remote sensing

Article

Remote Sensing Image Scene Classification Using Multi-Scale Completed Local Binary Patterns and Fisher Vectors

Longhui Huang [1], Chen Chen [2], Wei Li [1,*] and Qian Du [3]

[1] College of Information Science and Technology, Beijing University of Chemical Technology, 100029 Beijing, China; 15117950611@163.com

[2] Department of Electrical Engineering, University of Texas at Dallas, Dallas, TX 75080, USA; chenchen870713@gmail.com

[3] Department of Electrical and Computer Engineering, Mississippi State University, Starkville, MS 39762, USA; du@ece.msstate.edu

* Correspondence: liwei089@ieee.org; Tel.: +86-1814-6529-853

Academic Editors: Gonzalo Pajares Martinsanz, Xiaofeng Li and Prasad S. Thenkabail
Received: 18 February 2016; Accepted: 30 May 2016; Published: 8 June 2016

Abstract: An effective remote sensing image scene classification approach using patch-based multi-scale completed local binary pattern (MS-CLBP) features and a Fisher vector (FV) is proposed. The approach extracts a set of local patch descriptors by partitioning an image and its multi-scale versions into dense patches and using the CLBP descriptor to characterize local rotation invariant texture information. Then, Fisher vector encoding is used to encode the local patch descriptors (*i.e.*, patch-based CLBP features) into a discriminative representation. To improve the discriminative power of feature representation, multiple sets of parameters are used for CLBP to generate multiple FVs that are concatenated as the final representation for an image. A kernel-based extreme learning machine (KELM) is then employed for classification. The proposed method is extensively evaluated on two public benchmark remote sensing image datasets (*i.e.*, the 21-class land-use dataset and the 19-class satellite scene dataset) and leads to superior classification performance (93.00% for the 21-class dataset with an improvement of approximately 3% when compared with the state-of-the-art MS-CLBP and 94.32% for the 19-class dataset with an improvement of approximately 1%).

Keywords: remote sensing image scene classification; completed local binary patterns; multi-scale analysis; fisher vector; extreme learning machine

1. Introduction

Remote sensing is an effective tool for Earth observation, which has been widely applied in surveying land-use and land-cover classifications and monitoring their dynamic changes. With the improvement of spatial resolution, remote-sensing images present more detailed information such as spatial arrangement information and textural structures, which are of great help in recognizing different land-use and land-cover scene categories. The goal of image scene classification is to recognize the semantic categories of a given image based on some priori knowledge. Due to intra-class variations and wide range of illumination and scale changes, scene classification of high-resolution remote sensing images remains a challenging problem.

The last decade saw considerable efforts to employ computer vision techniques to classify aerial or satellite image scenes. The bag-of-visual-words (BOVW) model [1], which is one of the most popular approaches in image analysis and classification applications, provides an efficient approach to solve the problem of scene classification. The BOVW model, derived from document classification in text analysis, represents an image as a histogram of frequencies of a set of visual words by mapping the local features to a visual vocabulary. The vocabulary is pre-established by clustering the local features extracted from a collection of images. The traditional BOVW model ignores spatial and structural information, which severely limits its descriptive ability. To overcome this issue, a spatial pyramid matching (SPM) framework was proposed in [2]. This approach partitions an image into sub-regions, computes a BOVW histogram for each sub-region, and then concatenates the histograms from all sub-regions to form the SPM representation of an image. However, SPM only considers the absolute spatial arrangement, and the resulting features are sensitive to rotation variations. Thus, a spatial co-occurrence kernel, which is general enough to characterize a variety of spatial arrangements, was proposed in [3] to capture both the absolute and relative spatial layout of an image. In [4], a multi-resolution representation was incorporated into the bag-of-features model and two modalities of horizontal and vertical partitions were adopted to partition all resolution images into sub-regions to improve the SPM framework. In [5], a concentric circle-structured multi-scale BOVW model was presented to incorporate rotation-invariant spatial layout information into the original BOVW model.

The aforementioned BOVW variants focus on capturing the spatial layout information of scene images. However, the rich texture and structure information in high-resolution remote sensing images has not been fully exploited since they merely use the scale-invariant feature transform (SIFT) [6] descriptors to capture local features. There is also a great effort to evaluate various features and combinations of features for scene classification. In [7], a local structural texture similarity descriptor was applied to image blocks to represent structural texture for aerial image classification. In [8], semantic classifications of aerial images based on Gabor and Gist descriptors [9] were evaluated individually. In [10], four types of features consisting of raw pixel intensity values, oriented filter responses, SIFT-based feature descriptors, and self-similarity were used within the framework of unsupervised feature learning. In [11], global features extracted using the enhanced Gabor texture descriptor (EGTD) and local features extracted using the SIFT descriptor were fused in a hierarchical approach to improve the performance of remote sensing image scene classification.

Recently, deep learning has received great attention. Different from the afore-mentioned BOVW and its variants that are considered mid-level representations, deep learning is an end-to-end feature learning method (e.g., from an image to semantic label). Among deep learning-based networks, convolutional neural networks (CNNs) [12,13] may be the most popular for learning visual features in computer vision applications, such as remote sensing and large-scale visual recognition. K. Nogueira *et al.* [14] presented the PatreoNet, which has the capability to learn specific spatial features from remote sensing images without any pre-processing step or descriptor evaluation. AlexNet, proposed by Krizhevsky *et al.* [15], was the first to employ non-saturating neurons, GPU implementation of the convolution operation and dropout to prevent overfitting. GoogLeNet [16] deployed the CNN architecture and utilized filters of different sizes at the same layer to reduce the number of parameters of the network. However, CNNs have an intrinsic limitation, *i.e.*, the complicated pre-training process to adjust parameters.

In [17], multi-scale completed local binary patterns (MS-CLBP) features were utilized for remote sensing image classification. The extracted features can be considered global features in an image. However, the global feature representation may not able to characterize detailed structures and distinct objects. For example, some land-use and land-cover classes are defined mainly by individual objects, e.g., baseball fields and storage tanks. In this paper, we propose a local feature representation method based on patch-based MS-CLBP, which can be used to extract complementary features to global features. Specifically, the CLBP descriptor is applied to partition dense image patches and extract a set of local patch descriptors, which characterize the detailed local structure and texture information in high-resolution remote sensing images. Since the CLBP [18] operator belongs to a gray-scale and rotation-invariant texture operator, the extracted local descriptors are robust to rotation image transformations. Then, the Fisher kernel representation [19] is employed to encode the local descriptors into a discriminative representation (*i.e.*, Fisher vector (FV)). FV describes patch descriptors by their deviation from a “universal” generative Gaussian mixture model (GMM). To improve the discriminative power of the feature representation, multiple sets of parameters for the CLBP operator (*i.e.*, MS-CLBP) were utilized to generate multiple FVs. The final representation for an image is achieved by concatenating all the FVs. For classification, the kernel-based extreme learning machine (KELM) [20] is adopted for its efficient computation and good classification performance.

There are two main contributions from this work. First, a local feature representation method using patch-based MS-CLBP features and FV is proposed. The MS-CLBP operator is applied to the partitioned dense regions to extract a set of local patch descriptors, and then the Fisher kernel representation is used to encode the local descriptors into a discriminative representation of remote sensing images. Second, the two implementations of MS-CLBP are combined into a unified framework to build a more powerful feature representation. The proposed local feature representation method is evaluated using two public benchmark remote sensing image datasets. The experimental results verify the effectiveness of our proposed method as compared to state-of-the-art algorithms.

The remainder of the paper is organized as follows. Section 2 presents the related works including CLBP and the Fisher vector. Section 3 describes two implementations of MS-CLBP, patch-based MS-CLBP feature extraction, and the details of the proposed feature representation method. Section 4 provides the experimental results. Finally, Section 5 concludes the paper.

