Learning Rotation-Invariant Local Binary Descriptor

Yueqi Duan, Jiwen Lu, *Senior Member, IEEE*, Jianjiang Feng, *Member, IEEE*, and Jie Zhou, *Senior Member, IEEE*

***Abstract*— In this paper, we propose a rotation-invariant local binary descriptor (RI-LBD) learning method for visual recognition. Compared with hand-crafted local binary descriptors, such as local binary pattern and its variants, which require strong prior knowledge, local binary feature learning methods are more efficient and data-adaptive. Unlike existing learning-based local binary descriptors, such as compact binary face descriptor and simultaneous local binary feature learning and encoding, which are susceptible to rotations, our RI-LBD first categorizes each local patch into a rotational binary pattern (RBP), and then jointly learns the orientation for each pattern and the projection matrix to obtain RI-LBDs. As all the rotation variants of a patch belong to the same RBP, they are rotated into the same orientation and projected into the same binary descriptor. Then, we construct a codebook by a clustering method on the learned binary codes, and obtain a histogram feature for each image as the final representation. In order to exploit higher order statistical information, we extend our RI-LBD to the triple rotation-invariant co-occurrence local binary descriptor (TRICo-LBD) learning method, which learns a triple co-occurrence binary code for each local patch. Extensive experimental results on four different visual recognition tasks, including image patch matching, texture classification, face recognition, and scene classification, show that our RI-LBD and TRICo-LBD outperform most existing local descriptors.**

***Index Terms*— Rotation invariance, binary descriptor, feature learning, co-occurrence feature.**

I. Introduction

EXTRACTING distinctive features is one of the most active issues in computer vision which is widely applicable in many applications, e.g. face recognition [1]–[5], texture classification [6]–[8], object and scene recognition [9], [10], 3D reconstruction and many others. High quality representation and low computational cost are two essential properties for an effective feature descriptor. On one hand, it is crucial for a feature descriptor to be discriminative and robust, as real-world applications usually suffer from large intra-class variations. On the other hand, large amount of data and mobile devices with limited computational capabilities require efficient feature descriptors, which have low memory cost and high computational speed.

Over the past decade, local binary features have aroused extensive attention due to their robustness and efficiency, such as binary robust independent elementary feature (BRIEF) [11], binary robust invariant scalable keypoint (BRISK) [12], oriented FAST and rotated BRIEF (ORB) [13], fast retina keypoint (FREAK) [14], and LBP [1], [6] as well as its variants [10], [15]–[19]. Binary features deliver strong robustness over local changes and show high computational efficiency by substituting the Euclidean distance with the Hamming distance. However, most local binary features are hand-crafted, which require strong prior knowledge and are heuristic.

More recently, several learning-based local binary descriptors have been proposed to address the limitation by directly learning hash filters to project image patches into binary codes [2], [20]. Compared with hand-crafted methods, the learned binary codes deliver more properties, such as compact, energy-saving and evenly-distributed, which leads to stronger discriminative power. Moreover, learning-based local binary descriptors are more data-adaptive. However, these learning-based methods are sensitive to rotations, which are not applicable to databases with large rotation variations, such as texture classification, misaligned face recognition and scene classification, or to some real applications with unknown rotations in testing. To address the limitation, we propose a rotation-invariant local binary descriptor (RI-LBD) by jointly learning orientations for local patches and hash functions for feature projection. Fig. 1 illustrates the pipeline of our RI-LBD approach. Unlike existing local binary descriptor learning methods, our RI-LBD first classifies each local patch into a rotational binary pattern (RBP), and then jointly learns the rotational function for each pattern and the projection matrix in an unsupervised manner, where the rotational variations of an O-PDV are rotated into the same pattern. Then, we perform a clustering on the learned binary codes to construct a codebook, and extract a histogram feature with the codebook as the final representation of each image. As RI-LBD learns each feature from a single local patch, the higher order statistical information cannot be well exploited. To address this, we present a triple rotation-invariant co-occurrence local binary descriptor (TRICo-LBD) learning method, where triple adjacent O-PDVs are utilized to describe a single local region to capture the correlation among three co-occurred features. Extensive experimental results on four different visual recognition tasks including image patch matching, texture classification, face recognition and scene classification show that our RI-LBD and

Manuscript received July 20, 2016; revised March 27, 2017 and April 26, 2017; accepted May 11, 2017. Date of publication May 16, 2017; date of current version May 26, 2017. This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB1001001, in part by the National Natural Science Foundation of China under Grant 61672306, Grant 61572271, Grant 61527808, Grant 61373074, and Grant 61373090, in part by the National 1000 Young Talents Plan Program, in part by the National Basic Research Program of China under Grant 2014CB349304, in part by the Ministry of Education of China under Grant 20120002110033, and in part by the Tsinghua University Initiative Scientific Research Program. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. David Clausi. *(Corresponding author: Jiwen Lu.)*

The authors are with the Department of Automation, Tsinghua University, State Key Lab of Intelligent Technologies and Systems, and Tsinghua National Laboratory for Information Science and Technology (TNList), Beijing 100084, China (e-mail: duanyq14@mails.tsinghua.edu.cn; lujiwen@tsinghua.edu.cn; jfeng@tsinghua.edu.cn; jzhou@tsinghua.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2017.2704661

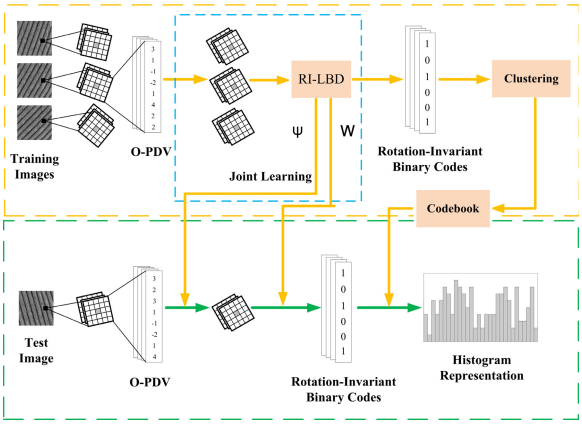


Fig. 1. The flowchart of our RI-LBD approach. In the training procedure, we first extract the ordered pixel difference vectors (O-PDV) for each training image, and learn the rotational function ψ for each rotational binary pattern (RBP) and the projection matrix $\mathbf{W}$ jointly in an unsupervised manner, so that the rotation variants of a local patch are mapped into the same binary codes. Then, we learn a codebook by clustering for feature encoding. In the testing procedure, we first extract O-PDVs for each test image, which are then rotated with the learned rotational function and projected into rotation-invariant binary descriptors with the projection matrix. Lastly, we construct a histogram feature from the binary codes with the codebook as the final representation.

TRICo-LBD are of wide applicability and outperform most existing state-of-the-art local descriptors.

II. Related Work

In this section, we briefly review three related topics: 1) binary feature descriptors, 2) feature learning, and 3) co-occurrence features.

A. Binary Feature Descriptors

Recently, binary feature descriptors have aroused increasing interest due to their robustness and efficiency. Earlier works include BRIEF [11], BRISK [12], ORB [13] and FREAK [14]. BRIEF directly compares the intensities of pairs of points to calculate binary vectors, which are fast to build and match. BRISK relies on a circular sampling pattern to achieve robustness. ORB obtains scale and orientation invariance by employing scale pyramids and orientation operators based on BRIEF. FREAK applies the retinal sampling grid for fast matching inspired by the human visual system, retina. However, these methods are susceptible to noise and transformation because only raw intensity comparisons are utilized. To address the limitation, several learning-based methods [21]–[25] have been proposed in recent years. For example, Trzcinski *et al.* presented BinBoost [23] by applying boosting for learning hash functions to obtain compact binary descriptors. They also proposed D-BRIEF [22] which learns discriminative projections by encoding similarity relationships. Balntas *et al.* presented binary online learned descriptor (BOLD) [25] which applies the LDA criterion by adapting the binary tests to each patch.

B. Feature Learning

In recent years, there has been great success of feature learning in visual analysis. A number of feature learning methods have been proposed in the literature [3], [26]–[29], and representative methods include restricted Boltzman machine [26], local quantized pattern (LQP) [29], discriminant face descriptor (DFD) [28], convolutional deep belief networks [27] and deep hidden identity features (DeepID) [3]. These learning-based features have achieved impressive performance in various computer vision tasks, yet convolutional neural networks [3], [27] outperform most of the others. However, convolutional neural networks require large number of labeled samples for feature learning because extensive parameters are usually required to estimate. Yet large amounts of labeled data are hard to collect for some practical applications, such as cross-modality face recognition, texture classification and facial age estimation. Therefore, several unsupervised local feature learning methods have been presented [2], [20]. Compact binary face descriptor (CBFD) [2] learns a hashing filter to project each local patch into compact binary codes in an unsupervised manner. Simultaneous local binary feature learning and encoding (SLBFLE) [20] jointly learns the binary feature and the codebook simultaneously with a one-stage procedure. However, these methods are susceptible to rotations, which limits their performance and applications.

C. Co-Occurrence Features

Compared with individual occurrence features, co-occurrence features capture the relationship between the related features and provide higher order statistical information. There are several co-occurrence features introduced in recent years, which can be categorized into two classes: holistic co-occurrence features [30], [31] and local co-occurrence features [10], [32]–[35]. Holistic co-occurrence features describe the relationship between visual semantic concepts, such as scenes and objects. For example, Rasiwasia and Vasconcelos [30] exploited the co-occurred natural scenes to model contextual relationships. Yuan *et al.* [31] mined co-occurrence patterns and integrated them through a boosting procedure considering both conjunction and disjunction forms. Instead of capturing related semantic spaces on the whole image, local co-occurrence features are extracted within adjacent local patches. For example, Ito *et al.* [32] proposed co-occurred heterogeneous features to describe various aspects of objects. Yang *et al.* [33] calculated pairwise statistics between ingredients to exploit spatial relationships. Nosaka *et al.* [34] introduced a co-occurrence adjacent local binary pattern (CoALBP) to exploit spatial relations among the adjacent LBPs. However, all the aforementioned local co-occurrence features are sensitive to rotations. More recently, several rotation-invariant local co-occurrence features have been proposed. Qi *et al.* [10] presented a pairwise rotation invariant co-occurrence local binary pattern (PRICoLBP) by using a pairwise transform invariance principle. They also proposed a globally rotation invariant multi-scale co-occurrence local binary pattern (MCLBP) by capturing the correlations among different scales. Unlike most existing co-occurrence features which are hand-crafted, we directly learn triple rotation-invariant co-occurrence local binary descriptors from raw pixels in this work.

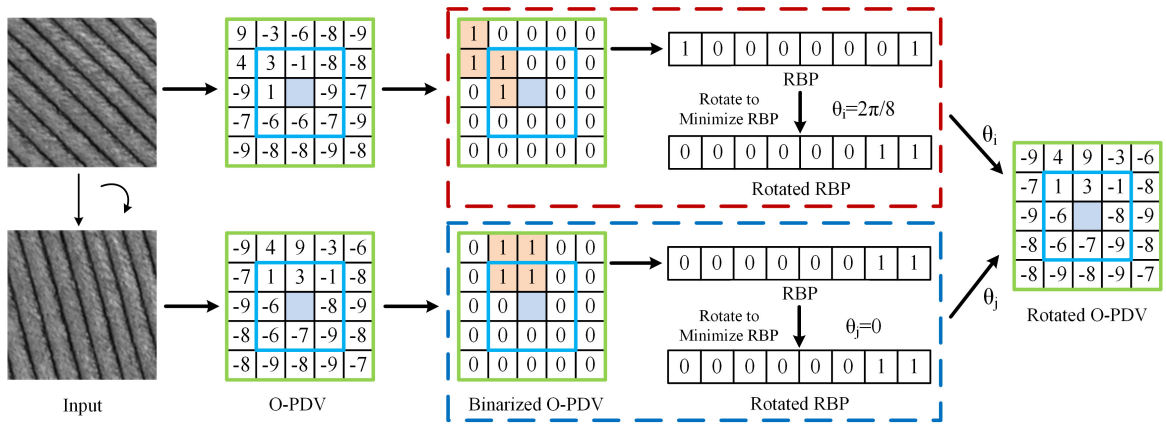


Fig. 2. An example of obtaining rotation invariance, where all the rotational variations of an O-PDV are rotated to the same benchmark orientation with the minimum energy of RBP.

