Hindawi
Mathematical Problems in Engineering
Volume 2018, Article ID 4945304, 14 pages
https://doi.org/10.1155/2018/4945304

Research Article

Mining Regional Co-Occurrence Patterns for Image Classification

Zhihang Ji,[1,2] **Sining Wu**,[1] **Fan Wang**,[1] **Lijuan Xu**,[1] **Yan Yang**,[1] **and Xiaopeng Hu**[1]

[1] *School of Computer Science and Technology, Dalian University of Technology, No. 2 Linggong Road, Dalian 116024, China*
[2] *Information Engineering College, Henan University of Science and Technology, No. 263 Kaiyuan Avenue, Luoyang 471023, China*

Correspondence should be addressed to Xiaopeng Hu; xphu@dlut.edu.cn

Received 30 April 2018; Revised 10 August 2018; Accepted 31 August 2018; Published 25 September 2018

Academic Editor: Bogdan Smolka

In the context of image classification, bag-of-visual-words mode is widely used for image representation. In recent years several works have aimed at exploiting color or spatial information to improve the representation. In this paper two kinds of representation vectors, namely, Global Color Co-occurrence Vector (GCCV) and Local Color Co-occurrence Vector (LCCV), are proposed. Both of them make use of the color and co-occurrence information of the superpixels in an image. GCCV describes the global statistical distribution of the colorful superpixels with embedding the spatial information between them. By this way, it is capable of capturing the color and structure information in large scale. Unlike the GCCV, LCCV, which is embedded in the Riemannian manifold space, reflects the color information within the superpixels in detail. It records the higher-order distribution of the color between the superpixels within a neighborhood by aggregating the co-occurrence information in the second-order pooling way. In the experiment, we incorporate the two proposed representation vectors with feature vector like LLC or CNN by Multiple Kernel Learning (MKL) technology. Several challenging datasets for visual classification are tested on the novel framework, and experimental results demonstrate the effectiveness of the proposed method.

1. Introduction

In computer vision research area, image classification refers to the ability of distinguishing images according to their visual content. During the past decade, that is a hot topic due to its extensive application. Since the existing gap between the semantic concept and visual information, it is still a very hard task to accomplish correctly. Then different strategy models have been proposed to fulfill the goal, such as BoVW (bag-of-visual-words) model [1, 2] and deep learning [3–5].

Recently, deep learning has attracted a lot of attention due to its excellent performance in multiple applications [6, 7]. For image classification task, deep learning has shown its outstanding results on most of the large-scale datasets. While this technology requires huge amounts of data to learn related features and parameters, it is unclear whether it is suitable for specific problems with relative small datasets.

Before the emergence of deep learning, BoVW model plays an important role in image classification tasks. Usually, most of the methods based on BoVW model use the well-known SIFT [8] or dense SIFT [9] descriptors as low-level features to represent images, which describe local regions of the image with intensity-level information. However these representations ignore the color cue and spatial information, which are often affected by illumination change, image rotating, position variance of object, and so on. That are considered to be some of the drawbacks for this original model.

In order to introduce color information, lots of methods have been put forward. A first set of methods consider the color information globally, as mentioned in [10]. Though those representations can extract the color information fast, no information about the spatial distribution of color is encoded, which has little discriminative power. A second set of methods are based on the fixed-size regions, which divide an image into fixed-size cells and extract color information from each separately. Usually those methods [11, 12] build a feature descriptor by fusing the texture cue with the color information and have a discriminative power. Finally, the

segmentation-based approaches are proposed to represent an image. In those methods, an image is divided into lots of regions by segmentation or clustering algorithm, then each region is encoded to a representative vector fusing color information and other cue, such as [13]. Due to including more information and keeping the region border well, those methods possess a more discriminative power.

In the meantime, some pioneering attempts to incorporate spatial information have been introduced. A dominant approach is the spatial pyramid match proposed by Lazebnik et al. [9], which describe the regions of an image in spatial position order. Further, several extensions of the SPM have been proposed. Sharma et al. [16] proposed a method to partition an image in learning way, addressing the problem of partitioning image uniformly. In a similar way, Jiang et al. [17] adopted Jensen-Shannon Tiling to learn the spatial BoVW representation which can partition an image adaptively. Different with SPM, which only considers the absolute spatial information, some research [18, 19] utilized the relative spatial relation, in which the property of co-occurrence was used often. Although those strategies are beneficial to boost the discriminative power further, there exists still some deficiency due to the color and spatial information is introduced separately. Thus, more discriminative methods are needed.

As we all know, the co-occurrence of features is helpful to the image classification task because it includes more information, such as relative relation between features. In this aspect, Li et al. [20] proposed a general framework called Markov stationary features (MSF) to extend histogram-based features. The MSF characterizes the spatial co-occurrence of histogram patterns by Markov chain models and encodes the spatial information in the feature representation. Although MSF is an innovative method to reflect the spatial cue by using the stationary distribution of the Markov chain models, it requires calculation of the higher-order transition matrix, which can be a prohibition for the histogram with large bins. Yi Yang et al. [21] introduced a novel representation termed spatial pyramid co-occurrence which captured both the absolute and relative spatial arrangement of the visual words and characterized a variety of spatial relationships. However the methods divided images by hard partitions without considering the locations of objects. The same object may be assigned into different partitions which reduced the accuracy of classification. Shulin Yang et al. [22] proposed a new representation for food images that calculated pairwise statistics at pixel levels. And then these statistics were accumulated in a multidimensional colorful histogram, which was used as a feature vector for a discriminative classifier. However the method chose pairwise colorful pixels randomly, which made the representation unsteady.

In order to make the best of the color and spatial information reasonably, two novel image representations are proposed which take into account color and spatial information at the same time. Both of them choose the superpixel as the basic unit and reflect the co-occurrence of the superpixels from different viewpoint. Then, an integrated image classification framework is constructed which combines the two proposed vectors with the traditional bag-of-feature or CNN feature by multiple kernel learning technology. Based on that, the scheme is evaluated on several of challenging datasets, and the experimental results show that the proposed methods are effective. Figure 1 illustrates the proposed framework.

The main contributions of our work lie in three aspects:

(a) First, due to the approximate uniform color distribution within each superpixel, the appearance is usually coherent. Based on this, we chose the mean color as a feature of a superpixel. Furthermore, we consider the regions, which are constructed by several of neighbored superpixels with similar mean color, as a block. Then a new representation for the image is designed, which not only characterizes the global distribution of the color blocks but also reflects the co-occurrence of the mean color within the block. We name the vector "Global Color Co-occurrence Vector" (GCCV).

(b) Second, using the color information of the pixels within the superpixel, another descripting method is proposed which is based on the co-occurrence of the colors. Different from the GCCV, this method describes the second-order distribution of the colors within the local pairs of neighboring superpixel. In this representation, the second-pooling technology is used [23], which maps all of the color co-occurrence information about the neighboring superpixels into a unified vector by a pooling-like procedure. The final vector obtained introduces the second-order information and is related to Riemannian manifold, which has a powerful discriminative ability. It is named as "Local Color Co-occurrence Vector" (LCCV).

(c) Third, compared to the traditional BoVW vector which only reflects the distribution of gray key-point descriptors, the two proposed ones include color and spatial information. These three kinds of representing vector have some complementary advantage. We adopt the multiple kernel learning technology to combine them together. By this way, we build a framework for image classification.

The rest of this paper is organized as follows: Related work is discussed in Section 2. In Section 3, the details of the proposed representation vectors are given. Then the performance of GCCV and LCCV on several of image datasets is shown in Section 4. Finally, we make a conclusion in Section 5.

