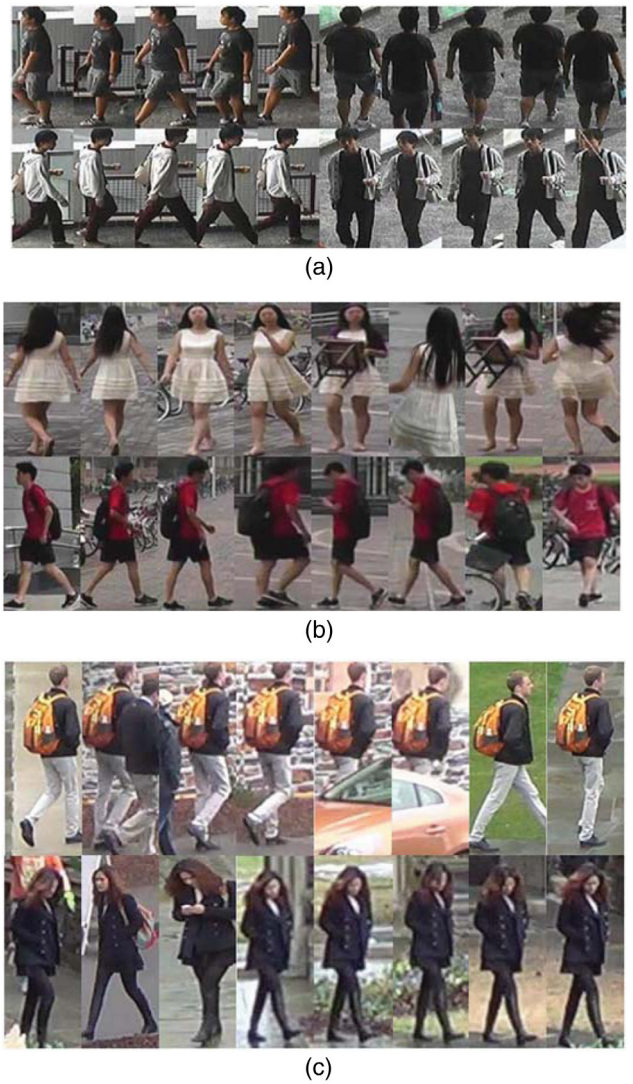

Fig. 6 Some examples of the (a) CUHK03, (b) Market1501, and (c) MARS datasets.

4.2.1 *Experiments on CUHK03*

It is the first person re-identification dataset that is large enough for deep learning. 1467 identities are collected from 5 different pairs of camera views. It provides the bounding boxes detected from deformable part models (DPM)[32] and manual labeling. Person detection quality is relatively good for this dataset.

Figure 7(a) and Table 1 provide the matching results of all compared algorithms. It can be seen that our method obviously outperforms all strong competitors, e.g., the top-1 ensembles by 4.1% at rank-1. It is worth mentioning that the methods we compare are relatively new in recent years, such as the top-2 JointRe-id, and our method is superior to them both in accuracy and efficiency. Figure 7(b) shows that our algorithm performs faster than the suboptimal algorithm ensembles.