III. Proposed Approach

In this section, we first present the proposed RI-LBD method, and then introduce the extended TRICo-LBD method. Lastly, we propose the feature representation using RI-LBD and TRICo-LBD.

A. RI-LBD Feature Learning

Rotation invariance is one of the most important contributions of the proposed RI-LBD. As the procedure of obtaining rotation invariance is relatively complicated, we first give an example for better illustration. The key idea is to rotate all the rotational variations of an O-PDV to the benchmark orientation, which minimizes the energy of RBP. Fig. 2 illustrates an example of obtaining rotation invariance. For each input image patch, we first compute its O-PDV. Then, through the binarization, we can obtain its RBP which describes the orientation of O-PDV. Finally, we rotate the O-PDV by θ to the benchmark orientation, which minimizes the energy of the corresponding RBP. As shown in Fig. 2, different rotational variations of an O-PDV are rotated to the same benchmark orientation.

In the following, we introduce the steps accordingly, which include ordered pixel difference vector, rotational binary pattern and rotation-invariant local binary feature learning.

1) Ordered Pixel Difference Vector: Firstly, we introduce an ordered pixel difference vector (O-PDV) as the input vector because it encodes lines and edges by measuring the differences between pixels.[1] Fig. 3 illustrates the approach to extract O-PDV. Similar to pixel difference vector (PDV), O-PDV calculates the differences between the central pixel and its neighbouring pixels at first. However, they are concatenated in order of scale and orientation, so that O-PDV is able to represent the rotation of the image patch by separately shifting the difference vector of each scale according to the angle. It is easy to prove that the length of the vector in r-scale is $8 \times r$, and the total length of the O-PDV is $d = 4 \times R \times (R+1)$.

[1]The conventional pixel difference vector (PDV) aligns the differences between central pixel and neighbouring pixels unordered, which cannot represent the rotation of the image.

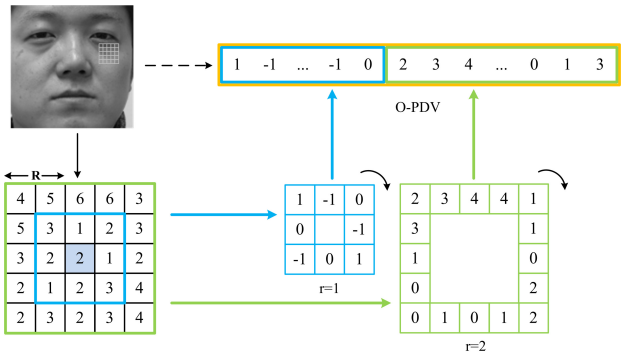


Fig. 3. An illustration of extracting an ordered pixel difference vector (O-PDV) from the original image in our approach. Given any pixel in the image, we first compute the differences between the central pixel and the neighbouring pixels in each scale, respectively. Then, for each scale, a difference vector is aligned clockwisely starting from the top-left. Lastly, difference vectors of different scales are concatenated from small to large into a longer vector, which becomes the O-PDV. For easy illustration, R is set as 2 in this figure.

Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N]$ be the N samples of O-PDVs from the training set. When there is a rotation on the image patch, the pixels with the same radius can be seen to locate on a circle and will move along the perimeter, which leads to a corresponding rotation on each scale of O-PDV respectively as aforementioned. With each patch orientation θ, we construct the corresponding rotation matrix $\mathbf{R}(\theta)$, which transforms the original O-PDV into a new rotated O-PDV according to the rotation on the image patch:

$$\mathbf{x}_n^{\theta} = \mathbf{R}(\theta)\mathbf{x}_n. \tag{1}$$

As $\mathbf{R}(\theta)$ rotates each scale of O-PDV respectively, it is a diagonal matrix with the rotation matrix for each scale $\mathbf{R}_r(\theta) \in \mathbb{R}^{8r \times 8r}$ as its diagonal:

$$\mathbf{R}(\theta) = \mathrm{diag}(\mathbf{R}_1(\theta), \cdots, \mathbf{R}_R(\theta)). \tag{2}$$

For each scale r, a one-bit circular shift corresponds to an angle of $\theta_r = 2\pi/8r$ rotation. Specifically, if the degree of rotation is not the multiple of θ_r, interpolation should be used instead of simple circular shifts. As each image patch

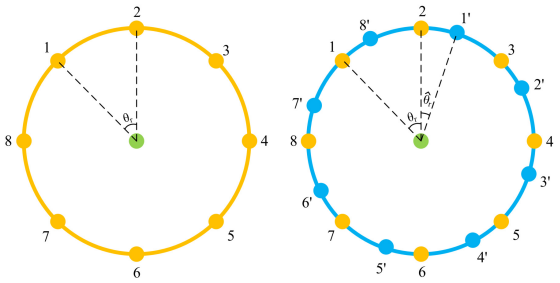


Fig. 4. An illustration of θ_r and $\hat{\theta}_r$. The circle on the left is the original patch, while the right represents the patch after rotation. We can see that $n_r = 1$ in this figure, so that $\theta = \theta_r + \hat{\theta}_r$.

is small, values in each scale can be seen as located on a circle. Therefore, we use two adjacent values in this scale to interpolate the new value after the rotation. Given any angle θ, we firstly calculate the times of one-bit circular shifts n_r and the remaining angle $\hat{\theta}_r$:

$$\theta = n_r\theta_r + \hat{\theta}_r, \tag{3}$$

where $n_r \in \mathbb{N}$ and $0 \leq \hat{\theta}_r < \theta_r$. An illustration of the angle decomposition is shown in Fig. 4.

Then, for the jth value in scale r of the new rotated O-PDV, $(j + n_r)\text{mod}(8 \times r)$th and $(j + n_r + 1)mod\ (8 \times r)$th values are the adjacent values in original O-PDV, where mod is the operation to obtain the least positive remainder in integer division.[2] Therefore, these two values are used to interpolate the new value, whose weights are $1 - \hat{\theta}_r/\theta_r$ and $\hat{\theta}_r/\theta_r$, respectively, and the representation of the rotation matrix for scale r is as follows:

$$R_{r,ij}(\theta) = \begin{cases} 1 - \hat{\theta}_r/\theta_r, & \text{if } i = (j + n_r)mod(8 \times r), \\ \hat{\theta}_r/\theta_r, & \text{if } i = (j + n_r + 1)mod(8 \times r), \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

In order to make the learned binary features rotation-invariant, all rotation variations of an O-PDV should be rotated into the same orientation. Let $x^r_{n,p}$ be the pth number in scale r of $\mathbf{x}_n$, and we define the energy of $\mathbf{x}_n$ as follows, which is highly related to the orientation of the O-PDV:

$$E_x(\mathbf{x}_n) = \sum_{r=1}^{R}\sum_{p=1}^{8r} 0.5 \times (\text{sgn}(x^r_{n,p}) + 1)2^{(1-p)/r}, \tag{5}$$

where $\text{sgn}(x) = 1$ if $x \geq 0$ and -1 otherwise. The binarization is to overcome the noises in the O-PDV, and the weights are designed to be the same for the same orientation of different scales, as shown in Fig. 5.

2) Rotational Binary Pattern: We discretize the orientation as $\Delta\theta = 2\pi/n_\theta$. As there are n_θ rotation variants of each patch, we obtain n_θ energies in total for each O-PDV through (5). We use this n_θ-dimensional energy vector to describe the rotational information of the local patch. It is easy to find that the angle between a local patch and its rotation variants can be estimated through the circular shifts of the energy vectors. In this paper, we fix n_θ to 24 which is large enough for rotation invariance. Then, we study the changing tendency of the energy vectors, and observe that the energies in a vector tend to keep increasing or decreasing with only a few extreme points. It is reasonable because most changes of pixels are smooth in a small patch. In order to exploit such information, we construct a 24-dimensional rotational binary pattern (RBP) $\mathbf{t}_n$ where its values equal to 1 if the energy is increasing at the next orientation and 0 otherwise with the circular assumption, and it is observed that over 90% 24-dimensional RBPs have less than or equal to four 0/1 bitwise shifts, which is called the uniform RBP.[3]

Similarly, we define the energy of RBP:

$$E_t(\mathbf{t}_n) = \sum_{p=1}^{n_\theta} t_{n,p} \times 2^{1-p}, \tag{6}$$

where $t_{n,p}$ is the pth value in RBP $\mathbf{t}_n$.

3) Rotation-Invariant Local Binary Feature Learning: As Fig. 6 illustrates that O-PDV and its RBP share the same angle under image rotation, we design a benchmark for rotational

[2]If the result of mod operation equals to zero, we change it into $8 \times r$.

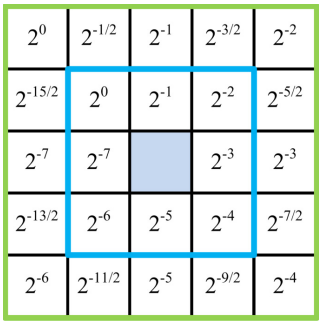


Fig. 5. An illustration of the weights to calculate the energy of O-PDV. The weights are designed to be the same for all the scales when sharing the same orientation.

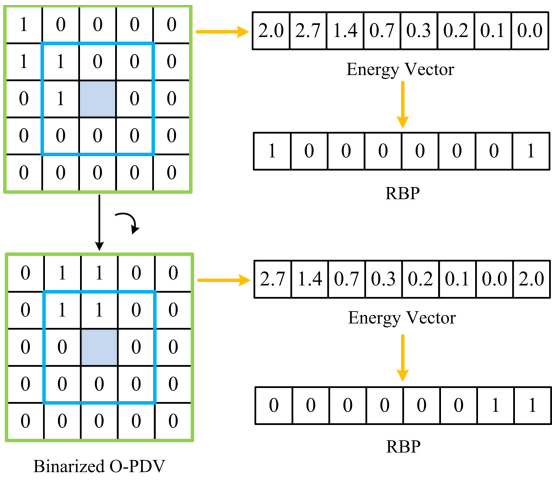


Fig. 6. We binarize and reconstruct the O-PDV into the patch form, and calculate the energy vector and its RBP. n_θ is set as 8 for easy illustration. When the local patch rotates $2\pi/8$ clockwise, both the energy vector and RBP shift accordingly, sharing the same angle.

[3]The number of bitwise shifts includes the changes between the last and the first values. In order to remove the noises, we will manually shift a bit if its both neighbours and at least one sub-neighbours are different from it.

variations of an O-PDV which has the minimum energy of its RBP, so that all rotational variants of an O-PDV share the same benchmark O-PDV. In order to obtain rotation invariance, an obvious idea is to rotate each O-PDV into the benchmark O-PDV at first.

However, the benchmark O-PDVs extracted from the training set may not be able to map into ideal binary codes which are compact and energy-saving. As the projection matrix is learned from millions of patches, for a single patch the benchmark orientation may not be the best option for good binary codes. A better mapping can also be obtained with the optimization of all the orientations. In order to exploit the implicit relationship between the orientations and the mapping, we take the benchmark O-PDVs as the initialization and design a joint learning method by adding the energy of RBPs into the objective function to obtain more data-adaptive orientations. More specifically, we jointly learn a rotational function ψ and mappings $\mathbf{w}_k$, where ψ provides a rotational angle for each RBP, i.e. $\theta_n = \psi(\mathbf{t}_n)$. As over 90% RBPs are uniform, the function ψ rotates each uniform RBP and its rotational variants into the same orientation to minimize the objective function, and simply applies the benchmark orientation to those non-uniform ones.

