remote sensing

MDPI

Article

The Fisher Kernel Coding Framework for High Spatial Resolution Scene Classification

Bei Zhao [1,2], Yanfei Zhong [1,*], Liangpei Zhang [1] and Bo Huang [2]

[1] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China; zhaoys@whu.edu.cn (B.Z.); zlp62@whu.edu.cn (L.Z.)

[2] Department of Geography and Resource Management, The Chinese University of Hong Kong, Sha Tin, Hong Kong; bohuang@cuhk.edu.hk

* Correspondence: zhongyanfei@whu.edu.cn; Tel./Fax: +86-27-6877-9969

Academic Editors: Josef Kellndorfer and Prasad S. Thenkabail

Received: 8 December 2015; Accepted: 14 February 2016; Published: 19 February 2016

Abstract: High spatial resolution (HSR) image scene classification is aimed at bridging the semantic gap between low-level features and high-level semantic concepts, which is a challenging task due to the complex distribution of ground objects in HSR images. Scene classification based on the bag-of-visual-words (BOVW) model is one of the most successful ways to acquire the high-level semantic concepts. However, the BOVW model assigns local low-level features to their closest visual words in the "visual vocabulary" (the codebook obtained by *k*-means clustering), which discards too many useful details of the low-level features in HSR images. In this paper, a feature coding method under the Fisher kernel (FK) coding framework is introduced to extend the BOVW model by characterizing the low-level features with a gradient vector instead of the count statistics in the BOVW model, which results in a significant decrease in the codebook size and an acceleration of the codebook learning process. By considering the differences in the distributions of the ground objects in different regions of the images, local FK (LFK) is proposed for the HSR image scene classification method. The experimental results show that the proposed scene classification methods under the FK coding framework can greatly reduce the computational cost, and can obtain a better scene classification accuracy than the methods based on the traditional BOVW model.

Keywords: fisher kernel; scene classification; Gaussian mixture model; feature coding; bag of visual words; high spatial resolution imagery

1. Introduction

A large amount of high spatial resolution (HSR) images are now available for precise land-use/land-cover investigation. The improvement of the spatial resolution of remote sensing images (less than 1 m) enables the analysis of the structure of ground objects. A lot of research has been undertaken on accurate ground object recognition (e.g., trees, buildings, roads) in HSR images [1–8]. However, the high-level semantic concepts, such as residential areas or commercial areas, cannot be acquired by these methods because of the so-called "semantic gap" between the low-level features and the high-level semantic concepts [9–12].

To bridge the semantic gap, scene classification methods based on the bag-of-visual-words (BOVW) model [13–18], part detectors [19,20], and neural networks [21–23] have been proposed, among which the BOVW model is one of the most popular approaches. In scene classification based on the BOVW model, the low-level features are extracted from the image by a local feature extraction method, e.g., mean/standard deviation statistics [9], the gray-level co-occurrence matrix [24], or scale invariant feature transform [25], and the low-level features are then assigned to their closest visual words in a "visual vocabulary", which is

a codebook learned from a large set of local low-level features with *k*-means clustering. The BOVW scene classification method then employs the statistical histogram of the visual words in the image to describe the image, and classifies it by a non-linear support vector machine (SVM) classifier [14,26]. Instead of classifying the histogram of visual words, the scene classification methods based on the probabilistic topic model [27,28], such as latent Dirichlet allocation (LDA) [9,27,29–31], are used to generate the latent topics of the visual words, and they then use the topics to represent the HSR image. To consider the spatial arrangement of the visual words in the images, different scene classification methods have been proposed with different spatial organization methods, such as the spatial pyramid matching (SPM) method [32,33], the pyramid of spatial relations method [15], and the concentric circle-structured multi-scale method [16]. The spatial relationship between visual words has also been taken into account by designing a spatial co-occurrence kernel for SVM [33,34]. However, all of these methods are designed based on the BOVW histogram description of HSR images, which loses a lot of details of the low-level features during the hard assignment to visual words.

To overcome this shortcoming, feature coding methods, e.g., sparse coding [35–40], use a coding vector to characterize each low-level feature. The coefficients of the low-level features are then reconstructed using multiple visual words instead of only one visual word. However, due to the complexity of HSR scene images, the feature coding methods all need a large codebook to code the complex low-level features precisely and obtain a satisfactory performance, which is computationally expensive. In order to decrease the size of the codebook, scene classification under the Fisher kernel (FK) coding framework [41,42] has been introduced for HSR images to characterize the low-level features with a gradient vector instead of a coding vector derived according to the distance.

Under the FK coding framework, a probabilistic generative model, such as the Gaussian mixture model (GMM), is employed to estimate the distribution of the low-level features, and the low-level features are then converted into mid-level features given the distribution of the low-level features by the gradient of the log-likelihood, which is called the FK coding procedure. The parameter space learned by the probabilistic generative model can be functionally viewed as the codebook of the low-level features. By converting the low-level features into the parameter space, the FK coding is able to preserve a lot of details of the low-level features in the coding process, which leads to a compact representation and a reduction in the size of the codebook.

In this paper, to further improve the performance of the scene classification, a local FK (LFK) coding scene classification method under the FK coding framework is proposed to incorporate the spatial information, where the local GMM (LGMM), a probabilistic generative model, is proposed to consider the spatial arrangement during estimation of the distribution of the low-level features, and the LFK coding is developed to code the spatial arrangement information into the representation. The scene classification methods developed under the FK coding framework, both with and without the incorporation of the spatial information, are called FK-S and FK-O, respectively. The contributions of this work consist of two main aspects:

(1) The introduction of a compact representation for HSR scene classification under the FK coding framework. By generating a compact representation by the use of a gradient vector instead of the count statistics in the BOVW model, the details of the low-level features can be preserved during the coding procedure, while the size of the codebook can be decreased to accelerate the speed of the codebook learning process for the HSR scene classification.
(2) The incorporation of spatial information into the scene classification under the FK coding framework, where the LGMM is able to incorporate the spatial information during the codebook learning of the low-level features, and LFK coding is correspondingly proposed to utilize this local information in the codebook.

The experimental results show that the proposed scene classification methods under the FK coding framework are able to greatly reduce the computational cost by the compact representation with a small codebook, and they can improve the performance of HSR scene classification.

The remainder of this paper is organized as follows. Section 2 describes scene classification under the FK coding framework for HSR imagery, which is followed by Sections 3 and 4 where the experimental datasets, the experimental scheme, the results, and analysis are reported. In Section 5, a discussion about the proposed method is conducted. Finally, conclusions are made in Section 6.

2. Fisher Kernel Coding Framework

To reduce the size of the codebook and preserve the details of the low-level features as much as possible, the FK coding framework is introduced to obtain compact descriptions for the scene classification of HSR images. Under the FK coding framework, a scene classification method is proposed to incorporate the spatial information of the HSR scenes. In the following parts, the FK coding framework is introduced for the representation of HSR images in Part A, while the scene classification methods under the FK coding framework, both with and without the incorporation of the spatial information (denoted by FK-S and FK-O, respectively), are described in Part B and C, respectively.

2.1. Fisher Kernel Coding Framework for the Representation of HSR Scenes

The Fisher kernel (FK) is a technique that combines the advantages of the generative and discriminative approaches by describing a signal with a gradient vector of its probability density function (PDF) with respect to the parameters of the PDF. Figure 1 shows the FK coding framework that is used to obtain the representation of the HSR imagery.

