Local Pattern Collocations Using Regional Co-occurrence Factorization

Qin Zou, *Member, IEEE*, Lihao Ni, Qian Wang, *Member, IEEE*, Zhongwen Hu, Qingquan Li, and Song Wang, *Senior Member, IEEE*

***Abstract*—Human vision benefits a lot from pattern collocations in visual activities such as object detection and recognition. Usually, pattern collocations display as the co-occurrences of visual primitives, e.g., colors, gradients, or textons, in neighboring regions. In the past two decades, many sophisticated local feature descriptors have been developed to describe visual primitives, and some of them even take into account the co-occurrence information for improving their discriminative power. However, most of these descriptors only consider feature co-occurrence within a very small neighborhood, e.g., 8-connected or 16-connected area, which would fall short in describing pattern collocations built up by feature co-occurrences in a wider neighborhood. In this paper, we propose to describe local pattern collocations by using a new and general regional co-occurrence approach. In this approach, an input image is first partitioned into a set of homogeneous superpixels. Then, features in each superpixel are extracted by a variety of local feature descriptors, based on which a number of patterns are computed. Finally, pattern co-occurrences within the superpixel and between the neighboring superpixels are calculated and factorized into a final descriptor for local pattern collocation. The proposed regional co-occurrence framework is extensively tested on a wide range of popular shape, color, and texture descriptors in terms of image and object categorizations. The experimental results have shown significant performance improvements by using the proposed framework over the existing popular descriptors.**

***Index Terms*—Co-occurrence matrix, feature descriptor, object recognition, painting classification, pattern collocation.**

I. Introduction

OVER the past decade, local feature descriptors have made significant success in addressing many problems in multimedia processing, i.e. image classification [1], [2], object categorization [3]–[5], and image/video analysis [6]–[8], etc. There are mainly three kinds of local feature descriptors – shape, color and texture descriptors. After extracting a set of local features in an image, a feature encoding step is usually applied to cast them into a fixed-length feature vector for image representation [9]. Two widely used feature encoding techniques are the Bag-of-Words (BoW) [10] and Fisher Vector (FV) [11].

BoW maps each local feature into a word in a pre-organized visual vocabulary, and counts the occurrence of each word to form a feature histogram. FV uses the linear combination of a number of Gaussian models to describe the distribution of local features in an image. In both encoding techniques, individual local features in the image are supposed to be independent and only their occurrences are considered. Consequently, the resulting feature vector does not take into account any spatial information of local descriptors and finally undermines its discriminative power in image classification or categorization.

A local feature computed at a specific position intrinsically carries a spatial tag, which enables the construction of local patterns. Similarly, local patterns carrying spatial tags may construct some higher-level patterns [3], [12]–[16], which we call local pattern collocations. The local pattern collocations are generally considered as providing higher-level information than the individual local patterns in image representation. This is similar to the human vision, where the cue of pattern collocation can substantially facilitate humans to detect or recognize many objects, e.g., the sunflowers which bloom into a brown pistil with yellow petals, the pandas which are born with adjacent black and white furs, and the human faces which have eyebrows grown at right above the eyes.

Many efforts have been made to describe pattern collocations in an image by following the clue of co-occurrence [17], [18]. In [17], a pairwise co-occurrence LBP [19] descriptor was proposed, where meaningful LBP pairs were defined based on a statistic analysis of LBPs within their 8- or 16- connected neighborhoods. The co-occurrences of the related LBP pairs were successfully used for texture description. In [18], color co-occurrence was analyzed among 8-connected neighborhoods. The color of each pixel was mapped into a color code according to a pre-computed color codebook, and the co-occurrences of color codes in the 8-connected neighborhoods are put together into a co-occurrence matrix. The effectiveness of the co-occurrence property was validated in image categorization by the above descriptors. However, these descriptors consider

Manuscript received March 22, 2016; revised July 17, 2016; accepted October 13, 2016. Date of publication October 24, 2016; date of current version February 14, 2017. This work was supported by the National Basic Research Program of China under Grant 2012CB725303, and the National Natural Science Foundation of China under Grant 61301277, Grant 61672376, and Grant 41371431. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Benoit Huet. *(Corresponding author: Qian Wang.)*

Q. Zou, L. Ni, and Q. Wang are with the State Key Laboratory of Software Engineering, School of Computer Science, Wuhan University, Wuhan 430072, China (e-mail: qzou@whu.edu.cn; lhni@whu.edu.cn; qianwang@whu.edu.cn).

Z. Hu and Q. Li are with the Shenzhen Key Laboratory of Spatial Smart Sensing and Service, Shenzhen University, Shenzhen 518060, China (e-mail: zwhoo@szu.edu.cn; liqq@szu.edu.cn).

S. Wang is with the Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29200 USA (e-mail: songwang@cec.sc.edu).

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/TMM.2016.2619912

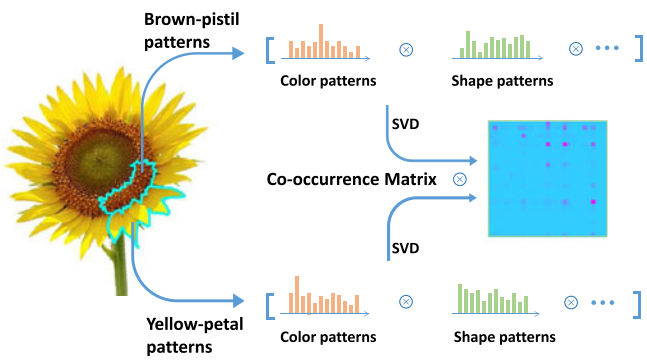


Fig. 1. Illustration of the proposed regional co-occurrence framework.

feature co-occurrence only in a small neighboring area, e.g., 8- or 16-connected neighborhoods, which would fall short in describing pattern collocations which hold co-occurrences over a larger neighborhood [20].

There have been research efforts trying to discover pattern collocations within a certain range in an image. In [21], semantic visual phrases were defined as the meaningful spatially co-occurrent patterns, which were extracted from a set of visual words by using a mining scheme. In addition, a top-down refinement procedure was applied on the discovered phrases to reduce the ones with ambiguous semantic meaning. In [22], first, delta visual phrases, i.e., visual phrases with high inter-class discrimination power, were constructed on frequently co-occurring visual patterns with similar spatial context. Then, the visual synset, i.e., visual phrases with high intra-class invariance power, was obtained by clustering delta visual phrases based on their class-probability distribution. The resulting semantic visual phrases from these methods can be used as a kind of pattern collocations, and describe the discriminative groups of object parts, e.g., the tire-and-door on a car, the nose-and-mouth on a human face, etc. However, it is often difficult to seek a balance between the discriminative power and the robustness in visual-phrase construction. In addition, the semantic visual phrases represent the structures sparsely in an image, which may neglect the pattern collocations in flat-feature regions, e.g., the grassland, sky, pavement, etc.

From the above discussions, we find that the co-occurrence and the spatial properties of local features are very important for encoding the local pattern collocations. Therefore, in this paper, we propose to compute co-occurrence between neighboring image regions, and densely encode the spatially co-occurrent patterns over the whole image. This leads to a new and general framework for the description of local pattern collocations, as illustrated in Fig. 1. Specifically, we employ superpixel segmentation techniques [23] to explicitly partition an image into a set of adjacent regions. Due to the homogeneous property of the superpixel, coherent appearance is usually observed within a superpixel. Then, we uniformly describe the patterns in each superpixel by quantizing the local features into a feature histogram, where each dimension of the histogram represents the occurrence of a pattern. Therefore, we can describe the co-occurrences of patterns in any two neighboring superpixels by computing a co-occurrence matrix of their feature histograms. We then use the co-occurrence matrix to study the pattern collocation. While local pattern collocation is the main focus of this paper, we also analyze the co-occurrences of different patterns between neighboring superpixel regions, and it leads to a regional co-occurrence framework for local pattern collocations.

Contributions:

1) Our main contribution is the design of a new framework for describing local pattern collocations using regional co-occurrence information. Under this framework, the problem of local pattern collocation description is addressed by computing the co-occurrence matrix of regionally packaged features over the image.
2) Based on the proposed framework, we develop a series of regional co-occurrence descriptors, e.g., regional shape co-occurrence descriptors with SIFT [24], [25] and SURF [26], regional color co-occurrence descriptors with CN [27], [28] and DD [29], and regional texture co-occurrence descriptors with LBPs. Boosted performances are obtained by using the new descriptors over the original ones in our experiments.
3) A factorization-based method is presented to fuse local pattern collocations built on different type of features, and the new ones are found to be additive to features encoded by other technique such as FV and DCNN [30].

The rest of this paper is organized as follows: Section II briefly reviews the related work. Section III describes the general framework for describing the local pattern collocations. Section IV introduces the implementations of popular shape, color and texture descriptors on the proposed framework. Section V demonstrates effectiveness of the proposed framework by experiments. Finally, Section VI concludes the paper.

II. Related Work

In the past two decades, many efforts have been made to analyze the co-occurrence of visual primitives for description of local patterns and pattern collocations [4], [6], [31]–[34]. In this section, we briefly overview them in a perspective of visual co-occurrence.

A. Co-occurrence-Based Local Patterns

In 1970s, the co-occurrence matrix was found to be effective in texture description [35], [36], where the co-occurrence matrix was computed based on the intensity gradient of gray images. In [37], the gray-level co-occurrence matrix was effectively fused with Gabor-filtering feature for textural classification. In [38], the co-occurrence matrix was computed on a series of distance parameters to increase its discrimination power. In [39], a texture descriptor was constructed by computing the co-occurrence matrix of LBPs. To handle color images, in [40], the traditional single-channel co-occurrence matrix was extended to the multi-channel co-occurrence matrix. Meanwhile, the co-occurrence among neighboring pixels was also found to be powerful for color image representation [18], [41], [42]. In [41], a color co-occurrence histogram, called color correlograms, was constructed by computing the co-occurrence

matrix between pixels in an 8-connected neighborhood. Similar method was used in [18], where the color co-occurrence histogram was applied to object detection, by examining the descriptor in each sliding window. In [42], color co-occurrence histogram was combined with other feature co-occurrence histogram such as HOG [43], to achieve an improved image classification performance. Note that, the above mentioned descriptors examine co-occurrence in an 8- or 16-pixel neighborhood. In this paper, we perform co-occurrence analysis on superpixel regions, instead of individual pixels and the proposed co-occurrence descriptor captures the pattern collocations in a much larger scale.