Let $b_{kn}^{\theta_n} = 0.5 \times (\text{sgn}(\mathbf{w}_k^T \mathbf{x}_n^{\theta_n}) + 1)$, and we can rewrite the objective function as follows:

$$\begin{aligned}\min_{\mathbf{w}_k,\psi} J &= J_1 + \lambda_1 J_2 + \lambda_2 J_3 \\ &= \sum_{n=1}^{N} E_t(\mathbf{t}_n^{\psi(\mathbf{t}_n)}) \\ &\quad + \lambda_1 \sum_{n=1}^{N}\sum_{k=1}^{K} ||(b_{kn}^{\psi(\mathbf{t}_n)} - 0.5) - \mathbf{w}_k^T \mathbf{x}_n^{\psi(\mathbf{t}_n)}||^2 \\ &\quad - \lambda_2 \sum_{n=1}^{N}\sum_{k=1}^{K} ||b_{kn}^{\psi(\mathbf{t}_n)} - \mu_k||^2, \end{aligned} \tag{7}$$

where N is the number of patches extracted from original images, K is the length of each binary feature, μ_k is the mean of the kth bit of all N input vectors, and λ_1 and λ_2 are parameters to balance the weight of different terms.

In (7), J_1 is to minimize the energy of each rotated O-PDV in order to obtain rotation-invariance, J_2 is to reduce the loss of the quantization between the original O-PDV and the learned binary codes, and J_3 is to maximize the variance of the learned binary codes to make them more independent.

Let $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_K] \in \mathbb{R}^{d\times K}$ be the projection matrix, and each rotated O-PDV $\mathbf{x}_n^{\psi(\mathbf{t}_n)}$ is mapped into a binary descriptor as follows:

$$\mathbf{b}_n^{\psi(\mathbf{t}_n)} = 0.5 \times (\text{sgn}(\mathbf{W}^T \mathbf{x}_n^{\psi(\mathbf{t}_n)}) + 1). \tag{8}$$

Then, the objective function (7) can be rewritten as follows:

$$\begin{aligned}\min_{\mathbf{W},\psi} J &= J_1 + \lambda_1 J_2 + \lambda_2 J_3 \\ &= E(\mathbf{X}_\psi) + \lambda_1 ||(\mathbf{B}_\psi - 0.5) - \mathbf{W}^T \mathbf{X}_\psi||_F^2 \\ &\quad - \lambda_2 \text{tr}((\mathbf{B}_\psi - \mathbf{U})^T (\mathbf{B}_\psi - \mathbf{U})), \end{aligned} \tag{9}$$

where $\mathbf{X}_\psi = [\mathbf{x}_1^{\psi(\mathbf{t}_1)}, \mathbf{x}_2^{\psi(\mathbf{t}_2)}, \cdots, \mathbf{x}_N^{\psi(\mathbf{t}_N)}]$ is the N samples of rotated O-PDVs, $E(\mathbf{X}_\psi) = \sum_{n=1}^{N} E(\mathbf{x}_n^{\psi(\mathbf{t}_n)})$ is the total energy of all rotated O-PDVs, $\mathbf{B}_\psi = 0.5 \times (\text{sgn}(\mathbf{W}^T \mathbf{X}_\psi + 1) \in \{0,1\}^{K\times N}$ is the matrix of all binary codes, $\mathbf{U} \in \mathbb{R}^{K\times N}$ is the mean matrix repeating the row vector of the mean of all binary bits.

We relax the non-linear sgn$(\cdot)$ function as its signed magnitude [36], [37] as it makes (9) an NP-hard problem. Therefore, J_3 can be rewritten as follows:

$$\begin{aligned} J_3 &= \text{tr}(\mathbf{W}^T \mathbf{X}_\psi \mathbf{X}_\psi^T \mathbf{W}) - 2 \times \text{tr}(\mathbf{W}^T \mathbf{X}_\psi \mathbf{M}^T \mathbf{W}) \\ &\quad + \text{tr}(\mathbf{W}^T \mathbf{M}\mathbf{M}^T \mathbf{W}), \end{aligned} \tag{10}$$

where $\mathbf{M} \in \mathbb{R}^{d\times N}$ consists of the mean vector of all O-PDVs repeated in rows.

The objective function in (9) is not convex for $\mathbf{W}$, $\mathbf{B}_\psi$ and ψ simultaneously. Therefore, we design the following iterative optimization method to update each of them when others are fixed:

Learning $\mathbf{B}_\psi$ Fixing W and ψ: fixing $\mathbf{W}$ and ψ, we can rewrite the objective function in (9) as follows:

$$\min_{\mathbf{B}_\psi} J(\mathbf{B}_\psi) = ||(\mathbf{B}_\psi - 0.5) - \mathbf{W}^T \mathbf{X}_\psi||_F^2. \tag{11}$$

The solution can be relaxed as follows since $\mathbf{B}_\psi$ is a binary matrix:

$$\mathbf{B}_\psi = 0.5 \times (\text{sgn}(\mathbf{W}^T \mathbf{X}_\psi) + 1). \tag{12}$$

Learning W fixing $\mathbf{B}_\psi$ and ψ: fixing $\mathbf{B}_\psi$ and ψ, the objective function in (9) can be rewritten as follows:

$$\begin{aligned} &\min_{\mathbf{W}} \; J(\mathbf{W}) = \text{tr}(\mathbf{W}^T \mathbf{Q} \mathbf{W}) - 2 \times \lambda_1 \text{tr}((\mathbf{B}_\psi - 0.5)\mathbf{X}_\psi^T \mathbf{W}) \\ &\text{subject to } \mathbf{W}^T \mathbf{W} = \mathbf{I}, \end{aligned} \tag{13}$$

where

$$\mathbf{Q} \triangleq \lambda_1 \mathbf{X}_\psi \mathbf{X}_\psi^T - \lambda_2 \times (\mathbf{X}_\psi \mathbf{X}_\psi^T - 2\mathbf{X}_\psi \mathbf{M}^T + \mathbf{M}\mathbf{M}^T). \tag{14}$$

We utilize the gradient descent method [38] to solve the objective function.

Learning ψ fixing W and $\mathbf{B}_\psi$: fixing $\mathbf{W}$ and $\mathbf{B}_\psi$, we update ψ for different uniform RBPs sequentially. More specifically, we first classify each RBP into a uniform pattern $\mathcal{C}_m$ or a non-uniform pattern $\mathcal{C}$. With the initialization of the benchmark orientation, all rotational variations of an RBP belong to the same pattern. Then, we sequentially learn an orientation for each uniform pattern $\mathcal{C}_m$. We fix the iteration step length to $\Delta\theta = 2\pi/24$, and ψ can be updated as follows:

$$\begin{aligned} \psi(\mathcal{C}_m) = \text{argmin} \sum_{\mathbf{t}_n \in \mathcal{C}_m} \{&J(\psi(\mathbf{t}_n) - \Delta\theta), J(\psi(\mathbf{t}_n)), \\ &J(\psi(\mathbf{t}_n) + \Delta\theta)\}. \end{aligned} \tag{15}$$

Algorithm 1 generalizes the procedure of our RI-LBD method in detail.

B. TRICo-LBD Feature Learning

While the proposed RI-LBD method learns rotation-invariant binary codes, each feature is learned from a single local patch which loses higher order statistical information. As co-occurrence features can exploit such information and show stronger discriminative power, we present a

Algorithm 1 RI-LBD

Input: Training set $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N]$, binary code length K, parameters λ_1 and λ_2, iteration number T, and convergence parameter ϵ

Output: Projection matrix $\mathbf{W}$

for $n = 1, 2, \cdots, N$ **do**
  Initialize $\psi(\mathbf{t}_n) = \arg\min_{\theta} E_t(\mathbf{t}_n^{\theta})$.
end for
Initialize $\mathbf{W}$ as the top K eigenvectors of $\mathbf{X}_{\psi}\mathbf{X}_{\psi}^T$ with the largest eigenvalues
for $t = 1, 2, \cdots, T$ **do**
  Update $\mathbf{B}_{\psi}$ fixing $\mathbf{W}$ and ψ using (12).
  Update $\mathbf{W}$ fixing $\mathbf{B}_{\psi}$ and ψ using (13).
  for $m = 1, 2, \cdots, M$ **do**
    Update $\psi(\mathcal{C}_m)$ with fixed $\mathbf{W}$ and $\mathbf{B}_{\psi}$ using (15).
  end for
  if $|\mathbf{W}^t - \mathbf{W}^{t-1}| < \epsilon$ and $t > 2$ **then**
    break
  end if
end for
return $\mathbf{W}$

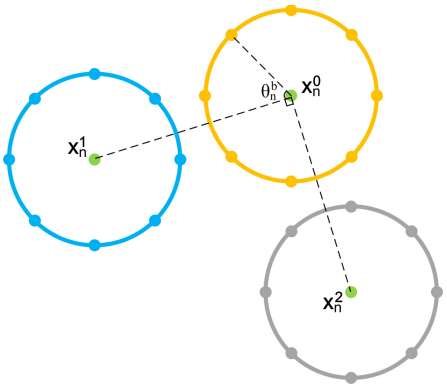


Fig. 7. An illustration of identifying co-occurrence O-PDVs based on the original point $\mathbf{x}_n$ and the orientation θ_n.

triple rotation-invariant co-occurrence local binary descriptor (TRICo-LBD) learning method to address the limitation, where triple adjacent O-PDVs are utilized to describe a single local region.

Suppose $[\mathbf{x}_n^0, \mathbf{x}_n^1, \mathbf{x}_n^2]_{co}$ is an input feature of triple co-occurrence O-PDVs, where $\mathbf{x}_n^0$ is the O-PDV at the original point and $\mathbf{x}_n^1$ and $\mathbf{x}_n^2$ are the target co-occurrence O-PDVs. In order to obtain rotation-invariant co-occurrence features, both global rotation invariance and local rotation invariance are required.

Global rotation invariance claims that under any image rotation, the same target points should be selected for a specific local patch, which needs a unique identification method for the targets. Let θ_n^b be the orientation of $\mathbf{x}_n^0$ to minimize the energy of its RBP, and then we identify the direction from $\mathbf{x}_n^1$ and $\mathbf{x}_n^2$ to $\mathbf{x}_n^0$ as θ_n and $\theta_n + \pi/2$, in order to obtain more information form the orthogonal aspect. Fig. 7 shows the method of identifying co-occurrence O-PDVs. In order to make the positions of co-occurrence O-PDVs on the grid, we discretize the orientation as $\Delta\theta = 2\pi/24$, and select the distance between the target O-PDVs and the original O-PDV as 3 pixels in infinite norm.

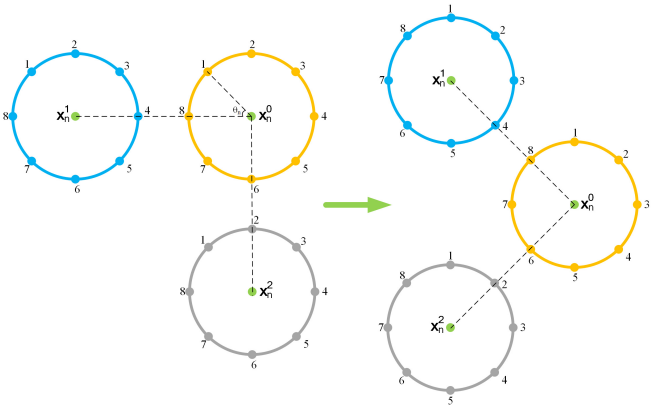


Fig. 8. An illustration of local rotation invariance. Given any orientation θ_n for the original O-PDV $\mathbf{x}_n^0$, the target O-PDVs $\mathbf{x}_n^1$ and $\mathbf{x}_n^2$ share the same orientation to maintain the correct relationship between the original O-PDV and the target O-PDVs. As the original O-PDV is rotation-invariant with the learned orientation, we can also yield the same triple co-occurrence feature in spite of the rotations.