2. Related Works

2.1. Completed Local Binary Patterns

Local binary patterns (LBP) [21,22] are an effective measure of spatial structure information of local image texture. Consider a center pixel and its gray value, t_c. Its neighboring pixels are equally spaced on a circle of radius r with the center at location t_c. If the coordinates of t_c are $(0,0)$ and m neighbors $\{t_i\}_{i=0}^{m-1}$ are considered, the coordinates of t_i are denoted as $(-r\sin(2\pi i/m), r\cos(2\pi i/m))$. Then, the LBP is calculated by thresholding the neighbors $\{t_i\}_{i=0}^{m-1}$ with the center pixel t_c to generate an m-bit binary number. The resulting LBP for t_c in decimal number can be expressed as follows:

$$LBP_{m,r}(t_c) = \sum_{i=0}^{m-1} s(t_i - t_c)\,2^i = \sum_{i=0}^{m-1} s(d_i)\,2^i,\ s(x) = \begin{cases} 1, & x \geqslant 0 \\ 0, & x < 0 \end{cases} \tag{1}$$

where $d_i = (t_i - t_c)$ represents the difference between the center pixel and each neighbor, which characterizes the spatial local structure at the center location. Further, the resulted d_i is robust to illumination changes and they are more efficient than the original image in pattern classification due to the fact that the central gray level t_c is removed. The difference vector d_i can be further decomposed into two components: the signs and magnitudes (absolute values of d_i, *i.e.*, $|d_i|$). However, the original LBP only uses the sign information of d_i while ignoring the magnitude information. In the improved CLBP [18], the signs and magnitudes are complementary, from which the difference vector d_i can be perfectly reconstructed. Figure 1 illustrates an example of the sign and magnitude components of the CLBP extracted from a sample block, where Figure 1a–d denote original 3 × 3 local structure, difference vector, sign vector and magnitude vector, respectively. Note that "0" is coded as "−1" in CLBP (as seen in Figure 1c). Two operators, CLBP-Sign (CLBP_S) and CLBP-Magnitude (CLBP_M), are used to encode these two components. CLBP_S is equivalent to the traditional LBP operator while the CLBP_M operator can be expressed as,

$$CLBP\_M_{m,r} = \sum_{i=0}^{m-1} f(|d_i|, c)2^i,\ f(x,y) = \begin{cases} 1, & x \geqslant y \\ 0, & x < y \end{cases} \tag{2}$$

where c is a threshold that is set to the mean value of $|d_i|$. Using Equations (1) and (2), two binary strings can be produced and denoted as CLBP_S and CLBP_M codes, respectively. Two ways to combine the CLBP_S and CLBP_M codes are presented in [18]. Here, the first way (concatenation) is used, in which the histograms of the CLBP_S and CLBP_M codes are calculated separately, and then the two histograms are concatenated. Note that there is also the CLBP-Center part which codes the values of the center pixels in the original CLBP. Here, only the CLBP_S and CLBP_M operators are considered for computational efficiency.

10	83	68
50	24	21
15	36	94

(a)

-14	59	44
26		-3
-9	12	70

(b)

-1	1	1
1		-1
-1	1	1

(c)

14	59	44
26		3
9	12	70

(d)

-1	1	1
-1		-1
-1	-1	1

(e)

Figure 1. (**a**) 3 × 3 sample block; (**b**) the local differences; (**c**) the sign component of CLBP; (**d**) the absolute value of local differences; (**e**) the magnitude component of CLBP.

Figure 2 presents an example of the CLBP_S and CLBP_M coded images corresponding to an input aerial scene (viaduct scene). It can be observed that CLBP_S and CLBP_M operators both can capture the spatial pattern and the contrast of local image texture, such as edges and corners.

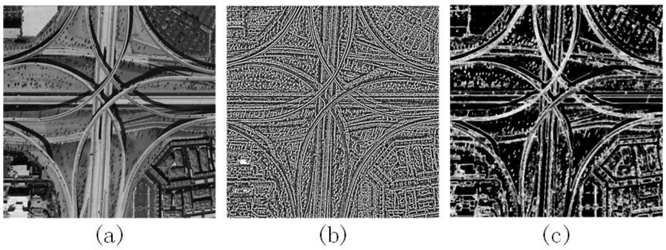


Figure 2. (**a**) Input image; (**b**) CLBP_S coded image; (**c**) CLBP_M coded image.

2.2. Fisher Vector

After local feature extraction (especially for patch-based feature extraction), the popular BOVW model is usually employed to encode features into histograms. However, the BOVW model has an intrinsic limitation, namely the computational cost in assignment of local features to visual words, which scales as the product of the number of visual words, the number of regions and the local feature dimensionality [23]. Several extensions to the basic BOVW model to build compact vocabularies have been proposed. The most appealing one is the Fisher kernel image representation [19,24], which uses high-dimensional gradient representation to represent an image. Due to informative representations with compact vocabularies, this representation contains more information than a simple histogram representation.

An FV is a special case of Fisher kernel construction. Let $X = \{x_t,\ t = 1 \ldots T\}$ be the set of local patch descriptors extracted from an image. A Gaussian mixture model (GMM) is trained on the training images using Maximum Likelihood (ML) estimation [25,26]. Let P denote the probability density function of the GMM with parameters $\lambda = \{\omega_i, \mu_i, \Sigma_i, i = 1 \ldots K\}$, where K is the number of components. ω_i, μ_i and Σ_i are the mixture weight, mean vector, and covariance matrix of the i^{th} Gaussian component, respectively. The image can be characterized by the gradient of the log-likelihood of the data on the model:

$$G_\lambda^X = \nabla_\lambda \log\mathrm{P}\left(X|\lambda\right) \tag{3}$$

The gradient describes the direction along which parameters are to be adjusted to best fit the data. Under an independence assumption, the covariance matrices are diagonal, *i.e.*, $\Sigma_i = \mathrm{diag}\left(\sigma_i^2\right)$. Then following [27], $L(X|\lambda) = \log P(X|\lambda)$ is written as,

$$L(X|\lambda) = \sum_{t=1}^{T} \log P(x_t|\lambda) \tag{4}$$

The probability density function of x_t generated by the GMM is

$$P(x_t|\lambda) = \sum_{i=1}^{k} \omega_i p_i(x_t|\lambda) \tag{5}$$

Let $\gamma_t(i)$ be the occupancy probability, *i.e.*, the probability of descriptor x_t generated by the i-th Gaussian.

$$\gamma_t(i) = P(i|\, x_t, \lambda) = \frac{\omega_i p_i(x_t|\lambda)}{\sum\limits_{j=1}^{k} \omega_j p_j(x_t|\lambda)} \tag{6}$$

with the Bayes formula mathematical derivations providing the following results,

$$\frac{\partial L(X|\lambda)}{\partial \omega_i} = \sum_{t=1}^{T} \left[\frac{\gamma_t(i)}{\omega_i} - \frac{\gamma_t(1)}{\omega_1} \right] \text{for } i \geqslant 2 \tag{7}$$

$$\frac{\partial L(X|\lambda)}{\partial \mu_i^d} = \sum_{t=1}^{T} \gamma_t(i) \left[\frac{x_t^d - \mu_i^d}{(\sigma_i^d)^2} \right] \tag{8}$$

$$\frac{\partial L(X|\lambda)}{\partial \sigma_i^d} = \sum_{t=1}^{T} \gamma_t(i) \left[\frac{(x_t^d - \mu_i^d)^2}{(\sigma_i^d)^3} - \frac{1}{\sigma_i^d} \right] \tag{9}$$

where d denotes the d^{th} dimension of a vector. The diagonal closed-form approximation in [27] is used to normalize the gradient vector by multiplying the square-root of the inverse of the Fisher information matrix, *i.e.*, $F_\lambda^{-1/2}$. Let f_{ω_i}, $f_{\mu_i^d}$, and $f_{\sigma_i^d}$ denote the diagonal of F_λ corresponding to $\partial L(X|\lambda)\,/\partial \omega_i$, $\partial L(X|\lambda)\,/\partial \mu_i^d$, and $\partial L(X|\lambda)\,/\partial \sigma_i^d$, respectively, and we have the following approximation,

$$f_{\omega_i} = T(\frac{1}{\omega_i} + \frac{1}{\omega_1}) \tag{10}$$

$$f_{\mu_i^d} = \frac{T\omega_i}{(\sigma_i^d)^2} \tag{11}$$

$$f_{\sigma_i^d} = \frac{2T\omega_i}{(\sigma_i^d)^2} \tag{12}$$

Thus, the normalized partial derivatives are ${f_{\omega_i}}^{-1/2} \partial L(X|\lambda)\,/\partial \omega_i$, ${f_{\mu_i^d}}^{-1/2} \partial L(X|\lambda)\,/\partial \mu_i^d$, and ${f_{\sigma_i^d}}^{-1/2} \partial L(X|\lambda)\,/\partial \sigma_i^d$. The final gradient vector (*i.e.*, FV) is just a concatenation of all the partial derivative vectors. Therefore, the dimensionality of FV is $(2D + 1) \times K$, where D denotes the size of the local descriptors.