2. Related Work

In the past few years, many efforts have been made to enrich the representation of images based on BoVW model, especially in introducing color cue [14, 25], spatial information [26, 27], co-occurrence [19, 28], and so on. In this section, we will review them briefly.

Color Cue. Original BoVW model describes the images with SIFT descriptor usually, in which only intensity information is used. In order to introduce color information, a lot of work has been achieved on creating new feature descriptors

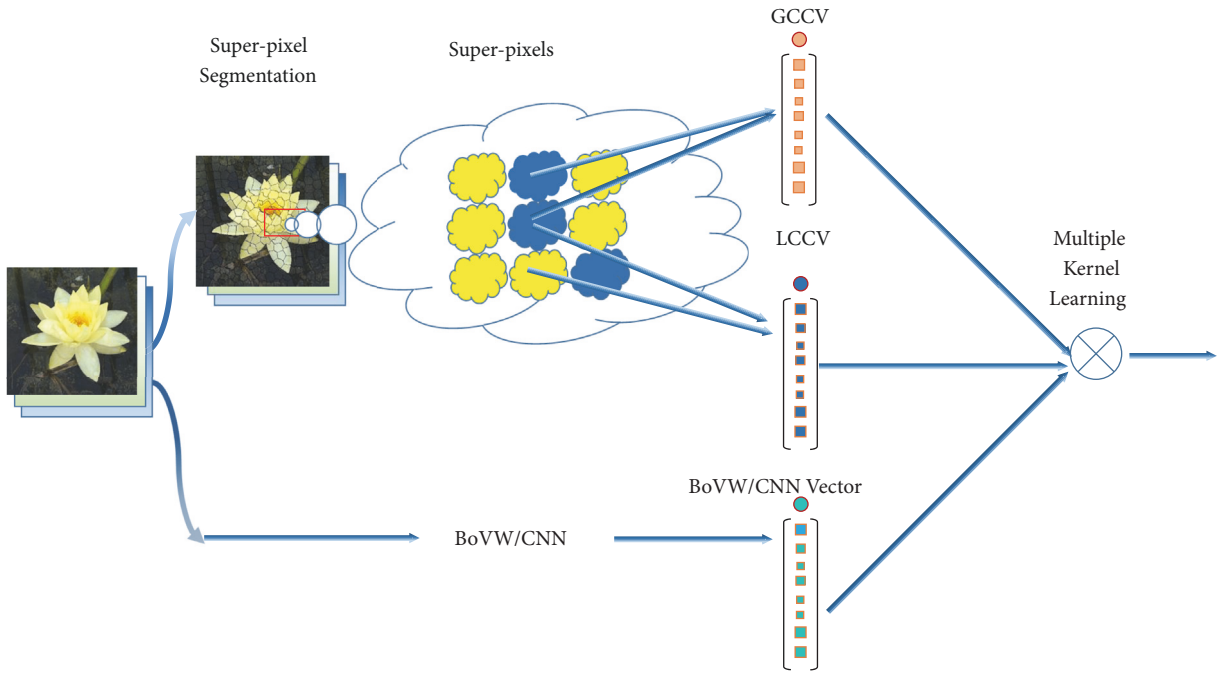

Figure 1: Overview of the proposed framework. (a) In the top two rows, superpixels are extracted from original images, and then the proposed representative vectors are built. (b) In the bottom row, the standard BoVW vector or CNN feature is extracted. (c) Finally, the three kinds of features are combined by MKL for the classification.

in which color channel information is encoded in various ways [14, 29, 30]. Koen E.A. van de Sande et al. [12] proposed CSIFT descriptor which extracted SIFT points on the opponent color channel while the intensity information had been eliminated. Later, the authors designed a new enhanced descriptor [31], TSIFT, which extracted SIFT points on intensity and opponent color channels. Different with the color feature descriptor, the research work [32] pointed out that the color histogram also was an effective way to describe images which was able to be used in object recognition and image classification. From analysis above-mentioned, the color information can play an important role in enhancing the classification accuracy. In this paper, we use the color information within superpixel regions to depict the image from local and global viewpoint, respectively.

Spatial Information. In the BoVW model, an image is represented as a histogram of quantized local features which loses spatial information about the patches. Recently, many methods have been proposed to incorporate spatial information into the BoVW representation. Some researchers [33, 34] encoded the absolute spatial information. However, these methods were not suitable for the condition where the image rotated its direction. Therefore, other methods which encoded the relative positional relationship between local patches were proposed. Saverese et al. [35] used relative distance of visual words to model the spatial correlations between quantized local descriptors. Different from directly using the distance information, Morioka and Satoh [36] proposed the Directional Local Pairwise Bases (DLPB) which introduced the directional information. Khan et al. [15] proposed a method to model global spatial distribution of visual words which exploits spatial orientations and distances of all pairs of similar descriptors in the image. Though these works make some advance, all of them are computationally expensive due to dealing with thousands of combinational pairings of key-point patches. In order to overcome this limitation, we choose the superpixel as the operation patch, the number of which in an image is less than the key-point descriptors. In addition, our method only exploits the local and similar superpixel. By this way, we can reduce the computation further.

Co-occurrence. As we all know, co-occurrence features are effective for object representation because it is far more informative than the occurrence of each event separately. So it is a common strategy to use co-occurrence to demonstrate spatial relationship and contextual information in image representation. In the past decade, several methods [19, 28] have been proposed to employ co-occurrence in visual classification tasks. According to whether the spatial neighboring relationship among features is used in computing the co-occurrence statistics, existing co-occurrence features can be sorted into two categories: global co-occurrence features and local co-occurrence features. Luo et al. [37] proposed a color edge co-occurrence histogram which was insensitive to object rotation, partial occlusion but not to scaling. In [28], Yuan *et al.* proposed to mine the co-occurrence statistics of SIFT descriptors for visual recognition. These works belong to the category of global co-occurrence features, while in some other literatures, the spatial co-occurrence was computed only within locally adjacent neighbors. In order to

augment the representative ability of co-occurrence further, in [38] the authors used a graphical model which modeled pairwise co-occurrence relationships of visual words and their spatial layout, achieving good performance in object and scene classification, respectively. Qi et al. [39] introduced a novel pairwise rotation invariant co-occurrence local binary pattern feature. Motivated by above works, we propose two types of co-occurrence representative vector, both of which are based on superpixel and fused with color information, but depict the image from different viewpoint separately.

3. The Framework with Proposed Representative Methods

In this section, we give the details of the proposed framework. First, the original LLC method [40] based on BoVW strategy is introduced briefly. And then, two novel representative methods exploiting superpixel are explained in detail. Finally, the fusion method of multiple kernel learning will be involved.

3.1. LLC Method for Image Classification. As mentioned above, BoVW strategy has been proven to be effective in image classification, which includes five key steps in its pipeline, such as feature extraction, building dictionary, feature coding, pooling operation, and classifier. Among them, feature coding is an important component and a lot of work has been devoted to it [41, 42]. From the hard voting [1] to the soft voting [43, 44] and then to those based on sparsity [41], the advance of encoding methods has been playing a great role in improving the classification performance [45]. Recent research has observed and validated that the locality is more essential than sparsity [40]. The locality-constrained linear coding (LLC) is an effective visual coding scheme which achieves state-of-the-art recognition performance. LLC utilizes the locality constraints to project each descriptor into its local coordinate system with low computational complexity. It aims to reconstruct visual descriptors with locality constraint and sparsity based on the following optimization criteria [40]:

$$\begin{aligned} \min_{C} \quad & \sum_{i=1}^{N} \left\| x_i - Bc_i \right\|^2 + \lambda \left\| d_i \odot c_i \right\|^2 \\ \text{s.t.} \quad & 1^T c_i = 1, \quad \forall i \end{aligned} \tag{1}$$

where B is a codebook learned from K-Means, $\odot$ denotes the elementwise multiplication, and d_i is the locality adaptor that gives different freedom for each basis vector proportional to its similarity to the input descriptor x_i, $C = [c_1, c_2, \cdots c_m]$ denotes the coding coefficients for X, and lambda is a constant to adjust the relationship between reconstruction errors and locality constraint. Because the LLC is based on SIFT descriptors which make use of the gradient cues, in our framework, it can be used to demonstrate the structure information of the images.