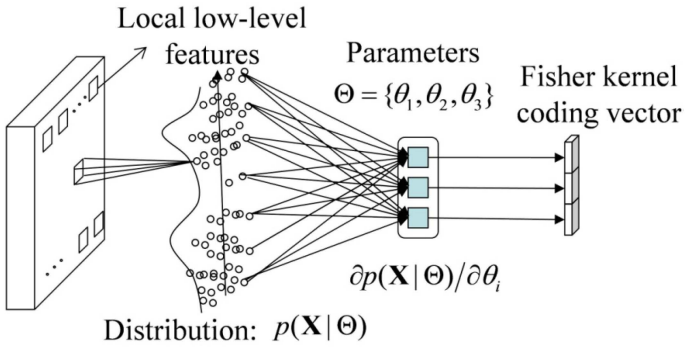


Figure 1. FK coding framework for the representation of HSR imagery.

We let p be the PDF of the local low-level features. The set of local low-level features in a HSR image $\mathbf{X} = \{\mathbf{X}_j\}_{j=1}^{n}$ can then be characterized by the gradient vector $\nabla_\Theta p(\mathbf{X}|\Theta)$, where n is the number of patches in the image, and Θ is the set of parameters of the PDF. The gradient vector describes the magnitude and direction that the parameters are modified to fit the data. To normalize the gradient vector, the Fisher information matrix is recommended, which measures the amount of information that $\mathbf{X}$ carries about the unknown Θ of the PDF, and can be written as:

$$\mathbf{F}_\Theta = E_\mathbf{X}\left[\left(\nabla_\Theta \log p(\mathbf{X}|\Theta)\right)\left(\nabla_\Theta \log p(\mathbf{X}|\Theta)\right)^T\right]. \quad (1)$$

The normalized gradient vector is then derived by:

$$G_\Theta^\mathbf{X} = \mathbf{F}_\Theta^{-1/2}\nabla_\Theta \log p(\mathbf{X}|\Theta). \quad (2)$$

Finally, the normalized gradient vector is used to represent the HSR image, and is classified by a discriminative classifier, such as SVM. Under this FK coding framework, the method of local low-level feature extraction, the probabilistic generative model, and the discriminative classifier can be changed according to the characteristics of these models and the HSR images.

2.2. Scene Classification without the Consideration of the Spatial Information (FK-O)

In this part, FK-O is introduced to classify HSR scenes without the consideration of the spatial information. Under the FK coding framework, the GMM is employed as the probabilistic generative model to estimate the PDF of the low-level features. The FK coding is then performed to obtain the coding vectors to represent the HSR scenes. Finally, the coding vectors of the training images are used to train the discriminative classifier, SVM, which is used to classify the coding vectors of the test images (Figure 2). The details are as follows.

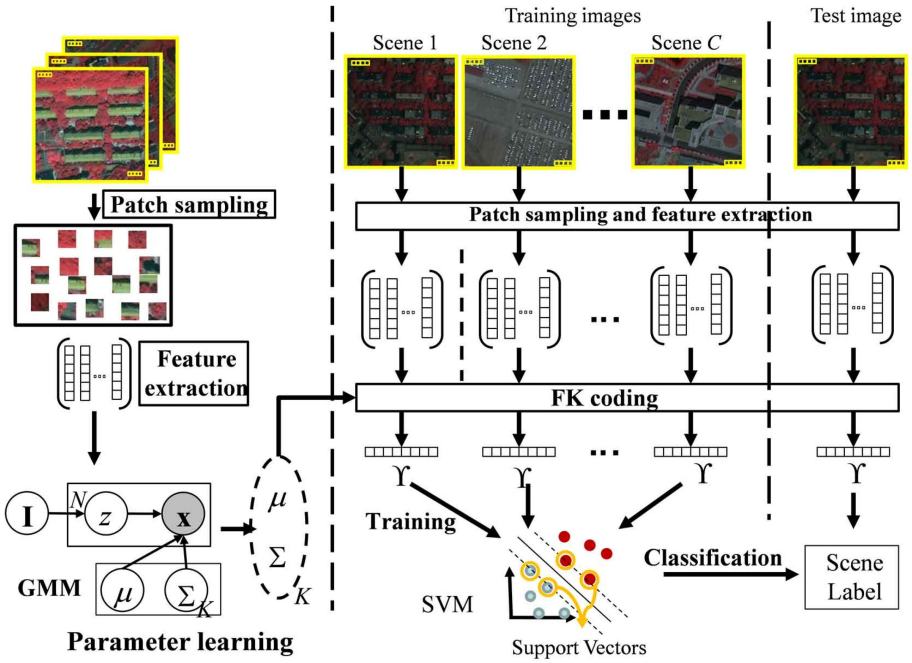


Figure 2. Procedure of the FK-O scene classification method.

2.2.1. Patch Sampling and Feature Extraction

For each scene image, the patches are evenly sampled from each region with a certain size and spacing (e.g., 8×8 pixels size and 4 pixels spacing), which are empirically selected to obtain a good scene classification performance. The local low-level features can then be extracted from the patches. To acquire the low-level features, there are many local descriptors, such as the descriptors based on the gray-level co-occurrence matrix [24] and scale invariant feature transform (SIFT) [25]. In this work, the mean/standard deviation statistics [9] are used to extract the low-level features because of their simplicity and performance in HSR scene classification.

We let $\mathbf{x}$ be the low-level features extracted from the patch, where $\mathbf{x}$ can be obtained by computing the mean and standard deviation features of this patch with Equation (3). In Equation (3), B is the number of spectral bands of the image, n is the number of pixels in the patch, and $v_{p,b}$ is the b-th band value of the p-th pixel in the patch

$$\begin{aligned} \mathbf{x} &= \left(x_1^{\mathrm{m}}, \cdots, x_B^{\mathrm{m}}, x_1^{\mathrm{std}}, \cdots, x_B^{\mathrm{std}}\right)^T \\ x_b^{\mathrm{m}} &= \textstyle\sum_{p=1}^{n} v_{p,b}/n, x_b^{std} = \sqrt{\textstyle\sum_{p=1}^{n} \left(v_{p,b} - x_b^{\mathrm{m}}\right)^2/n} \end{aligned} . \tag{3}$$

2.2.2. Fisher Kernel Coding and Scene Classification

To obtain a compact representation of the HSR scene, the FK coding method is introduced to code the low-level features into mid-level coding vectors, without losing too many details. Before the FK coding, the distribution of the low-level features should be estimated by the GMM. We let $\mathbf{x}_j$ be the

low-level feature of the j-th patch, and the sets of patches used to learn the parameters of the GMM $\Theta = \{\alpha_k, \mu_k, \Sigma_k\}_{k=1}^{K}$ can then be denoted by $\mathbf{I} = \{\mathbf{x}_j\}_{j=1}^{N}$, where $\{\alpha_k\}_{k=1}^{K}$ are the priors of the Gaussians, $\boldsymbol{\mu}_k = \{\mu_{k,d}\}_{d=1}^{\mathrm{D}}$ and $\boldsymbol{\Sigma}$ are the mean and covariance matrix of the k-th Gaussian component, D ($D = 2B$) is the dimension of the features, and K is the number of Gaussian components. For the FK coding, the covariance matrix $\boldsymbol{\Sigma}$ of each cluster is usually approximated by a diagonal matrix $\boldsymbol{\sigma}$, where the diagonal elements are the variances of the features of the pixels in the cluster. We let d be the index of the components of the features, $\boldsymbol{\sigma}_k^2 = \mathrm{diag}(\sigma_{k,1}^2, \sigma_{k,2}^2, \cdots, \sigma_{k,D}^2)$.