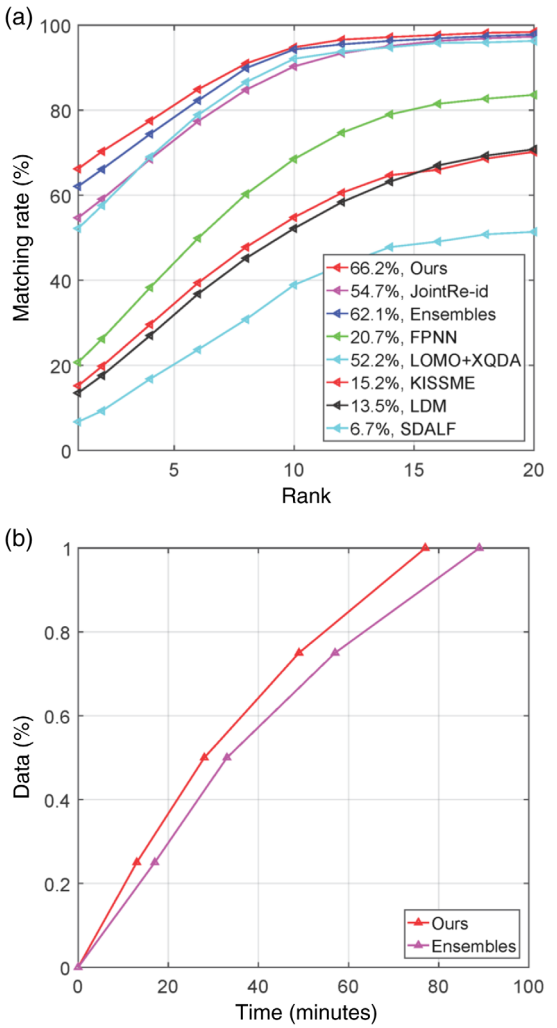

Fig. 7 (a) CMC and (b) time consuming of CUHK03 dataset.

Table 1 Top r rank matching accuracy (%) on CUHK03.

Method	Rank = 1	Rank = 10	Rank = 15	Rank = 20
Ours	66.2	94.8	97.4	98.4
JointRe-id[30]	54.5	92.1	96.7	97.3
Ensembles[33]	62.1	94.3	96.6	97.8
FPNN[20]	20.7	68.5	81.2	83.6
LOMO + XQDA[9]	52.2	92.1	95.6	96.3
KISSME[10]	15.2	54.8	65.0	70.2
LDM[34]	13.5	52.2	65.7	70.8
SDALF[14]	6.7	38.9	48.2	51.4

4.2.2 *Experiments on market1501*

It contains a large number of identities and each identity has several images from six disjoint cameras. This dataset also includes 2793 false alarms from DPM as distractors to mimic the real scenario. Quality of the bounding boxes is worse than CUHK03. Later in the ICCV 2015 release version, 500k distractors are integrated to make this dataset large scale.

Table 2 Top r rank matching accuracy (%) on Market1501.

Method	Rank = 1	Rank = 5	Rank = 10	Rank = 20
Ours	87.53	95.28	97.14	98.37
APR[25]	84.29	93.20	95.19	97.00
PIE[36]	78.06	90.76	94.41	96.52

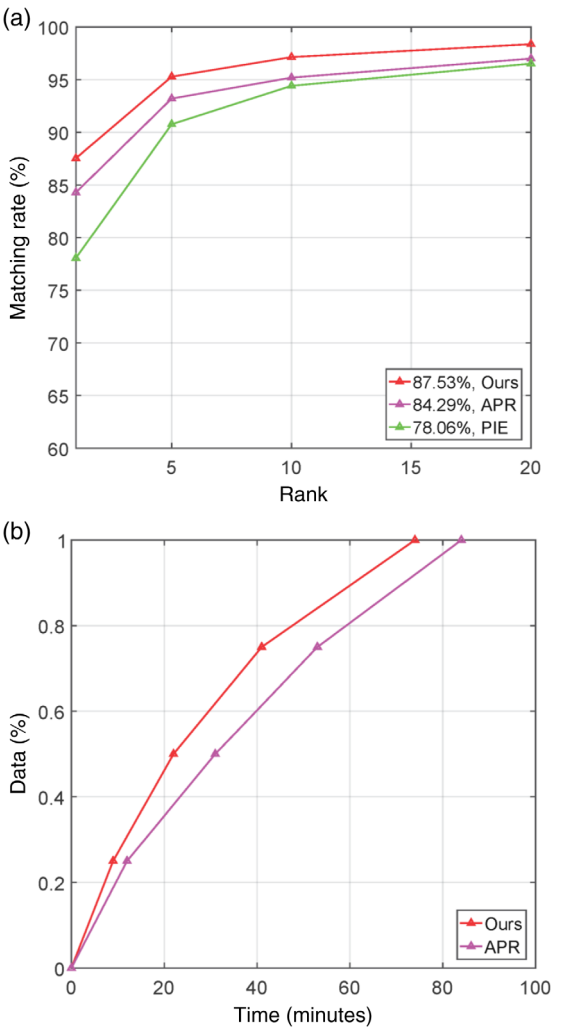

Fig. 8 (a) CMC and (b) time consuming of Market1501 dataset.