3.2. Global Color Co-Occurrence Vector. Color histogram is one of the most well-known technologies to express the color information of an image and widely used in some image recognition tasks. This method describes the global distribution of the color information. However, it does not consider the spatial relation between the colors, which can be regarded as a defect of this method. In the image classification task, the contents of images can vary significantly so that it is not suitable to use the color histogram directly.

In the meantime, we also know that human vision system always select a few elements of attention on some low-level local visual structures and suppresses irrelevant materials in recognition procedure [46]. On the other hand, the images often include various universal local structures and share some similar local structures when belonging to the same class. It is deserved to study how to represent an image using the local structures.

Based on the analysis above, we propose a novel representation which describes the color and structure information effectively. The idea is motivated by the research in [24], where Liu et al. designed a new descriptor named microstructure descriptor (MSD) which can characterize the local structures in the image, as shown in Figure 2.

In [24], an image $f(x, y)$ is denoted as $f(x, y) = c, c \in \{1, 2, \cdots M\}$, where c represents the color value of point (x, y) and M indicates number of principle colors in the dataset. In a 3×3 block of $f(x, y)$, $P_0 = (x_0, y_0)$ denotes the center position of the block, and let $f(P_0) = c_0$. The eight nearest pixels of P_0 are indicated by $P_i = (x_i, y_i)$ and let $f(P_i) = c_i$ $i = 1, 2, \cdots 8$. Taking this 3×3 block as filter and moving it from left-to-right and top-to-bottom throughout the image, the microstructures in the image can be detected and described as follows:

$$H(c_0) = \frac{N\{f(p_0) = c_0 \wedge f(p_i) = c_0 \parallel |p_i - p_0| = 1\}}{8\overline{N}\{f(p_0) = c_0\}} \quad \text{where } i \in \{1, 2, \cdots 8\} \tag{2}$$

Here, the symbol N denotes the cooccurring number of values c_0 and c_i, and $\overline{N}$ denotes the occurring number of the color c_0.

The MSD not only includes the color information but also embeds the spatial relationship between the pixels, which can be considered as a descriptor about the local visual structure. However, there exists some drawback with this method. As shown in Figure 2, the layout of these local structures are different but have the same MSD expression. At the same time, based on the pixel, this descriptor needs a great computation cost. So it is not suitable to be used directly in image classification task.

Here, we propose an improved method which overcomes the shortcoming of MSD and strengthens its robustness. The proposed method chooses the superpixel as basic unit and takes the mean color within it as its color value due to its almost coherent appearance. For image classification task, it is difficult to depict all the color information for every image precisely, because there are so many different colors in the images. Therefore we build a color codebook for the superpixels, which can be used to express the color information concisely. In the procedure of building color

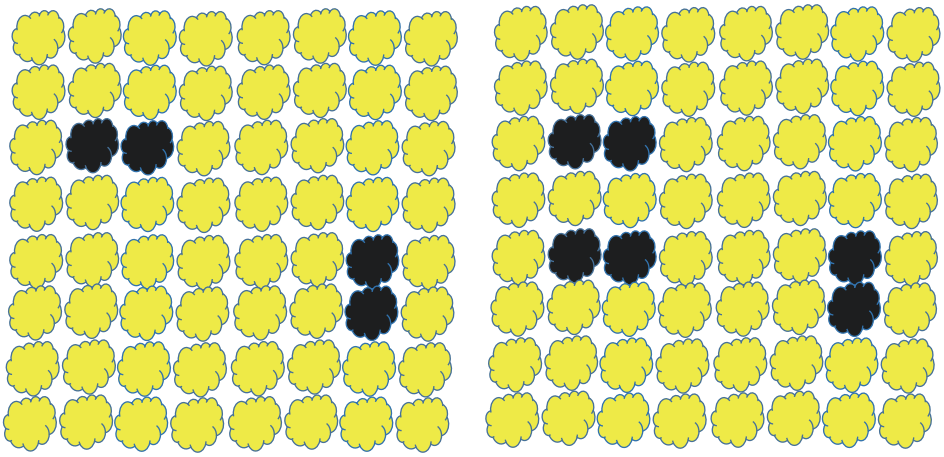

Figure 2: An illustration of the microstructures in the image. (a) In [24], a patch denotes a pixel, while in our work it denotes a superpixel. (b) The two images include different microstructures separately. However, we will obtain the same descriptor when adopting the method in [24]. The proposed method in our work can deal with this problem.

codebook, we randomly select some images from training set and collect all the mean color values of the superpixels in the selected images. Finally, the clustering method is applied to build the color codebook. After obtaining the color codebook, every superpixel can be assigned a color value in the color codebook that is closest to its original one, where the Euclid distance is used as the measurement metric. In Section 4, we will discuss the influence about the size of the codebook. Based on the superpixel and color codebook, we propose our representative method, which describes the distributions of colors and the spatial layout of them. Like MSD, the global distribution about structures with color c_i can be written as

$$H(c_i) = \frac{\sum_{j=1}^{K_i} \left(\left(f\left(sp_j\right) = c_i \wedge f\left(sp_{nj}\right) = c_i \right) / N\left\{ Neighborhood\left(f\left(sp_j\right) = c_i \right) \right\} \right)}{K_i} \frac{N(sp)}{K_i} \quad (3)$$

$$where\ sp_{nj} \in Neighborhood\left(sp_j\right)$$

Here, $H(c_i)$ and K_i denote, respectively, the distribution and the number of the superpixels whose color is c_i in an image. $N(sp)$ denotes the number of all the superpixels in an image. And $N\{Neighborhood(f(sp_j) = c_i)\}$ shows the number of superpixels which are in the neighborhood of a superpixel whose color is c_i. The same as the MSD, we exploit the co-occurrence of the superpixel within a neighborhood but our method overcomes the drawback mentioned above by introducing the ratio between $N(sp)$ and K_i. By the same way for every color, we define the GCCV as

$$GCCV(img) = \left[H(c_1), H(c_2), \cdots, H(c_M) \right] \quad (4)$$

Here, M denotes the color number in the color codebook. For an image, the value of $N(sp)$ is controlled and set as the same value and we can omit it in the computation process. This representative vector has the benefit which not only describes the global distribution of the color structures but also reflects the latent spatial relation within it. Algorithm 1 shows the creating procedure in detail.

3.3. Local Color Co-Occurrence Vector. In the preceding section, we propose the GCCV to represent an image which utilizes the color and spatial information. This method describes the image roughly because it only chooses the mean color value in a superpixel region as the basic feature and neglects the detailed information in that area.

In order to overcome this defect, we need a more powerful representation. As we all know, there are many strategies to depict the pattern of a region in an image, such as texture description, color histogram, and local gradient descriptor. Unlike the GCCV in Lab space, here we propose another method to depict superpixel in RGB space, which can be treated as a complement to the GCCV. Figure 3 shows the building procedure.

As we all know, *RGB* histogram is often used in computer vision research area, which is a combination of three $1D$ histograms based on the R, G, B channels. But this representation is often effected by illumination change and not suitable for the image classification task. Therefore we use the rghistogram [12] to depict the superpixel region. The

Input: (a) Input image I

(b) Parameters of superpixel R

(c) Color vocabulary D

Output: GCCV Vector

(1) generate N superpixels

$$[sp_1, \cdots, sp_N] \longleftarrow Superpixel(I, R)$$

(2) according to the color vocabulary D, assigning color value for every superpixel sp_i.