Local rotation invariance requires the same rotation angle for all triple occurrence O-PDVs, as illustrated in Fig. 8. Therefore, we share the learned orientation for each local patch with all the triple co-occurrence O-PDVs and obtain rotation invariance for the triple co-occurrence feature.

Similarly, the objective function of learning TRICo-LBD is formulated:

$$\begin{aligned}\min_{\mathbf{W}_c, \psi} G &= J_1(\psi) + \sum_{c=0}^{2}(\lambda_1 J_2(\mathbf{W}_c, \psi) + \lambda_2 J_3(\mathbf{W}_c, \psi)) \\ &= E(\mathbf{X}_{\psi}^0) + \sum_{c=0}^{2}(\lambda_1 ||(\mathbf{B}_{\psi}^c - 0.5) - \mathbf{W}_c^T \mathbf{X}_{\psi}^c||_F^2 \\ &\quad - \lambda_2 \mathrm{tr}((\mathbf{B}_{\psi}^c - \mathbf{U}_c)^T(\mathbf{B}_{\psi}^c - \mathbf{U}_c))), \end{aligned} \tag{16}$$

where $\mathbf{X}_{\psi}^c = [(\mathbf{x}_1^c)^{\psi(\mathbf{t}_1^0)}, (\mathbf{x}_2^c)^{\psi(\mathbf{t}_2^0)}, \cdots, (\mathbf{x}_N^c)^{\psi(\mathbf{t}_N^0)}]$ is the N samples of rotated co-occurrence O-PDVs, $\mathbf{B}_{\psi}^c$ is the matrix of all co-occurrence binary codes, $\mathbf{U}_c$ is the mean matrix of co-occurrence binary bits.

Similarly, we also use the iterative optimization method to update each with the others fixed to learn $\mathbf{W}_c$, ψ and $\mathbf{B}_{\psi}^c$.

Learning $\mathbf{B}_{\psi}^c$ fixing $\mathbf{W}_c$ and ψ: fixing $\mathbf{W}_c$ and ψ, the objective function in (16) can be rewritten as follows:

$$\min_{\mathbf{B}_{\psi}^c} J(\mathbf{B}_{\psi}) = ||(\mathbf{B}_{\psi}^c - 0.5) - \mathbf{W}_c^T \mathbf{X}_{\psi}^c||_F^2. \tag{17}$$

The solution can be relaxed as:

$$\mathbf{B}_{\psi}^c = 0.5 \times \ (\mathrm{sgn}(\mathbf{W}_c^T \mathbf{X}_{\psi}^c) + 1). \tag{18}$$

Learning $\mathbf{W}_c$ fixing $\mathbf{B}_{\psi}^c$ and ψ: fixing $\mathbf{B}_{\psi}^c$ and ψ, the objective function in (16) can be rewritten as follows:

$$\begin{aligned}\min_{\mathbf{W}_c} \ & J(\mathbf{W}_c) = \mathrm{tr}(\mathbf{W}_c^T \mathbf{Q}_c \mathbf{W}_c) \\ & \quad - 2 \times \lambda_1 \mathrm{tr}((\mathbf{B}_{\psi}^c - 0.5)(\mathbf{X}_{\psi}^c)^T \mathbf{W}_c) \\ \text{subject to } & \mathbf{W}_c^T \mathbf{W}_c = \mathbf{I}, \end{aligned} \tag{19}$$

Algorithm 2 TRICo-LBD

Input: Training set $\mathbf{X}_c = [\mathbf{x}_1^c, \mathbf{x}_2^c, \cdots, \mathbf{x}_N^c]$, binary code length K, parameters λ_1 and λ_2, iteration number T, and convergence parameter ϵ

Output: Projection matrix $\mathbf{W}_c$

for $n = 1, 2, \cdots, N$ **do**
  Initialize $\psi(\mathbf{t}_n) = \arg\min_{\theta} E_t(\mathbf{t}_n^{\theta})$.
end for
Initialize $\mathbf{W}_c$ as the top K eigenvectors of $\mathbf{X}_{\psi}^c(\mathbf{X}_{\psi}^c)^T$ with the largest eigenvalues
for $t = 1, 2, \cdots, T$ **do**
  Update $\mathbf{B}_{\psi}^c$ fixing $\mathbf{W}_c$ and ψ using (18).
  Update $\mathbf{W}_c$ fixing $\mathbf{B}_{\psi}^c$ and ψ using (19).
  for $m = 1, 2, \cdots, M$ **do**
    Update $\psi(\mathcal{C}_m)$ with fixed $\mathbf{W}_c$ and $\mathbf{B}_{\psi}^c$ using (21).
  end for
  if $|\mathbf{W}^t - \mathbf{W}^{t-1}| < \epsilon$ and $t > 2$ **then**
    break
  end if
end for
return $\mathbf{W}$

where

$$\mathbf{Q}_c \triangleq \lambda_1 \mathbf{X}_{\psi}^c(\mathbf{X}_{\psi}^c)^T - \lambda_2 \times (\mathbf{X}_{\psi}^c(\mathbf{X}_{\psi}^c)^T - 2\mathbf{X}_{\psi}^c\mathbf{M}_c^T + \mathbf{M}_c\mathbf{M}_c^T). \quad (20)$$

In (20), $\mathbf{M}_c$ represents the mean matrix of $\mathbf{X}_c$ repeated in rows. We also utilize the gradient descent method to solve $\mathbf{W}_c$.

Learning ψ fixing $\mathbf{W}_c$ and $\mathbf{B}_{\psi}^c$: fixing $\mathbf{W}$ and $\mathbf{B}_{\psi}^c$, we also sequentially learn an orientation for each uniform pattern $\mathcal{C}_m$:

$$\psi(\mathcal{C}_m) = \text{argmin} \sum_{\mathbf{t}_n \in \mathcal{C}_m} \{J(\psi(\mathbf{t}_n^0) - \Delta\theta), J(\psi(\mathbf{t}_n^0)), J(\psi(\mathbf{t}_n^0) + \Delta\theta)\}. \quad (21)$$

Algorithm 2 generalizes the procedure of our TRICo-LBD method in detail.

C. Feature Representation Based on RI-LBD and TRICo-LBD

For RI-LBD, we first rotate each O-PDV from the test set to minimize the energy. Then, the rotated O-PDV is projected into a low-dimensional binary vector using the projection matrix $\mathbf{W}$. With a codebook learned from the training set by an unsupervised clustering method,[4] all binary codes are represented as a histogram feature, which is the final representation. Fig. 9 illustrates the approach of feature representation based on RI-LBD.[5]

For TRICo-LBD, the orientation for each co-occurrence O-PDV from the test set is determined by minimizing the energy of the original O-PDV. Similarly, with the learned projection matrices and the codebook, each test image is finally represented as a histogram feature.

[4]We use K-means method to learn the dictionary for simplicity.

[5]The operation of division is only executed for the images with good alignments, otherwise the image is managed as a whole.

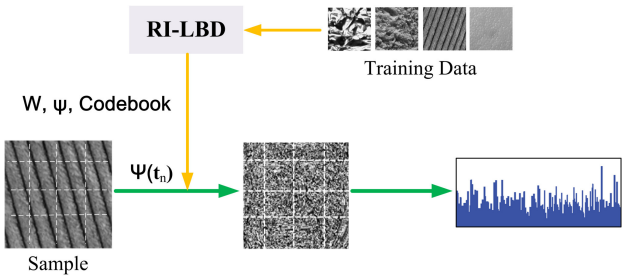


Fig. 9. The flow-chart of feature representation approach based on RI-LBD. We first divide each training image into several non-overlapped regions, and jointly learn the rotational function ψ and the feature mapping $\mathbf{W}$ to project each image patch into binary codes. Then, a codebook is learned for each region. For each sample image, we first rotate each image patch with the learned rotational function, and then utilize the learned projection matrix and the codebook to obtain histogram feature for each block. Lastly, The histogram features are concatenate into a longer feature for the final representation.

IV. Discussion

A. Advantages of Joint Learning

In our work, we jointly learn the rotational function for each RBP and the projection matrix in order to better exploit the implicit relationship between the orientations and the projection matrix. Compared with the two-stage method which aligns the orientation first and then learns the projection matrix, there are two key reasons for the proposed joint learning method. 1) The benchmark orientation which minimizes the energy of RBP is not robust in a small range. If we only consider the pixels from a specific scale, the benchmark orientation for the scale is fixed due to the observation of the uniform pattern of LBP. However, these orientations from different scales are similar but possibly different, which leads to an uncertainty of the benchmark orientation for the local patch. Therefore, the orientations which are close to the benchmark orientation should be considered, and we need to learn for a better one. 2) The benchmark patches may not be able to be projected into discriminative binary codes which are compact and energy saving. As the projection matrix is learned from millions of patches, for a single patch the benchmark orientation may not be the best option for good binary codes; a better mapping can also be obtained with the optimization of all the orientations.

B. Comparison of LBP and RBP

Both LBP [1] and RBP are binary patterns where most of them are uniform, yet there are two main differences between them. 1) LBP describes the pixels on a circle of radius from the central pixel and the uniform pattern only exists on a specific scale, while RBP represents the whole local patch. 2) Each bit of LBP only represents a single pixel, while a RBP bit contains the rotational information of the local patch.

C. Comparison of Canonical Direction and Energy Minimization

Both canonical direction and the proposed energy minimization methods are effective approaches to obtain rotation invariance, where RLBP [19] and SIFT [39] are two conventional local feature methods with canonical direction. RLBP defines the dominant direction through the index of the

TABLE I

95% ERROR RATES (ERR) (%) COMPARED WITH THE STATE-OF-THE-ART BINARY DESCRIPTORS ON BROWN DATASET, WHERE BOOSTED SSC, BRISK, ORB AND BRIEF ARE UNSUPERVISED BINARY FEATURE AND LDAHASH AND D-BRIEF ARE SUPERVISED. THE FAMOUS REAL-VALUED FEATURE SIFT IS PROVIDED FOR REFERENCE

Train Test	Yosemite Notre Dame	Yosemite Liberty	Notre Dame Yosemite	Notre Dame Liberty	Liberty Notre Dame	Liberty Yosemite	Average ERR
SIFT [39] (128 bytes)	28.09	36.27	29.15	36.27	28.09	29.15	31.17
Boosted SSC [40] (16 bytes)	72.20	71.59	76.00	70.35	72.95	77.99	73.51
BRISK [12] (64 bytes)	74.88	79.36	73.21	79.36	74.88	73.21	75.81
ORB [13] (32 bytes)	54.57	59.15	54.96	59.15	54.57	54.96	56.23
BRIEF [11] (32 bytes)	54.57	59.15	54.96	59.15	54.57	54.96	56.23
LDAHash [21] (16 bytes)	51.58	49.66	52.95	49.66	51.58	52.95	51.40
D-BRIEF [22] (4 bytes)	43.96	53.39	46.22	51.30	43.10	47.29	47.54
RI-LBD (3 bytes)	53.86	56.88	53.47	55.34	52.98	54.27	54.47

neighbour with the maximum difference between the central pixel, and SIFT applies a more complicated procedure. In fact, the proposed energy minimization approach can also be seen as a simple canonical direction method, like RLBP. Instead of simply taking the neighbour with the maximum difference as the canonical direction which is susceptible to illumination, the canonical direction of RI-LBD can be defined as the orientation where pixels in the clockwise direction are smaller than the central pixel and the ones in the counter-clockwise direction are larger. Then, we rotate this canonical orientation into the top left corner with the largest weight which gradually descends clockwisely. Compared with SIFT which suffers from heavy computational cost and works on a few detected keypoints, the energy minimization method only needs vector multiplications, which is more suitable for the dense sampling situation.

V. EXPERIMENTS

We evaluate our RI-LBD and TRICo-LBD on four different visual recognition tasks including image patch matching, texture classification, face recognition and scene classification to show the effectiveness of the proposed method.