For large-scale image identification, a logical idea to tackle this challenge of person re-id is deep learning due to adequate training samples. Both APR[35] and PIE[36] in Table 2 are methods based on deep learning. The results of the experiment show that the method proposed in this paper is better than the two methods above. Figure 8(a) shows the experimental comparison curve. Figure 8(b) shows that our algorithm performs faster than the suboptimal algorithm APR.

4.2.3 *Experiments on MARS*

The dataset (motion analysis and re-identification set) is an extension version of the Market1501 dataset. It is the first large scale video-based person re-id dataset. Since all bounding boxes and trajectories are generated automatically, it contains distractors and each identity may have more than one trajectory. This dataset is much larger than the previous two datasets, remaining challenging to evaluate the performance of the proposed method.

The comparison with the state-of-the-art algorithms on MARS is shown in Table 3. On MARS, we obtain

Table 3 Top *r* rank matching accuracy (%) on MARS.

Method	Rank = 1	Rank = 5	Rank = 10	Rank = 20
Ours	74	92	95	97
SPTF-Net[37] + Eucl.	71	89	93	96
CNN embedding[21] + Eucl.	59	77	81	87
CNN embedding[21] + XQDA	65	82	86	89

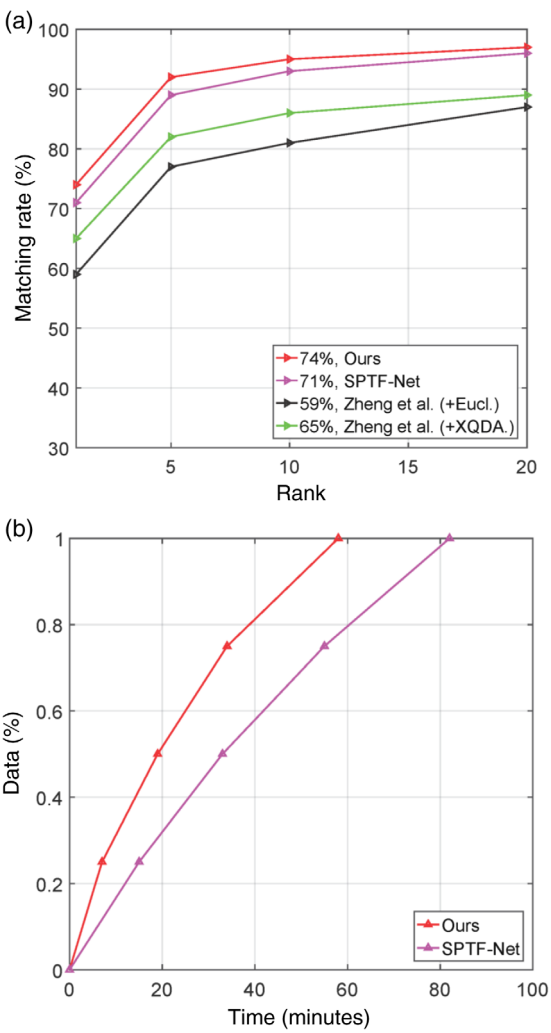

Fig. 9 (a) CMC and (b) time consuming of MARS dataset.