B. Co-occurrence-Based Pattern Collocations

Pattern collocations were often exploited based on the co-occurrence of local patterns, where the local patterns are documented on bag-of-features or LBPs. To discover semantic visual phrases, data-mining strategies have been widely employed [21], [22], [44]–[47]. In [21], [22], visual phrases are defined as meaningful spatial co-occurrent patterns of visual words, and significant spatial co-occurrent patterns are discovered by using frequent itemset mining techniques. The resulting pattern collocations can effectively describe object parts such as door-and-tire of a car, and nose-and-mouth of a human face. However, pattern collocations that are made up of similar patterns in the image may be neglected, especially those in a flat-feature region, e.g., a grassland, a pavement or a part of the sky. To address this issue, several strategies have been developed to encode more spacial information for co-occurrence pattern extraction, e.g., taking into account the contextual information [44], [48], expanding the searching to the whole image [49], using geometry-preserving visual phrases [50], and performing a hierarchical spatial partitioning to the image [51]. Besides SIFT, other primitive feature descriptors such as color histogram, HOG and LBP are also studied for co-occurrence pattern extraction [17], [20], [52], [53]. In [17], the co-occurrence pattern extracted from pairwise LBPs can achieve rotation invariance, which is desired in applications such as material recognition and flower species classification. In [53], co-occurrences of LBP were examined on five types of textons. However, the texton holds a size of only 2×2 pixels, which would be too small to encode pattern collocations over a larger area. In our cases, we propose to compute co-occurrence between neighboring image regions, based on which we can perform a dense encoding of the spatially co-occurrent patterns over the whole image.

Note that, some other methods also directly considered the regional information [54], [55]. In [54], a probabilistic method was proposed to model the relation between spatial co-occurrence of visual words associated with superpixels. However, in this method each superpixel should be assigned a scene label for training, which would be a labor-intensive task. In our method, the co-occurrences of superpixels are encoded automatically without any human interactions. In [55], the co-occurrences of appearance and shape features were examined in rectangles. In our experiments, we will demonstrate that the superpixels will be a better choice than rectangles.

Considering feature relations in an even larger range, [56] exploits the neighborhood information among images. To learn a discriminant hashing function, it extracts local discriminative information among content-similar images using an regularization method built on maximum-entropy principle. While [56] examines the neighborhood information between different images in the context of image retrieval, our work studies the neighborhood within an image for image classification.

III. A Regional Co-occurrence Framework

In this section, we begin with a definition to the local pattern collocation, and then describe the calculation of regional co-occurrence matrix, and at last introduce the details on using factorization for better co-occurrence information.

A. Local Pattern Collocation

Suppose $\mathcal{R}_a$ and $\mathcal{R}_b$ are two neighboring regions, $\mathcal{F}_a = \{f_{a1}, f_{a2}, \ldots, f_{am}\}$ are m features in region $\mathcal{R}_a$, and $\mathcal{F}_b = \{f_{b1}, f_{b2}, \ldots, f_{bn}\}$ are n features in region $\mathcal{R}_b$. Let's further suppose a number of patterns can be extracted by encoding the features in regions $\mathcal{R}_a$ and $\mathcal{R}_b$

$$\mathcal{P}_a = \Psi(\mathcal{F}_a) \tag{1}$$

and

$$\mathcal{P}_b = \Psi(\mathcal{F}_b) \tag{2}$$

where $\mathcal{P}_a$ and $\mathcal{P}_a$ denote the patterns in $\mathcal{R}_a$ and $\mathcal{R}_b$, respectively, and $\Psi(\cdot)$ denotes a function of feature encoding, e.g., the bag-of-words. Without loss of generality, (1) and (2) can be rewritten as

$$\begin{aligned}\mathcal{P}_a &= [p_{a1}, p_{a2}, \ldots, p_{ak}] \\ &= [n_{a1}, n_{a2}, \ldots, n_{ak}][\vec{b}_1, \vec{b}_2, \ldots, \vec{b}_k]\end{aligned} \tag{3}$$

and

$$\begin{aligned}\mathcal{P}_b &= [p_{b1}, p_{b2}, \ldots, p_{bk}] \\ &= [n_{b1}, n_{b2}, \ldots, n_{bk}][\vec{b}_1, \vec{b}_2, \ldots, \vec{b}_k]\end{aligned} \tag{4}$$

where $[\vec{b}_1, \vec{b}_2, \ldots, \vec{b}_k]$ denote the k pattern bases, which construct a $k \times k$ identity matrix, as shown in (5)

$$[\vec{b}_1, \vec{b}_2, \ldots, \vec{b}_k] = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ & & \ldots & \\ 0 & 0 & \ldots & 1 \end{bmatrix}_{k \times k}. \tag{5}$$

The n_{ai} and n_{bi} denote the frequency of pattern base b_i in $\mathcal{P}_a$ and $\mathcal{P}_a$, respectively. Then, a local pattern collocation is defined as a pair of patterns with each coming from one of the regions

$$\begin{aligned}\mathcal{C}_{ij} &= [p_{ai}, p_{bj}] \\ &= [n_{ai}\vec{b}_i, n_{bj}\vec{b}_j]\end{aligned} \tag{6}$$

where $1 \leq i \leq k$, and $1 \leq j \leq k$. Note that, different local feature descriptors can produce different $\mathcal{F}_a$ and $\mathcal{F}_b$. In addition, given the specific $\mathcal{F}_a$ and $\mathcal{F}_b$, different encoding algorithms

will result in different $\mathcal{P}_a$ and $\mathcal{P}_b$, and hence different pattern collocations.

B. Regional Co-occurrence Matrix

As $\mathcal{P}_a$ and $\mathcal{P}_b$ are the patterns in two neighboring regions $\mathcal{R}_a$ and $\mathcal{R}_b$, respectively, as stated in (3) and (4), all pattern collocations between $\mathcal{R}_a$ and $\mathcal{R}_b$ can be described as

$$\begin{aligned}\mathcal{C} &= [p_{a1}, p_{a2}, \ldots, p_{ak}] \otimes [p_{b1}, p_{b2}, \ldots, p_{bk}] \\ &= \begin{bmatrix} [p_{a1}, p_{b1}] & [p_{a1}, p_{b2}] & \ldots & [p_{a1}, p_{bk}] \\ [p_{a2}, p_{b1}] & [p_{a2}, p_{b2}] & \ldots & [p_{a2}, p_{bk}] \\ \ldots & \ldots & \ldots & \ldots \\ [p_{ak}, p_{b1}] & [p_{ak}, p_{b2}] & \ldots & [p_{ak}, p_{bk}] \end{bmatrix}\end{aligned} \tag{7}$$

where '$\otimes$' computes the Kronecker product [57]. Given that the pattern bases $[\vec{b}_1, \vec{b}_2, \ldots, \vec{b}_k]$ can be treated as constant, the pattern collocations in (7) can be simplified as

$$\mathcal{C} = \begin{bmatrix} [n_{a1}, n_{b1}] & [n_{a1}, n_{b2}] & \ldots & [n_{a1}, n_{bk}] \\ [n_{a2}, n_{b1}] & [n_{a2}, n_{b2}] & \ldots & [n_{a2}, n_{bk}] \\ \ldots & \ldots & \ldots & \ldots \\ [n_{ak}, n_{b1}] & [n_{ak}, n_{b2}] & \ldots & [n_{ak}, n_{bk}] \end{bmatrix}. \tag{8}$$

Since a pattern collocation is highly related to the frequency of the involved pair of patterns, we propose to describe each pattern collocation by multiplying the occurrence frequencies of the involved patterns

$$[n_{ai}, n_{bj}] \Rightarrow [n_{ai} \cdot n_{bj}]. \tag{9}$$

Thus, the pattern collocations in (8) can be rewritten as a co-occurrence matrix

$$\mathcal{C} = [n_{a1}, n_{a2}, \ldots, n_{ak}]^\top [n_{b1}, n_{b2}, \ldots, n_{bk}]. \tag{10}$$

Note that, the co-occurrence matrix reflects the probability of co-occurrence of a pair of patterns between the two specified regions. The probabilities of pattern co-occurrence regarding to two neighboring regions may potentially reflect the pattern collocation.

C. Regional Co-occurrence Factorization

In the above discussion, we only consider local pattern collocation in one feature space, i.e., one same kind of features. In fact, multiple local feature descriptors can be employed to extract features from an image, e.g., the shape descriptors such as SIFT and SURF, and the color descriptors such as CN and DD, where pattern collocations may be built by patterns upon different features. In this subsection, we propose a regional co-occurrence factorization scheme for the description of local pattern collocations in multiple feature spaces.