For image patch matching, we used a single learned binary vector to represent a image patch instead of a histogram feature to directly show the efficiency of the learned binary codes. For the other three applications, we compared our method with two local binary feature representation methods including LBP [1] and CBFD [2], where LBP is a widely-used binary feature and CBFD is a state-of-the-art binary code learning method. We also compared our methods with the state-of-the-art rotation-invariant local binary descriptor PRICoLBP [10] on texture classification and scene classification. Moreover, in order to evaluate the effectiveness of joint learning, we compared the two-stage learning method with the proposed joint learning method in each dataset. The two-stage scheme rotates each image patch to minimize its RBP at first (as a preprocessing) and then learns or tests with the rotated local patches. Without joint learning, it may lose the implicit relationship between the orientation and the mapping.

A. Image Patch Matching

In this section, we evaluate the proposed RI-LBD on the Brown dataset [41], which includes Liberty, Notre Dame and Yosemite and each contains more than 400,000 image patches. For each dataset, there are 20,000 training pairs and 10,000 test pairs where half of them are matched pairs and the others are mismatched pairs.

In the experiments, we simply employed a single binary code to describe a image patch instead of the histogram feature to evaluate the effectiveness of the learned binary descriptor. More specifically, we applied a relatively large R so that each O-PDV could cover a whole image patch, and each image patch was projected into a binary code as the feature representation. We compared the proposed RI-LBD with several conventional features including real-valued descriptor SIFT [39], unsupervised binary descriptors Boosted SSC [40], BRIEF [11], ORB [13] and BRISK [12], and supervised binary descriptors LDA-HASH [21] and D-BRIEF [22]. Three parameters λ_1, λ_2 and binary code length K were fixed to 0.001, 0.01 and 24. As the Brown dataset suffers from small rotational variants, we simply learned the rotational function ψ without initialization.

Table I shows the 95% error rates (ERR) of the Brown dataset, and we can see that our RI-LBD achieves 54.47% error rate when the recall rate is 95%. As an unsupervised binary code learning method, the proposed RI-LBD outperforms all state-of-the-art unsupervised binary descriptors, with over 10 times shorter binary code length. Also, it obtains comparable results with the supervised method LDAHash with more than 5 times smaller storage space.

B. Texture Classification

In this section, we evaluate the proposed RI-LBD and TRICo-LBD on three widely used texture databases, including the Brodatz album [42], the KTH-TIPS [43] dataset, the CUReT [44] dataset and the Outex_TC12 dataset [45]. The Brodatz album is a well-known benchmark dataset for texture classification, which contains 111 classes with 9 images for each class. We used 1 image in each class as the training set for feature learning, 3 as the gallery set and 5 as the probe set. The KTH-TIPS dataset contains 10 texture classes, and each class consists of 81 samples which are captured under nine scales, three different poses and three different illumination directions. In each category, we utilized 10 samples as the training set, 40 as the gallery set and 31 as the probe set. For the CUReT dataset, we used the same subset as [10], [46], [47] which

TABLE II

ACCURACY (%) OF THE KTH-TIPS DATASET VERSUS VARYING λ_1 AND λ_2

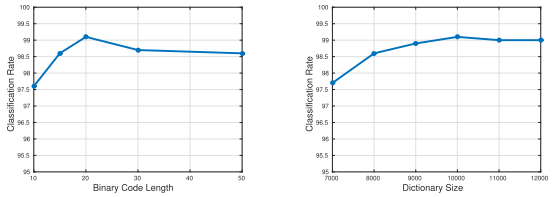
Parameters		Acc
$\lambda_1 = 0.01$	$\lambda_2 = 10$	98.6
$\lambda_1 = 0.01$	$\lambda_2 = 1$	99.0
$\lambda_1 = 0.01$	$\lambda_2 = 0.1$	98.5
$\lambda_1 = 0.001$	$\lambda_2 = 10$	98.6
$\lambda_1 = 0.001$	$\lambda_2 = 1$	**99.1**
$\lambda_1 = 0.001$	$\lambda_2 = 0.1$	98.3
$\lambda_1 = 0.0001$	$\lambda_2 = 10$	97.5
$\lambda_1 = 0.0001$	$\lambda_2 = 1$	98.3
$\lambda_1 = 0.0001$	$\lambda_2 = 0.1$	98.6

includes 61 classes with 92 images for each category, and 9 of them were set as the training set, 46 of them as the gallery set and 37 of them as the probe set. The Outex_TC12 dataset is a popular benchmark for rotation invariance evaluation, which includes 24 classes of textures, varying from illuminations and rotations. Both Outex_TC12_000 and Outex_TC12_001 contain 200 samples per category, where we utilized 10 samples as the training set, 20 samples as the gallery set and 170 samples as the probe set.

1) Parameter Analysis: In RI-LBD, each O-PDV was mapped into a K-bit rotation-invariant binary descriptor with the learned rotational function ψ and projection $\mathbf{W}$, which then encoded into histogram representation with the codebook. We first tested the classification rate of KTH-TIPS with random sampling, and then applied these parameters on KTH-TIPS with another sampling as well as other experiments. In our experiments, neighbourhood radius size R was set as 3 to establish a 48-dimensional O-PDV for each pixel, and we examined the classification accuracy with different λ_1 and λ_2 by fixing the binary code length K as 20 and the dictionary size as 10000. We also applied the whitened PCA (WPCA) method to reduce the dimension of the feature into 50 to reduce the redundancy, and we used the SVM with RBF kernel for the texture classification. The preserved dimension is relatively low due to the repetitive pattern of the texture image. Tab. II shows the results versus λ_1 and λ_2 , and they were selected as 0.001 and 1, respectively. Moreover, when the parameters λ_1 and λ_2 are very large, the approach degenerates into the two-stage scheme, where the rotational function ψ is useless in the situation. Through Tab. II we can observe that the classification rate decreases when λ_1 and λ_2 are relatively large, which illustrates the effectiveness of joint learning. The direct comparison of the proposed joint learning method and the two-stage scheme is presented in all the following experiments.

Then, we tested the binary code length K with the dictionary size fixed as 10,000, and Fig. 10 (a) shows that the best result was obtained with the binary length set as 20. Similarly, Fig. 10 (b) shows that the best dictionary size was 10,000 with the binary code length fixed to 20.

For TRICo-LBD, λ_1 and λ_2 and the dictionary size are the same as RI-LBD, and the code length for each co-occurred O-PDV was 15 so that the total binary feature length K is 45.



(a) (b)

Fig. 10. Classification rates (%) of the KTH-TIPS dataset versus different (a) binary code length (bit) and (b) dictionary size.

TABLE III

TEXTURE CLASSIFICATION RESULTS (%) OF DIFFERENT METHODS ON BRODATZ, KTH-TIPS AND CUReT DATABASES

Method	Brodatz	KTH-TIPS	CUReT
LBP [1]	91.6	92.2	96.3
CBFD [2]	96.3	99.0	98.3
CLBP-S/M/C [48]	94.8	98.4	**98.9**
LBPHF-S-M [18]	95.3	97.8	97.6
CoALBP [34]	94.2	97.0	98.0
LBPV [16]	93.8	95.5	94.0
MSLBP [6]	91.6	92.2	96.3
Liu *et al.* [49]	94.2	–	98.5
VZ-Patch [46]	92.9	92.4	98.0
Caputo *et al.* [47]	95.0	94.8	98.5
Lazebnik *et al.* [7]	88.2	91.3	72.5
PRICoLBP [10]	96.9	98.4	98.4
RI-LBD (Two-Stage)	95.2	98.3	96.7
RI-LBD (Joint)	97.8	**99.3**	98.6
TRICo-LBD	**98.1**	**99.3**	**98.9**

2) Comparison With the State-of-the-Art Methods: Table III tabulates the classification rates of the proposed RI-LBD and the state-of-the-art methods. The methods for comparison are selected from the recently published papers which followed the same protocol for fair comparison. We can observe that there are few methods which perform well on all the three widely used texture databases, yet RI-LBD achieves the best results on Brodatz album and KTH-TIPS database, and obtains very competitive performance on CUReT database. Moreover, the error rate is reduced to less than half on KTH-TIPS dataset (from 1.6% to 0.7%), which is significant improvement. RI-LBD outperforms all LBP variants such as MSLBP, CoALBP, LBPV, LBPHF-S, LBPHF-S-M and PRICoLBP because it is a learning-based method which is more data-adaptive and delivers stronger discriminative power, and overcomes some of the bag-of-words methods, which proves its effectiveness. For TRICo-LBD, the performance is a little higher than RI-LBD because it exploits higher order co-occurred information. As texture images suffer from severe repetitive patterns, the advantage of co-occurred information is not significant.

Table IV shows the experimental results of different local binary descriptors on Outex_TC12. Following [55], we evaluated the average accuracy of Outex_TC12_000 and Outex_TC12_001 under varying neighborhood sizes. In Table IV, CLBP [48], CLBC [53], BRINT [54], LBPV [16] and MRELBP [8] are the combined descriptors of LBP and other complementary features, while the others are single

TABLE IV

TEXTURE CLASSIFICATION RESULTS (%) OF DIFFERENT LOCAL BINARY DESCRIPTORS ON OUTEX_TC12 (OUTEX_TC12_000 AND OUTEX_TC12_001) UNDER VARYING NEIGHBORHOOD SIZES. THE RESULTS ARE THE AVERAGE SCORES OF OUTEX_TC12_000 AND OUTEX_TC12_001

Method	Outex_TC12		
	5×5	7×7	9×9
LBP [6]	82.07	86.79	89.64
CBFD [2]	63.27	68.84	72.66
LTP [15]	86.46	90.88	92.08
RLBP [50]	83.45	88.01	90.73
NTLBP [51]	82.21	88.28	91.61
DLBP [52]	74.11	81.24	85.53
CLBP [48]	94.48	95.67	95.78
CLBC [53]	93.98	95.17	95.75
BRINT [54]	94.29	96.28	97.16
LBPV [16]	89.59	94.16	95.80
MRELBP [8]	**96.24**	-	**99.03**
RI-LBD	90.06	92.85	94.12
RI-LBD (Combine)	95.04	**96.90**	97.77
TRICo-LBD	92.38	94.96	95.63

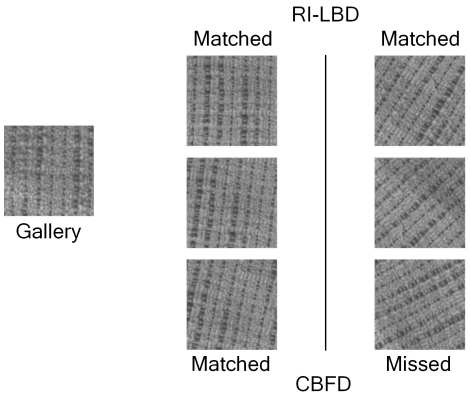


Fig. 11. Some classification examples of the Outex_TC12 dataset. All six probe images are correctly matched for RI-LBD, while the ones of large rotations are missed by CBFD.

local binary descriptors. We observe that the proposed RI-LBD and TRICo-LBD outperform other single binary descriptors, especially for more than 20% improvement on accuracy compared with CBFD. CBFD is an effective local binary descriptor learning method, which obtains nearly 100% classification rate for the part of probe images with small rotations. However, the performance of CBFD suffers from severe deterioration with larger rotations, because the learned binary codes for the rotation variations of a local patch are totally different. On the contrary, RI-LBD learns the orientation for each local patch, where rotation variations are rotated to the same orientation and achieve rotation invariance. Fig. 11 shows some classification examples on Outex_TC12. We can observe that the proposed RI-LBD classifies the correct texture pattern even under large rotations. We combined the proposed RI-LBD with LBP to conduct a fair comparison with combined features, which obtains outstanding performance.

In addition, we tested the VGG-16 [56] to evaluate deep learning approaches by finetuning on the Outex_TC12 dataset, which obtains an accuracy of 88.20%. Deep learning presents stronger discriminative power compared with the proposed RI-LBD and TRICo-LBD. However, as CNN features are not robust to rotations, probe images with large rotations are misclassified. Instead, the proposed methods are rotation-invariant, which present better performance.