Given the low-level features $\mathbf{X} = \{\mathbf{x}_j\}_{j=1}^{n}$ in an image, where $\mathbf{x}_j = \{x_{j,d}\}_{d=1}^{D}$, and n is the number of patches in the image, the image can then be described by the normalized gradient vector (Equation (4)) under the FK coding framework.

$$G_{\Theta}^{\mathbf{X}} = \mathbf{F}_{\Theta}^{-1/2}\left(\frac{1}{n}\sum_{j=1}^{n}\nabla_{\Theta}\log p(\mathbf{x}_j|\Theta)\right), \Theta = \{\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k | k = 1, \cdots K\}. \tag{4}$$

The FK coding vector with respect to $\bar{}_k$ and $\boldsymbol{\text{œ}}_k$ can be derived as shown in Equations (5) and (6), respectively, where the posterior probability $\tau_{j,k}$ can be obtained by Equation (7),

$$\Upsilon_{\mu_{k,d}} = \frac{1}{n\sqrt{\alpha_k}}\sum_{j=1}^{n}\tau_{j,k}\left(x_{j,d} - \mu_{k,d}\right)/\sigma_{k,d}, \tag{5}$$

$$\Upsilon_{\sigma_{k,d}} = \frac{1}{n\sqrt{2\alpha_k}}\sum_{j=1}^{n}\tau_{j,k}\left((x_{j,d} - \mu_{k,d})^2/\sigma_{k,d}^2 - 1\right), \tag{6}$$

$$\tau_{j,k} = \frac{\alpha_k p(\mathbf{x}_j|\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k)}{\sum_{k=1}^{K}\alpha_k p(\mathbf{x}_j|\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k)}. \tag{7}$$

The Fisher vector of an image can be written as $\Upsilon = (\Upsilon_{\mu}, \Upsilon_{\sigma}) \in \Re^{2KD}$, where $\Upsilon_{\mu} = (\Upsilon_{\mu_{1,1}}, \cdots, \Upsilon_{\mu_{1,D}}, \cdots, \Upsilon_{\mu_{K,1}}, \cdots, \Upsilon_{\mu_{K,D}})$ and $\Upsilon_{\sigma} = (\Upsilon_{\sigma_{1,1}}, \cdots, \Upsilon_{\sigma_{1,D}}, \cdots, \Upsilon_{\sigma_{K,1}}, \cdots, \Upsilon_{\sigma_{K,D}})$. From Equations (5) and (6), it can be seen that the low-level features are coded by the gradient between the low-level features and the parameters of the Gaussian components, which infers that the coding vector can preserve the details of the low-level features as much as possible, compared to the traditional feature coding method based on the distance. In addition, in order to improve the performance, L_2-normalization and power normalization are recommended by Perronnin *et al.* [41]. After the FK coding, each image can be represented by an FK coding vector Υ.

Finally, the coding vectors of the training images are used to train an SVM classifier [43], while the coding vector of the test image is classified by the trained SVM. During the training of the SVM classifier, the histogram intersection kernel (HIK) is adopted due to its performance in image classification [44]. The HIK is defined as shown in Equation (8), where q is the index of the component of the coding vector,

$$k(\boldsymbol{\gamma}, \boldsymbol{\gamma}^i) = \sum_{q}\min(\gamma_q, \gamma_q^i). \tag{8}$$

2.3. Scene Classification with the Consideration of the Spatial Information (FK-S)

In order to consider the spatial information, the scene classification method under the FK coding framework for HSR scenes, FK-S, is proposed in this part. The procedure of the proposed method is shown in Figure 3.

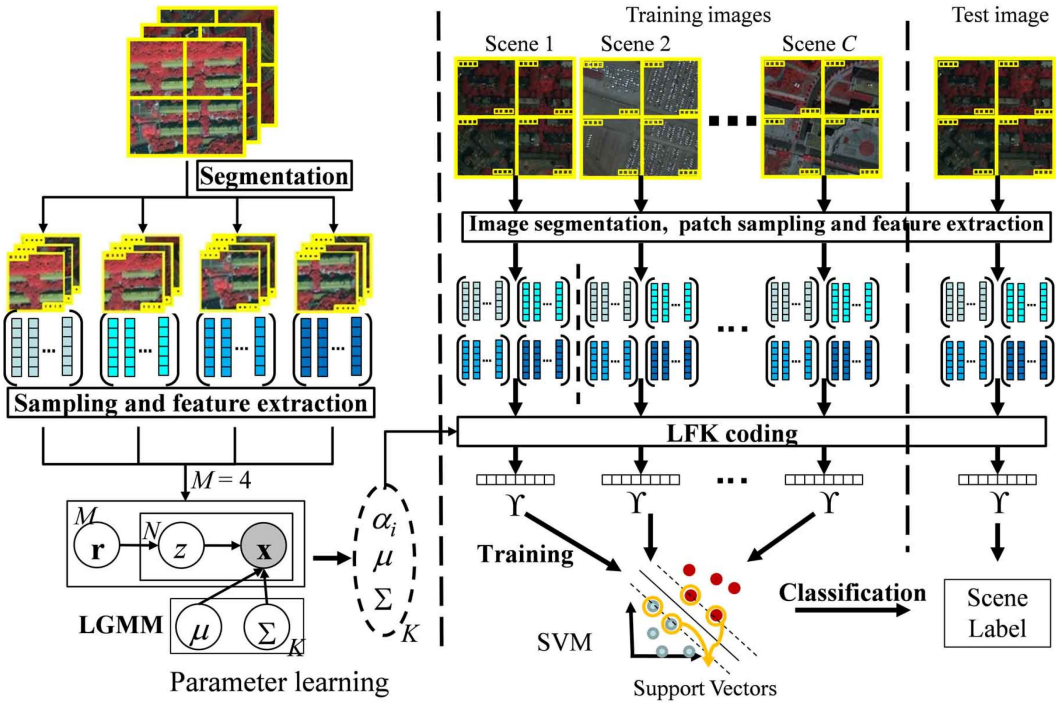


Figure 3. Procedure of the FK-S scene classification method.

Instead of the GMM, FK-S uses the LGMM to estimate the distribution of the low-level features by considering the difference between different regions of the HSR scenes, while LFK is developed to code the HSR images to adapt to the change brought about by the change of distribution estimation method. The details of FK-S are described in the following parts.

2.3.1. Image Segmentation, Patch Sampling, and Feature Extraction

For each scene image, chessboard segmentation is used to split the whole image into multiple regions, while the patches are evenly sampled from each region with a certain size and spacing. The local low-level features can then be extracted from the patches. Figure 4 shows the multiple regions of an image produced by chessboard segmentation with different numbers of regions M, where i is the index of the regions, j is the index of the patches, and $\mathbf{x}_{i,j}$ is the low-level feature extracted from the j-th patch in the i-th region.