Let $\mathcal{R}_a$ and $\mathcal{R}_b$ be two neighboring regions, as illustrated by Fig. 2, a number of T feature descriptors are applied to extract local features, $\mathcal{P}_i^a = [p_{i1}^a, p_{i2}^a, \ldots, p_{im}^a]$ be m patterns in region $\mathcal{R}_a$ w.r.t. feature descriptor $\mathcal{F}_i$, and $\mathcal{P}_j^b = [p_{j1}^b, p_{j2}^b, \ldots, p_{jn}^b]$ be n patterns in region $\mathcal{R}_b$ w.r.t. feature descriptor $\mathcal{F}_j$, with $1 \leq i, j \leq T$, then the patterns in the two regions can be formulated

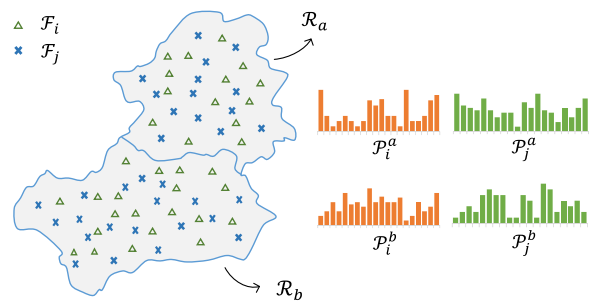


Fig. 2. Illustration of regional local patterns.

as

$$\mathcal{P}^a = [\mathcal{P}_1^a, \mathcal{P}_2^a, \ldots, \mathcal{P}_T^a] \tag{11}$$

and

$$\mathcal{P}^b = [\mathcal{P}_1^b, \mathcal{P}_2^b, \ldots, \mathcal{P}_T^b]. \tag{12}$$

Let's further suppose $N_i^a = [n_1^a, n_2^a, \ldots, n_m^a]$ be the occurrence frequency of patterns in $\mathcal{P}_i^a$, and $N_j^b = [n_1^b, n_2^b, \ldots, n_n^b]$ be the occurrence frequency of patterns in $\mathcal{P}_j^b$, we first compute the co-occurrence matrix of different kinds of patterns, e.g., between shape patterns and color patterns, within region $\mathcal{R}_a$ and region $\mathcal{R}_b$, respectively

$$\mathcal{C}^a = [\Phi(\Phi(\ldots\Phi(N_1^a \otimes N_2^a) \otimes \ldots) \otimes N_T^a)] \tag{13}$$

$$\mathcal{C}^b = [\Phi(\Phi(\ldots\Phi(N_1^b \otimes N_2^b) \otimes \ldots) \otimes N_T^b)] \tag{14}$$

where $\Phi(\cdot)$ reshapes a matrix to a one-row vector. Then the pattern collocations in multiple feature spaces in $\mathcal{R}_a$ and $\mathcal{R}_b$ can be computed by

$$\mathcal{C}^{ab} = \mathcal{C}^a \otimes \mathcal{C}^b. \tag{15}$$

The dimension of $\mathcal{C}^a$ and $\mathcal{C}^b$ is $M = \prod_{i=1}^T \mathsf{dim}(\mathcal{P}_i^b)$, where $\mathsf{dim}(\cdot)$ returns the dimension value of a vector. The resultant $\mathcal{C}^{ab}$ would have a very large dimension, and may contain redundant information. Therefore, before computing $\mathcal{C}^{ab}$, we make a factorization to $\mathcal{C}^a$ and $\mathcal{C}^b$.

Suppose $\mathcal{C}_{M\times N} = [\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_N]$ contains N co-occurrence matrices computed by (13) and (14) on a set of training images, each $\mathcal{C}_i$ corresponds to a local region, and has been reshaped to an $M \times 1$ vector, then we use singular value decomposition (SVD) to factorize $\mathcal{C}$

$$(\mathrm{U}, \Sigma, \mathrm{V}) = \mathsf{SVD}(\mathcal{C}) \tag{16}$$

and get $\mathcal{C} = \mathrm{U}\Sigma\mathrm{V}^\top$, where U is an $M \times M$ dimension matrix of left singular vectors, Σ is an $M \times N$ dimension diagonal matrix of singular values, and V is an $N \times N$ dimension matrix of right singular vectors. In practices, most of the singular values are very small or zero, we only consider the L largest singular values, in which $L = \sum_1^T \mathsf{dim}(\mathcal{P}_i^b)$. Therefore, the reduced space is now represented by $\hat{\Sigma}$ that has a dimension of $L \times L$, $\hat{\mathrm{V}}$ that has dimension $N \times L$, and $\hat{\mathrm{U}}$ that has dimension $M \times L$. Then for each $\mathcal{C}_i$ from a region, we get the new vector by

$$\hat{\mathcal{C}}_i = (\mathcal{C}_i^\top \hat{\mathrm{U}} \hat{\Sigma}^{-1})^\top. \tag{17}$$

Consequently, the $\mathcal{C}^a$ and $\mathcal{C}^b$ in (15) can be factorized into $\hat{\mathcal{C}}^a$ and $\hat{\mathcal{C}}^b$. Both of them have a dimension of $L \times 1$, and the co-occurrence matrix $\hat{\mathcal{C}}^{ab}$ can be calculated by (18), which has a dimension of $L \times L$

$$\hat{\mathcal{C}}^{ab} = \hat{\mathcal{C}}^a \otimes (\hat{\mathcal{C}}^b)^\top. \tag{18}$$

D. Regional Co-occurrence Descriptor

To apply the regional co-occurrence factorization to an image, we first partition the image into a number of non-overlapping neighboring regions. Specifically, we employ superpixel techniques for image partitioning.

Suppose $\{S_i | i = 1, 2, \ldots, M\}$ be a number of M superpixels of an input image I, and $\hat{\mathcal{C}}_i$ be the patterns in S_i as produced by using (17), then for each superpixel S_i, we calculate its regional co-occurrence matrix by (19)

$$\Gamma_i = \sum_{S_j \in \mathsf{NB}(S_i)} (\hat{\mathcal{C}}_i \otimes \hat{\mathcal{C}}_j^\top) \tag{19}$$

where $\mathsf{NB}(S_i)$ returns the 8 nearest neighboring superpixels of S_i. Then the regional co-occurrence matrix of image I can be computed by

$$\Psi_I = \mathsf{Norm}(\sum_{1}^{M} \Gamma_i) \tag{20}$$

where $\mathsf{Norm}(\cdot)$ denotes an L_1 normalization of an matrix, i.e., each element is divided by the sum of all elements in the matrix. To enhance the generalization power of Ψ_I, we make it symmetric and take its upper triangle matrix as the final descriptor for local pattern collocations, as formulated by (21)

$$\hat{\Psi}_I = \mathsf{Triu}(\Psi_I^\mathsf{T} + \Psi_I) \tag{21}$$

where $\mathsf{Triu}(\cdot)$ takes the upper triangle of the matrix.

IV. Implementations

Sections III-B and III-C provide two ways to describe local pattern collocations. In this section, we give details of the implementation to both of them.

A. Region Partitioning

In this paper, we use SLIC—a popular superpixel technique [23] for region partitioning, as it can generate similar-size homogeneous-appearance regions. The homogeneous property of the resulting regions provides possibilities of a precise description of the co-occurrent patterns, which would bring more discriminative power to the co-occurrence matrix in the later procedure. An illustration is given in Fig. 3. In addition, the size of the superpixel, i.e., $sSize$, and the compactness of the superpixel boundary can be controlled with SLIC parameters. Fig. 4 shows an example for region partitioning with different $sSize$. A smaller superpixel commonly contains less features in it. For the superpixel size, we will investigate its influence on the performance of the corresponding local pattern collocation descriptors. Also, we will study the necessity of superpixel segmentation by comparing it with the standard grid-based partitioning.

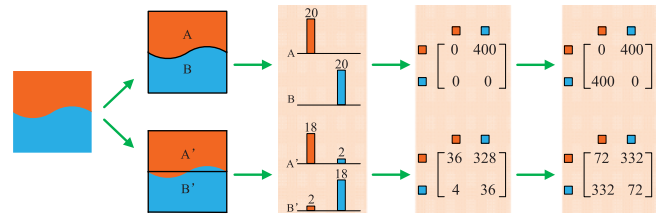


Fig. 3. An illustration to the effect of different partitioning strategies: superpixels v.s. rectangles. Column 1: an original region consisting of two subregions with homogeneous color feature contained. Column 2: region partitioning using superpixels (at top) and rectangles (at bottom). Column 3: color feature histograms in subregions A and B. Column 4: the regional co-occurrence matrix computed on the color feature histograms. Column 5: the symmetrized regional co-occurrence matrix.

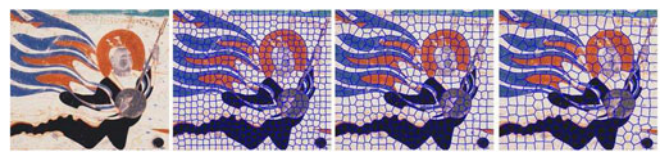


Fig. 4. Superpixels when using different $sSize$. From left to right, the original image, and superpixels with $sSize = 200$, 400, and 600, respectively.

We can see that the value in the (normalized) co-occurrence matrix reflects the probability of co-occurrence of a pair of feature value between the two specified regions. These probabilities of feature co-occurrence regarding to two separated regions may potentially reflect the collocation pattern of the considered feature and contain more discriminative information for classification.

B. Feature Extraction

We try up a number of popular shape, color and texture descriptors for local feature extraction.

Shape descriptors: The shape features in an image are usually expressed by the gradients. Based on image gradients, some shape descriptors have been developed and successfully applied in image matching and object categorization. In this work, we employ two popular local shape descriptors.

i) *SIFT*—Scale Invariant Feature Transform Descriptor [24]. SIFT produces 128-dimension features based on the gradients at particular interest points, and are invariant to image scale and rotation. They are also robust to changes in illumination, noise, and minor changes in viewpoint. We employ the implementations provided by VLFeat toolbox [58].

ii) *SURF*—Speed-Up Robust Feature Descriptor [26]. SURF performs in a similar way as SIFT. It uses an integer approximation of the determinant of Hessian blob detector, which can be computed with 3 integer operations using a pre-computed integral image. The 64-dimension features are based on the sum of the Haar wavelet response around the points of interest, which can also be computed with the aid of the integral image.

Color descriptors: Color representation plays an important role in image and object categorization. We use three popular color descriptors to extract color features in the image.

TABLE I
PROPERTIES OF DIFFERENT LBPS

Types	Number of neighbors	Uniform	Rotation invariant	Dimension
LBP11	8	Yes	No	59
LBP12	8	No	Yes	10
LBP13	8	Yes	Yes	36
LBP21	16	Yes	No	243
LBP22	16	No	Yes	18
LBP23	16	Yes	Yes	4116

i) *rgHist*—Normalized RGB Histograms Descriptor [5]. In the normalized RGB color model, the components r and g describe the color information in the image since b is redundant as $r + g + b = 1$ through the normalization. Owing to the normalization, r and g are scale-invariant and thereby invariant to light intensity changes.

ii) *CN*—Color Names Descriptor [27], [28]. Color names are linguistic labels when human perceives color in nature scenes, where a total number of eleven basic color names are defined. In [28], the eleven color names were automatically learnt from Google images, in which the color space was partitioned into eleven regions. Then, an 11-dimensional local color descriptor can be constructed by counting the occurrence of each color name over a local region of the image.

iii) *DD*—Discriminative Color Descriptor [29]. Based on the information theory, it clusters color values into a number of categories by minimizing their mutual information in image representation. In this way, the color space can be partitioned into a number of regions. Similar to CN, DD describes the occurrence of colors in an specific image region. The resulted descriptor was claimed to hold good photometric invariance and high discriminative power.