TABLE V

CLASSIFICATION RATES (%) RI-LBD COMPARISON WITH DIFFERENT METHODS ON THE ROTATED TEXTURE DATABASES

Method	LBP	CBFD	PRICoLBP	RI-LBD
Brodatz	91.6	96.3	96.9	97.8
rot-Brodatz	85.1	84.9	91.8	93.5
ΔAcc	6.5	11.4	5.1	**4.3**
KTH-TIPS	92.2	99.0	98.4	99.3
rot-KTH-TIPS	86.9	87.4	95.6	95.2
ΔAcc	5.3	11.6	**2.8**	**4.1**
CUReT	96.3	98.3	98.4	98.6
rot-CUReT	90.7	83.6	93.9	94.1
ΔAcc	5.6	14.7	**4.5**	**4.5**

TABLE VI

FEATURE DIMENSION AND FEATURE EXTRACTION TIME (ms) AND OF RI-LBD AND TRICO-LBD COMPARED WITH DIFFERENT LOCAL BINARY FEATURES

Method	Feature Dimension	Time
LBP [6]	210	87.2
LTP [15]	420	231.8
RLBP [50]	210	488.6
NTLBP [51]	388	332.3
DLBP [52]	14,150	565.3
CLBP [48]	3,552	127.9
CLBC [53]	4,168	202.9
BRINT [54]	1,296	248.8
MRELBP [8]	800	416.6
RI-LBD	6,000	317.8
TRICo-LBD	10,000	865.4

3) Rotation Invariance: In order to investigate the rotation invariance of the proposed RI-LBD method, we added arbitrary rotation variations on Brodatz, KTH-TIPS and CUReT to establish new rot-Brodatz, rot-KTH-TIPS and rot-CUReT databases. We compared the proposed RI-LBD with LBP, CBFD and PRICoLBP, and experimental results are shown in Table V. We can see that both CBFD and RI-LBD perform well on the original datasets as they only contain small natural rotations. However, the classification rate of CBFD descends heavily on rotated datasets, yet a relatively good performance is still achieved by RI-LBD. The proposed RI-LBD achieves comparable robustness to rotations with PRICoLBP, which proves the effectiveness on rotation invariance of RI-LBD.

4) Computational Time: We designed an experiment on the Outex_TC12 dataset to evaluate computational time, where we set the dictionary sizes for RI-LBD and TRICo-LBD as 6000 and 10000, respectively. The dictionary sizes are set relatively small because the contents of texture images are simple and repetitive. Our hardware configuration comprises of a 2.8-GHz CPU and a 15G RAM. Table VI shows the feature dimension and feature extraction time of different methods. We observe that the computational cost of the

proposed RI-LBD is comparable to other local binary features, and TRICo-LBD is more time consuming to achieve better performance. Though, the feature dimension can be further reduced by applying WPCA, and the computational time is still acceptable.

C. Face Recognition

We compare our RI-LBD and TRICo-LBD methods with several state-of-the-art descriptors on two widely used face databases including LFW [57] and FERET [58]. The followings describe the details and the results.

The LFW dataset [57] consists of 5749 subjects with total 13233 face images, which were captured from the web in wild conditions. We evaluated our RI-LBD method with the unsupervised setting in our expeiment. We followed the standard evaluation protocol on "View 2" dataset [57], including 6000 pairs with half of them matched and the others mismatched. They were divided into 10 folds with 300 positive pairs and 300 negative pairs for each fold. Assuming that the deviation is not too large, each face image was firstly aligned with a conventional 2D affine transformation and then cropped into 128×128 to remove the background information.

The FERET dataset [58] contains 13539 face images of 1565 subjects in different gender, age and ethnicity. We followed the standard FERET evaluation protocol [58] with six subsets including the *training*, *fa*, *fb*, *fc*, *dup1* and *dup2*. According to the provided eye coordinates, all face images were firstly aligned and cropped into 128×128 pixels. We utilized the *training* set for feature learning, the *fa* set as the gallery set, and the others as the probe set.

In our experiments, we first divided face images into 8×8 non-overlapped regions as they were with good alignment. Following the parameter analysis, R was fixed as 3, λ_1 and λ_2 were set as 0.001 and 1 respectively, and binary length K was 20. As images were divided into sub-regions in face recognition, the dictionary size was 600 for each region, so that the final representation of a face image was a 38400-dimensional feature vector ($38400 = 600 \times 8 \times 8$). We directly learned the rotational function for each uniform RBP without initialization, because the rotations were small for aligned face images. For the two-stage scheme, the rotational angle was limited from $-\pi/12$ to $\pi/12$. Lastly, WPCA was used to reduce the feature dimension to 1000, and we applied the nearest neighbour (NN) method for face recognition.

1) Comparison With the State-of-the-Art Methods: Table VII tabulates the mean verification rate and the area under ROC on LFW dataset and Fig. 12 shows the ROC curve of our RI-LBD and TRICo-LBD compared with the state-of-the-art methods. Table VIII shows the rank-one recognition rate on FERET dataset. We see that our methods achieved a very competitive result on LFW, and obtained the best recognition rates on all four subsets of FERET. Although the faces have been already pre-aligned, our RI-LBD and TRICo-LBD outperformed the learning-based local face descriptor such as DFD and CBFD because of the small misalignments. The property of rotation-invariant is further proved in the following subsection. Multi-directional multi-level dual-cross patterns (MDML-DCPs) [64] applies the first derivative of Gaussian operator and exploits both holistic-level and component-level facial features, which presents strong robustness for face description. Compared with the carefully designed state-of-the-art face descriptor MDML-DCPs, the proposed RI-LBD and TRICo-LBD are general local binary descriptor learning methods, which mainly focus on rotation invariance and obtain comparable results with MDML-DCPs on the FERET dataset. PAF delivered an excellent result on LFW dataset, yet it combines local Gabor filters for face representation and it also requires strong prior

TABLE VII

MEAN VERIFICATION RATE (VR) (%) AND AREA UNDER ROC (AUC) (%) COMPARED WITH THE STATE-OF-THE-ART ON LFW

Method	VR	AUC
LBP [1]	69.45	75.47
CBFD [2]	84.21	88.65
POEM [59]	75.22	–
PEM (LBP) [60]	81.10	–
PEM (SIFT) [60]	81.38	–
DFD [28]	84.02	–
High-dim LBP [61]	84.08	–
PAF [62]	–	94.05
RI-LBD (Two-Stage)	82.38	86.54
RI-LBD (Joint)	84.75	90.14
RI-LBD (Combine)	**87.38**	**95.36**
TRICo-LBD	85.93	93.69

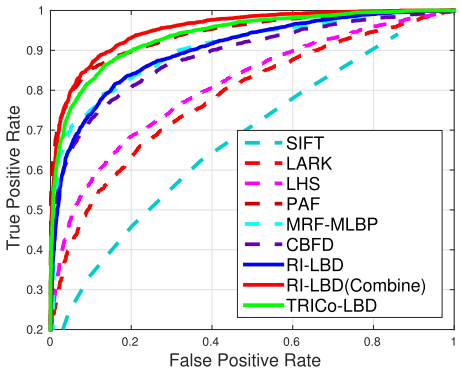


Fig. 12. ROC curves of different methods on the unsupervised setting of LFW database.

TABLE VIII

RANK-ONE RECOGNITION RATES (%) COMPARED WITH THE STATE-OF-THE-ART ON FERET

Method	fb	fc	dup1	dup2
LBP [1]	93.0	51.0	61.0	50.0
CBFD [2]	99.8	**100.0**	93.5	93.2
GV-LBP-TOP [17]	98.4	99.0	82.0	81.6
GV-LBP [17]	98.1	98.5	80.9	81.2
LQP [29]	99.8	94.3	85.5	78.6
POEM [59]	97.0	95.0	77.6	76.2
s-POEM [63]	99.4	**100.0**	91.7	90.2
DFD [28]	99.4	**100.0**	91.8	92.3
MDML-DCPs + WPCA [64]	99.8	**100.0**	**96.1**	**95.7**
RI-LBD (Two-Stage)	99.1	99.4	92.8	93.2
RI-LBD (Joint)	**100.0**	**100.0**	93.9	94.2
TRICo-LBD	**100.0**	**100.0**	**95.4**	**95.2**

TABLE IX
RECOGNITION RATES (%) OF CBFD AND RI-LBD ON THE ROTATED FACE DATABASES

Method	CBFD	RI-LBD
LFW	84.21	84.75
rot-LFW	72.83	81.47
ΔAcc	11.38	**3.28**
FERET (dup1,dup2)	93.5, 93.2	93.9, 94.2
rot-FERET (dup1,dup2)	85.2, 84.7	91.3, 90.7
ΔAcc	8.3, 8.5	**2.6**, **3.5**

TABLE X
CROSS-DATASET EVALUATION ON DIFFERENT TRAINING SETS (%)

Training Set	LFW		FERET			
	VR	AUC	fb	fc	dup1	dup2
LFW	84.75	90.14	99.2	99.8	93.2	93.1
FERET	83.78	88.53	100.0	100.0	93.9	94.2

knowledge to design a pose-adaptive filter. However, our method is an unsupervised local binary feature learning method, which directly learns from raw pixels and does not require such prior knowledge. We evaluated our method combined with Gabor filter, and obtained higher AUC than PAF.

2) Rotation Invariance: We also added arbitrary rotation variations ranging from $-\pi/12$ to $\pi/12$ on LFW and FERET datasets similarly, and the comparison of CBFD and RI-LBD are shown in Table IX. Note that as the rotated face images were not aligned, we directly learned on the whole image without devision. The two recognition rates for FERET dataset represent the performance on *dup1* and *dup2* subsets, respectively. Experimental results illustrate that, the improvement of the proposed RI-LBD is small on the pre-aligned original datasets, but it presents an outstanding performance with the arbitrary rotation on the image, which proves the property of rotation-invariant of the proposed method.

3) Cross-Dataset Evaluation: In real applications, large variations always exist between the training set and the test set due to varying environments. In order to further evaluate the generalization ability of the proposed method, we designed a cross-dataset experiment by utilizing different face databases for training and testing. Firstly, we took the "View 1" subset of LFW as the training set to train the orientation θ_n, the projection $\mathbf{W}$ and the codebook, which were used to evaluate on the FERET dataset. Then, FERET was used for training, and tested on LFW with the unsupervised setting. As the unconstrained face images in LFW greatly differ from the face images in FERET which are captured under controlled condition, this experiment is especially designed to evaluate the efficiency of RI-LBD under different conditions. Table X shows that the best results were achieved by utilizing the same training set and test set, and there was a small decline when they were different. Though, the result of the cross-dataset evaluation is still comparable with the state-of-the-art face descriptors, which proves RI-LBD has a strong generalization ability.

TABLE XI
COMPUTATIONAL TIME (ms) OF THE PROPOSED RI-LBD COMPARED WITH DIFFERENT FEATURES

Method	Feature Dimension	Time
LBP [1]	3,776	22.9
SIFT [39]	8,192	63.7
CBFD [2]	32,000	227.3
RI-LBD	38,400	389.2

4) Computational Time: We compared the computational time with the real-valued feature SIFT and two local binary descriptors LBP and CBFD. Table XI shows the feature dimension and the computational time of our RI-LBD and other methods. In this specific application of face recognition, we have divided each face into 64 subregions to obtain a concatenated descriptor, which is the main reason of the high dimensionality ($38400 = 600 \times 64$). With higher feature dimensions, both CBFD and RI-LBD deliver stronger discriminative power than LBP and SIFT. Moreover, while RI-LBD obtains rotation invariance during the learning procedure, it still achieves comparable efficiency compared with CBFD.