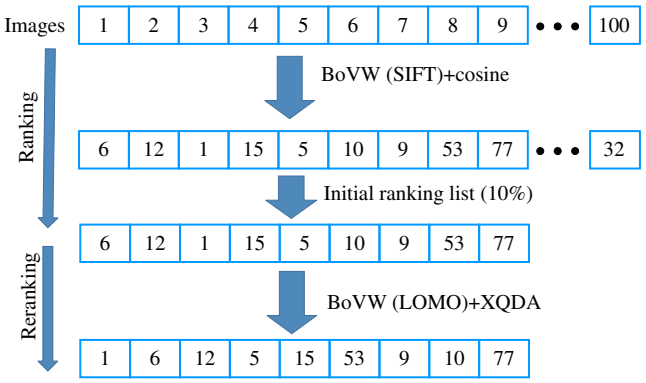

Fig. 10 The schematic diagram of the reranking.

Table 4 Person re-id matching rates (%) at different ranks on CUHK03.

	CUHK03			
Method	Rank at 1	10	15	20
BoVW (SIFT) + cosine	44.3	58.6	67.8	71.5
BoVW (LOMO) + cosine	50.3	65.6	72.4	78.1
BoVW (SIFT) + XQDA	57.2	78.8	83.7	85.3
BoVW (LOMO) + XQDA	61.5	89.7	92.3	93.8
BoVW (SIFT and LOMO) + reranking	70.4	96.2	97.9	98.7

Table 5 Person re-id matching rates (%) at different ranks on Market1501.

	Market1501			
Method	Rank at 1	5	10	20
BoVW (SIFT) + cosine	52.6	60.7	64.8	68.5
BoVW (LOMO) + cosine	59.8	68.1	72.5	74.3
BoVW (SIFT) + XQDA	69.7	76.2	79.8	81.2
BoVW (LOMO) + XQDA	81.6	89.7	92.3	94.1
BoVW (SIFT and LOMO) + reranking	89.8	96.8	97.3	98.6

Table 6 Person re-id matching rates (%) at different ranks on MARS.

	Market1501			
Method	Rank at 1	5	10	20
BoVW (SIFT) + cosine	41	52	61	66
BoVW (LOMO) + cosine	47	61	69	73
BoVW (SIFT) + XQDA	56	72	78	82
BoVW (LOMO) + XQDA	69	88	90	91
BoVW (SIFT and LOMO) + reranking	81	93	96	98