Texture descriptors: Local Binary Patterns (LBP) [19] has been found to be a powerful descriptor for texture classification. Generally, an LBP feature is created by comparing the center pixel to each of its 8 (or 16) neighbors in an image window (which is called a cell), following the pixels along a circle in clockwise or counter-clockwise direction. If the center pixel's value is greater than the neighbor's value, it creates a '0', and otherwise a '1'. This gives an 8-digit (or 16-digit) binary number. Then a histogram is computed over all the cells over an image by calculating the occurring frequency of each binary number. This feature histogram is often normalized to earn a higher descriptive power.

A series of LBP descriptors are employed to extract texture features in our work. We try LBPs with different number of neighbors, uniform or non-uniform, etc., which are listed in Table I.

C. Feature Encoding

Once a set of local features are extracted from a region, the feature encoding is applied on them to build local patterns. For features from each type of LBP, we directly project them into a histogram with their pre-defined pattern tables [19]. For SIFT, SURF, CN, DD and *rg*Hist features, we follow a standard BoW and quantize each type of features in one single region into a feature histogram. Thus, each value in the feature histogram represents a pattern. Before building the feature histogram, a feature codebook is constructed for each type of features by using K-means clustering. Then a hard-assignment strategy is used for feature quantization. Specifically, the K-means is initialized with k centers. In the experiments, we will study how the value of k influences the performance of the final descriptor in the context of image categorization.

Algorithm 1: Regional Co-Occurrence Factorization.

Input:
A set of images: $\{I_i | i = 1, 2, \ldots, N\}$;
A set of local feature descriptors: $\{\mathcal{F}_i | i{=}1, 2, \ldots, T\}$;
The superpixel size: $sSize$;
The number of centers for K-means: k;

Output:
The regional co-occurrence matrices of all images: $\hat{\Psi}$;

1: Partitioning each I_i by using SLIC superpixel algorithm with $sSize$;
2: Extracting features for each superpixel with each $\mathcal{F}_i$;
3: Constructing a feature codebook for each type of features with K-means clustering;
4: Building the local patterns for each type of features within each superpixel;
5: Computing a pattern co-occurrence matrix for each superpixel with (13);
6: Collecting all co-occurrence matrices from step 5, and factorizing them with (16);
7: Computing a factorized pattern co-occurrence matrix for each superpixel with (17);
8: Calculating a regional co-occurrence matrix for each superpixel with (19);
9: Computing the regional co-occurrence matrix $\hat{\Psi}_i$ for each image with (21);
10: **return** $\hat{\Psi}{=}\{\hat{\Psi}_i | i = 1, 2, \ldots, N\}$.

D. Descriptors Construction

After local patterns are extracted, the final regional co-occurrence descriptors can be constructed using the regional co-occurrence framework as introduced in Section III. To be specific, we summarize in Algorithm 1 the pipeline for constructing the regional co-occurrence descriptors.

V. EXPERIMENTS AND RESULTS

In this section, first, we introduce five image datasets for evaluation, including two real-object datasets and three art-painting datasets. Then, we evaluate and validate the effectiveness of the proposed framework through extensive experimental comparisons. After that, we study the influence of parameters on the performances of the proposed regional co-occurrence descriptors.[1] At last, we use three more datasets to test the additivity property

[1]Codes are available at https://sites.google.com/site/qinzoucn/documents

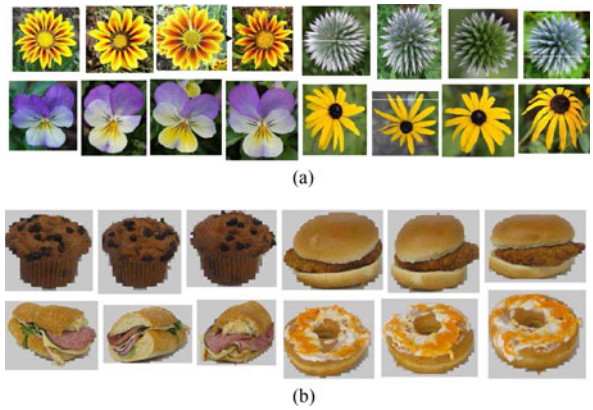


Fig. 5. Sample images from two natural image datasets. (a) Flowers102. (b) PFID61.

of the proposed descriptors to some state-of-the-art methods. Note that, the main purpose of the experiments is to validate that the proposed framework is effective over traditional shape, color and texture descriptors.

A. Datasets

Flowers102: The Oxford Flowers 102 dataset contains 8189 images [59], which are divided into 102 categories with 40 to 250 images per category. Sample images are shown in Fig. 5(a). Image masks are used to extract features in the region of interest. In the experiments, we randomly select 20 images each category for training and the rest for test. The procedure is repeated thirty times and the results are averaged.

PFID61: The Pittsburgh Food Image Dataset [60] contains 1098 fast-food images from 61 food categories, with masked foreground. Each food category has three different instances. For each food instance, six images are collected from six viewpoints, with 60 degrees apart. The experimental protocol was used in [33], [61] with a 3-fold cross validation strategy, where 12 images from two instances were used for training and 6 images from the third instance for test. This procedure is repeated three times, with a different instance serving as the test set. The results are averaged.

ArtMv6: The dataset contains 3185 paintings belonging to 6 different art movements, i.e., Renaissance, Baroque, Rococo, Romanticism, Impressionism and Cubism, from more than 600 authors [62]. Among these genres, some are easy to discern, e.g., the Cubism and Renaissance, and some are mixed and hard to separate, e.g., Baroque and Renaissance. The images are acquired from various sources, and lack cohesion in acquisition conditions. The classifiers are trained with 600 images which are specified and uniformly covering all 6 artistic genres.

PT-91: The dataset consists of 4266 images of 91 artists [63]. The artists in the dataset come from different eras. Fig. 6(b) shows example painting images of four different artists from this dataset. There are variable number of images per artist ranging from 31 to 56. The large number of images and artist categories make the problem of computational painting categorization extremely challenging. A protocol attached specifies 1991 images for training, and the other 2275 images for test.

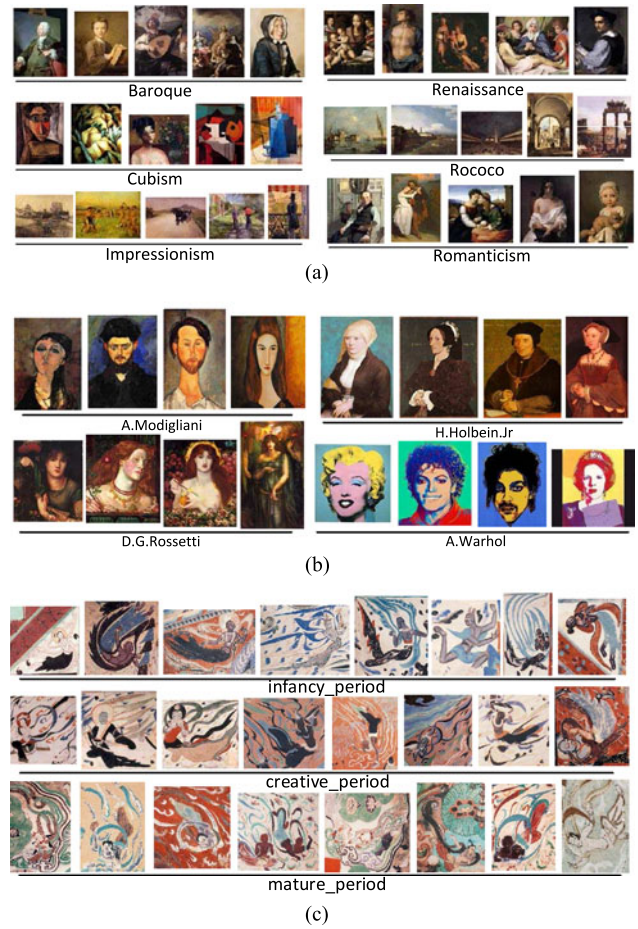


Fig. 6. Samples images from three painting datasets. (a) ArtMv6. (b) PT-91. (c) DH660.

DH660: It contains 660 *Flying-Apsaras* painting images from Mogao Grottoes in Dunhuang, China [64]. These 660 images have been categorized into three classes according to the art eras in the development of the *Flying-Apsaras* art, with 220 from the infancy period, 220 from the creative period, and 220 from the mature period of the *Flying-Apsaras* art. Samples of the collected images are shown in Fig. 6(c). For each of the three categories, half data are taken for training, and the remaining half are taken for testing.

B. Effectiveness

In order to find to what extent the proposed regional co-occurrence framework works, we make comparisons by using and not using the proposed framework for the local feature descriptors mentioned in Section IV-B. First, we visually examine the effectiveness of regional co-occurrence framework in image representation. Then, we quantitatively analyze the improvement brought by the proposed framework over the original feature descriptors.