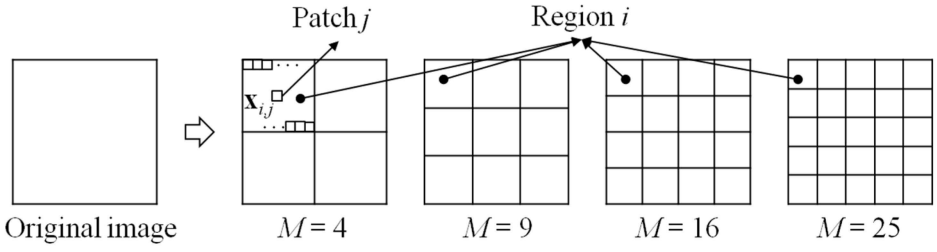


Figure 4. Image segmentation by chessboard segmentation with different numbers of regions.

As in FK-O, FK-S also employs the mean/standard deviation statistics to extract the low-level features (Equation (3)). We let $x_{i,j,d}$ be the d-th component of $\mathbf{x}_{i,j}$, and $\mathbf{x}_{i,j} = \left(x_{i,j,1}, x_{i,j,2}, \cdots, x_{i,j,D}\right) \in \Re^D$, where D is the dimension of the low-level features. All the regions of the images can then be denoted as $\mathbf{R} = \{\mathbf{r}_i\}_{i=1}^{M}$, where M is the number of regions, $\mathbf{r}_i = \left\{\mathbf{x}_{i,1}, \mathbf{x}_{i,2}, \cdots, \mathbf{x}_{i,n_i}\right\}$ is the set of low-level features in the i-th region, and n_i is the number of patches in the i-th region.

2.3.2. Learning the Parameters of the Local Gaussian Mixture Model (LGMM)

Considering that the traditional GMM (Figure 5a) generates all the features $\mathbf{x}$ in the whole scope of the images from Gaussians with the same priors $P(z \mid \mathbf{I})$ (also known as mixing weights), which ignores the spatial arrangement of the HSR images during the estimation of the distribution of the low-level features, the LGMM (Figure 5b) is used to learn the distribution of the low-level features, where the features $\mathbf{x}$ in the different regions are generated from Gaussians with different priors $\{P(\mathrm{z}|\mathbf{r}_i)\}_{i=1}^{M}$. In particular, for the i-th region $\mathbf{r}_i$, the identities of the Gaussians z are generated from the priors $P(z \mid \mathbf{r}_i)$, and the features $\mathbf{x}$ in this region can then be extracted from the Gaussians identified by the corresponding $\mathbf{z}$. Due to the different treatment of different regions, the LGMM is able to estimate different sets of priors of Gaussians for different regions $\{P(\mathrm{z}|\mathbf{r}_i)\}_{i=1}^{M}$, which reflects the different distributions of low-level features in the different regions. Therefore, the distribution of the low-level features estimated by the LGMM can take into account the spatial arrangement of the low-level features.

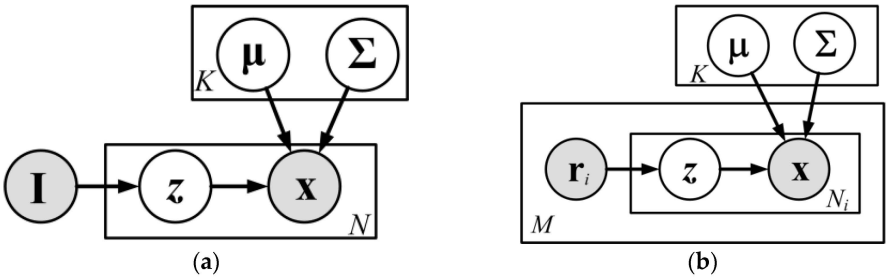


Figure 5. Graphical model representations. (**a**) GMM; (**b**) LGMM.

We let $\mathrm{z}_{i,j}$ be the latent value of the low-level feature $\mathbf{x}_{i,j}$ in $\mathbf{r}_i$, and the probability of pixel $\mathbf{x}_{i,j}$ being drawn from the k-th Gaussian ($\mathrm{z}_{i,j} = k$) is described in Equation (9), where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the mean vector and the covariance matrix of the k-th Gaussian, respectively,

$$p(\mathbf{x}_{i,j}|\mathbf{r}_i, z_{i,j} = k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi)^{\frac{D}{2}}|\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}_{i,j} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_{i,j} - \boldsymbol{\mu}_k)\right). \tag{9}$$

In order to learn the parameters of the distribution of the low-level features for the HSR scenes, a number of images are randomly selected from the HSR image dataset, and should be divided into M regions by chessboard segmentation (Figure 3). All the low-level features of patches in the same region of all the selected images are collected and form a new set of features $\mathbf{r}_\mathbf{i}' = \{\mathbf{x}_{i,1}, \mathbf{x}_{i,2}, \cdots, \mathbf{x}_{i,N_i}\}$, where $i \in \{1, 2, \cdots, M\}$, $N_i = \sum_l n_{l,i}$, and $n_{l,i}$ is the number of patches in the i-th region of the l-th selected image. Assuming that all the local low-level features are independent, the log-likelihood of all the features can be formulated by Equation (10), where $\alpha_{i,k} = P(\mathrm{z}_{i,j} = \mathrm{k}|\mathbf{r}_i)$. The log-likelihood of all the features is then parameterized by $\Theta = \{\{\alpha_{i,k}\}_{i=1}^{M}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$.

$$L(\Theta; \mathbf{X}) = \log p(\mathbf{X}|\mathbf{R}, \Theta) = \sum_{i=1}^{M}\sum_{j=1}^{N_i} \log(\sum_{k=1}^{K} \alpha_{i,k} p(\mathbf{x}_{i,j}|\mathbf{r}_i, z_{i,j} = k, \mu_k, \Sigma_k)). \tag{10}$$

The expectation-maximization (EM) algorithm is employed to estimate the parameters of the LGMM, as in the GMM. The EM algorithm begins with an initial estimate $\Theta^{(0)}$ and repeats the following two steps:

E-step. Compute the expected value $Q(\Theta|\Theta^{(t)}) = E_{z|\mathbf{X},\mathbf{R},\Theta^{(t)}}\left[L\left(\Theta; \mathbf{X}, \mathbf{z}\right)\right]$ of the log-likelihood function with respect to the conditional distribution $P(\mathrm{z}_{i,j} = \mathrm{k}|\mathbf{r}_i, \mathbf{x}_{i,j}, \Theta^{(t)})$, according to Equation (11).

$$Q(\Theta|\Theta^{(t)}) = \sum_{i=1}^{M}\sum_{j=1}^{N}\sum_{k=1}^{K} P(z_{i,j}=k|\mathbf{r}_i,\mathbf{x}_{i,j},\Theta^{(t)})\log(\alpha_{i,k}P(\mathbf{x}_{i,j}|\mathbf{r}_i,z_{i,j}=k,\mu_k,\Sigma_k)). \tag{11}$$