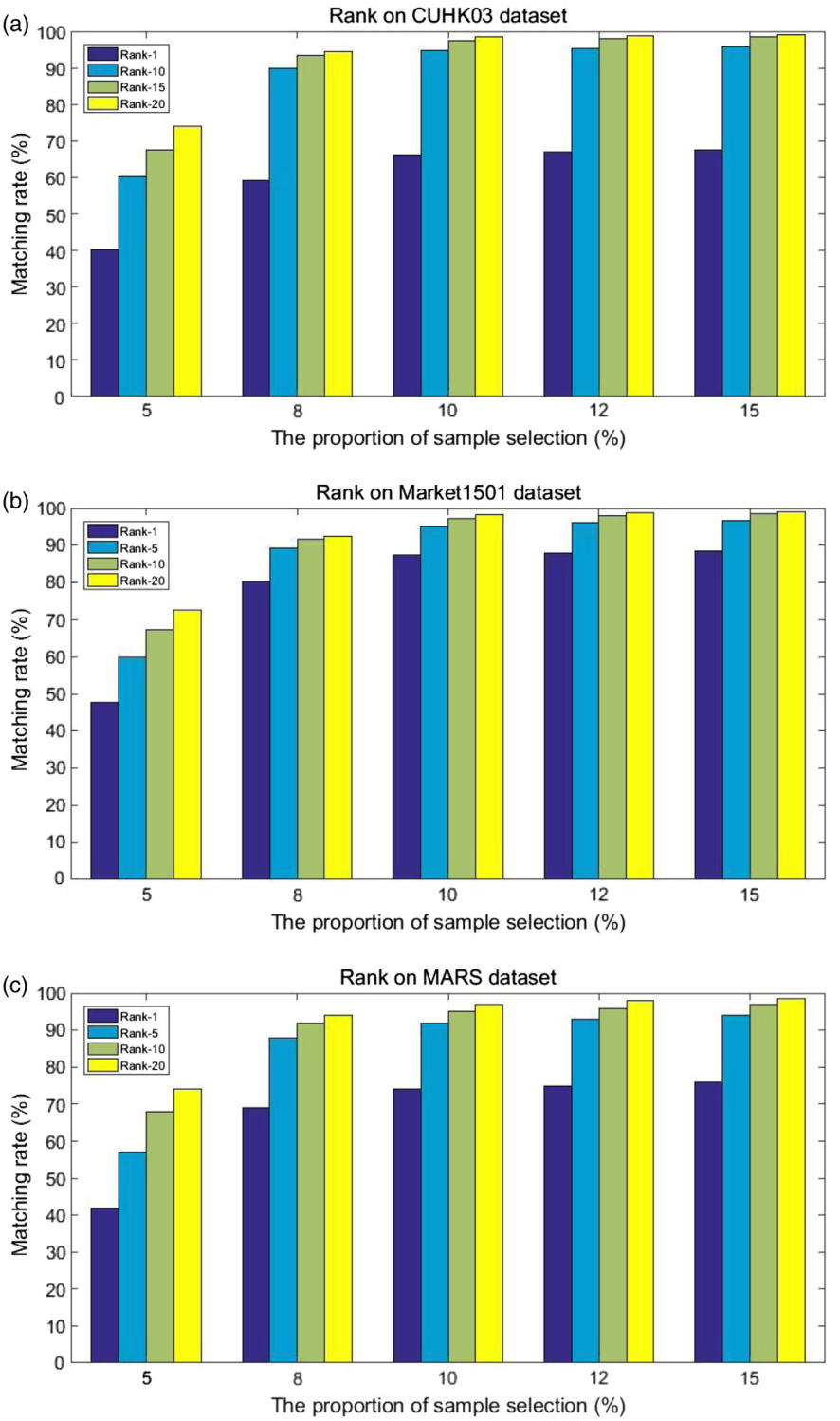

Fig. 11 The effect of proportion of sample selection on experimental results: (a) CUHK03, (b) Market1501, and (c) MARS.

rank-1 = 74%, rank-5 = 92%, rank-10 = 95%, and rank-20 = 97%. It can be clearly seen from Fig. 9(a) that our method is better than other state-of-the-art methods in each level. Our method outperforms the top-1 SPTF-Net by 3% at rank-1, and the top-2 Zhang et al. (XQDA) by 9%. Figure 9(b) shows that our algorithm performs noticeably faster than the suboptimal algorithm SPTF-Net.

4.3 *Reranking*

Reranking is an important step to improve the matching rate of person re-id, especially in large-scale datasets. In many cases, once metric often does not accurately query the object image due to the large variance among images of the same person and the subtle differences among different persons. A logical idea to tackle this problem is reranking the initial ranking list.

In this paper, we make a reranking setting to further improve the accuracy of person re-identification. We rank the matching results of the BoVW histogram (SIFT) with cosine distance at first. Then in the second phase, other visual "words" are utilized to rerank the initial list with the XQDA algorithm. In addition, to reduce the noise of irrelevant samples, our reranking only focuses on a few samples of the initial ranking list.

The framework is shown in Fig. 10. The process can be divided into two phases. In the first phase, BoVW (SIFT) + cosine is implemented, then we obtain the initial ranking list. In the second phases, BoVW (LOMO) + XQDA is implemented, and we obtain the final result from the top 10% of the initial ranking list.

Tables 4–6 illustrate the experimental results of different features and different metric methods in this paper. We do lots of experiments on CUHK03, Market1501, MARS datasets. We can see that BoVW (SIFT and LOMO) + reranking achieves the best results.