Visual comparisons: An excellent descriptor for image representation is supposed to hold high inter-class discrimination power, meanwhile to hold high intra-class invariance power. Then the descriptors constructed by the proposed framework

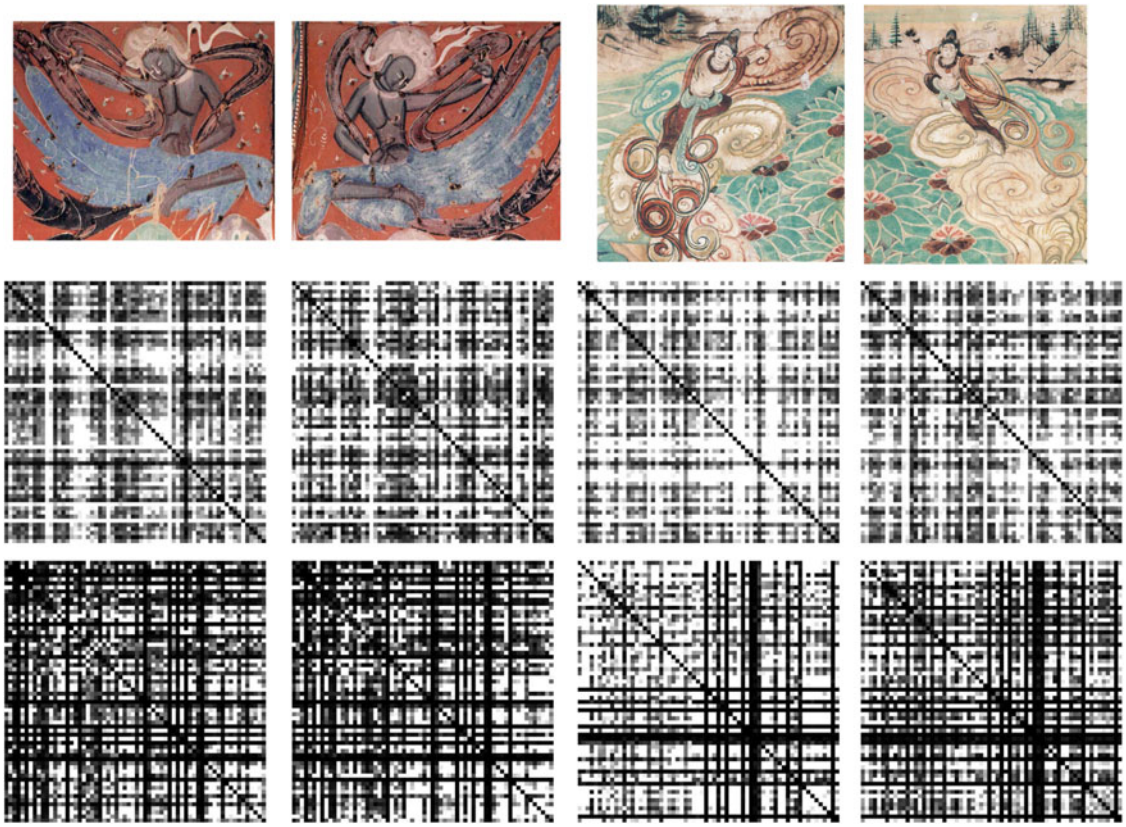


Fig. 7. Visual comparison. Top row: two pairs of original images. Middle row: regional co-occurrence matrix from RSC. Bottom row: regional co-occurrence matrix from RCC.

are desired to produce feature vectors (or matrices) with small distance on the same-category images, and produce feature vectors (or matrices) with large distance on the different-category images. In this experiment, two pairs of images from DH660 dataset are used, as shown in the top row of Fig. 7. The pair on the left are from one category—'infancy period', and the pair on the right are from another category—'mature period'. It can be seen that, each pair of images have similar color collocations, and curve structures, but images between the pairs are very different in these attributes, which we state as the difference in art styles. For all four images, the proposed regional SIFT co-occurrence descriptor (RSC) and the regional CN11 co-occurrence descriptor (RCC) are used to extract the shape co-occurrence matrices, and the color co-occurrence matrices, respectively, which are shown in the middle row and bottom row of Fig. 7. Due to the limitation of display, we only use 64-words codebooks for RSC and RCC. From Fig. 7 we can see, for both RSC and RCC, matrices within one pair are similar, and matrices between the two pairs are dissimilar, which visually shows the inter-class discriminative power and intra-class invariance power of RSC and RCC.

Quantitative evaluations: We also make quantitative comparisons by using and not using the regional co-occurrence framework. For shape descriptors—SIFT and SURF, a dense sampling strategy is employed. Specifically, multi-scale dense SIFTs are extracted from each input image, with a sampling step of 4, and a number of scales of 5. For SURF, a sampling step of 2 is used. Then, SIFT and SURF will extract a similar number of features on one same image. For color descriptors - *rg*Hist, CN, and DD, features are extracted with a sampling step of 5, and a number of scales of 2. Note that, CN produces 11-dimensional features, and is named as CN11. Two versions of DD are used, i.e., 25 dimensions and 50 dimensions, which are named as DD25 and DD50, respectively. For texture descriptors, five LBPs are tested and features (patterns) are computed at each pixel in the image. The classification accuracy is used as a metric for performance evaluation, which is defined by (22)

$$\text{Accuracy} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}. \tag{22}$$

We use 'Yes' to denote the case using the regional co-occurrence framework, and 'No' to denote the case not using it. To make fair comparisons, we uniformly set $k = 128$ to run the K-means for feature-codebook construction. In the case of 'No', the shape and color features are encoded by the bag-of-words framework. While in the case of 'Yes', a number of regional co-occurrence descriptors are formed, e.g., the RSC, RCC, etc. Due to the intrinsic property of LBPs, patterns are directly extracted from image patches by using the corresponding LBP descriptors. Thus feature histogram can be constructed in an image/region without a feature codebook. Note that, LBP23 is excluded from regional co-occurrence computation since it produces 4116-dimensional features, which incurs an intolerable computing cost.

TABLE II
PERFORMANCES OF LOCAL DESCRIPTORS BY USING AND NOT-USING THE REGIONAL CO-OCCURRENCE FRAMEWORK (RC)

	descriptor	using RC ?	Flowers102	PFID61	DH660	ArtMv6	PT-91	Ave. Gain
Shape	SIFT	No	47.07	36.52	69.61	52.32	27.60	
		Yes	58.39	**41.8**	80.16	62.58	**36.18**	9.20
	SURF	No	60.56	40.71	81.76	54.83	26.07	
		Yes	**61.72**	41.26	83.17	62.21	34.29	3.74
Color	CN11	No	43.66	26.87	83.45	53.37	16.04	
		Yes	50.72	36.70	**91.24**	**64.17**	20.92	7.79
	DD25	No	47.20	26.41	79.29	52.05	16.57	
		Yes	51.02	35.70	89.92	63.51	19.91	7.71
	DD50	No	48.07	26.87	80.12	53.23	16.62	
		Yes	52.15	36.16	90.54	64.06	21.05	7.81
	*rg*Hist	No	41.99	26.87	68.79	47.89	12.79	
		Yes	49.86	32.06	88.69	61.09	20.70	10.81
Texture	LBP11	No	26.47	23.41	66.77	55.51	17.45	
		Yes	46.22	30.51	73.99	58.03	27.52	9.33
	LBP12	No	15.08	16.58	61.61	49.28	8.92	
		Yes	27.73	23.04	66.23	54.33	13.85	6.74
	LBP13	No	20.92	17.12	63.14	52.72	10.29	
		Yes	32.90	23.50	65.68	55.31	17.10	5.90
	LBP21	No	35.40	26.05	71.93	52.66	24.79	
		Yes	50.56	33.79	77.95	57.49	32.88	8.37
	LBP22	No	18.92	16.21	62.26	53.36	10.55	
		Yes	32.72	24.95	66.36	56.22	15.12	6.81

The results for all the shape, color and texture descriptors, using and not using the regional co-occurrence framework, are listed in Table II. It can be seen from Table II that, all the 11 descriptors equipped with the proposed regional co-occurrence framework achieve boosted performances in image categorization on the five datasets, which demonstrates the effectiveness of the proposed framework. Note that, the regional CN11 co-occurrence descriptor (RCC) achieves an accuracy over 90% on the DH660 dataset, which is 7.79% higher than that obtained by the original CN11 (not using the regional co-occurrence framework). It is because that, paintings in DH660 are very rich in colors, and the color collocations have become a distinct painting style for the Dunhuang paintings. It in turn demonstrates that the regional co-occurrence framework can effectively improve the discriminative power of original descriptors (not using the regional co-occurrence framework), and better describe the pattern collocations.

C. Robustness

We examine the robustness of the proposed framework by using different feature-combination strategies. Fig. 8 shows the results of different methods on all five datasets. In the charts in Fig. 8, Bar 1 and Bar 2 denote the regional CN11 co-occurrence descriptor RCC and the regional SIFT co-occurrence descriptor RSC, respectively. Note that, here RCC and RSC are constructed on feature codebooks with 256 visual words. In Bar 3, 'RCC+RSC' denotes the method that concatenates the RCC and RSC feature matrices. In Bar 4, '[RCC+RSC]' denotes the method that factorizes the RCC features and RSC features before computing the final co-occurrence matrix, as described in Algorithm 1. In Bar 5, the IFV [11] is equipped with 256 Gaussian models. In Bar 6 and 7, the IFV is used in a late fusion manner. The difference lies in that, in '[RCC+RSC]+IFV', RCC and RSC features are early fused, where a factorization process is applied before co-occurrence computation, while in 'RCC+RSC+IFV', all kernels are produced independently. It can be seen from Fig. 8 that, the combination of RCC and RSC achieves higher performances than any single descriptor, and '[RCC+RSC]' outperforms 'RCC+RSC' on all five datasets. It is because that, the local pattern collocations constructed by different types of features have been encoded by an early fusion in the factorization step [as defined by (18)], which endows '[RCC+RSC]' with more discriminative power than 'RCC+RSC', as color patterns and shape patterns are encoded independently in 'RCC+RSC'. It can also be seen from Bar 6 and Bar 7 that, traditional descriptors wrapped by the proposed framework are additive to IFV in object and painting classification.