D. Scene Classification

In this section, we evaluate the proposed methods on three widely used scene classification datasets, including Scene-15 [65], MIT Indoor-67 [66] and SUN397 [67]. Scene-15 is a widely used database for scene classification, which contains totally 15 indoor and outdoor categories, which include kitchen, office, bedroom, living room, store, industrial, inside cite, tall building, highway, street, open country, coast, forest, mountain and suburb, with 201 to 410 images per class. In our experiments, we first resized the images with the minimum dimension of 256 pixels, and followed the standard evaluation protocol [65] by randomly selecting 30 images per class as the training set, 100 images as the gallery set and the rest as the probe set. MIT Indoor-67 [66] is a popular indoor scene classification dataset, which contains 16520 images of 67 indoor scenes. We approximately used 30 images per class for training, 50 images as the gallery set and 20 images as the probe set. SUN397 [67] is a large-scale scene dataset, which consists of 108754 images of 394 scene categories, with at least 100 images per class. We trained and tested the proposed approaches on ten partitions, where 50 training images and 50 testing images have been used. We repeated for 10 times and took the average as the classification rate. We set λ_1 and λ_2 as 0.0001 and 0.1, respectively. The binary length was 20, the codebook size was 15000, and the dimension preserved in WPCA was 70. SVM with RBF kernel was used for classification.

Table XII shows the results of different methods on the Scene-15 dataset, where TRICo-LBD obtains the highest classification rate, while PRICoLBP [10] and our RI-LBD achieve comparable performance. The proposed TRICo-LBD exploits spatial co-occurrence patterns, providing higher order statistical information in binary feature description. Compared with simple applications such as texture classification which contains many repetitive patterns, the co-occurred information

TABLE XII

CLASSIFICATION RESULTS (%) ON SCENE-15 DATABASE OF DIFFERENT METHODS

Method	Acc
LBP [1]	47.9 ± 1.1
CBFD [2]	74.2 ± 0.6
16 channel weak features [65]	45.3 ± 0.5
Bag of SIFT (200 Codebooks) [65]	72.2 ± 0.6
Bag of SIFT (800 Codebooks) [65]	74.8 ± 0.3
CENTIST [9]	73.3 ± 1.0
BSC [30]	72.5 ± 0.3
Kernel Codebook [68]	75.6 ± 0.7
PRICoLBP [10]	79.2 ± 0.7
RI-LBD (Two-Stage)	76.8 ± 0.4
RI-LBD (Joint)	78.9 ± 0.6
TRICo-LBD	$\mathbf{82.0 \pm 0.9}$

TABLE XIII

COMPARISON OF ACCURACY (%) WITH DIFFERENT SCENE CLASSIFICATION METHODS ON THE MIT INDOOR-67 DATASET

Method	Accuracy
ROI [66]	26.05
DSFL [69]	52.24
BOP [70]	46.10
IFV [70]	60.77
IFV+BOP [70]	63.10
Mode-Seeking [71]	64.03
Mode-Seeking+IFV [71]	66.87
PlaceNet [72]	68.24
MOP-CNN [73]	68.90
HybridNet [72]	70.80
CFV (VGG-19) [74]	81.00
CS (VGG-19) [75]	**82.24**
RI-LBD	58.73
RI-LBD + DDML	72.98
TRICo-LBD	61.44
TRICo-LBD + DDML	**78.17**

TABLE XIV

COMPARISON OF ACCURACY (%) WITH DIFFERENT SCENE CLASSIFICATION METHODS ON THE SUN397 DATASET

Method	Accuracy
S-manifold [76]	28.90
OTC [77]	34.56
contextBoW+semantic [78]	35.60
Xiao *et al.* [67]	38.00
DeCAF [79]	40.94
MOP-CNN [73]	51.98
HybridNet [72]	53.86
PlacesNet [72]	54.23
RI-LBD	34.97
RI-LBD + DDML	53.82
TRICo-LBD	38.04
TRICo-LBD + DDML	**56.41**

is more effective in complicated scene classification with a 3.1% of improvement.

Table XIII and Table XIV show the experimental results of different scene classification methods on the MIT Indoor-67 dataset and the SUN397 dataset, respectively. Among the listed approaches, DeCAF [79], PlaceNet [72], MOP-CNN [73], HybridNet [72], CFV [74] and CS [75] are deep learning methods, while the others are conventional methods. As unsupervised feature learning methods, the proposed RI-LBD and TRICo-LBD obtain comparable performance with conventional scene classification methods. In order to conduct a fair comparison with the supervised deep learning approaches, we have exploited the label information by applying the discriminative deep metric learning (DDML) [80] method to learn discriminative similarity measure function. We observe that the proposed methods obtain encouraging performance compared with the CNN methods. As RI-LBD and TRICo-LBD still learn binary codes in an unsupervised manner, the performance can be further improved by learning supervised projections.

TABLE XV

RECOGNITION ACCURACY (%) OF RI-LBD AND TRICO-LBD WITH THE SAME PARAMETERS COMPARED WITH THE ORIGINAL ACCURACIES. IN THIS TABLE, RI REPRESENTS RI-LBD AND TRI REPRESENTS TRICO-LBD

Dataset	RI	RI (ori)	TRI	TRI (ori)
Brodatz	97.8	97.8	98.1	98.1
KTH-TIPS	99.3	99.3	99.3	99.3
CUReT	98.6	98.6	98.9	98.9
Outex_TC12	92.9	92.9	95.0	95.0
LFW	83.1	84.8	83.8	85.9
FERET (Average)	95.6	97.0	95.9	97.7
Scene-15	76.3	78.9	80.1	82.0
MIT Indoor-67	55.7	58.7	59.3	61.4
SUN397	34.4	35.0	36.8	38.0

E. Robustness Analysis

In this paper, we conducted experiments on four visual recognition tasks, which include image patch matching, texture classification, face recognition and scene classification. As each visual recognition task presents different properties, we simply fixed the parameters within each application.

In order to evaluate the robustness of the proposed RI-LBD and TRICo-LBD to parameters, we conducted an experiment by fixing parameters for all the tasks. As image patch matching aims to directly match local patches through the learned binary codes, and the other three applications need to recognize holistic images using histogram features, we tested the last three image recognition tasks by fixing the same parameters as the results of the subsection of *Parameter Analysis*. More specifically, we set $R = 3$, $\lambda_1 = 0.001$ and $\lambda_2 = 1$ for both approaches. For RI-LBD, the binary length K and dictionary size was set as 20 and 10000, respectively, and 45 and 15000 for TRICo-LBD.

Table XV shows the experimental results of the fixed approaches compared with the original accuracies. As we applied the parameters used for texture classification, the results remain the same on Brodatz, KTH-TIPS, CUReT and Outex_TC12, while the performance suffers from a reasonable drop on other datasets. For face recognition, the main reason is that we do not divide the facial images into 8×8 regions to take the advantage of face alignment. For scene classification, the codebook size is slightly small for the description of the relatively complicated images.

F. Discussion

The above experiments suggest the following four observations:

1) Our RI-LBD obtains rotation invariance by jointly learning the rotational function ψ for each RBP to minimize the energy and the projection matrix $\mathbf{W}$ to make the learned binary codes compact and energy-saving. Compared with the two-stage scheme, RI-LBD exploits the implicit relationship between the orientation and the mapping, which achieves higher classification accuracy.
2) RI-LBD outperforms most state-of-the-art hand-crafted local binary features on all the proposed applications as it learns binary codes in a data-driven way and presents more properties. Therefore, RI-LBD is more data-adaptive and shows stronger discriminative ability.
3) The proposed RI-LBD shows strong efficiency and generalization ability through the cross-dataset evaluation, which can apply to the applications where large variations exist between the training set and the test set.
4) The proposed TRICo-LBD presents stronger discriminative power because it exploits higher order statistical information to encode the co-occurred patterns. Such information is more effective in complicated applications such as face recognition and scene classification than simple texture classification.

VI. Conclusion

In this paper, we have proposed a rotation-invariant local binary descriptor (RI-LBD) learning method for visual recognition. Specifically, we jointly learn the rotational function for each rotational binary pattern (RBP) and the projection matrix to obtain rotation-invariant binary codes, which can apply to more applications with rotation variations. The proposed RI-LBD method outperforms most state-of-the-art methods on four different applications, which proves the effectiveness of our method. Moreover, we have developed a TRICo-LBD method by exploiting co-occurred patterns to provide higher order statistical information and have obtained large improvement in complicated applications. To further improve the classification ability, it is interesting to apply our methods to deep learning framework in the future.

References

[1] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.

[2] J. Lu, V. E. Liong, X. Zhou, and J. Zhou, "Learning compact binary face descriptor for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 10, pp. 2041–2056, Oct. 2015.

[3] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proc. CVPR*, 2014, pp. 1891–1898.

[4] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Proc. NIPS*, 2014, pp. 1988–1996.

[5] Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," in *Proc. CVPR*, 2015, pp. 2892–2900.

[6] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[7] S. Lazebnik, C. Schmid, and J. Ponce, "A sparse texture representation using local affine regions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1265–1278, Aug. 2005.

[8] L. Liu, S. Lao, P. W. Fieguth, Y. Guo, X. Wang, and M. Pietikäinen, "Median robust extended local binary pattern for texture classification," *IEEE Trans. Image Process.*, vol. 25, no. 3, pp. 1368–1381, Mar. 2016.

[9] J. Wu and J. M. Rehg, "Centrist: A visual descriptor for scene categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1489–1501, Aug. 2011.

[10] X. Qi, R. Xiao, C.-C. Li, Y. Qiao, J. Guo, and X. Tang, "Pairwise rotation invariant co-occurrence local binary pattern," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 11, pp. 2199–2213, Nov. 2014.

[11] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "BRIEF: Binary robust independent elementary features," in *Proc. ECCV*, 2010, pp. 778–792.

[12] S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: Binary robust invariant scalable keypoints," in *Proc. ICCV*, 2011, pp. 2548–2555.

[13] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to sift or surf," in *Proc. ICCV*, 2011, pp. 2564–2571.

[14] A. Alahi, R. Ortiz, and P. Vandergheynst, "FREAK: Fast retina keypoint," in *Proc. CVPR*, 2012, pp. 510–517.

[15] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1635–1650, Jun. 2010.

[16] Z. Guo, L. Zhang, and D. Zhang, "Rotation invariant texture classification using LBP variance (LBPV) with global matching," *Pattern Recognit.*, vol. 43, no. 3, pp. 706–719, 2010.

[17] Z. Lei, S. Liao, M. Pietikäinen, and S. Z. Li, "Face recognition by exploring information jointly in space, scale and orientation," *IEEE Trans. Image Process.*, vol. 20, no. 1, pp. 247–256, Jan. 2011.

[18] G. Zhao, T. Ahonen, J. Matas, and M. Pietikäinen, "Rotation-invariant image and video description with local binary pattern features," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1465–1477, Apr. 2012.

[19] R. Mehta and K. O. Egiazarian, "Rotated local binary pattern (RLBP)—Rotation invariant texture descriptor," in *Proc. ICPRAM*, 2013, pp. 497–502.

[20] J. Lu, V. E. Liong, and J. Zhou, "Simultaneous local binary feature learning and encoding for face recognition," in *Proc. ICCV*, 2015, pp. 3721–3729.

[21] C. Strecha, A. M. Bronstein, M. M. Bronstein, and P. Fua, "LDAHash: Improved matching with smaller descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 66–78, Jan. 2012.

[22] T. Trzcinski and V. Lepetit, "Efficient discriminative projections for compact binary descriptors," in *Proc. ECCV*, 2012, pp. 228–242.

[23] T. Trzcinski, M. Christoudias, P. Fua, and V. Lepetit, "Boosting binary keypoint descriptors," in *Proc. CVPR*, 2013, pp. 2874–2881.

[24] S. Zhang, Q. Tian, Q. Huang, W. Gao, and Y. Rui, "USB: Ultrashort binary descriptor for fast visual matching and retrieval," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3671–3683, Aug. 2014.