In addition, the proportion of our sample selection from the initial ranking list will also have an impact on the results of the experiment. Figure 11 shows the effect of proportion of sample selection from initial ranking list on experimental results. As we can see from Fig. 11, the matching rate has increased significantly when the proportion is not up to 10%. However, the matching rate is growing very slowly after the proportion reaches 10%. Therefore, we select 10% samples from the initial list to rerank, which is the result of the compromise between time consumption and matching accuracy.

5 Conclusion

We have presented a bag of local features for person re-identification. The proposed BoVW model integrates both local SIFT and LOMO into a histogram. The CDL algorithm is used to simplify the features and eliminate the redundant information to some extent. The final feature representation is simple and effective, especially for large-scale person re-identification. The results of our extensive experiments indicated that the proposed approach outperforms many state-of-the-art methods both in terms of accuracy and efficiency.

There are two interesting ways to extend this work in the future. We intend to embed the deep features in the BoVW model proposed in this paper. An end-to-end network architecture will be designed to further improve the accuracy of person re-identification. In addition, we would like to improve the efficiency of large-scale person search using indexing structures or hashing techniques. Finally, we hope that our algorithm can play an important role in practical applications.

Acknowledgments

The research was supported by the National Natural Science Foundation of China (Nos. 61471110 and 61733003), National Key R&D Program of China (No. 2017YFC0805000/5005), and Fundamental Research Funds for the Central Universities (Nos. N172608005 and N160413002).

References

1. Z. Liu, J. Chen, and Y. Wang, "A fast adaptive spatio-temporal 3D feature for video-based person re-identification," in *IEEE Int. Conf. on Image Processing (ICIP)*, pp. 4294–4298 (2016).
2. F. Wang et al., "Joint learning of single-image and cross-image representations for person re-identification," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 1288–1296 (2016).
3. S. Li, M. Shao, and Y. Fu, "Cross-view projective dictionary learning for person re-identification," in *Int. Conf. on Artificial Intelligence*, pp. 2155–2161 (2015).
4. B. Ma, Y. Su, and F. Jurie, *Discriminative Image Descriptors for Person Re-identification*, Springer, London (2014).
5. H. Liu et al., "End-to-end comparative attention networks for person re-identification," *IEEE Trans. Image Process.* **26**(7), 3492–3506 (2017).
6. R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 3586–3593 (2013).
7. I. Gonzlez-Daz et al., "Neighborhood matching for image retrieval," *IEEE Trans. Multimedia* **19**(3), 544–558 (2017).
8. L. Zheng et al., "Scalable person re-identification: a benchmark," in *IEEE Int. Conf. on Computer Vision (ICCV)*, pp. 1116–1124 (2015).
9. S. Liao et al., "Person re-identification by local maximal occurrence representation and metric learning," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 2197–2206 (2015).
10. M. Kstinger et al., "Large scale metric learning from equivalence constraints," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 2288–2295 (2012).
11. L. Bazzani, M. Cristani, and V. Murino, "Symmetry-driven accumulation of local features for human characterization and re-identification," *Comput. Vision Image Understanding* **117**(2), 130–144 (2013).
12. M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Image Process.* **54**, 4311–4322 (2006).
13. X. Liu et al., "Semi-supervised coupled dictionary learning for person re-identification," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 3550–3557 (2014).
14. Q. Zhang and B. Li, "Discriminative k-SVD for dictionary learning in face recognition," in *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp. 2691–2698 (2010).
15. S. Gu et al., "Projective dictionary pair learning for pattern classification," in *Int. Conf. on Neural Information Processing Systems*, pp. 793–801 (2014).
16. S. Li, M. Shao, and Y. Fu, "Person re-identification by cross-view multi-level dictionary learning," *IEEE Trans. Pattern Anal. Mach. Intell.* PP(99), 1–1 (2017).
17. H. Jgou, M. Douze, and C. Schmid, "Improving bag-of-features for large scale image search," *Int. J. Comput. Vision* **87**(3), 316–336 (2010).
18. J. Wang et al., "Notice of violation of IEEE publication principles bag-of-features based medical image retrieval via multiple assignment and visual words weighting," *IEEE Trans. Med. Imaging* **30**(11), 1996–2011 (2011).
19. L. Zhu et al., "Weighting scheme for image retrieval based on bag-of-visual-words," *IET Image Process.* **8**(9), 509–518 (2014).
20. W. Li et al., "Deepreid: deep filter pairing neural network for person re-identification," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 152–159 (2014).
21. L. Zheng et al., "MARS: a video benchmark for large-scale person re-identification," in *European Conf. on Computer Vision*, pp. 868–884, Springer International Publishing (2016).
22. Y. Yang et al., "Salient color names for person re-identification," *Lect. Notes Comput Sci.* **8689**(9), 536–551 (2014).
23. R. Zhao, W. Oyang, and X. Wang, "Person re-identification by saliency learning," *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(2), 356–370 (2017).
24. S. Wu et al., "An enhanced deep feature representation for person re-identification," in *IEEE Winter Conf. on Applications of Computer Vision (WACV)*, pp. 1–8 (2016).
25. E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 3908–3916 (2015).
26. T. Xiao et al., "Learning deep feature representations with domain guided dropout for person re-identification," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1249–1258 (2016).
27. L. An et al., "Person reidentification with reference descriptor," *IEEE Trans. Circuits Syst. Video Technol.* **26**(4), 776–787 (2016).
28. Y. Xie et al., "Adaptive metric learning and probe-specific reranking for person reidentification," *IEEE Signal Process. Lett.* **24**(6), pp. 853–857 (2017).
29. D J Jobson et al., "A multiscale retinex for bridging the gap between color images and the human observation of scenes," *IEEE Trans. Image Process.* **6**(7), 965–976 (1997).
30. S. Liao et al., "Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes," in *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp. 1301–1306 (2010).
31. T. Ojala and I. Harwood, "A comparative study of texture measures with classification based on feature distributions," *Pattern Recognit.* **29**(1), 51–59 (1996).