D. Impact of Parameters

Experiments are conducted on Flowers102 and ArtMv6 to examine the influences of parameters sSize and k on performances of the descriptors using the proposed regional co-occurrence framework. Note that sSize is a parameter initializing the superpixel size, and k is the number of centers for K-means clustering. Fig. 9(a) and (b) show the impact of k, and Fig. 9(c) and (d) show the impact of sSize. Note that, when $s\text{Size} = 1$, the region becomes 1×1 pixel, and the proposed regional co-occurrence descriptors degenerate to the co-occurrence descriptors, as proposed in [18]. It can be seen from Fig. 9, a higher k generally leads to higher performances, which reflects the fact that a larger codebook commonly brings higher discriminative power. It can also be seen that, a too small or too large sSize will lead to decreased performances. It is because that, overly small regions will be less capable of constructing the local pattern collocations, and overly large regions will enclose local pattern

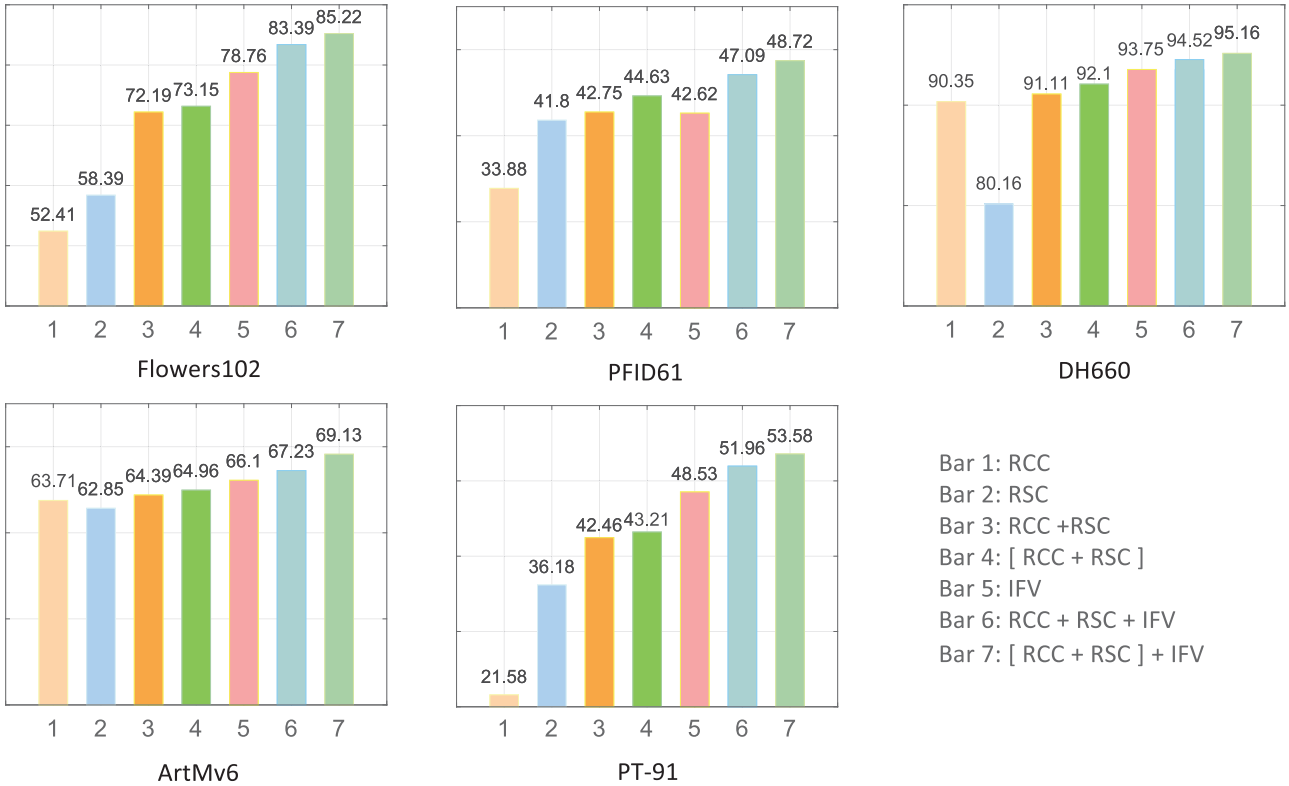


Fig. 8. Comparison of different feature-combination strategies on five datasets. Note that "[RCC+RSC]" denotes the feature fusion by using factorization.

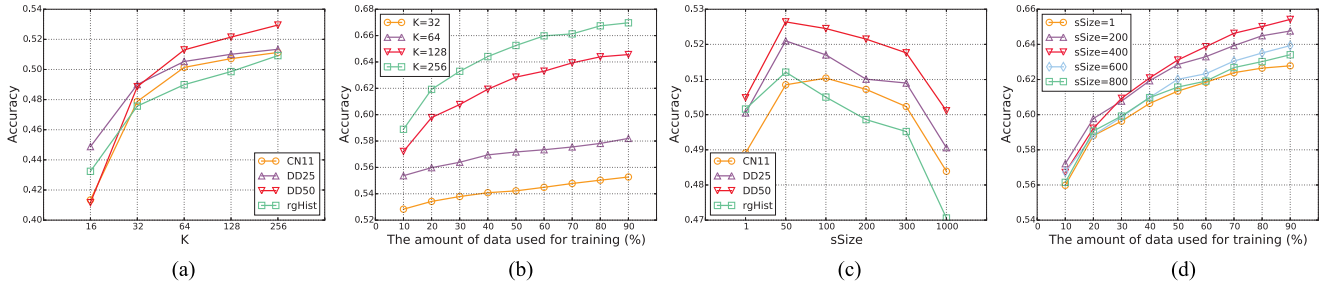


Fig. 9. Performances under different k and sSize. (a) Different k for four regional color co-occurrence descriptors on Flowers102. (b) Different k for regional SIFT co-occurrence descriptor on ArtMv6. (c) Different sSize for four regional color co-occurrence descriptors on Flowers102. (d) Different sSize for regional SIFT co-occurrence descriptors on ArtMv6. Note that, sSize $= 200$ for (a) and (b), and $k = 128$ for (c) and (d).

collocations within a same region, and undermine the performance of the proposed regional co-occurrence framework.

In addition, to examine the advancement of a superpixel-based image partitioning, we compare RSC by using superpixels and square-grids partitioning strategies, respectively. From Fig. 10(a) and (b) we can see that, the region partitioning using superpixels uniformly produces higher classification accuracy for RSC under different k. Furthermore, we examine how the compactness of superpixels influences the final performance. In Fig. 10(d), six different superpixel segmentations on a flower image have been shown which are obtained with compactness value of 0, 5, 10, 20, 100 and 1000, respectively. We fix sSize $=$ 200 and run the classification by changing the compactness. It can be found from Fig. 10(c) that, a compactness value close to 0 or 1000 leads to decreased performances, while the compactness value equal to 10 or 20 generates relatively better results. This is because, too large a compactness will generate superpixels similar to grids, and too small a compactness will make the superpixels too sharp to form enough meaningful neighborhoods, both of which undermine the performance of the proposed regional co-occurrence descriptors.

E. Additivity to the State-of-the-Art Methods

To further evaluate the performance of the proposed method in handling fine-grained dataset and large and complex dataset, we use three more datasets for evaluation.

Pascal'12: The PASCAL VOC Challenge 2012 [65] dataset contains 28,952 images of 20 different object categories. Each image is provided with bounding boxes indicating one or more

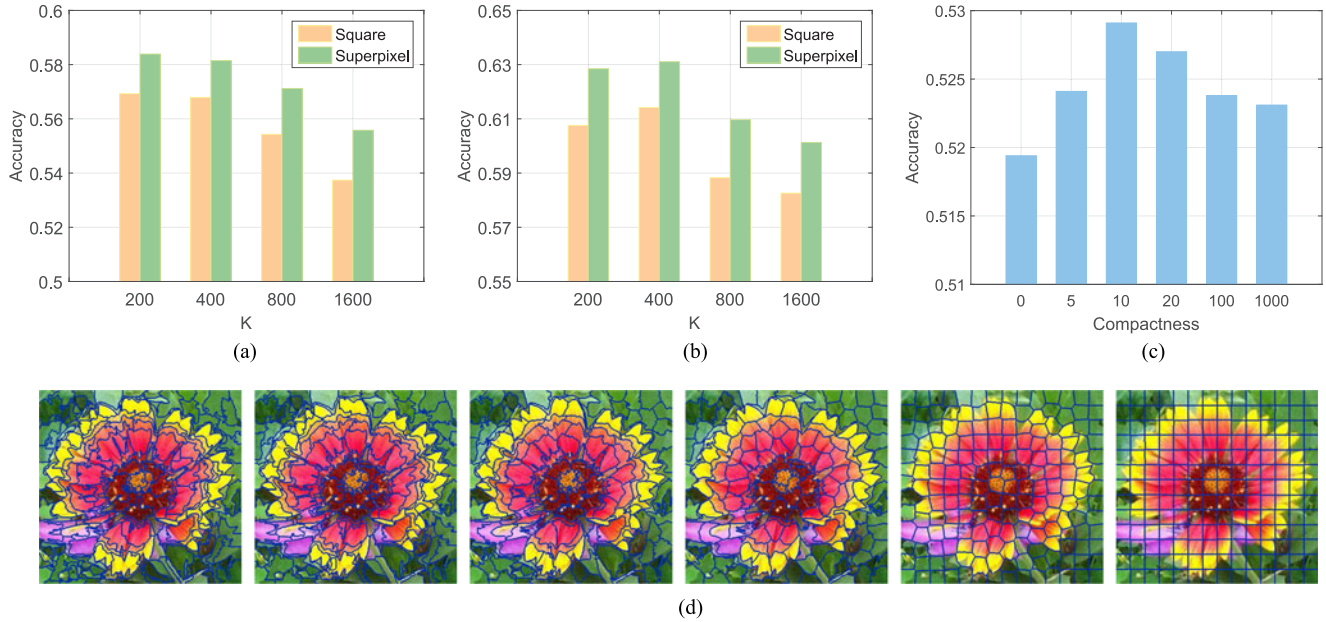


Fig. 10. Performance under different partitioning strategies. (a) and (b) show the results of RSC using square-grid segmentation and superpixel segmentation on Flowers102 and ArtMv6, respectively. (c) shows the results of RCC on Flowers102 using superpixels under different compactness values. (d) shows superpixels with compactness of 0, 5, 10, 20, 100, and 1000, respectively.

objects within it. The data has been split into 50% for training/validation and 50% for testing. The distributions of images and objects by class are approximately equal across the training/validation and testing sets.

Birds200: The Caltech-UCSD Birds 200-2011 [66] consists of 11,788 bird images which are from 200 breeds. There are about 50% images used for training and the remain for test. The training and test samples have been specified by the usage protocol along with the dataset. We use the foreground mask provided by the dataset for regional co-occurrence feature extraction, and the subimage cropped by the minimum bounding rectangle of the mask for neural-network-based feature learning.

Cats&Dogs: The Cats&Dogs dataset [67] is a collection of 7,349 images of cats and dogs of 37 different breeds, where 25 are dogs and 12 are cats. Images are divided into training, validation, and test sets, in a similar manner to the PASCAL VOC data. The dataset contains about 200 images for each breed.