(3) for $i = 1 : 1 : N$

(1) $j \longleftarrow op\{$read the color value of superpixel $i\}$

(2) accumulate the number of supper-pixel with color j.

$$Sum\_C_j \longleftarrow Sum\_C_j + 1$$

(3) $N(Neighbor(sp_i)) \longleftarrow op\{$count the number of neighbor of superpixel $i\}$

(4) $N(SC(sp_i)) \longleftarrow$

$op\{$count the number of neighbor with the same color as superpixel $i\}$

(5) according to the formula (3) compute the value

$$ratio_j \longleftarrow \frac{N(SC(sp_i))}{N(Nbor(sp_i))}$$

and $\text{sum\_}ratio_j \longleftarrow \text{sum\_}ratio_j + ratio_j$

end

(4) for $k = 1 : 1 : M$

$$H(c_k) \longleftarrow \frac{\text{sum\_}ratio_k}{Sum\_C_k} \times \frac{1}{Sum\_C_k}$$

end

return $GCCV = [H(c_1), H(c_2), \cdots H(c_M)]$

Algorithm 1: Global Color Cooccurrence Vector.

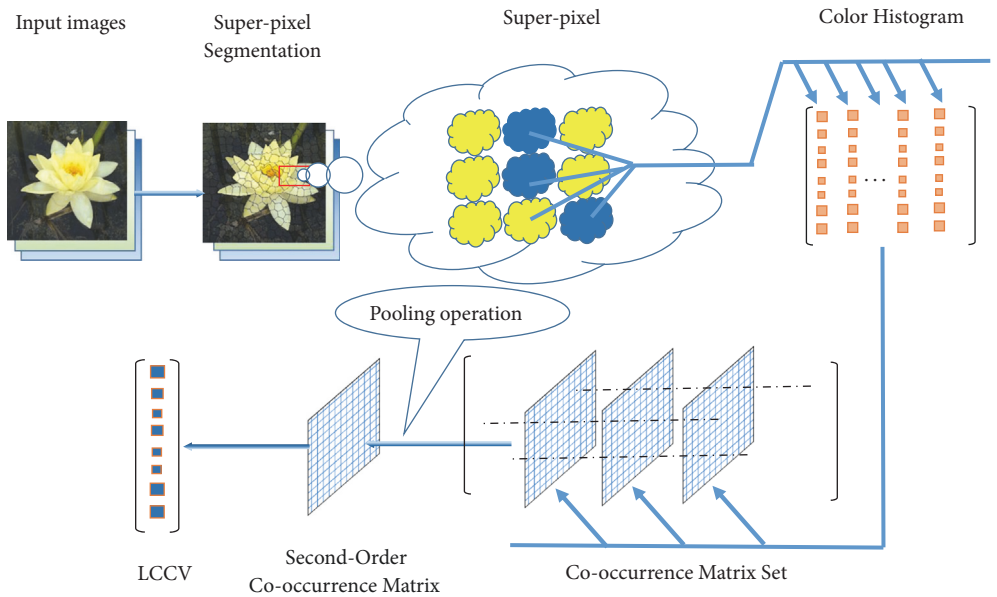

Figure 3: The procedure of building LCCV descriptor.

rghistogram is based on the normalized *RGB* color model which utilizes the chromaticity components r and g to describe the color information in the image (b is redundant as $r + g + b = 1$), where

$$\begin{pmatrix} r \\ g \\ b \end{pmatrix} = \begin{pmatrix} \frac{R}{R+G+B} \\ \frac{G}{R+G+B} \\ \frac{B}{R+G+B} \end{pmatrix} \tag{5}$$

Due to the normalization, r and g are scale-invariant and thereby invariant to light intensity changes, shadows [12]. And besides that, in order to enhance its invariance further, a codebook in *rg* color space is built and then the similar points in the rg color space can be substituted approximately by one corresponding "color word", which will alleviate the negative influence of the photometric changes.

Additionally, as mentioned above, the co-occurrence has the property to reflect the spatial relation between superpixels. Therefore another represent vector is proposed which utilizes the color histogram to depict the pattern of the superpixel and adopt the co-occurrence to represent the spatial relation implicitly between them. Here, it is important to note that only the co-occurrence between neighboring superpixels is considered.

Suppose SP_a and SP_b are two neighboring superpixels in an image, while $rgHist_a = \{rg_{a1}, rg_{a2} \cdots, rg_{aM}\}$ and $rgHist_b = \{rg_{b1}, rg_{b2} \cdots, rg_{bM}\}$ are the rghistograms, respectively. Then the co-occurrence matrix of these two superpixels can be described as

$$C_{ab} = rgHist_a^T \cdot rgHist_b \quad (where\ rgHist\ representing\ row\ vector) \tag{6}$$

which can be used to depict the co-occurrence of color distribution in a neighboring area. To some extent, if the histogram is considered as the first-order color distribution information, then the co-occurrence can provide the second-order information, which is valuable for the image classification [47, 48].

Based on the discussion mentioned above, the creating procedure of Local Color Co-occurrence Vector can be represented as follows: First, for a superpixel, we need to compute multiple co-occurrence matrices between each of its neighboring superpixels and itself. Second, we perform the same operation for every superpixel in an image according to formula (6). Third, all the co-occurrence matrices are integrated by second-pooling way, namely, computing the mean value of all the co-occurrence matrix at the same position in average pooling operation (the reason for choosing average pooling operation will be discussed in Section 4.2). This can be formularized as

$$\begin{aligned} C_{mean} &= \frac{1}{\sum_{i=1}^{N} K_i} \sum_{i=1}^{N} \sum_{j}^{K_i} C_i^j \\ &= \frac{1}{\sum_{i=1}^{N} K_i} \sum_{i=1}^{N} \sum_{j}^{K_i} rgHist_i^T \cdot rgHist_{ij} \end{aligned} \tag{7}$$

where N denotes the total number of the superpixels in an image and K_i is the number of superpixels neighboring the ith one. The symbol $rgHist_i$ denotes the rghistogram of the ith superpixel and $rgHist_{ij}$ indicates the rghistogram of the jth superpixel neighboring the ith one.

C_{mean} is a symmetric and positive-definite (SPD) matrix, which does not embed in an Euclidean space but a smooth Riemannian manifold. In the Log-Euclidean framework [49], the SPD matrices form a commutative Lie group which is equipped with an Euclidean structure. This framework enables us to compute the logarithms of SPD matrices, which can be flexibly and efficiently handled with common Euclidean operations.

Based on above analysis, C_{mean} can be vectorized by this way,

$$LCCV\ (I) = Vectorize\left(Triu\left(logm\left(C_{mean}\right)\right)\right) \tag{8}$$

where $logm(\cdot)$ denotes the *logm* transformation for SPD matrix and $Triu(\cdot)$ operation unpacks the upper elements of the matrix then $Vectorize(\cdot)$ concatenates it into a vector. Algorithm 2 shows the procedure of creating LCCV in detail.

3.4. Feature Fusion with Multiple Kernel Learning. In this paper, we use multiple kernel learning (MKL) technology to integrate various kinds of image features. Based on the MKL, we can train a SVM with a combined kernel, which can be used to fuse the different kinds of features with adaptive weight. The combined kernel can be expressed as follows:

$$K_{comb}(x, y) = \sum_{i=1}^{K} \alpha_i K_i(x, y) \quad where\ \alpha_i \geq 0 \text{ and } \sum_{i=1}^{K} \alpha_i = 1 \tag{9}$$

where α_i is the weight to combine sub-kernel $K_i(x, y)$. Here, we need to choose the types of kernels for each feature, depending on their discriminative power and computational cost, while the weight can be estimated automatically by MKL technology from the training data. By this way, we can fuse the proposed features with other features to obtain an optimal classifier.