32. S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," *Adv. Neural Inf. Process. Syst.* **15**(2), 561–568 (2003).
33. S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Learning to rank in person re-identification with metric ensembles," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 1846–1855 (2015).
34. M. Guillaumin, J. Verbeek, and C. Schmid, "Is that you? Metric learning approaches for face identification," in *IEEE Int. Conf. on Computer Vision*, pp. 498–505 (2010).
35. Y. Lin et al., "Improving person re-identification by attribute and identity learning," CoRR abs/1703.07220 (2017).
36. L. Zheng et al., "Pose invariant embedding for deep person re-identification," CoRR abs/1701.07732 (2017).
37. L. Chen et al., "Deep spatial-temporal fusion network for video-based person re-identification," in *IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1478–1485 (2017).

Yixiu Liu received his BE degree from Northeastern University at Qinhuangdao, Qinhuangdao, China, in 2016 and is currently working toward his PhD at the School of Information Science and Engineering, Northeastern University, Shenyang, China. His research interests are in the area of computer vision, including person re-identification and pedestrian tracking.

Yunzhou Zhang received his BS and MS degrees in mechanical and electronic engineering from the National University of Defense Technology, Changsha, China, in 1997 and 2000, respectively, and his PhD in pattern recognition and intelligent system from Northeastern University, Shenyang, China, in 2009, where he is currently a professor at the Faculty of Robot Science and Engineering, Northeastern University, China. His research interests include intelligent robot, image processing, and sensor networks.

Jianning Chi received his BS and MS degrees from Northeastern University, China, and his PhD in computer science from the University of Saskatchewan, Canada. He currently works at the College of Robot Science and Engineering, Northeastern University. His research interests are in the area of image processing including image description, multiscale methods and sparse representations of images, spatial and frequency image filtering, and image segmentation.