In Equation (11), $\tau_{k.i,j}^{(t)} = P(\mathrm{z}_{i,j}=\mathrm{k}|\mathbf{r}_i,\mathbf{x}_{i,j},\Theta^{(t)})$ can be calculated by Equation (12),

$$\tau_{k,i,j}^{(t)} = P(z_{i,j}=k|\mathbf{r}_i,\mathbf{x}_{i,j},\Theta^{(t)}) = \frac{\alpha_{i,k}^{(t)}p(\mathbf{x}_{i,j}|\mathbf{r}_i,z_{i,j}^{(t)}=k,\boldsymbol{\mu}_k^{(t)},\boldsymbol{\Sigma}_k^{(t)})}{\sum_{k=1}^{K}\alpha_{i,k}^{(t)}p(\mathbf{x}_{i,j}|\mathbf{r}_i,z_{i,j}^{(t)}=k,\boldsymbol{\mu}_k^{(t)},\boldsymbol{\Sigma}_k^{(t)})}. \tag{12}$$

M-step. Maximize $Q(\Theta|\Theta^{(t)})$ with the constraint that $\sum_{k=1}^{K}\alpha_{i,k} = 1$ to obtain the update equation of parameters $\Theta = \{\{\alpha_{i,k}\}_{i=1}^{M},\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\}_{k=1}^{K}$. To solve this problem, Lagrange multipliers $\{\lambda_i\}_{i=1}^{M}$ are introduced into the objective function $Q(\Theta|\Theta^{(t)})$. The new objective function $\Phi(\Theta|\Theta^{(t)})$ can then be rewritten as:

$$\Phi(\Theta|\Theta^{(t)}) = \sum_{i=1}^{M}\lambda_i(1-\sum_{k=1}^{K}\alpha_{i,k}) + \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{K}\tau_{k,i,j}^{(t)}\log(\alpha_{i,k}P(\mathbf{x}_{i,j}|\mathbf{r}_i,z_{i,j}=k,\mu_k,\Sigma_k)). \tag{13}$$

To obtain the updated equation of $\alpha_{i,k}^{(t+1)}$, $\mu_k^{(t+1)}$, and $\Sigma_k^{(t+1)}$, the objective functions are obtained by isolating the terms with $\alpha_{i,k}$, $\mu_k^{(t+1)}$, and $\Sigma_k^{(t+1)}$, respectively, and can be written as:

$$\Phi(\alpha_{i,k}|\Theta^{(t)}) = -\lambda_i\alpha_{i,k} + \sum_{j=1}^{N_i}\tau_{k,i,j}^{(t)}\log(\alpha_{i,k}), \tag{14}$$

$$\Phi(\mu_k|\Theta^{(t)}) = -\sum_{i=1}^{M}\sum_{j=1}^{N_i}\tau_{k,i,j}^{(t)}(\mathbf{x}_{i,j}-\mu_k^{(t+1)})^{T}\Sigma_k^{-1}(\mathbf{x}_{i,j}-\mu_k^{(t+1)}), \tag{15}$$

$$\Phi(\Sigma_k|\Theta^{(t)}) = -\sum_{i=1}^{M}\sum_{j=1}^{N}\tau_{k,i,j}^{(t)}(\log|\Sigma_k| + (\mathbf{x}_{i,j}-\mu_k^{(t+1)})^{T}\Sigma_k^{-1}(\mathbf{x}_{i,j}-\mu_k^{(t+1)})). \tag{16}$$

By maximizing the objective functions, the updated equations of $\alpha_{i,k}^{(t+1)}$, $\mu_k^{(t+1)}$, and $\Sigma_k^{(t+1)}$ can be obtained as shown in Equations (17)–(19), respectively:

$$\alpha_{i,k}^{(t+1)} = \sum_{j=1}^{N_i}\tau_{k,i,j}^{(t)}/N_i, \tag{17}$$

$$\boldsymbol{\mu}_k^{(t+1)} = \sum_{i=1}^{M}\sum_{j=1}^{N_i}\tau_{k,i,j}^{(t)}\mathbf{x}_{i,j}/\sum_{i=1}^{M}\sum_{j=1}^{N_i}\tau_{k,i,j}^{(t)}, \tag{18}$$

$$\boldsymbol{\Sigma}_k^{(t+1)} = \sum_{i=1}^{M}\sum_{j=1}^{N_i}\tau_{k,i,j}^{(t)}(\mathbf{x}_{i,j}-\boldsymbol{\mu}_k^{(t+1)})(\mathbf{x}_{i,j}-\boldsymbol{\mu}_k^{(t+1)})^{T}/\sum_{i=1}^{M}\sum_{j=1}^{N_i}\tau_{k,i,j}^{(t)}. \tag{19}$$

The EM algorithm is terminated when the last two values of the log-likelihood are close enough (below some preset convergence threshold) or the number of iterations reaches the preset number. Similarly, assuming that the components of the feature vectors are independent, the covariance matrix $\boldsymbol{\Sigma}$ of each Gaussian can be replaced with a diagonal matrix $\boldsymbol{\sigma}^2$. Equations (9) and (19) can then be rewritten as Equations (20) and (21), respectively, where $\boldsymbol{\sigma}_k^2 = \mathbf{diag}(\sigma_{k,1}^2,\sigma_{k,2}^2,\cdots,\sigma_{k,D}^2)$,

$$P(\mathbf{x}_{i,j}|\mathbf{r}_i,z_{i,j}=k,\boldsymbol{\mu}_k,\boldsymbol{\sigma}_k) = (2\pi)^{-D/2}\prod_{d=1}^{D}\sigma_{k,d}^{-1/2}exp\left(-\sum_{d=1}^{D}\frac{(x_{i,j,d}-\mu_{k,d})^2}{2\sigma_{k,d}}\right), \tag{20}$$

$$\sigma_{k,d}^{2\ (t+1)} = \sum_{i=1}^{M}\sum_{j=1}^{N_i}\tau_{k,i,j}^{(t)}(x_{i,j,d}-\mu_{k,d}^{(t+1)})^2/\sum_{i=1}^{M}\sum_{j=1}^{N_i}\tau_{k,i,j}^{(t)}. \tag{21}$$

The M sets of features are then used to learn the parameters $\Theta^* = \{\{\alpha_{i,k}\}_{i=1}^{M},\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\}_{k=1}^{K}$ by the use of the LGMM.

2.3.3. Local Fisher Kernel (LFK) Coding and Scene Classification

To incorporate the spatial information contained in the parameters obtained by the LGMM, an LFK coding method is proposed under the FK coding framework.