3. Proposed Feature Representation Method

Inspired by the success of CLBP and FV in computer vision applications, we propose an effective image representation approach for remote sensing image scene classification based on patch-based MS-CLBP features and FV. The patch-based MS-CLBP is applied as the local patch descriptors and then the FV is chosen as the encoding strategy to generate a high-dimensional representation of an image.

3.1. Two Implementations of Multi-Scale Completed Local Binary Patterns

CLBP features computed from a single-scale may not be able to detect the dominant texture features from an image. A possible solution is to characterize the image texture in multiple resolutions, *i.e.*, MS-CLBP. There are two implementations for the MS-CLBP descriptor [17].

In the first implementation, the radius of a circle r is altered to change the spatial resolution. The multi-scale analysis is accomplished by combining the information provided by multiple operators of varying (m, r). For simplicity, the number of neighbors is fixed to m and different values of r are tuned to achieve the optimal combination. An example of a 3-scale (three r values) CLBP operator is illustrated in Figure 3. The CLBP_S and CLBP_M histogram features extracted from each scale are concatenated to form an MS-CLBP representation. One disadvantage of this multi-scale analysis implementation is that the computational complexity increases due to multiple resolutions.

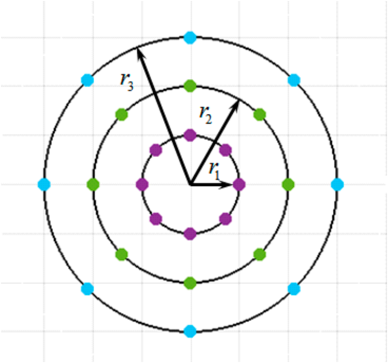


Figure 3. An example of the first implementation of a 3-scale CLBP operator ($m = 8$, $r_1 = 1$, $r_2 = 2$, and $r_3 = 3$).

In the second implementation, the original image is down-sampled using the bicubic interpolation to obtain multiple images at different scales. The value of scale is between 0 and 1 (here, 1 denotes the original image). Then, the CLBP_S and CLBP_M operators of fixed radius and the number of neighbors are applied to the multiple-scale images. The CLBP_S and CLBP_M histogram features extracted from each scale image are concatenated to form an MS-CLBP representation. An example of the second implementation of the MS-CLBP descriptor is shown in Figure 4.

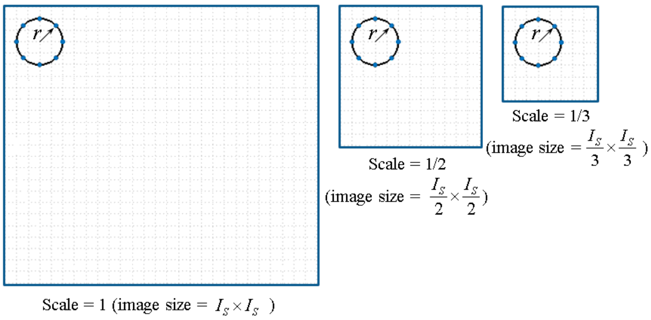

r
r
r
Scale = 1/3
(image size = $\frac{I_S}{3} \times \frac{I_S}{3}$)
Scale = 1/2
(image size = $\frac{I_S}{2} \times \frac{I_S}{2}$)
Scale = 1 (image size = $I_S \times I_S$)

Figure 4. An example of the second implementation of a 3-scale CLBP operator ($m = 8$, $r = 2$).

3.2. Patch-Based MS-CLBP Feature Extraction

Given an image, the CLBP [18] operator with a parameter set (m, r) is applied to generate two CLBP coded images with one corresponding to the sign component (*i.e.*, CLBP_S coded image) and the other the magnitude component (*i.e.*, CLBP_M coded image). Two complementary components of CLBP (CLBP_S and CLBP_M) can capture the spatial patterns and contrast of local image texture, such as edges and corners. Then, the CLBP coded images are partitioned into $B \times B$ overlapped patches in an image grid. For simplicity, the overlap between two patches is half of the patch size (*i.e.*, $B/2$) in both horizontal and vertical directions. To incorporate spatial structures of an image at different scales (or sizes) and create more patch descriptors, here the second implementation of MS-CLBP is employed by resizing the original image to different scales (e.g., 1/2 and 1/3 of the original image). Specifically, the CLBP operator with the same parameter set is applied to the multi-scale images to generate patch-based CLBP histogram features. For patch i, two occurrence histograms (*i.e.*, the nonparametric statistical estimate) are computed from the sign component (CLBP_S) and the magnitude component (CLBP_M). A histogram feature vector denoted by $\mathbf{h}_i$ is formed by concatenating the two histograms. If M patches are extracted from the multi-scale images, a feature matrix denoted by $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_M]$ is generated to represent the original image. Each column of the matrix $\mathbf{H}$ is a histogram feature vector for a patch. The proposed patch-based CLBP feature extraction method is illustrated in Figure 5.

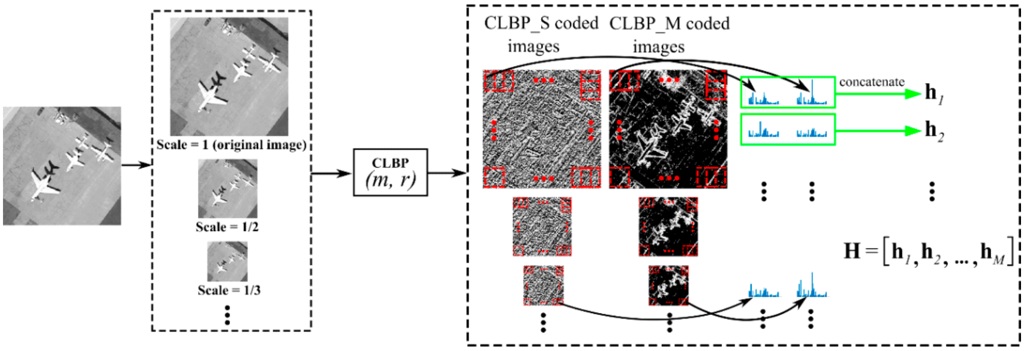


Figure 5. Patch-based CLBP feature extraction.

As noted in [21], LBP features computed from a single scale may not be able to represent intrinsic texture features. Therefore, different parameter sets (m, r) are utilized for the CLBP operator to achieve the first implementation of the MS-CLBP as described in [17]. Specifically, the number of neighbors (m) is fixed and multiple radii (r) are used in the patch-based CLBP feature extraction as shown in Figure 5. If q parameter sets (*i.e.*, $\{(m, r_1), (m, r_2), ..., (m, r_q)\}$) are considered, a set of q feature matrices denoted by $\left\{\mathbf{H}_{(m,r_1)}, \mathbf{H}_{(m,r_2)}, ..., \mathbf{H}_{(m,r_q)}\right\}$ can be obtained for an image. It is worth noting that the proposed patch-based MS-CLBP feature extraction effectively unifies the two implementations of the MS-CLBP [17].

3.3. A Fisher Kernel Representation

Fisher kernel representation [19] is an effective patch aggregation mechanism to characterize a sample of low-level features, and it exhibits superior performance over the BOVW model. Therefore, the Fisher kernel representation is employed to encode the dense local patch descriptors.