Sample images from the above three datasets are shown in Fig. 11, and the information of all eight datasets used in our experiments has been summarized in Table III. To be mentioned, the Birds200 and Cats&Dogs are fine-grained image datasets. Besides running the proposed regional co-occurrence descriptors, we also run two state-of-the-art methods AlexNet [30] and GoogLeNet [68] on the Pascal'12, Birds200 and Cats&Dogs. The AlexNet and GoogLeNet are two implementations of the deep convolutional neural network (DCNN). For both of them, we use the models pre-trained on the ImageNet dataset in two strategies. In strategy one, an image from a target dataset is input into the model, and the output of the last full-connection layer, i.e., the layer before the prediction layer, is taken as image features. In strategy two, the pre-trained model is firstly fine-tuned on the target dataset, and then feature extraction is conducted using the strategy one. Features generated from the former are named as AlexNet$_{fc}$ and GoogLeNet$_{fc}$, while the features from

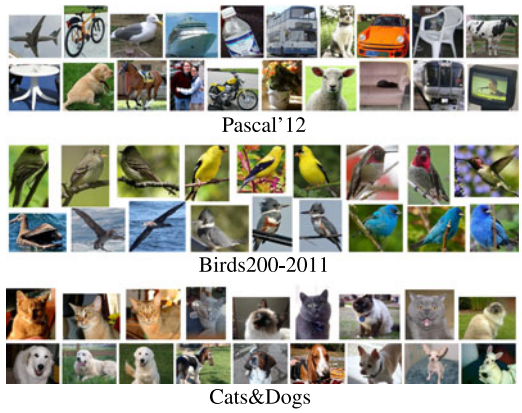


Fig. 11. Samples images from three more datasets.

TABLE III
INFORMATION OF ALL EIGHT DATASETS

Dataset	Number of images	Foreground	Image size	Categories
Flowers102	8189	mask	origin	102
PFID61	1098	rectangle	origin	61
ArtMv6	3185	-	256	6
PT-91	4266	-	300	91
DH660	660	-	origin	3
Pascal'12	28952	rectangle	256	20
Birds200	11788	mask	256	200
Cats&Dogs	7349	-	500	37

*Note: the image size refers to the wider side of the image.

TABLE IV
COMBINING AND COMPARISON WITH THE STATE-OF-THE-ART METHODS

Descriptor	Pascal'12	Birds200	Cats&Dogs
CN11	13.59	7.57	10.38
RCC	23.34	17.93	15.83
RSC	41.52	14.56	25.70
RSC+RCC	46.57	24.94	26.07
[RSC+RCC]	47.25	27.41	27.57
IFV	55.37	40.66	46.15
IFV+RCC	59.06	46.93	49.00
AlexNet$_{fc}$	71.71	54.25	77.18
AlexNet$_{fc}$+RCC	72.19	54.93	77.69
AlexNet$_{ft}$	72.99	55.94	78.76
AlexNet$_{ft}$+RCC	73.46	56.58	79.16
GoogLeNet$_{fc}$	76.98	66.32	84.39
GoogLeNet$_{fc}$+RCC	77.81	66.94	85.04
GoogLeNet$_{ft}$	79.81	77.03	88.20
GoogLeNet$_{ft}$+RCC	80.34	77.48	88.94
AlexNet$_{ft}$+RSC+RCC	73.80	57.31	79.37
AlexNet$_{ft}$+[RSC+RCC]	74.50	58.13	79.71
GoogLeNet$_{ft}$+RSC+RCC	81.02	78.05	89.33
GoogLeNet$_{ft}$+[RSC+RCC]	81.45	78.55	89.59

the later are named as AlexNet$_{ft}$ and GoogLeNet$_{ft}$. To test the additivity of the proposed RCC and RSC features, we combine them with the DCNN features for classification. The results are shown in Table IV. It can be seen from Table IV that, comparing with CN11 (using BoW with 500 codes), the RCC, i.e., the regional CN11 co-occurrence descriptor (using RC with 128 codes), achieves an improvement of about 10%, 10% and 5% in classification accuracy on the three datasets, respectively. It indicates that the proposed regional co-occurrence framework is very effective in encoding the local color features. The performance of '[RCC+RSC]' is uniformly higher than that of 'RCC+RSC' on the three datasets. It is because that the co-occurrence factorization in '[RCC+RSC]' has successfully improved the discriminative power. From Table IV we can also see, DCNN obtain good results and the fine-tuned features outperform the direct full-connection features produced by the model pre-trained on ImageNet. When combining the proposed features with AlexNet or GoogLeNet, improved results are obtained, e.g., 'GoogLeNet$_{ft}$+[RCC+RSC]' outperforms 'GoogLeNet$_{ft}$' by 1.64%, 1.52% and 1.39% on the three datasets, respectively, which indicates a good additivity property of the proposed method.

It is worth noting that, in the experiments RCC and RSC features are later fused with DCNN features by a concatenation operation. As the same in RCC and RSC, the fused features are trained and tested by a multi-class SVM equipped with a χ^2 kernel. The Caffe toolkit [69] is employed to fine-tune the DCNN models, where two blocks of GTX TITAN-X-12G GPU are used. In DCNN fine-tuning, it takes about 2.5 hours for AlexNet and 5 hours for GoogLeNet with a task of 50,000 iterations. In regional co-occurrence feature factorization, it takes about 2 hours for a parallel-implemented SVD to decompose a matrix constructed by 200,000 sampled co-occurrence features on a 2.4 GHz CPU with 6 cores.

VI. CONCLUSION

In this work, a regional co-occurrence framework was proposed to describe local pattern collocations for image and object categorizations. With this framework, a range of regional co-occurrence descriptors were developed. In addition, a factorization-based method was proposed to fuse different types of features and build pattern collocations in multiple feature spaces. With the multi-feature based patterns, pattern collocations between neighboring regions can be constructed by computing a co-occurrence matrix at each local region. In the experiments, five datasets of real objects and three datasets of art paintings were used for performance evaluation. Results from visual and quantitative comparisons demonstrated the effectiveness and robustness of the proposed framework. Comparison results also showed that the matrix factorization had effectively contributed to the improvement of discriminative power, and the proposed regional co-occurrence descriptors were highly additive to the state-of-the-art methods such as IFV and DCNN.

ACKNOWLEDGMENT

The authors would like to thank Dr. X. Qi for helpful discussions.

REFERENCES

[1] C. Luo, B. Ni, S. Yan, and M. Wang, "Image classification by selective regularized subspace learning," *IEEE Trans. Multimedia*, vol. 18, no. 1, pp. 40–50, Jan. 2016.
[2] S. Liao, X. Li, H. T. Shen, Y. Yang, and X. Du, "Tag features for geo-aware image classification," *IEEE Trans. Multimedia*, vol. 17, no. 7, pp. 1058–1067, Jul. 2015.
[3] C. Galleguillos, A. Rabinovich, and S. Belongie, "Object categorization using co-occurrence, location and appearance," in *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, Jun. 2008, pp. 1–8.
[4] L. Wu, Y. Hu, M. Li, N. Yu, and X.-S. Hua, "Scale-invariant visual language modeling for object categorization," *IEEE Trans. Multimedia*, vol. 11, no. 2, pp. 286–294, Feb. 2009.
[5] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1582–1596, Sep. 2010.
[6] S. M. Assari, A. R. Zamir, and M. Shah, "Video classification using semantic concept co-occurrences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 2529–2536.
[7] Y.-G. Jiang, Q. Dai, T. Mei, Y. Rui, and S.-F. Chang, "Super fast event recognition in internet videos," *IEEE Trans. Multimedia*, vol. 17, no. 8, pp. 1174–1186, Aug. 2015.
[8] W. Zhou *et al.*, "Towards codebook-free: Scalable cascaded hashing for mobile image search," *IEEE Trans. Multimedia*, vol. 16, no. 3, pp. 601–611, Apr. 2014.
[9] Q. Wang *et al.*, "Searchable encryption over feature-rich data," *IEEE Trans. Dependable Secure Comput.*, to be published, doi: 10.1109/TDSC.2016.2593444.
[10] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2003, vol. 2, pp. 1470–1477.
[11] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 143–156.
[12] F. Monay, P. Quelhas, J.-M. Odobez, and D. Gatica-Perez, "Integrating co-occurrence and spatial contexts on patch based scene segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshop*, Jun. 2006, pp. 14–21.
[13] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, and H.-J. Zhang, "Image classification with kernelized spatial-context," *IEEE Trans. Multimedia*, vol. 12, no. 4, pp. 278–287, Jun. 2010.

[14] L. Ladicky, C. Russell, P. Kohli, and P. H. Torr, "Graph cut based inference with co-occurrence statistics," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 239–253.

[15] D. Zhang *et al.*,"Survey on context-awareness in ubiquitous media," *Multimedia Tools Appl.*, vol. 67, no. 1, pp. 179–211, 2013.

[16] X. J., D. Y., and C. G., "Detection evolution with multi-order contextual co-occurrence," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 1798–1805.

[17] X. Qi, R. Xiao, J. Guo, and L. Zhang, "Pairwise rotation invariant co-occurrence local binary pattern," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 158–171.

[18] P. Chang and J. Krumm, "Object recognition with color cooccurrence histograms," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2009, vol. 2, p. 504.

[19] 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.

[20] Q. Zou, X. Qi, Q. Li, and S. Wang, "Discriminative regional color co-occurrence descriptor," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2015, pp. 696–700.

[21] J. Yuan, Y. Wu, and M. Yang, "Discovery of collocation patterns: From visual words to visual phrases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2007, pp. 1–8.

[22] Y.-T. Zheng, M. Zhao, S.-Y. Neo, T.-S. Chua, and Q. Tian, "Visual synset: Towards a higher-level visual representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2008, pp. 1–8.

[23] R. Achanta *et al.*, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[24] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[25] S. Hu, Q. Wang, J. Wang, Z. Qin, and K. Ren, "Securing SIFT: Privacy-preserving outsourcing computation of feature extractions over encrypted image data," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3411–3425, Jul. 2016.

[26] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 404–417.

[27] R. Benavente, M. Vanrell, and R. Baldrich, "Parametric fuzzy sets for automatic color naming," *J. Opt. Soc. Amer.*, vol. 25, no. 10, pp. 2582–2593, 2008.

[28] J. van de Weijer, C. Schmid, J. Verbeek, and D. Larlus, "Learning color names for real-world applications," *IEEE Trans. Image Process.*, vol. 18, no. 7, pp. 1512–1524, Jul. 2009.

[29] R. Khan *et al.*, "Discriminative color descriptors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 2866–2873.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. Conf.*, 2012, pp. 1097–1105.