Given the low-level features $\mathbf{R} = \{\mathbf{r}_i\}_{i=1}^{M}, \mathbf{r}_i = \{\mathbf{x}_{i,j}\}_{j=1}^{n_i}$ in an image, the LFK coding vector of the image can then be described by Equation (22) under the FK coding framework, where n_i is the number of patches in the i-th region of the image,

$$G_{\Theta}^{\mathbf{R}} = \mathbf{F}_{\Theta}^{-1/2}\left(\frac{1}{M}\sum_{i=1}^{M}\frac{1}{n_i}\sum_{j=1}^{n_i}\nabla_{\Theta}\log p(\mathbf{x}_{i,j}|\Theta)\right), \Theta = \left\{\{\alpha_{i,k}\}_{i=1}^{M}, \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k\right\}_{k=1}^{K}. \quad (22)$$

The LFK coding vector with respect to $\alpha_{i,k}$, $\boldsymbol{\mu}_k$, and $\boldsymbol{\sigma}_k$ can be derived as shown in Equations (23)–(25), respectively, where the posterior probability $\tau_{i,j,k}$ can be obtained by Equation (12) with the parameters $\Theta^* = \{\{\alpha_{i,k}\}_{i=1}^{M}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$,

$$\Upsilon_{\alpha_{i,k}} = \frac{1}{\sqrt{n_i\left(\frac{1}{\alpha_{i,k}}+\frac{1}{\alpha_{i,1}}\right)}}\sum_{j=1}^{n_i}\left(\frac{\tau_{i,j,k}}{\alpha_{i,k}}-\frac{\tau_{i,j,1}}{\alpha_{i,1}}\right), \quad k \geqslant 2. \quad (23)$$

$$\Upsilon_{\mu_{k,d}} = \frac{1}{\sqrt{\sum_{i=1}^{M} n_i\alpha_{i,k}}}\sum_{i=1}^{M}\sum_{j=1}^{n_i}\tau_{i,j,k}\left(\frac{x_{i,j,d}-\mu_{k,d}}{\sigma_{k,d}}\right), \quad (24)$$

$$\Upsilon_{\sigma_{k,d}} = \frac{1}{\sqrt{2\sum_{i=1}^{M} n_i\alpha_{i,k}}}\sum_{i=1}^{M}\sum_{j=1}^{n_i}\tau_{i,j,k}\left(\frac{(x_{i,j,d}-\mu_{k,d})^2}{\sigma_{k,d}^2}-1\right), \quad (25)$$

Finally, the LFK coding vector of an image can be written as $\Upsilon = (\Upsilon_{\alpha}, \Upsilon_{\mu}, \Upsilon_{\sigma}) \in \Re^{2KD+M(K-1)}$, where $\Upsilon_{\alpha} = (\Upsilon_{\alpha_{1,2}}, \cdots, \Upsilon_{\alpha_{1,K}}, \cdots, \Upsilon_{\alpha_{M,2}}, \cdots, \Upsilon_{\alpha_{M,K}})$, $\Upsilon_{\mu} = (\Upsilon_{\mu_{1,1}}, \cdots, \Upsilon_{\mu_{1,D}}, \cdots, \Upsilon_{\mu_{K,1}}, \cdots, \Upsilon_{\mu_{K,D}})$, and $\Upsilon_{\sigma} = (\Upsilon_{\sigma_{1,1}}, \cdots, \Upsilon_{\sigma_{1,D}}, \cdots, \Upsilon_{\sigma_{K,1}}, \cdots, \Upsilon_{\sigma_{K,D}})$. It is worth noting that the LFK coding vector with respect to the priors $\Upsilon_{\alpha_{i,k}}$ contains the spatial information obtained by the LGMM, and the number of components of $\Upsilon_{\alpha_{i,k}}$ $M(K-1)$ should be kept at less than 50% of the dimension of the LFK coding vector, $2KD+M(K-1)$, to ensure that the spatial information is less important than the low-level feature information in the coding vector. Therefore, the number of regions M should be less than $2KD/(K-1) \approx 2D$. In addition, when M is a small number, the importance of the spatial information decreases, and we recommend that M should be set as larger than 1. For example, when the number of bands of the images $B = 3$, then $D = 2B = 6$, and $1 < M < 2D = 12$. Between $M = 4$ and $M = 9$, we recommend $M = 9$, because it can explore more spatial information for the HSR scene images.

As in FK-O, L_2-normalization and power normalization are recommended to improve the performance of FK-S. After the LFK coding, each image can be represented by an LFK coding vector Υ^{I}. Finally, the coding vectors of the training images are used to train an SVM classifier with HIK, while the coding vector of the test image is classified by the trained SVM.

Both FK-O and FK-S are developed under the FK coding framework, where the low-level features are coded by the gradient between the low-level features and the parameters of the Gaussian components, which leads to the ability to preserve more of the details of the low-level features than the traditional feature coding method based on the distance.

3. Datasets and Experimental Scheme

In order to test the performance of the scene classification methods developed under the FK coding framework for HSR imagery, namely FK-O and FK-S, the commonly used UC Merced (UCM)

land-use dataset [33] (Figure 6), a Google dataset (Figure 7), and an IKONOS dataset (Figure 8) were used to conduct the scene classification experiments. The BOVW model, SPM [32], LDA [9], and LDA with a hybrid strategy (P-LDA) [30] were employed as the comparison methods, where the classifier of BOVW was SVM with a radial basis function (RBF) kernel. For the UCM dataset, the accuracies published in the previous works [15–17,19,22,33–35] are also reported.

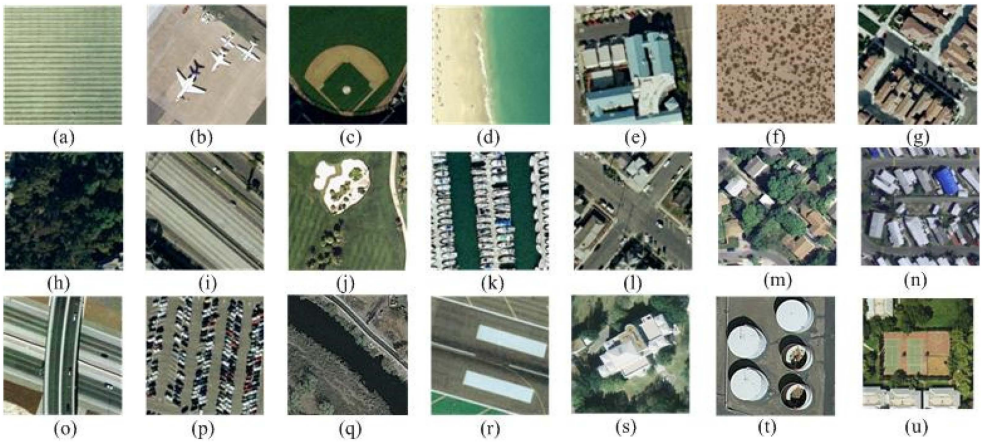


Figure 6. UCM dataset. (**a**–**u**) agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis courts.

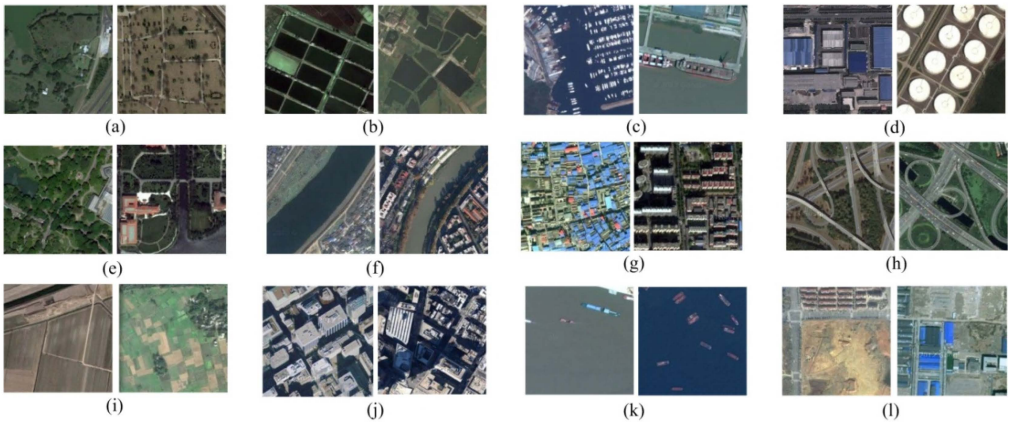


Figure 7. Google dataset. (**a**–**l**) meadow, pond, harbor, industrial, park, river, residential, overpass, agriculture, commercial, water, and idle land.