4. Experiments and Results

In this section, we present the datasets used and the implementation details first. Then, the effectiveness of the proposed representation and framework will be evaluated and validated through extensive experimental comparisons. All of the algorithms are implemented by the Matlab software with the third-party library of VLfeat [50] and MatConvNet [51].

4.1. Datasets and Experimental Setting. In the experimental procedure, we evaluate and validate the effectiveness of our proposed method in Flowers17 [52], Flowers102 [53], Caltech101 [54], and Caltech256 [55] separately. In order to compare our methods with other work, we inherit the same settings from the state-of-the-art algorithms. For LLC

Input: (a) Input image I
(b) Parameters of superpixel R
Output: LCCV Vector
(1) generate N superpixels

$$[sp_1, \cdots, sp_N] = Superpixel(I, R)$$

(2) $rghist(sp_i) \longleftarrow op\{$compute rghistogram for every superpixel $sp_i\}$.
(3) for $i = 1 : 1 : N$
(1) $\{Neig(sp_i)\} \longleftarrow op\{$search the neighbors of superpixel $i\}$.
(2) $\{C_i^j\} \longleftarrow op\left\{\begin{array}{c} for\ every\ superpixel\ j, \\ compute\ the\ co-occurrence\ matrix \\ between\ superpixel\ j\ and\ i\ based\ on\ the\ rghistogram. \end{array}\right\}$.
where $j \in \{Neig(sp_i)\}$
end
(4) compute the mean co-occurrence matrix C_{mean}

$$C_{mean} \longleftarrow \frac{1}{\sum_{i=1}^{N} K_i} \sum_{i=1}^{N} \sum_{j}^{K_i} C_i^j$$

(5) $LCCV \longleftarrow op\{$vectorize upper triangle matrix of logm(C_{mean}) $\}$
return $LCCV$

Algorithm 2: Local Color Cooccurrence Vector.

representation, SIFT features are extracted from densely located patches centered at every 4 pixels on gray images and the spatial bin sizes are multiscale with the set of $\{4, 6, 8, 10\}$ pixels. Then, we use K-Means algorithm to construct codebook. The codebook size is 1024 for Flowers17, 2048 for Flowers102 and Caltech101, and 4096 for Caltech256. We apply a 3-layer $\{1\times1, 2\times2, 4\times4\}$ SPM for enhancing the spatial context. After that, L2-norm normalization is performed. Finally, linear SVM classifier is adopted to train and test images. For GCCV and LCCV representation, SVM classifier with χ^2 kernel is adopted.

4.2. The Key Parameters for GCCV and LCCV. In our approach, several of parameters have to be set carefully which have an important influence on the classification results. The first are about color space quantization which are related to the GCCV and LCCV, respectively. The second are related to superpixel which control the size of the superpixels region and the strength of the spatial regularization.

(a) The Size of the Color Codebook. As mentioned above, in the process of creating the GCCV and LCCV vector, both of them need to build a color codebook, so the size of the codebook is an important parameter for the two vectors. Because the GCCV and LCCV depend on the Lab color space and RGB color space, respectively, we need to quantize the corresponding color space separately. Figures 4 and 5 demonstrate the performance of the GCCV and LCCV with the different color codebook size separately. From it, we can know how the codebook size affects the performance. Comparing the trade-off between the computation cost and classification results, we choose the 30 and 50 as the best size for GCCV and LCCV to Flowers17 dataset, 50 and 60 as the best size to Flowers102 dataset, and 70 and 80 to Caltech101 dataset, 80 and 80 to Caltech256 dataset.

(b) Max Pooling versus Average Pooling. In the LCCD, for each superpixel, the co-occurrence information of the colors is aggregated together to get the corresponding pooled feature, as shown in Figure 3. Usually, two pooling methods have been used in image representation: average pooling and max pooling, here which mean getting the average value or max value at the same position in the co-occurrence matrix, respectively. Figure 6 shows the contrast result which come from average pooling and max pooling, respectively. From it, we can find the average pooling is the best choice, as we do in Section 3.3. The reason is that the distribution of the co-occurrence information is dense, only less than 30% elements equaling to zero in the representation vector, while the max pooling is always popular in conjunction with sparse information [56].

(c) The Region Size of the Superpixel. Both of the GCCV and LCCV are based on superpixel, and the region size of the superpixel has an important effect on them. Figures 7 and 8 reflect the influence separately.

For the GCCV, which expects to express the global distribution of color information combining some spatial cue implicitly, the homogeneity of the color in a superpixel is necessary. So the smaller regional size is, the easier it is to satisfy that need. But if it is set to its lower limit value, few pixels per superpixel, it will increase the computation cost and complexity of the spatial relation and at the same time lose its discrimination ability. Hence, we assign the value 20 to the region size of the SLIC superpixel for all of the datasets.

For the LCCV, which makes use of the co-occurrences of the color information between the neighboring superpixels, the average pooling is the best choice as the above-mentioned and Figure 7 shows the impact of the region size about superpixel on the accuracy.

Next, we will further explain the relationship between the region size and the average pooling operation. Given

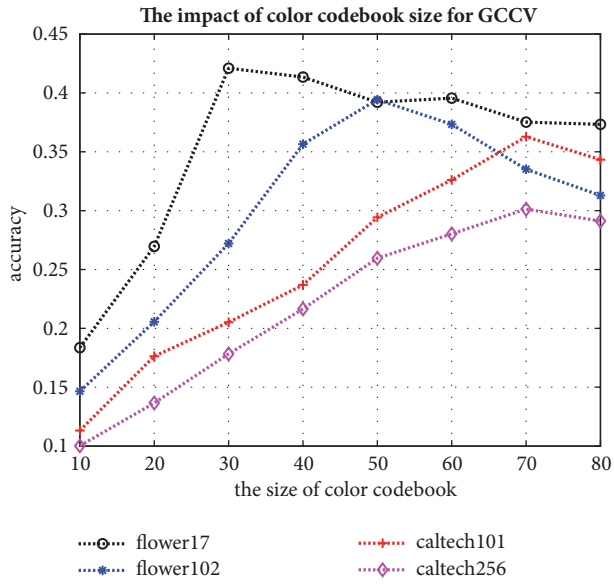

Figure 4: The impact of color codebook size of GCCV to four different datasets.

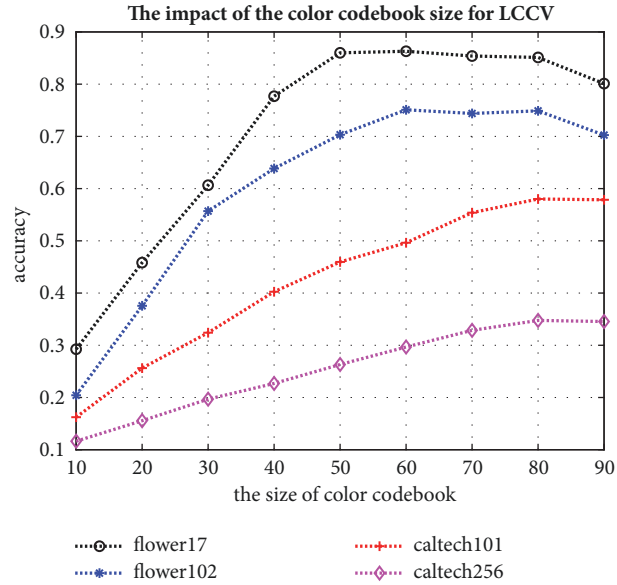

Figure 5: The impact of color codebook size of LCCV to four different datasets.