Given N_T training images with N_T feature matrices, $\left\{\mathbf{H}^{[1]}, \mathbf{H}^{[2]}, ..., \mathbf{H}^{[N_T]}\right\}$ representing the local patch descriptors (*i.e.*, patch-based CLBP features) of each image are obtained using the feature extraction method illustrated in Figure 5. Since q parameter sets (*i.e.*, $\{(m, r_1), (m, r_2), ..., (m, r_q)\}$) are employed for the CLBP operator, each image yields q feature matrices denoted by $\left\{\mathbf{H}^{[j]}_{(m,r_1)}, \mathbf{H}^{[j]}_{(m,r_2)}, ..., \mathbf{H}^{[j]}_{(m,r_q)}\right\}$, where $j \in [1, 2, ..., N_T]$. For each CLBP parameter set, the corresponding feature matrices of the training data are used to estimate the GMM parameters via the Expectation-Maximization (EM) algorithm. Therefore, for q CLBP parameter sets, q GMMs are created. After estimating the GMM parameters, q FVs are obtained for an image. Then, the q FVs are simply concatenated as the final feature representation. Figure 6 shows the detailed procedure for generating FVs. As illustrated in Figure 6, the stacked FVs ($\mathbf{f}$) from the q CLBP parameter sets serve as the final feature representation of an image before being fed into a classifier.

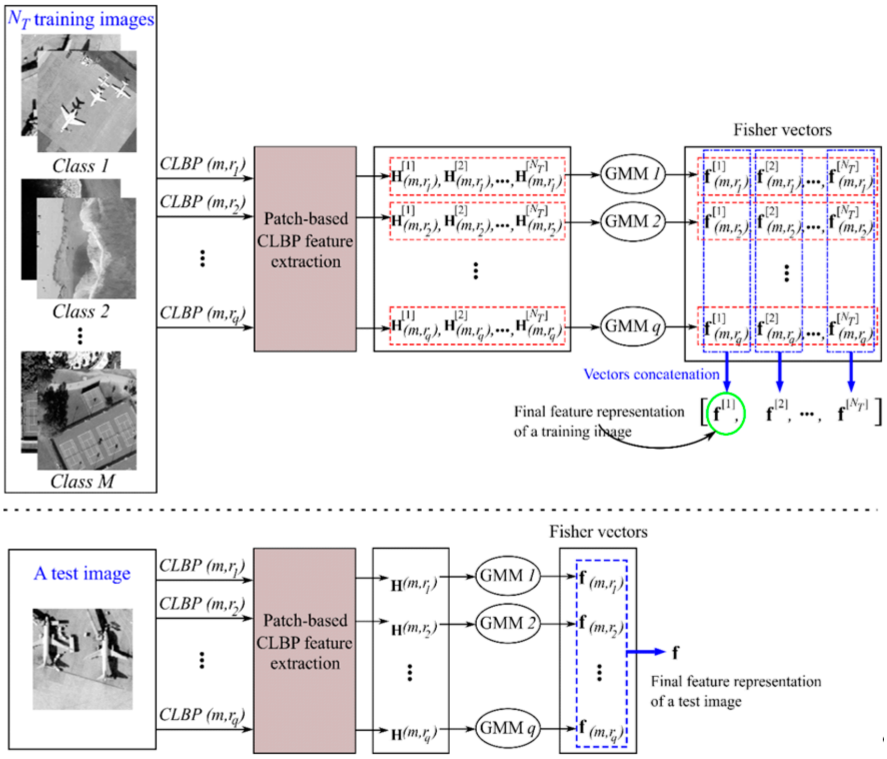

N_T training images
Class 1
Class 2
Class M
CLBP (m,r_1)
CLBP (m,r_2)
CLBP (m,r_q)
Patch-based CLBP feature extraction
$\mathbf{H}^{[1]}_{(m,r_1)}, \mathbf{H}^{[2]}_{(m,r_1)}, \dots, \mathbf{H}^{[N_T]}_{(m,r_1)}$
GMM 1
GMM 2
GMM q
Fisher vectors
Vectors concatenation
Final feature representation of a training image
$[\mathbf{f}^{[1]}, \mathbf{f}^{[2]}, \dots, \mathbf{f}^{[N_T]}]$
A test image
$\mathbf{f}_{(m,r_1)}$
Final feature representation of a test image

Figure 6. Fisher vector representation.

4. Experiments

Two standard public domain datasets are used to demonstrate the effectiveness of the proposed image representation method for remote sensing land-use scene classification. In the experiments, KELM with a radial basis function (RBF) kernel is employed for classification due to its generally excellent classification performance and low computational cost. The classification performance of the proposed method is compared with the state-of-the-art in the literature.

4.1. Experimental Data and Setup

The first dataset is the well-known UC-Merced land-use dataset [28]. It is the first public ground truth land-use scene image dataset that consists of 21 land-use classes and each class contains 100 images with a size of 256 × 256 pixels. The images were manually extracted from aerial orthoimagery downloaded from the United States Geological Survey (USGS) National Map. This is a challenging dataset due to a variety of spatial patterns in those 21 classes. Sample images of each land-use class are shown in Figure 7. To facilitate a fair comparison, the same experimental setting reported in [28] is followed. Five-fold cross-validation is performed, in which the dataset is randomly partitioned into five equal subsets. There are 20 images from each land-use class in a subset. Four subsets are used for training and the remaining subset for testing. The classification accuracy is the average over the five cross-validation evaluations.

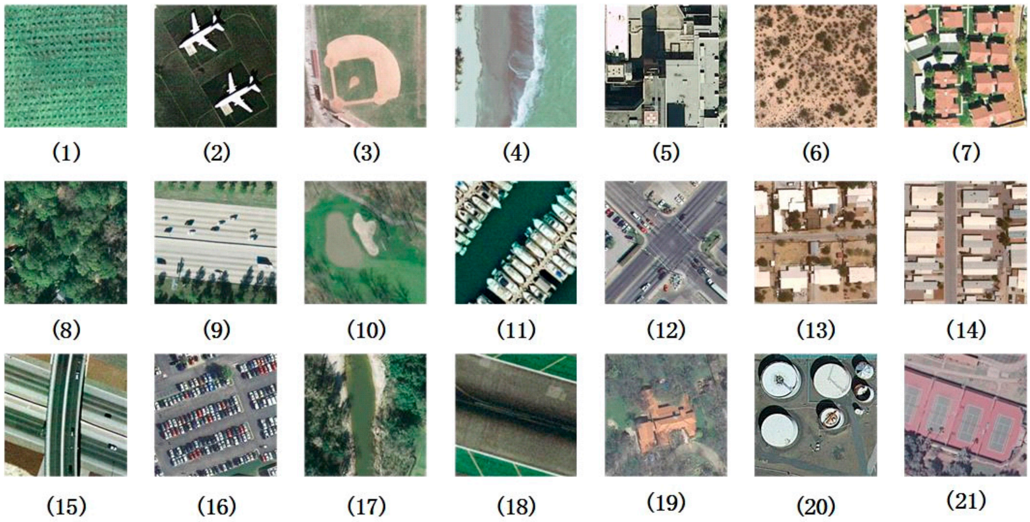


Figure 7. Examples from the 21-class land-use dataset: (**1**) agricultural; (**2**) airplane; (**3**) baseball diamond; (**4**) beach; (**5**) buildings; (**6**) chaparral; (**7**) dense residential; (**8**) forest; (**9**) freeway; (**10**) golf course; (**11**) harbor; (**12**) intersection; (**13**) medium density residential; (**14**) mobile home park; (**15**) overpass; (**16**) parking lot; (**17**) river; (**18**) runway; (**19**) sparse residential; (**20**) storage tanks; (**21**) tennis courts.

The second dataset used in our experiments is the 19-class satellite scene dataset [29]. It consists of 19 classes of high-resolution satellite scenes. There are 50 images with sizes of 600 × 600 pixels for each class. The images are extracted from large satellite images on Google Earth. An example of each class is shown in Figure 8. The same experimental setup in [30] was used. Here, 30 images are randomly selected per class as training data and the remaining images as testing data. The experiment is repeated 10 times with different realizations of randomly selected training and testing images. Classification accuracy is averaged over the 10 trials.