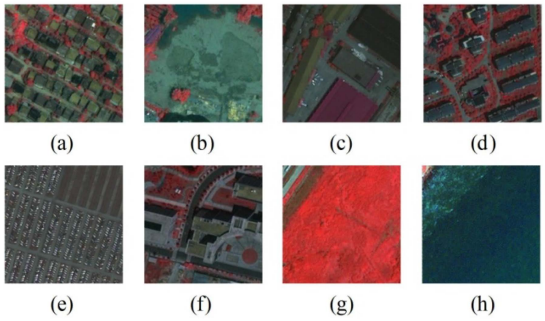


Figure 8. Wuhan IKONOS dataset. (**a**–**h**) dense residential, idle, industrial, medium residential, parking lot, commercial, vegetation, and water.

3.1. Experimental Datasets

UCM dataset: the UCM dataset contains 21 land-use classes (Figure 6), namely agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis courts. In the UCM dataset, each class consists of 100 aerial orthophotographs with 256 × 256 pixels and a 1 ft resolution, which were extracted from large images in the USGS National Map Urban Area image collection for various urban areas around the US.

Google dataset: the Google dataset was acquired from Google Earth (Google Inc., Cambridge, MA, USA) and mainly covers urban areas in China. This dataset contains meadow, pond, harbor, industrial, park, river, residential, overpass, agriculture, commercial, water, and idle land classes (Figure 7). Each class contains 200 images with a 2 m spatial resolution and a size of 200 × 200 pixels.

Wuhan IKONOS dataset: The HSR images in the Wuhan IKONOS dataset were acquired over the city of Wuhan in China by the IKONOS sensor in June 2009. The spatial resolutions of the panchromatic images and the multispectral images are 1 m and 4 m, respectively. All the images in the Wuhan IKONOS dataset were obtained by Gram–Schmidt pan-sharpening with ENVI 4.7 software. In the Wuhan IKONOS dataset, eight scene classes are defined, namely dense residential, idle, industrial, medium residential, parking lot, commercial, vegetation, and water (Figure 8). Each class contains 30 images with a size of 150 × 150 pixels, a 1 m spatial resolution, and blue, green, red, and near-infrared bands. A large image with a size of 6150 × 8250 pixels and a 1 m resolution was used for the annotation experiment (Figure 9a).

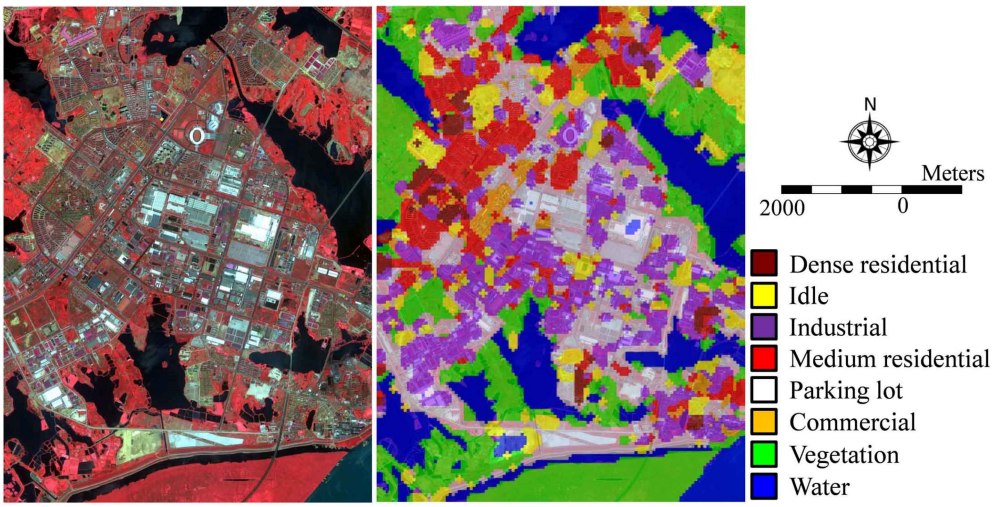


Figure 9. Large image annotation using the Wuhan IKONOS dataset. (**a**) false-color image of the large image with 6150 × 8250 pixels; (**b**) annotated large image.

3.2. Experimental Scheme

In the experiments, the BOVW, LDA, and P-LDA scene classification methods employed the mean and standard deviation statistics as the low-level feature extractor, in the same way as the FK-O and FK-S scene classification methods. For the SPM scene classification method, not only the mean and standard deviation statistics, but also the SIFT descriptor, were used to extract the low-level features, and are denoted by SPM-MeanStd and SPM-SIFT, respectively. For SPM, it was found that the number of pyramid levels is better set as one, rather than two, in the experiments with the three datasets. Therefore, the accuracies acquired by SPM with one level of pyramid were used for the comparison. During the low-level feature extraction using the mean and standard deviation statistics or SIFT, the size and spacing of the sampling patch were empirically set. For the three datasets, different sizes and

spacings of sampling patches were tested by the use of SPM. The results (Figure 10) showed that it is best to set the patch size and spacing as 8 × 8 pixels and four pixels, respectively.

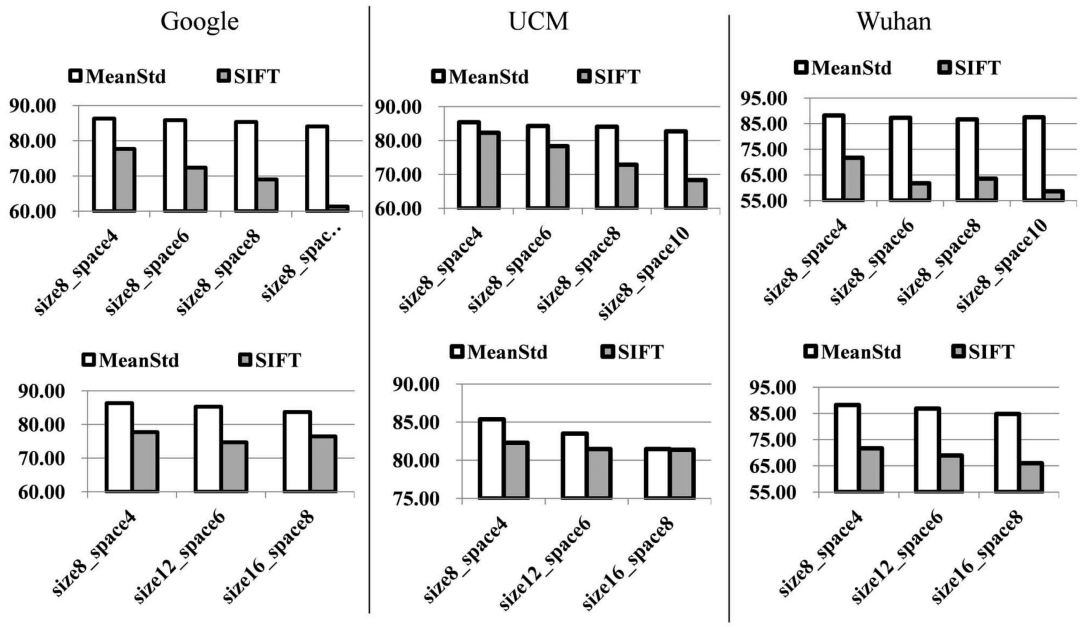


Figure 10. Classification performance with different patch sizes and spacing. The top and bottom rows show the classification accuracies when varying the patch spacing from four to ten pixels, with the patch size as 8 × 8 pixels, and when varying the patch size from 8 × 8 to 16 × 16 pixels, with the patch spacing as 50% of the size, respectively.