two neighboring superpixels SP_i and SP_j, f_i and f_j are two variables corresponding to one bin value of the color histogram, respectively, which are satisfied with the binomial distribution $B(N_i, \alpha_i)$ and $B(N_j, \alpha_j)$ and independent identical distribution, as the same assumption with [56]. Here, N_i and N_j are the number of pixels in a superpixel which is in proportion to the region size, while α_i and α_j are the mean value. According to the analysis in [56], the variance of the pooling variable is the key factor. The lower the variance, the stronger its discriminative ability. In this case, the pooling variable $Corr_{ij}$, which is the corresponding element in color co-occurrence matrix, is the product of two bin values coming from different color histogram, i.e., $Corr_{ij} = f_i * f_j$. Because the variances of f_i and f_j are in proportion to $1/N_i$ and$1/N_j$, the variance of $Corr_{ij}$ is related to the $1/(N_i * N_j)$, $1/N_i$ and $1/N_j$. So, the variance of the pooling result will decreasing as the size of the superpixel increases and the LCCV will get more discriminative power. But if the size of superpixel increases further and the number of superpixels of an image will decrease, the discriminative ability of LCCV will decline due to the weakness of spatial information between the superpixels. This is in accordance with our experiment results.

4.3. The Results and Analysis

4.3.1. Validation of Our Methods. In this subsection, we are to demonstrate experiment results on four datasets to further

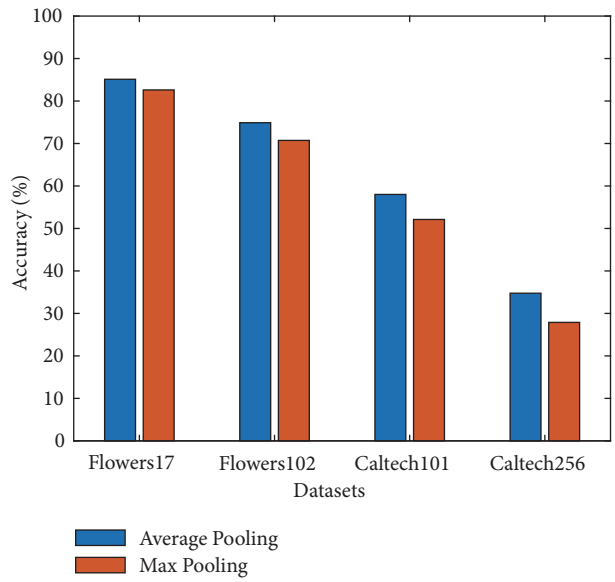

Figure 6: Comparison for predictions of LCCD when adapting average pooling operation and max pooling operation for four different datasets.

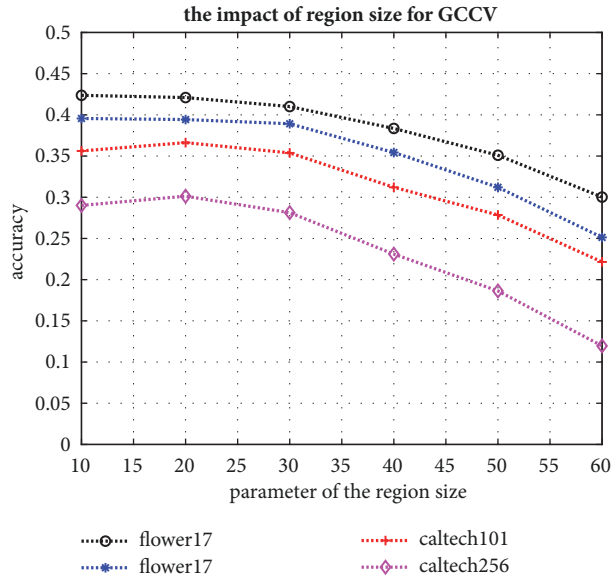

Figure 7: The impact of region size for GCCV to four different datasets.

analyze the validation of our methods. Table 1 contrastively shows the classification accuracy of our proposed methods.

(a) Experimental Results for Flowers17 and Flowers102. The Flowers17 dataset has 1360 images with 17 classes of flowers. Each class has 80 images. In our experiments, for each class, 40 and 20 images are used for training and validation with the rest for testing. The Flowers102 dataset has 8,189 images with 101 classes of flowers. Each class has a minimum of 40 images. In our experiments, 15 and 15 images are used for training and validation with the rest for testing. For these two datasets, we resize images to ensure its size not larger than 480∗480.

Because the two datasets are specific to flower object which have rich color information, the color cue plays a key role in the recognizing progress. Here, we chose the original LLC method as the comparison object to examine the effectiveness of our proposed. With the setting mentioned above, LLC achieves 71.26% and 70.29% accuracy on the two datasets, respectively, which chooses the gray dense SIFT as the feature here. The proposed method GCCV, which mainly reflects the global color distribution of the image, achieves 41.10% and 39.43% accuracy. The second proposed methods LCCV, which mainly make use of the local color co-occurrence of the image, outperform LLC by a margin

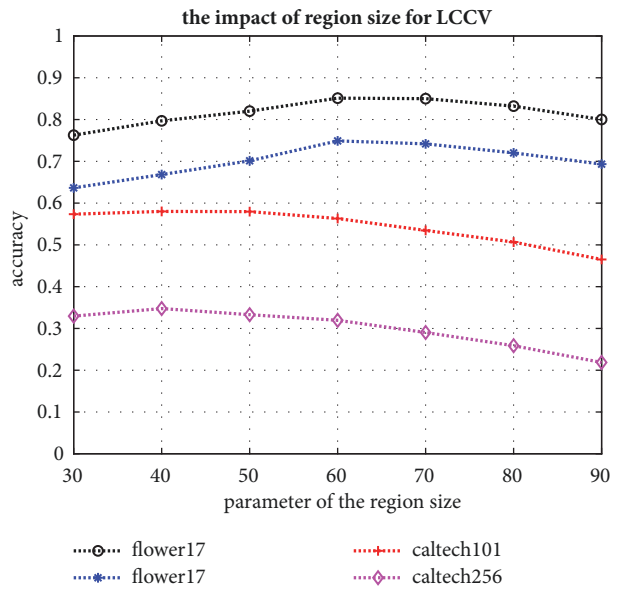

Figure 8: The impact of region size for LCCV to four different datasets.

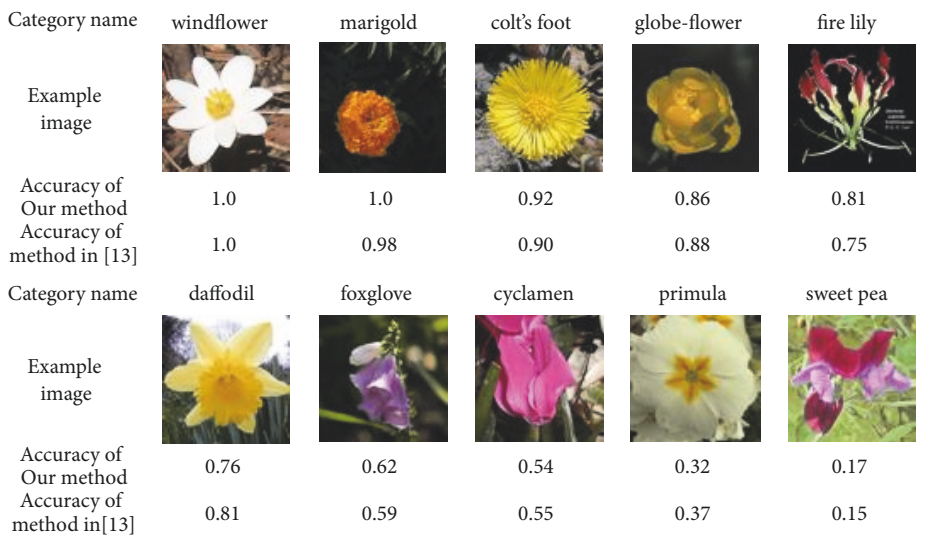

Figure 9: Example images with classification result from the Flower102 dataset.

about 15% and 5%. This confirms that LCCV is effective which combines the spatial and color information by the second-pooling way. In addition, we try to combine the GCCV with LCCV using MKL technology, which can advance the accuracy further. In the same way, the combination of GCCV, LCCV, and LLC can gain better result. Some example images with the classification results are illustrated in Figure 9.