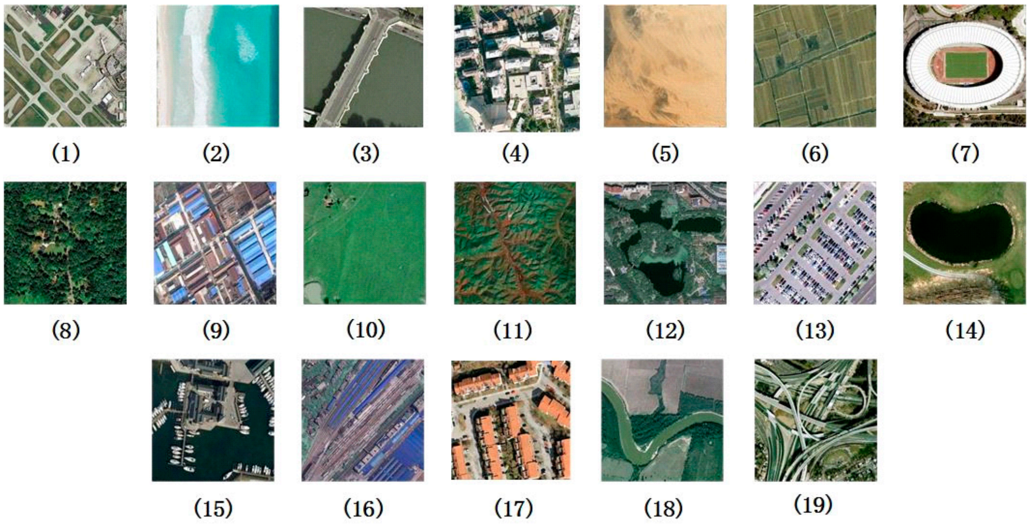


Figure 8. Examples from the 19-class satellite scene dataset: (**1**) airport; (**2**) beach; (**3**) bridge; (**4**) commercial; (**5**) desert; (**6**) farmland; (**7**) football field; (**8**) forest; (**9**) industrial; (**10**) meadow; (**11**) mountain; (**12**) park; (**13**) parking; (**14**) pond; (**15**) port; (**16**) railway station; (**17**) residential; (**18**) river; (**19**) viaduct.

Note that the original images in these two datasets are color images; the images are converted from the RGB color space to the YCbCr color space, and the Y component (luminance) is used for scene classification.

4.2. Parameters Setting

The number of neighbors (m) in the CLBP operator has a direct impact on the dimensionality of the FV since patch-based CLBP features are used as local patch descriptors. A large value of m will increase the feature dimensionality and then increase the computational complexity. Based on the parameter tuning results in [17], $m = 8$ is empirically chosen for both the 21-class land-use dataset and the 19-class satellite scene dataset as it balances the classification performance and computational complexity. In addition, the parameter settings in [17] are adopted for the MS-CLBP descriptor. Specifically, 6 radii (*i.e.*, $r = [1:6]$) are used for the MS-CLBP descriptor, resulting 6 parameters sets $\{(m = 8, r_1 = 1), ..., (m = 8, r_6 = 6)\}$.

Then, the number of scales is studied for the first implementation of the MS-CLBP operator for generating multi-scale images and the number of Gaussians (K) in the GMM. For the 21-class land-use dataset, 80 images are randomly selected per class for training and the remaining images for testing. For the 19-class satellite scene dataset, 30 images per class are randomly selected for training and the remaining images for testing. Different numbers of Gaussians are examined along with different choices of multiple scales including $\{1, 1/[1:2], ..., 1/[1:6]\}$. For instance, $1/[1:2]$ indicates that scale = 1 (original image) and scale = 1/2 (down-sampled image at half of the size of the original image) are used to generate two images with two scales. Figures 9 and 10 present the classification results with different numbers of Gaussians in the GMM and different numbers of scales for the two datasets, respectively.

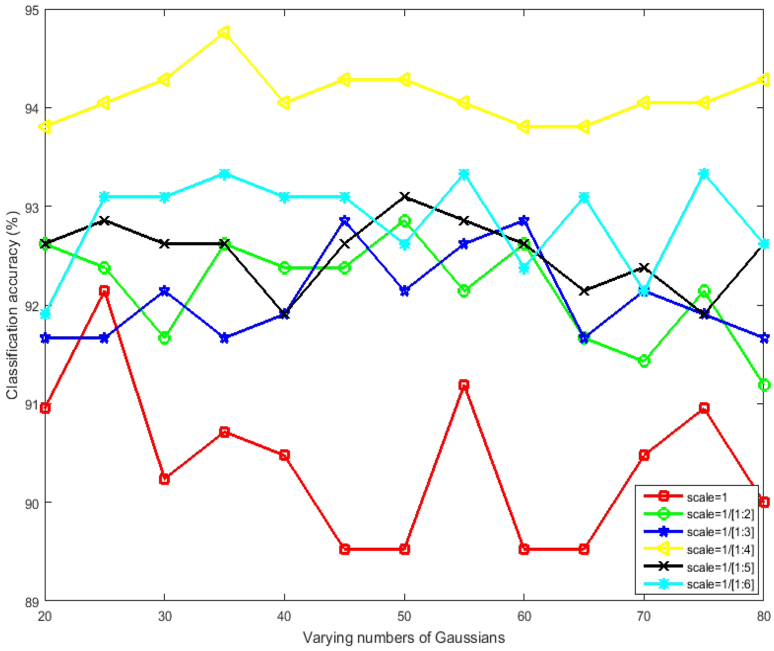


Figure 9. Classification accuracy (%) *versus* varying numbers of Gaussians and scales for our proposed method for the 21-class land-use dataset.

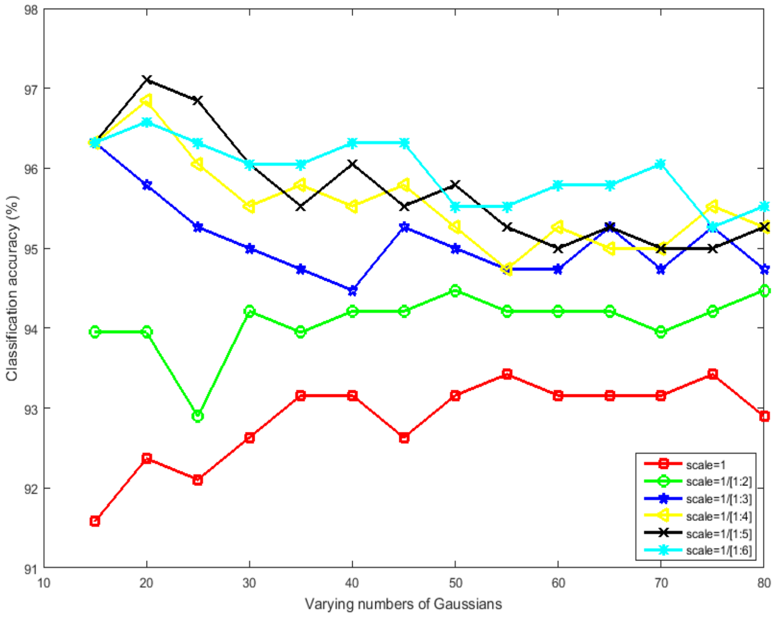


Figure 10. Classification accuracy (%) *versus* varying numbers of Gaussians and scales for our proposed method for the 19-class satellite scene dataset.

Thus, the optimal number of Gaussians for the 21-class land-use dataset is 35 and the optimal multiple scales is 1/[1 : 4] simultaneously considering classification accuracy and computational complexity. Similarly, the optimal number of Gaussians for the 19-class satellite scene dataset is 20 and the optimal multiple scale is 1/[1 : 4].

Since the proposed method extracts dense local patches, the size of the patch ($B \times B$) is determined empirically. The classification accuracies with varying patch sizes are illustrated in Figure 11. It is obvious that $B = 32$ achieves the best classification performance for the 21-class land-use dataset. The size of the images in the 19-class dataset is 600 × 600 pixels, which is about twice the size of the images in the 21-class dataset. Therefore, the patch size is set a $B = 64$ for the 19-class dataset.

In addition, to gain computational efficiency, principal component analysis (PCA) [31,32] is employed to reduce the dimensionality of FV features. The PCA projection matrix was calculated using the features of the training data, and the principal components that accounted for 95% of the total variation of the training features are considered in our experiments.