[31] D. Crandall and J. Luo, "Robust color object detection using spatial-color joint probability functions," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2004, pp. 379–385.

[32] G. Chen, Y. Ding, J. Xiao, and T. X. Han, "Detection evolution with multi-order contextual co-occurrence," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 1798–1805.

[33] B. Ni, M. Xu, J. Tang, S. Yan, and P. Moulin, "Omni-range spatial contexts for visual classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 3514–3521.

[34] Y. Jiang, J. Meng, and J. Yuan, "Randomized visual phrases for object search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 3100–3107.

[35] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.

[36] L. S. Davis, S. A. Johns, and J. Aggarwal, "Texture analysis using generalized co-occurrence matrices," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 3, pp. 251–259, Jul. 1979.

[37] D. A. Clausi and H. Deng, "Design-based texture feature fusion using Gabor filters and co-occurrence probabilities," *IEEE Trans. Image Process.*, vol. 14, no. 7, pp. 925–936, Jul. 2005.

[38] A. Gelzinis, A. Verikas, and M. Bacauskiene, "Increasing the discrimination power of the co-occurrence matrix-based features," *Pattern Recog.*, vol. 40, no. 9, pp. 2367–2372, 2007.

[39] X. Sun, J. Wang, R. Chen, M. F. She, and L. Kong, "Multi-scale local pattern co-occurrence matrix for textural image classification," in *Proc. Int. Joint Conf. Neural Netw.*, 2012, pp. 1–7.

[40] C. Palm, "Color texture classification by integrative co-occurrence matrices," *Pattern Recog.*, vol. 37, no. 5, pp. 965–976, 2004.

[41] J. Huang, S. Kumar, M. Mitra, W. Zhu, and R. Zabih, "Image indexing using color correlograms," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 1997, pp. 762–768.

[42] S. Ito and S. Kubota, "Object classification using heterogeneous co-occurrence features," in *Proc. 11th Eur. Conf. Comput. Vis.*, 2010, pp. 209–222.

[43] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2005, pp. 886–893.

[44] J. Yuan and Y. Wu, "Context-aware clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2008, pp. 1–8.

[45] S. Zhang, Q. Tian, G. Hua, Q. Huang, and W. Gao, "Generating descriptive visual words and visual phrases for large-scale image applications," *IEEE Trans. Image Process.*, vol. 20, no. 9, pp. 2664–2677, Sep. 2011.

[46] J. Yuan, M. Yang, and Y. Wu, "Mining discriminative co-occurrence patterns for visual recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 2777–2784.

[47] B. Fernando, E. Fromont, and T. Tuytelaars, "Mining mid-level features for image classification," *Int. J. Comput. Vis.*, vol. 108, no. 3, pp. 186–203, 2014.

[48] H. Wang, J. Yuan, and Y. Wu, "Context-aware discovery of visual co-occurrence patterns," *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 1805–1819, Apr. 2014.

[49] J. Gao, Y. Hu, J. Liu, and R. Yang, "Unsupervised learning of high-order structural semantics from images," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep.-Oct. 2009, pp. 2122–2129.

[50] Y. Zhang, Z. Jia, and T. Chen, "Image retrieval with geometry-preserving visual phrases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 809–816.

[51] Y. Yang and S. Newsam, "Spatial pyramid co-occurrence for image classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1465–1472.

[52] T. Watanabe, S. Ito, and K. Yokoi, "Co-occurrence histograms of oriented gradients for pedestrian detection," in *Proc. 3rd Pacific Rim Symp. Adv. Image Video Technol.*, 2009, pp. 37–47.

[53] G.-H. Liu and J.-Y. Yang, "Image retrieval based on the texton co-occurrence matrix," *Pattern Recog.*, vol. 41, no. 12, pp. 3521–3527, 2008.

[54] B. Miçušík and J. Koşeck, "Semantic segmentation of street scenes by superpixel co-occurrence and 3D geometry," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Sep.-Oct. 2009, pp. 625–632.

[55] T. Mita, T. Kaneko, B. Stenger, and O. Hori, "Discriminative feature co-occurrence selection for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 7, pp. 1257–1269, Jul. 2008.

[56] J. Tang, Z. Li, M. Wang, and R. Zhao, "Neighborhood discriminant hashing for large-scale image retrieval," *IEEE Trans. Image Process.*, vol. 24, no. 9, pp. 2827–2840, Sep. 2015.

[57] H. V. Henderson and S. R. Searle, "The vec-permutation matrix, the vec operator and Kronecker products: A review," *Linear Multilinear Algebra*, vol. 9, no. 4, pp. 271–288, 1980.

[58] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," 2008. [Online]. Available: http://www.vlfeat.org/

[59] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. 6th Indian Conf. Comput. Vis., Graph. Image Process.*, 2008, pp. 722–729.

[60] M. Chen *et al.*, "PFID: Pittsburgh fast-food image dataset," in *Proc. IEEE Int. Conf. Image Process.*, Nov. 2009, pp. 289–292.

[61] S. Yang, M. Chen, D. Pomerleau, and R. Sukthankar, "Food recognition using statistics of pairwise local features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2010, pp. 2249–2256.

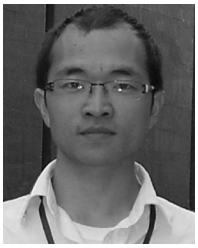
[62] R. Condorovici, C. Florea, R. Vranceanu, and C. Vertan, "Perceptually-inspired artistic genre identification system in digitized painting collections," in *Proc. Scand. Conf. Image Anal.*, 2013, pp. 687–696.

[63] F. S. Khan, S. Beigpour, J. van de Weijer, and M. Felsberg, "Painting-91: A large scale database for computational painting categorization," *Mach. Vision Appl.*, vol. 25, no. 6, pp. 1385–1397, 2014.

[64] Q. Zou, Y. Cao, Q. Li, C. Huang, and S. Wang, "Chronological classification of ancient paintings using appearance and shape features," *Pattern Recog. Lett.*, vol. 49, pp. 146–154, 2014.

[65] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes challenge 2012 (VOC2012) results," 2012. [Online]. Available: http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html

[66] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200–2011 dataset," California Inst. Technol., Pasadena, CA, USA, Tech. Rep. CNS-TR-2011-001, 2011.
[67] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar, "Cats and dogs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 3498–3505.
[68] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2015, pp. 1–9.
[69] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," *CoRR*, 2014. [Online]. Available: http://arxiv.org/abs/1408.5093



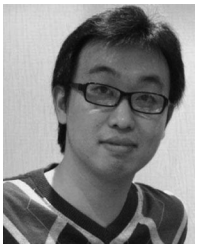
Qin Zou (M'13) received the B.E. degree in information science and the Ph.D. degree in photogrammetry and remote sensing (computer vision) from Wuhan University, Wuhan, China, in 2004 and 2012, respectively.

From 2010 to 2011, he was a visiting student with the Computer Vision Laboratory, University of South Carolina, Columbia, SC, USA. He is an Assistant Professor with the Department of Computer Science, Wuhan University. His research interests include computer vision, machine learning, and multimedia processing.

Prof. Zou is a Member of the ACM.

Lihao Ni is working toward the M.S. degree at the School of Computer Science, Wuhan University, Wuhan, China.

He actively participates in scientific research and was the recipient of the first prize in the "National Undergraduate Internet of Things Competition, China" in 2015. His research interests include pattern recognition and biometric identification.



Qian Wang (S'08–M'11) received the B.S. degree from Wuhan University, Wuhan, China, in 2003, the M.S. degree from Shanghai Institute of Microsystem and Information Technology (SIMIT), Chinese Academy of Sciences, Shanghai, China, in 2006, and the Ph.D. degree from Illinois Institute of Technology, Chicago, IL, USA, in 2012, all in electrical engineering.

He is a Professor with the School of Computer Science, Wuhan University. His research interests include wireless network security and privacy, cloud computing security, and applied cryptography.

Prof. Wang is a Member of the ACM. He is an expert under the National "1000 Young Talents Program" of China. He is a corecipient of the Best Paper Award from the 2011 IEEE International Conference on Network Protocols.

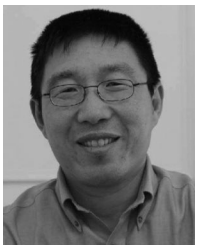
Zhongwen Hu received the B.Sc. degree in remote sensing and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2008 and 2013, respectively.

He is currently an Assistant Professor with the Key Laboratory for Geo-Environmental Monitoring of Coastal Zone, National Administration of Surveying, Mapping and Geoinformation, Shenzhen Key Laboratory of Spatial Smart Sensing and Services, and the College of Life Sciences and Oceanography, Shenzhen University, Shenzhen, China. His research interests include image segmentation and coastal remote sensing.

Qingquan Li received the Ph.D. degree in GIS and photogrammetry from the Wuhan Technical University of Surveying and Mapping, Wuhan, China, in 1998.

From 1988 to 1996, he was an Assistant Professor with Wuhan University, Wuhan, China, where he became an Associate Professor in 1996 and has been a Professor since 1998. He is currently the President and a Professor of Shenzhen University, Shenzhen, China; a Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, and the Director of the Shenzhen Key Laboratory of Spatial Smart Sensing and Service, Shenzhen, China. His research interests include intelligent transportation systems, 3D and dynamic data modeling, and pattern recognition.

Prof. Li is an academician of the International Academy of Sciences for Europe and Asia, an expert in Modern Traffic with the National 863 Plan, and an Editorial Board Member of the *Surveying and Mapping Journal* and the *Wuhan University Journal—Information Science Edition*.



Song Wang (S'00–M'02–SM'12) received the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign (UIUC), Champaign, IL, USA, in 2002.

From 1998 to 2002, he worked as a Research Assistant with the Image Formation and Processing Group, Beckman Institute, UIUC. In 2002, he joined the Department of Computer Science and Engineering, University of South Carolina, Columbia, SC, USA, where he is currently a Professor. His research interests include computer vision, medical image processing, and machine learning.

Prof. Wang is a Senior Member of the IEEE Computer Society. He is currently serving as the Publicity/Web Portal Chair of the Technical Committee of Pattern Analysis and Machine Intelligence of the IEEE Computer Society and an Associate Editor of *Pattern Recognition Letters*.