(b) Experimental Results for Caltech101 and Caltech256. The Caltech101 dataset has 102 object classes, including a background class, with high intraclass appearance variability. The number of images per category varies from 31 to 800 images.

In our experiments, for each class, 20 images are used for training and validation and 20 for testing. The Caltech256 dataset has 257 categories, including a cluster class. The number of images per category varies from 80 to 827. In our experiments, for each class, 50 images are used for training and validation and 20 for testing. For these two datasets, we resize images to ensure its size not larger than $300*300$.

Different from the Flowers dataset, the color of the images in Caltech is distributed in a large range and is not suitable to be chosen as the key discriminative element. Even so, the color information can play a subsidiary role. The result in Table 1 shows that the LLC method achieves 74.14% and 49.59% accuracy for Caltech101 and Caltech256, and the GCCV and LCCV only achieve 36.28% and 58.01% for Caltech101 and 30.13% and 34.76% for Caltech256, respectively. But when we combine the LLC, GCCV, and LCCV with MKL technology, 77.02% and 51.32% accuracy are achieved. This can show that the color and spatial relative information can

Table 1: Comparison with the LLC features in different ways on experimental datasets.

	LLC	GCCV	LCCV	GCCV+LCCV	GCCV+LCCV+LLC
Flowers17	71.26	42.10	86.00	88.62	90.13
Flowers102	70.29	39.43	75.09	77.61	78.35
Caltech101	74.14	36.28	58.01	62.13	77.02
Caltech256	49.59	30.13	34.76	37.29	51.32

Table 2: Comparison with closely related works on experimental datasets.

	GCCV+LCCV+LLC	CA(SIFT, {CN, HUE}) [14]	SPS [15]	RCC [13]
Flowers17	90.13	89.00	81.64	90.06
Flowers102	78.35	71.63	68.54	76.15
Caltech101	77.02	57.26	60.66	75.35
Caltech256	51.32	30.18	36.82	50.61

enrich the representation of the image and improving the discriminative ability.

4.3.2. Comparison with Closely Related Works. Here, we compare our method with [13–15]. All of them concern modeling color and spatial information into the image representation. Table 2 shows the details of the comparisons. For those datasets, our representation provides the best classification results. Our method holds different advantages over other methods. For example, David A. Rojas Vigo [14] introduces the color information both in the feature detection and in extraction stages but ignores the spatial information. On the contrary, only the spatial information between patterns is considered in [15]. Although authors [13] pay attention to color and spatial information at the same time, they only focus on the information in local area and ignore the global distribution.

4.3.3. Comparison to State of the Art. To further evaluate the performance of our methods, we fuse them with CNN features. Here we choose the VGGNet16 [4] pretrained model, which is trained on ImageNet dataset, as feature extractors. In the step of extracting features, an image from 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. Here, we still use the MKL technology to fuse the different features and set the same parameter value of the GCCV and LCCV as above. The results are shown in Table 3. From the table we can see that, comparing with original CNN, the (GCCV+LCCV+CNN) achieves an improvement in accuracy on the four datasets, respectively. It indicates that the combination of GCCV and LCCV captures some information which can enrich the CNN feature further and is useful for improving the recognition performance.

4.3.4. Summary for the Result Analysis. From the results above, especially Table 1, we can know that GCCV representation is not suitable to be used alone, which only demonstrates the color distribution of the patterns globally. The LCCV describes the second-order color information of the patterns with implicit spatial relationship, which gain good result when it is used alone to deal with datasets like Flower17 and Flower102. When the datasets like Caltech101 and Caltech256 are given, where the color information is not the important cue for the object, the accuracy of classification is still low. So, for the datasets like Caltech, structural information like BoVW representation, i.e., LLC, can be introduced. Experiment results show the combination of GCCV, LCCV, and LLC (or CNN) is a powerful representation.

Table 3: Comparison with the original CNN features on testing datasets by the MKL way.

	CNN	GCCV+LCCV+CNN
Flowers17	87.99	92.13
Flowers102	79.53	82.25
Caltech101	92.00	93.01
Caltech256	76.53	77.47

5. Conclusion

In this paper, two representation vectors for image classification are proposed, both of which are capable of modeling the colors distribution efficiently based on co-occurrence. In order to express the color distribution of the images adequately, the vectors are designed from the different viewpoints. The first one, named GCCV, describes the global distribution of colors in large scale with embedding the spatial information implicitly. As a complementary, the LCCV records the co-occurrence of color distribution within a local area, which reflects the second-order color information in detail. Fusing the proposed vectors with other representation like BoVW vectors or CNN vectors by MKL technology, we can obtain more discriminative representation. Experiments demonstrate that our approach succeeds in introducing spatial and color information into image representation and it outperforms other state-of-the-art methods.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proceedings of the European Conference on Computer Vision (ECCV '4)*, pp. 1–22, 2004.

[2] J. Deng, A. C. Berg, K. Li, and L. Fei-Fei, "What does classifying more than 10,000 image categories tell us?" in *Proceedings of the European Conference on Computer Vision (ECCV 2010)*, pp. 71–84, 2010.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 26th Annual Conference on Neural Information Processing Systems (NIPS '12)*, pp. 1097–1105, December 2012.

[4] K. Simonyan and A. Zisserman, *Very Deep Convolutional Networks for Large-Scale Image Recognition*, 2014, https://arxiv.org/abs/1409.1556.

[5] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[6] L. Deng, J. Li, J.-T. Huang et al., "Recent advances in deep learning for speech research at Microsoft," in *Proceedings of the 2013 38th IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2013*, pp. 8604–8608, Canada, May 2013.

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Proceedings of the Conference on Neural Information Processing Systems (NIPS 2015)*, pp. 91–99, 2015.

[8] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the 7th IEEE International Conference on Computer Vision (ICCV '99)*, vol. 2, pp. 1150–1157, Kerkyra, Greece, September 1999.

[9] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, pp. 2169–2178, June 2006.

[10] M. J. Swain and D. H. Ballard, "Color indexing," *International Journal of Computer Vision*, vol. 7, no. 1, pp. 11–32, 1991.

[11] J. Zhang, Y. Barhomi, and T. Serre, "A new biologically inspired color image descriptor," in *Proceedings of the European Conference on Computer Vision (ECCV 2012)*, pp. 312–324, 2012.

[12] K. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1582–1596, 2010.

[13] Q. Zou, L. Ni, Q. Wang, Z. Hu, Q. Li, and S. Wang, "Local Pattern Collocations Using Regional Co-occurrence Factorization," *IEEE Transactions on Multimedia*, vol. 19, no. 3, pp. 492–505, 2017.

[14] D. A. Rojas Vigo, F. S. Khan, J. Van De Weijer, and T. Gevers, "The impact of color on bag-of-words based object recognition," in *Proceedings of the 2010 20th International Conference on Pattern Recognition, ICPR 2010*, pp. 1549–1553, Turkey, August 2010.