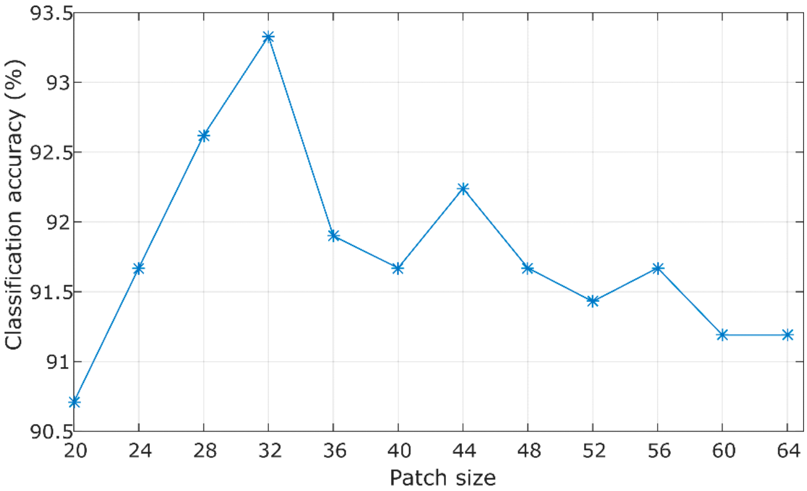


Figure 11. Classification accuracy (%) *versus* varying patch sizes for the 21-class land-use dataset.

4.3. FV Representation vs. BOVW Model

To verify the advantage of FV as compared to the BOVW model, the MS-CLBP+BOVW is applied to both the 21-class land-use dataset and the 19-class satellite scene dataset and the performance is compared with our approach. The same parameters are used for the MS-CLBP feature. In the BOVW model, 30,000 patches are randomly selected from all patches and K-means clustering is employed to generate 1024 visual words as a typical setting. The classification performance of the proposed method and MS-CLBP+BOVW is evaluated over each category of the two datasets as shown in Figures 12 and 13, respectively. As can be seen from Figure 12, the proposed method provides better performance than MS-CLBP+BOVW in almost all categories except two, medium density residential and parking lot, and two categories (agricultural and forest) have equal performance. In Figure 13, the proposed method achieves greater accuracy than all classes except beach and industrial for the 19-class satellite scene dataset.

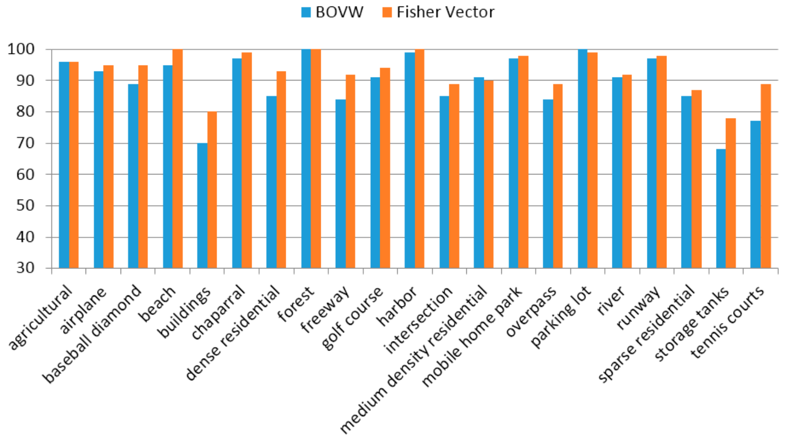


Figure 12. Per-class accuracy (%) of the proposed method and MS-CLBP+BOVW on the 21-class land-use dataset.

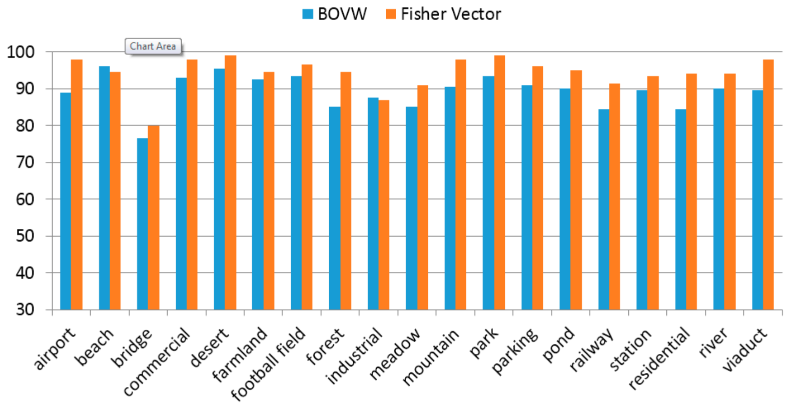


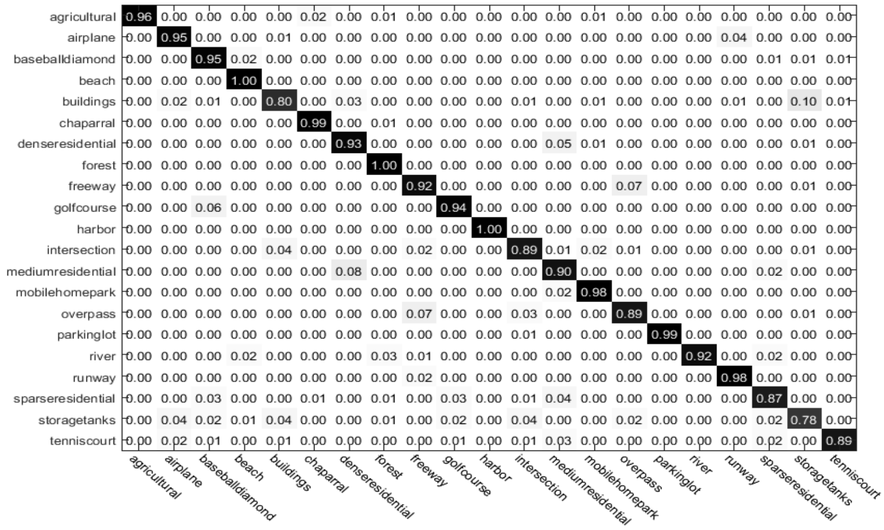
Figure 13. Per-class accuracy (%) of the proposed method and MS-CLBP+BOVW on the 19-class satellite scene dataset.

4.4. Comparison to the State-of-the-Art Methods

In this section, the effectiveness of the proposed image representation method is evaluated by comparing its performance with previously reported performance in the literature. Specifically, the proposed method is compared with the MS-CLBP descriptor [17] applied to an entire remote sensing image to obtain a global feature representation. The comparison results are reported in Table 1. From the comparison results, the proposed method achieves superior classification performance over other existing methods, which demonstrates the effectiveness of the proposed image representation for remote sensing land-use scene classification. The improvement of the proposed method over the global representation developed in [17] is 2.4%. This improvement is mainly due to the proposed local feature representation framework which unifies the two implementations of the MS-CLBP descriptor. Moreover, the proposed approach is an approximately 4.7% improvement over the MS-CLBP + BOVW method, which verifies the advantage of the Fisher kernel representation as compared to the BOVW model. Figure 14 shows the confusion matrix of the proposed method for the 21-class land-use dataset. The diagonal elements of the matrix denote the mean class-specific classification accuracy (%). We find an interesting phenomenon from Figure 14 that diagonal elements for beach and forest are extremely large but diagonal elements for storage tank is relatively small. The reasons are that images of beach and forest present rich texture and structures information; within-class similarity for the beach and forest categories is high but relatively low for category of storage tank; and some images of storage tank are similar to other class such as buildings.

Table 1. Comparison of classification accuracy (%) forthe 21-class land-use dataset.