[25] V. Balntas, L. Tang, and K. Mikolajczyk, "BOLD—Binary online learned descriptor for efficient image matching," in *Proc. CVPR*, 2015, pp. 2367–2375.

[26] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.

[27] G. B. Huang, "Learning hierarchical representations for face verification with convolutional deep belief networks," in *Proc. CVPR*, 2012, pp. 2518–2525.

[28] Z. Lei, M. Pietikäinen, and S. Z. Li, "Learning discriminant face descriptor," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 289–302, Feb. 2014.

[29] S. U. Hussain, T. Napoléon, and F. Jurie, "Face recognition using local quantized patterns," in *Proc. BMVC*, 2012, pp. 1–12.

[30] N. Rasiwasia and N. Vasconcelos, "Holistic context modeling using semantic co-occurrences," in *Proc. CVPR*, 2009, pp. 1889–1895.

[31] J. Yuan, M. Yang, and Y. Wu, "Mining discriminative co-occurrence patterns for visual recognition," in *Proc. CVPR*, 2011, pp. 2777–2784.

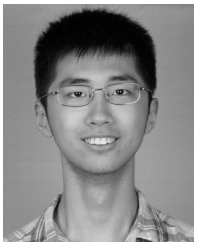
[32] S. Ito and S. Kubota, "Object classification using heterogeneous co-occurrence features," in *Proc. ECCV*, 2010, pp. 701–714.

[33] S. Yang, M. Chen, D. Pomerleau, and R. Sukthankar, "Food recognition using statistics of pairwise local features," in *Proc. CVPR*, 2010, pp. 2249–2256.

[34] R. Nosaka, Y. Ohkawa, and K. Fukui, "Feature extraction based on co-occurrence of adjacent local binary patterns," in *Proc. Adv. Image Video Technol.*, 2011, pp. 82–91.

[35] X. Qi, L. Shen, G. Zhao, Q. Li, and M. Pietikäinen, "Globally rotation invariant multi-scale co-occurrence local binary pattern," *Image Vis. Comput.*, vol. 43, pp. 16–26, Nov. 2015.

[36] Y. Gong and S. Lazebnik, "Iterative quantization: A procrustean approach to learning binary codes," in *Proc. CVPR*, 2011, pp. 817–824.
[37] J. Wang, S. Kumar, and S. F. Chang, "Semi-supervised hashing for scalable image retrieval," in *Proc. CVPR*, 2010, pp. 3424–3431.
[38] Z. Wen and W. Yin, "A feasible method for optimization with orthogonality constraints," *Math. Programm.*, vol. 142, no. 1, pp. 397–434, 2013.
[39] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[40] G. Shakhnarovich, "Learning task-specific similarity," Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., Massachusetts Inst. Technol., Cambridge, MA, USA, 2005.
[41] M. Brown, G. Hua, and S. Winder, "Discriminative learning of local image descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 43–57, Jan. 2011.
[42] P. Brodatz, *Textures: A Photographic Album for Artists and Designers*, vol. 66. New York, NY, USA: Dover, 1966.
[43] E. Hayman, B. Caputo, M. Fritz, and J. O. Eklundh, "On the significance of real-world conditions for material classification," in *Proc. ECCV*, 2004, pp. 253–266.
[44] K. J. Dana, B. van Ginneken, S. K. Nayar, and J. J. Koenderink, "Reflectance and texture of real-world surfaces," *ACM Trans. Graph.*, vol. 18, no. 1, pp. 1–34, Jan. 1999.
[45] T. Ojala, T. Mäenpää, M. Pietikäinen, J. Viertola, J. Kyllonen, and S. Huovinen, "Outex—New framework for empirical evaluation of texture analysis algorithms," in *Proc. ICPR*, vol. 1. 2002, pp. 701–706.
[46] M. Varma and A. Zisserman, "A statistical approach to material classification using image patch exemplars," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 2032–2047, Nov. 2009.
[47] B. Caputo, E. Hayman, M. Fritz, and J.-O. Eklundh, "Classifying materials in the real world," *Image Vis. Comput.*, vol. 28, no. 1, pp. 150–163, 2010.
[48] Z. Guo and D. Zhang, "A completed modeling of local binary pattern operator for texture classification," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1657–1663, Jan. 2010.
[49] L. Liu and P. W. Fieguth, "Texture classification from random features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 3, pp. 574–586, Mar. 2012.
[50] J. Chen, V. Kellokumpu, G. Zhao, and M. Pietikäinen, "RLBP: Robust local binary pattern," in *Proc. BMVC*, 2013, pp. 1–12.
[51] A. Fathi and A. R. Naghsh-Nilchi, "Noise tolerant local binary pattern operator for efficient texture analysis," *Pattern Recognit. Lett.*, vol. 33, no. 9, pp. 1093–1100, Jul. 2012.
[52] S. Liao, M. W. K. Law, and A. C. S. Chung, "Dominant local binary patterns for texture classification," *IEEE Trans. Image Process.*, vol. 18, no. 5, pp. 1107–1118, May 2009.
[53] Y. Zhao, D.-S. Huang, and W. Jia, "Completed local binary count for rotation invariant texture classification," *IEEE Trans. Image Process.*, vol. 21, no. 10, pp. 4492–4497, Oct. 2012.
[54] L. Liu, Y. Long, P. W. Fieguth, S. Lao, and G. Zhao, "BRINT: Binary rotation invariant and noise tolerant texture classification," *IEEE Trans. Image Process.*, vol. 23, no. 7, pp. 3071–3084, Jul. 2014.
[55] L. Liu, P. Fieguth, Y. Guo, X. Wang, and M. Pietikäinen, "Local binary features for texture classification: Taxonomy and experimental study," *Pattern Recognit.*, vol. 62, pp. 135–160, Feb. 2017.
[56] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015, pp. 1–14.
[57] G. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," College Inf. Comput. Sci., Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep. 07-49, 2007.
[58] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090–1104, Oct. 2000.
[59] N.-S. Vu and A. Caplier, "Enhanced patterns of oriented edge magnitudes for face recognition and image matching," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 1352–1365, Mar. 2012.
[60] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang, "Probabilistic elastic matching for pose variant face verification," in *Proc. CVPR*, 2013, pp. 3499–3506.
[61] D. Chen, X. Cao, F. Wen, and J. Sun, "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification," in *Proc. CVPR*, 2013, pp. 3025–3032.
[62] D. Yi, Z. Lei, and S. Z. Li, "Towards pose robust face recognition," in *Proc. CVPR*, 2013, pp. 3539–3545.
[63] N.-S. Vu, "Exploring patterns of gradient orientations and magnitudes for face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 2, pp. 295–304, Feb. 2013.
[64] C. Ding, J. Choi, D. Tao, and L. S. Davis, "Multi-directional multi-level dual-cross patterns for robust face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 3, pp. 518–531, Mar. 2016.
[65] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. CVPR*, 2006, pp. 2169–2178.
[66] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *Proc. CVPR*, 2009, pp. 413–420.
[67] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *Proc. CVPR*, 2010, pp. 3485–3492.
[68] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J.-M. Geusebroek, "Visual word ambiguity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1271–1283, Jul. 2010.
[69] Z. Zuo, G. Wang, B. Shuai, L. Zhao, Q. Yang, and X. Jiang, "Learning discriminative and shareable features for scene classification," in *Proc. ECCV*, 2014, pp. 552–568.
[70] M. Juneja, A. Vedaldi, C. Jawahar, and A. Zisserman, "Blocks that shout: Distinctive parts for scene classification," in *Proc. CVPR*, 2013, pp. 923–930.
[71] C. Doersch, A. Gupta, and A. A. Efros, "Mid-level visual element discovery as discriminative mode seeking," in *Proc. NIPS*, 2013, pp. 494–502.
[72] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proc. NIPS*, 2014, pp. 487–495.
[73] Y. Gong, L. Wang, R. Guo, and S. Lazebnik, "Multi-scale orderless pooling of deep convolutional activation features," in *Proc. ECCV*, 2014, pp. 392–407.
[74] M. Cimpoi, S. Maji, and A. Vedaldi, "Deep filter banks for texture recognition and segmentation," in *Proc. CVPR*, 2015, pp. 3828–3836.
[75] G.-S. Xie, X.-Y. Zhang, S. Yan, and C.-L. Liu, "Hybrid CNN and dictionary-based models for scene recognition and domain adaptation," *IEEE Trans. Circuits Syst. Video Technol.*, Dec. 2015, doi: 10.1109/TCSVT.2015.2511543.
[76] R. Kwitt, N. Vasconcelos, and N. Rasiwasia, "Scene recognition on the semantic manifold," in *Proc. ECCV*, 2012, pp. 359–372.
[77] R. Margolin, L. Zelnik-Manor, and A. Tal, "OTC: A novel local descriptor for scene classification," in *Proc. ECCV*, 2014, pp. 377–391.
[78] Y. Su and F. Jurie, "Improving image classification using semantic attributes," *Int. J. Comput. Vis.*, vol. 100, no. 1, pp. 59–77, 2012.
[79] J. Donahue *et al.*, "DeCAF: A deep convolutional activation feature for generic visual recognition," in *Proc. ICML*, 2014, pp. 647–655.
[80] J. Hu, J. Lu, and Y.-P. Tan, "Discriminative deep metric learning for face verification in the wild," in *Proc. CVPR*, 2014, pp. 1875–1882.



Yueqi Duan received the B.E. degree from the Department of Automation, Tsinghua University, Beijing, China, in 2010, where he is currently pursuing the Ph.D. degree with the Department of Automation. His research interests include visual recognition, feature learning, and binary descriptor.

Jiwen Lu received the B.Eng. degree in mechanical engineering and the M.Eng. degree in electrical engineering from the Xi'an University of Technology, Xi'an, China, in 2003 and 2006, respectively, and the Ph.D. degree in electrical engineering from Nanyang Technological University, Singapore, in 2012. From 2011 to 2015, he was a Research Scientist with the Advanced Digital Sciences Center, Singapore. He is currently an Associate Professor with the Department of Automation, Tsinghua University, Beijing, China. His current research interests include computer vision, pattern recognition, and machine learning. He has authored or co-authored over 150 scientific papers in these areas, where 42 were the IEEE Transactions papers. He was a recipient of the National 1000 Young Talents Plan Program in 2015. He is the Workshop Chair/Special Session Chair/Area Chair for over ten international conferences. He serves as an Associate Editor of the *Pattern Recognition Letters*, the *Neurocomputing*, and the IEEE ACCESS, a Managing Guest Editor of the *Pattern Recognition* and the *Image and Vision Computing*, a Guest Editor of the *Computer Vision and Image Understanding*, and an elected member of the Information Forensics and Security Technical Committee of the IEEE Signal Processing Society.

Jianjiang Feng received the B.S. and Ph.D. degrees from the School of Telecommunication Engineering, Beijing University of Posts and Telecommunications, China, in 2000 and 2007, respectively. From 2008 to 2009, he was a Post-Doctoral Researcher with the PRIP Laboratory, Michigan State University. He is currently an Associate Professor with the Department of Automation, Tsinghua University, Beijing. His research interests include fingerprint recognition and computer vision. He is an Associate Editor of the *Image and Vision Computing*.

Jie Zhou received the B.S. and M.S. degrees from the Department of Mathematics, Nankai University, Tianjin, China, in 1990 and 1992, respectively, and the Ph.D. degree from the Institute of Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology, Wuhan, China, in 1995. From 1995 to 1997, he served as a Post-Doctoral Fellow with the Department of Automation, Tsinghua University, Beijing, China, where he has been a Full Professor since 2003. He has authored over 100 papers in peer-reviewed journals and conferences. Among them, over 40 papers have been published in top journals and conferences, such as the IEEE PAMI, TIP, and CVPR. His current research interests include computer vision, pattern recognition, and image processing. He received the National Outstanding Youth Foundation of China Award. He is an Associate Editor of the *International Journal of Robotics and Automation* and two other journals.