[15] R. Khan, C. Barat, D. Muselet, and C. Ducottet, "Spatial histograms of soft pairwise similar patches to improve the bag-of-visual-words model," *Computer Vision and Image Understanding*, vol. 132, pp. 102–112, 2015.

[16] G. Sharma and F. Jurie, "Learning discriminative spatial representation for image classification," in *Proceedings of the 2011 22nd British Machine Vision Conference, BMVC 2011*, pp. 1–11, UK, September 2011.

[17] L. Jiang, W. Tong, D. Meng, and A. G. Hauptmann, "Towards efficient learning of optimal spatial Bag-of-Words representations," in *Proceedings of the 2014 4th ACM International Conference on Multimedia Retrieval, ICMR 2014*, pp. 121–128, UK, April 2014.

[18] A. Bolovinou, I. Pratikakis, and S. Perantonis, "Bag of spatio-visual words for context inference in scene classification," *Pattern Recognition*, vol. 46, no. 3, pp. 1039–1053, 2013.

[19] F. Xiao and Y. J. Lee, "Discovering the spatial extent of relative attributes," in *Proceedings of the 15th IEEE International Conference on Computer Vision, ICCV 2015*, pp. 1458–1466, Chile, December 2015.

[20] J. Li, W. Wu, T. Wang, and Y. Zhang, "One step beyond histograms: image representation using markov stationary features," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008)*, pp. 1–8, 2008.

[21] Y. Yang and S. Newsam, "Spatial pyramid co-occurrence for image classification," in *Proceedings of the 2011 IEEE International Conference on Computer Vision, ICCV 2011*, pp. 1465–1472, IEEE, Barcelona, Spain, November 2011.

[22] S. Yang, M. Chen, D. Pomerleau, and R. Sukthankar, "Food recognition using statistics of pairwise local features," in *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2010*, pp. 2249–2256, USA, June 2010.

[23] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu, "Free-form region description with second-order pooling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 6, pp. 1177–1189, 2015.

[24] G.-H. Liu, Z.-Y. Li, L. Zhang, and Y. Xu, "Image retrieval based on micro-structure descriptor," *Pattern Recognition*, vol. 44, no. 9, pp. 2123–2133, 2011.

[25] B. Gecer, G. Azzopardi, and N. Petkov, "Color-blob-based COSFIRE filters for object recognition," *Image and Vision Computing*, vol. 57, pp. 165–174, 2017.

[26] J. Sánchez, F. Perronnin, and T. De Campos, "Modeling the spatial layout of images beyond spatial pyramids," *Pattern Recognition Letters*, vol. 33, no. 16, pp. 2216–2223, 2012.

[27] M. Malinowski and M. Fritz, *A pooling Approach to Modelling Spatial Relations for Image Retrieval and Annotation*, 2014, https://arxiv.org/abs/1412.2133.

[28] J. Yuan, M. Yang, and Y. Wu, "Mining discriminative co-occurrence patterns for visual recognition," in *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011*, pp. 2777–2784, USA, June 2011.

[29] B. Fernando, E. Fromont, D. Muselet, and M. Sebban, "Discriminative feature fusion for image classification," in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2012*, pp. 3434–3441, USA, June 2012.

[30] S. Tahery and M. S. Drew, "A novel colour hessian and its applications," *Electronic Imaging*, vol. 2017, no. 18, pp. 171–176, 2017.

[31] K. van de Sande and T. Sande, "Illumination-invariant descriptors for discriminative visual object categorization," Technical Report University of Amsterdam, 2012.

[32] G.-H. Liu and J.-Y. Yang, "Content-based image retrieval using color difference histogram," *Pattern Recognition*, vol. 46, no. 1, pp. 188–198, 2013.

[33] L. Xie, Q. Tian, M. Wang, and B. Zhang, "Spatial pooling of heterogeneous features for image classification," *IEEE Transactions on Image Processing*, vol. 23, no. 5, pp. 1994–2008, 2014.

[34] O. A. B. Penatti, F. B. Silva, E. Valle, V. Gouet-Brunet, and R. D. S. Torres, "Visual word spatial arrangement for image retrieval and classification," *Pattern Recognition*, vol. 47, no. 2, pp. 705–720, 2014.

[35] S. Savarese, J. Winn, and A. Criminisi, "Discriminative object class models of appearance and shape by correlatons," in *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2006*, pp. 2033–2040, USA, June 2006.

[36] N. Morioka and S. I. Satoh, "Learning directional local pairwise bases with sparse coding," in *Proceedings of the 2010 21st British Machine Vision Conference, (BMVC 2010)*, pp. 1–11, UK, September 2010.

[37] J. Luo and D. Crandall, "Color object detection using spatial-color joint probability functions," *IEEE Transactions on Image Processing*, vol. 15, no. 6, pp. 1443–1453, 2006.

[38] A. Bosch, A. Zisserman, and X. Muñoz, "Scene classification using a hybrid generative/discriminative approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 4, pp. 712–727, 2008.

[39] X. Qi, R. Xiao, C.-G. Li, Y. Qiao, J. Guo, and X. Tang, "Pairwise rotation invariant co-occurrence local binary pattern," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 11, pp. 2199–2213, 2014.

[40] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '10)*, pp. 3360–3367, IEEE, June 2010.

[41] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '09)*, pp. 1794–1801, 2009.

[42] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *Proceedings of the 11th European Conference on Computer Vision (ECCV '10)*, pp. 143–156, Crete, Greece, 2010.

[43] J. C. van Gemert, J. Geusebroek, C. J. Veenman, and A. W. Smeulders, "Kernel codebooks for scene categorization," in *Proceedings of the European Conference on Computer Vision (ECCV 2008)*, pp. 696–709, 2008.

[44] L. Liu, L. Wang, and X. Liu, "In defense of soft-assignment coding," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV '11)*, pp. 2486–2493, 2011.

[45] Y. Huang, Z. Wu, L. Wang, and T. Tan, "Feature coding in image classification: a comprehensive study," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 493–506, 2014.

[46] R. Desimone, "Visual attention mediated by biased competition in extrastriate visual cortex," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 353, no. 1373, pp. 1245–1255, 1998.

[47] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the Fisher vector: theory and practice," *International Journal of Computer Vision*, vol. 105, no. 3, pp. 222–245, 2013.

[48] P. Li, J. Xie, Q. Wang, and W. Zuo, "Is Second-Order Information Helpful for Large-Scale Visual Recognition?" in *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2089–2097, Venice, October 2017.

[49] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, "Geometric means in a novel vector space structure on symmetric positive-definite matrices," *SIAM Journal on Matrix Analysis and Applications*, vol. 29, no. 1, pp. 328–347, 2006/07.

[50] A. Vedaldi and B. Fulkerson, "Vlfeat: an open and portable library of computer vision algorithms," in *Proceedings of the International Conference on Multimedia (MM '10)*, pp. 1469–1472, October 2010.

[51] A. Vedaldi and K. Lenc, "Matconvnet: Convolutional neural networks for matlab," in *Proceedings of the 23rd ACM International Conference on Multimedia*, pp. 689–692, Australia, October 2015.

[52] M.-E. Nilsback and A. Zisserman, "A visual vocabulary for flower classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '06)*, pp. 1447–1454, IEEE, June 2006.

[53] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proceedings of the 6th Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP '08)*, pp. 722–729, December 2008.

[54] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories," *Computer Vision and Image Understanding*, vol. 106, no. 1, pp. 59–70, 2007.

[55] G. Griffin, A. Holub, and P. Perona, *Caltech-256 Object Category Dataset*, 2007.

[56] Y.-L. Boureau, J. Ponce, and Y. LeCun, "A theoretical analysis of feature pooling in visual recognition," in *Proceedings of the International Conference on Machine Learning (ICML 2010)*, pp. 111–118, 2010.