Method	Accuracy(Mean ± std)
BOVW [28]	76.8
SPM [28]	75.3
BOVW + Spatial Co-occurrence Kernel [28]	77.7
Color Gabor [28]	80.5
Color histogram (HLS) [28]	81.2
Structural texture similarity [7]	86.0
Unsupervised feature learning [33]	81.7 ± 1.2
Saliency-Guided unsupervised feature learning [34]	82.7 ± 1.2
Concentric circle-structured multiscale BOVW [5]	86.6 ± 0.8
Multifeature concatenation [35]	89.5 ± 0.8
Pyramid-of-Spatial-Relatons (PSR) [36]	89.1
MCBGP + E-ELM [37]	86.52 ± 1.3
ConvNet with specific spatial features [38]	89.39 ± 1.10
gradient boosting randomconvolutional network [39]	94.53
GoogLeNet [40]	92.80 ± 0.61
OverFeatConvNets [40]	90.91 ± 1.19
MS-CLBP [17]	90.6 ± 1.4
MS-CLBP + BOVW	89.27 ± 2.9
The Proposed	**93.00 ± 1.2**



	agricultural	airplane	baseballdiamond	beach	buildings	chaparral	denseresidential	forest	freeway	golfcourse	harbor	intersection	mediumresidential	mobilehomepark	overpass	parkinglot	river	runway	sparseresidential	storagetanks	tenniscourt
agricultural	0.96	0.00	0.00	0.00	0.00	0.02	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
airplane	0.00	0.95	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00
baseballdiamond	0.00	0.00	0.95	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01
beach	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
buildings	0.00	0.02	0.01	0.00	0.80	0.00	0.03	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.10	0.01
chaparral	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
denseresidential	0.00	0.00	0.00	0.00	0.00	0.00	0.93	0.00	0.00	0.00	0.00	0.00	0.05	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00
forest	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
freeway	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.92	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.00	0.01	0.00
golfcourse	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.94	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
harbor	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
intersection	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.02	0.00	0.00	0.89	0.01	0.02	0.01	0.00	0.00	0.00	0.00	0.01	0.00
mediumresidential	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.90	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00
mobilehomepark	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.98	0.00	0.00	0.00	0.00	0.00	0.00	0.00
overpass	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.03	0.00	0.00	0.89	0.00	0.00	0.00	0.00	0.01	0.00
parkinglot	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.99	0.00	0.00	0.00	0.00	0.00
river	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.92	0.00	0.02	0.00	0.00
runway	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.98	0.00	0.00	0.00
sparseresidential	0.00	0.00	0.03	0.00	0.00	0.01	0.00	0.01	0.00	0.03	0.00	0.01	0.04	0.00	0.00	0.00	0.00	0.00	0.87	0.00	0.00
storagetanks	0.00	0.04	0.02	0.01	0.04	0.00	0.00	0.01	0.00	0.02	0.00	0.04	0.00	0.00	0.02	0.00	0.00	0.00	0.02	0.78	0.00
tenniscourt	0.00	0.02	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.03	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.89

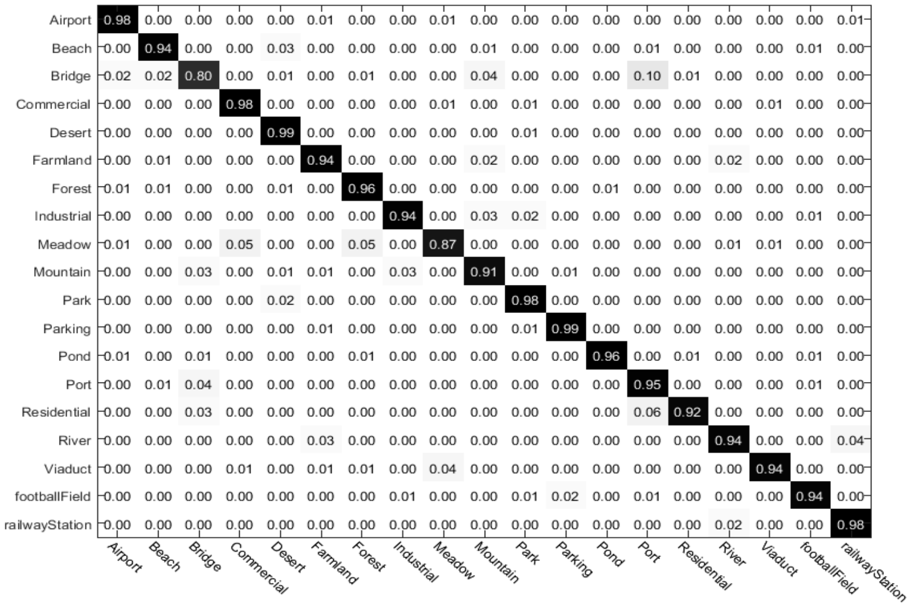
Figure 14. Confusion matrix of proposed method for the 21-class land-use dataset.

When compared with CNNs, it can be found that the classification accuracy of CNNs is close to that of our method. Even though the performance of some CNNs is better than the proposed method, they need a pre-training process with a large amount of external data. Thus our method is still competitive in terms of limited requirement for external data.

The comparison results for the 19-class satellite scene dataset are listed in Table 2. It indicates that the proposed method outperforms other existing methods and achieves the best performance. The proposed method provides about 7% improvement over the method in [31] which utilized a combination of multiple sets of features, indicating the superior discriminative power of the proposed feature representation. The confusion matrix of the proposed method for the 19-class satellite scene dataset is shown in Figure 15. From diagonal elements of the matrix, the classification accuracy for bridges is relatively small because some texture information in the images of bridges is similar to those in the images of ports.

Table 2. Comparison of classification accuracy (%) for the 19-class satellite scene dataset.

Method	Accuracy (Mean ± std)
Bag of colors [25]	70.6 ± 1.5
Tree of c-shapes [25]	80.4 ± 1.8
Bag of SIFT [25]	85.5 ± 1.2
Multifeature concatenation [25]	90.8 ± 0.7
LTP-HF [23]	77.6
SIFT + LTP-HF + Color histogram [23]	93.6
MS-CLBP [1]	93.4 ± 1.1
MS-CLBP + BOVW	89.29 ± 1.3
The Proposed	**94.32 ± 1.2**



	Airport	Beach	Bridge	Commercial	Desert	Farmland	Forest	Industrial	Meadow	Mountain	Park	Parking	Pond	Port	Residential	River	Viaduct	footballField	railwayStation
Airport	0.98	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
Beach	0.00	0.94	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00
Bridge	0.02	0.02	0.80	0.00	0.01	0.00	0.01	0.00	0.00	0.04	0.00	0.00	0.00	0.10	0.01	0.00	0.00	0.00	0.00
Commercial	0.00	0.00	0.00	0.98	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
Desert	0.00	0.00	0.00	0.00	0.99	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Farmland	0.00	0.01	0.00	0.00	0.00	0.94	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00
Forest	0.01	0.01	0.00	0.00	0.01	0.00	0.96	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
Industrial	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.94	0.00	0.03	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
Meadow	0.01	0.00	0.00	0.05	0.00	0.00	0.05	0.00	0.87	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00
Mountain	0.00	0.00	0.03	0.00	0.01	0.01	0.00	0.03	0.00	0.91	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Park	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.98	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Parking	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pond	0.01	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.96	0.00	0.01	0.00	0.00	0.01	0.00
Port	0.00	0.01	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.95	0.00	0.00	0.00	0.01	0.00
Residential	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.92	0.00	0.00	0.00	0.00
River	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.94	0.00	0.00	0.04
Viaduct	0.00	0.00	0.00	0.01	0.00	0.01	0.01	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.94	0.00	0.00
footballField	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.02	0.00	0.01	0.00	0.00	0.00	0.94	0.00
railwayStation	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.98

Figure 15. Confusion matrix of proposed method for the 19-class satellite scene dataset.

5. Conclusions

In this paper, an effective image representation method for remote sensing image scene classification was introduced. The proposed representation method is based on multi-scale local binary patterns features and Fisher vectors. The MS-CLBP was applied to the partitioned dense regions of an image to extract a set of local patch descriptors, which characterize the detailed structure and texture information in high-resolution remote sensing images. The Fisher vector was employed to encode the local descriptors into a high-dimensional gradient representation, which can enhance the discriminative power of feature representation. Experimental results on two land-use scene datasets demonstrated that the proposed image representation approach obtained superior performance as compared to the existing methods for scene classification, with an obvious improvement such as 3% for the 21-class land-use dataset compared with the state-of-the-art MS-CLBP and 1% for the 19-class satellite scene dataset. In future work, combining global and local feature representations for remote sensing image scene classification will be investigated.

Acknowledgments: This work was supported by the National Natural Science Foundation of China under Grants No. NSFC-61571033, 61302164, and partly by the Fundamental Research Funds for the Central Universities under Grants No. BUCTRC201401, BUCTRC201615, XK1521.

Author Contributions: Longhui Huang, Chen Chen and Wei Li provided the overall conception of this research, and designed the methodology and experiments. Longhui Huang and Chen Chen carried out the implementation of the proposed algorithm, conducted the experiments and analysis, and wrote the manuscript. Wei Li and Qian Du reviewed and edited the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

LBP	Local binary patterns
CLBP	Completed local binary patterns
MS-CLBP	Multi-scale completed local binary patterns
FV	Fisher vector
ELM	Extreme learning machine
KELM	Kernel-based extreme learning machine
BOVW	Bag-of-visual-words
SPM	Spatial pyramid matching
SIFT	Scale-invariant feature transform
EGTD	Enhanced Gabor texture descriptor
GMM	Gaussian mixture model
CLBP_S	Completed local binary patterns sign component
CLBP_M	Completed local binary patterns magnitude component
RBF	Radial basis function
USGS	United States Geological Survey
PCA	Principal component analysis

References

1. Yang, J.; Jiang, Y.-G.; Hauptmann, A.G.; Ngo, C.-W. Evaluating bag-of-visual-words representations in scene classification. In Proceedings of the International Workshop on Workshop on Multimedia Information Retrieval, the 15th ACM International Conference on Multimedia, Augsburg, Bavaria, Germany, 23–28 September 2007; pp. 197–206.
2. Lazebnik, S.; Schmid, C.; Ponce, J. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New York, NY, USA, 17–22 June 2006; pp. 2169–2178.
3. Yang, Y.; Newsam, S. Spatial pyramid co-occurrence for image classification. In Proceedings of the International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 1465–1472.
4. Zhou, L.; Zhou, Z.; Hu, D. Scene classification using a multi-resolution bag-of-features model. *Pattern Recognit.* **2013**, *46*, 424–433. [CrossRef]
5. Zhao, L.-J.; Tang, P.; Huo, L.-Z. Land-use scene classification using a concentric circle-structured multiscale bag-of-visual-words model. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 4620–4631. [CrossRef]
6. Lowe, D.G. Distinctive image features from scale-invariant key points. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
7. Risojevic, V.; Babic, Z. Aerial image classification using structural texture similarity. In Proceedings of the IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Bilbao, Spain, 14–17 December 2011; pp. 190–195.

8. Risojević, V.; Momić, S.; Babić, Z. Gabor descriptors for aerial image classification. In *Adaptive and Natural Computing Algorithms*; Springer: Berlin, Germany; Heidelberg, Germany, 2011; pp. 51–60.
9. Oliva, A.; Torralba, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vis.* **2001**, *42*, 145–175. [CrossRef]
10. Zheng, X.; Sun, X.; Fu, K.; Wang, H. Automatic annotation of satellite images via multifeature joint sparse coding with spatial relation constraint. *IEEE Geosci. Remote Sens. Lett.* **2013**, *10*, 652–656. [CrossRef]
11. Risojevic, V.; Babic, Z. Fusion of global and local descriptors for remote sensing image classification. *IEEE Geosci. Remote Sens. Lett.* **2013**, *10*, 836–840. [CrossRef]
12. Goodfellow, I.; Courville, A.; Bengio, Y. *Deep Learning. Book in Preparation for MIT Press*; The MIT Press: Cambridge, MA, USA, 2016.
13. Bengio, Y. Learning deep architectures for AI. *Found. Trends Mach. Learn.* **2009**, *2*, 1–127. [CrossRef]
14. Yue, J.; Zhao, W.; Mao, S.; Liu, H. Spectral—Spatial classification of hyperspectral images using deep convolutional neural networks. *Remote Sens. Lett.* **2015**, *6*, 468–477. [CrossRef]
15. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Image net classification with deep convolutional neural networks. In *Neural Information Processing Systems*; Neural Information Processing Systems Foundation, Inc.: San Diego, CA, USA, 2012; pp. 1106–1114.
16. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
17. Chen, C.; Zhang, B.; Su, H.; Li, W.; Wang, L. Land-use scene classification using multi-scale completed local binary patterns. *Signal Image Video Process.* **2015**, *10*, 1–8. [CrossRef]
18. Guo, Z.; Zhang, L.; Zhang, D. A Completed modeling of local binary pattern operator for texture classification. *IEEE Trans Image Process.* **2010**, *19*, 1657–1663. [PubMed]
19. Perronnin, F.; Sánchez, J.; Mensink, T. *Improving the Fisher Kernel for Large-Scale Image Classification. Computer Vision—ECCV 2010 Lecture Notes in Computer Science*; Springer-Verlag: Berlin, Germany, 2010; pp. 143–156.
20. Huang, G.-B.; Zhou, H.; Ding, X.; Zhang, R. Extreme Learning Machine for Regression and Multiclass Classification. *IEEE Trans. Syst. Man Cybern.* **2012**, *42*, 513–529. [CrossRef] [PubMed]
21. Ojala, T.; Pietikainen, M.; Maenpaa, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 971–987. [CrossRef]
22. Li, W.; Chen, C.; Su, H.; Du, Q. Local binary patterns and extreme learning machine for hyperspectral imagery classification. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 3681–3693. [CrossRef]
23. Krapac, J.; Verbeek, J.; Jurie, F. Modeling spatial layout with fisher vectors for image categorization. In Proceedings of International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 1487–1494.
24. Sánchez, J.; Perronnin, F.; Mensink, T.; Verbeek, J. Image classification with the fisher vector: Theory and practice. *Int. J. Comput. Vis.* **2013**, *105*, 222–245. [CrossRef]
25. Liu, C. Maximum likelihood estimation from incomplete data via EM-type Algorithms. In *Advanced Medical Statistics*; World Scientific Publishing Co.: Hackensack, NJ, USA, 2003; pp. 1051–1071.
26. Jaakkola, T.S.; Haussler, D. Exploiting generative models in discriminative classifiers. *Adv. Neural Inf. Process. Syst.* **1999**, *11*, 487–493.
27. Perronnin, F.; Dance, C. Fisher kernels on visual vocabularies for image categorization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8.
28. Yang, Y.; Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 3–5 November 2010; pp. 270–279.
29. Dai, D.; Yang, W. Satellite image classification via two-layer sparse coding with biased image representation. *IEEE Geosci. Remote Sens. Lett.* **2011**, *8*, 173–176. [CrossRef]
30. Sheng, G.; Yang, W.; Xu, T.; Sun, H. High-resolution satellite scene classification using a sparse coding based multiple feature combination. *Int. J. Remote Sens.* **2012**, *33*, 2395–2412. [CrossRef]
31. Ren, J.; Zabalza, J.; Marshall, S.; Zheng, J. Effective feature extraction and data reduction in remote sensing using hyperspectral imaging [applications corner]. *IEEE Sign. Process. Mag.* **2014**, *31*, 149–154. [CrossRef]

32. Chen, C.; Li, W.; Tramel, E.W.; Fowler, J.E. Reconstruction of hyperspectral imagery from random projections using multi hypothesis prediction. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 365–374. [CrossRef]
33. Cheriyadat, A.M. Unsupervised feature learning for aerial scene classification. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 439–451. [CrossRef]
34. Zhang, F.; Du, B.; Zhang, L. Saliency-guided unsupervised feature learning for scene classification. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 2175–2184. [CrossRef]
35. Shao, W.; Yang, W.; Xia, G.-S.; Liu, G. A hierarchical scheme of multiple feature fusion for high-resolution satellite scene categorization. In *Lecture Notes in Computer Science Computer Vision Systems*; Springer: Berlin, Germany; Heidelberg, Germany, 2013; pp. 324–333.
36. Chen, S.; Tian, Y. Pyramid of spatial relatons for scene-level land use classification. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 1947–1957. [CrossRef]
37. Cvetković, S.; Stojanović, M.B.; Nikolić, S.V. Multi-channel descriptors and ensemble of extreme learning machines for classification of remote sensing images. *Sign. Process.* **2015**, *39*, 111–120. [CrossRef]
38. Keiller, N.; Waner, O.; Jefersson, A.; Dos, S. Improving spatial feature representation from aerial scenes by using convolutional networks. In Proceedings of the SIBGRAPI Conference on Graphics, Patterns and Images, Salvador, 26–29 August 2015; pp. 44–51.
39. Zhang, F.; Du, B.; Zhang, L. Scene classification via a gradient boosting random convolutional network framework. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 1793–1802. [CrossRef]
40. Keiller, N.; Otavio, P.; Jefersson, S. Towards better exploiting convolutional neural networks for remote sensing scene classification. **2016**, *ArXiv E-Prints*, arXiv:1602.01517, http://arxiv.org/abs/1602.01517.