For the BOVW, SPM, LDA, and P-LDA methods, the number of cluster centers was set to 1000, which was optimally selected from 200, 400, 600, 800, 1000, and 1200 by considering the scene classification accuracy and the computational complexity. For LDA and P-LDA, the topic numbers were optimally chosen from 30 to 100 with a step size of 10 during the scene classification. The topic numbers of P-LDA were set to 100 for the three datasets to obtain the best classification accuracies. For LDA, the number of topics in each scene class was automatically optimized in the procedure of scene classification according to the perplexity index [9]. The parameters of the SVM classifier were tuned to obtain the best accuracy.

The codebooks were obtained by GMM, LGMM, or *k*-means with 1050, 960, and 192 images randomly selected from the UCM dataset, the Google dataset, and the Wuhan IKONOS dataset, respectively. For FK-O and FK-S, the number of Gaussians K of the GMM and the LGMM was varied between 8, 16, 32, 64, and 128. For the LGMM, the number of regions M was varied between 4, 9, 16, and 25. By varying these parameters, the best classification accuracies were used for the comparison. In the scene classification, 80, 100, and 24 images per class were randomly selected to train the SVM classifier from the UCM dataset, the Google dataset, and the Wuhan IKONOS dataset, respectively, while the rest of the images were used to test the performance. The classification performance was quantitatively evaluated by the classification accuracy, as defined in Equation (26), where N_c is the number of correctly classified images in the test images, and N_t is the total number of test images. The scene classification experiments were repeated 20 times to generate the mean and standard deviation of the accuracies,

$$Acc = N_c/N_t. \tag{26}$$

An annotation experiment was also performed to test the performance of the proposed scene classification method with a large HSR image (Figure 9a), using the Wuhan IKONOS dataset. During the annotation of the large image, the large image was split into a set of scene images, where the image

size and spacing were set to 150 × 150 pixels and 100 pixels, respectively. Therefore, there were 50 overlapping pixels between two adjacent images. All the labeled images in the Wuhan IKONOS dataset were used to train the FK-S model, which was employed to classify the scene images obtained from the large image. For the overlapping pixels between adjacent images, their class labels were determined by the majority voting rule. The large annotation maps were evaluated visually by overlaying the annotation maps on the original image (with 60% transparency).

4. Results and Accuracies

The FK-O method obtained the highest classification accuracies when the number of Gaussians K was set to 128, 64, and 32 for the UCM dataset, the Google dataset, and the Wuhan IKONOS dataset, respectively, while the FK-S method obtained the best performance when K was set to 128, 128, and 32 for the UCM dataset, the Google dataset, and the Wuhan IKONOS dataset, respectively. For all the datasets, when the number of regions M was set to 9, FK-S acquired the best accuracy. The classification accuracies of the different methods for the three image datasets are reported in Table 1. Here, it can be seen that the feature coding methods under the FK coding framework, namely FK-O and FK-S, acquired accuracies of 91.38 ± 1.54(%) and 91.63 ± 1.49(%) for the UCM dataset, 90.16 ± 0.82(%) and 90.40 ± 0.84(%) for the Google dataset, and 89.67 ± 4.19(%) and 90.71 ± 4.41(%) for the Wuhan IKONOS dataset, respectively.

(1) Comparison between the feature coding methods under the FK coding framework and the traditional methods based on the BOVW. When compared to the traditional BOVW method, scene classification based on FK-O and FK-S improved the classification accuracy by about 19% for the UCM dataset and about 9%–10% for the Google dataset and the Wuhan IKONOS dataset. In contrast to the SPM-MeanStd method, FK-O and FK-S increased the accuracy by about 6%, 4%, and 2% for the UCM dataset, the Google dataset, and the Wuhan IKONOS dataset, respectively. Compared to the LDA and P-LDA methods, FK-O and FK-S improved the accuracy by more than 9%, 8%, and 5% for the UCM dataset, the Google dataset, and the Wuhan IKONOS dataset, respectively.

Table 1. Classification accuracies (%) of the different methods.

	UCM	Google	Wuhan IKONOS
BOVW	72.05 ± 1.41	81.10 ± 1.37	80.75 ± 5.16
SPM	85.38 ± 1.85	86.31 ± 0.90	88.21 ± 4.29
LDA	81.92 ± 1.12	60.32 ± 1.20	77.34 ± 6.23
P-LDA	81.27 ± 2.01	81.81 ± 1.05	84.69 ± 4.74
FK-Linear	87.70 ± 1.72	87.53 ± 0.51	78.23 ± 4.25
LFK-Linear	88.69 ± 2.01	88.42 ± 0.96	79.69 ± 5.32
FK-O	91.38 ± 1.54	90.16 ± 0.82	89.67 ± 4.19
FK-S	91.63 ± 1.49	90.40 ± 0.84	90.71 ± 4.41

(2) Comparison between before and after considering the spatial information. For all the datasets, the FK-S scene classification method obtained slightly higher classification accuracies than the FK-O scene classification method, which suggests that considering the spatial information during the parameter learning and coding can improve the classification performance.

(3) Comparison between the linear kernel and HIK kernel of SVM. The FK-O (FK-S) scene classification method with HIK kernel increased the accuracy by about 2%, 2%, and 10% when compared to the FK-Linear (LFK-Linear) classification method with linear kernel for the UCM dataset, the Google dataset, and the Wuhan IKONOS dataset, respectively.

(4) Comparison of the codebook size. For FK-O and FK-S, the size of the codebook is the number of Gaussian components K. The sizes of BOVW, FK-O, and FK-S codebooks corresponding to the accuracies in Table 1 are recorded in Table 2, where the codebook sizes of FK-O are 128, 64, and 32, while the codebook sizes of FK-S are 128, 128, and 32 for the UCM dataset, the Google dataset, and the

Wuhan IKONOS dataset, respectively. The codebook size of BOVW is 1000 for all the datasets. By the use of a PC with a 2.5 GHz Intel Core i5-3210M processor, the cost times of the different methods are reported in Table 2, which infers that the cost times of FK-O and FK-S are less than those of BOVW. Table 2 also indicates that the cost times of FK-S are greater than those of FK-O. This evidence infers that scene classification under the FK coding framework can reduce the size of the codebook and the computational cost, to obtain a more compact representation of the scenes.

Table 2. Sizes of codebook and cost times of the different methods.

Datasets	BOVW		FK-O		FK-S	
	Size	Time (s)	Size	Time (s)	Size	Time (s)
UCM	1000	11,544	128	8840	128	9247
Google	1000	6528	64	2942	128	5510
Wuhan	1000	881	32	119	32	309

(5) Comparison with the state-of-the-art. The published classification accuracies of different methods for the UCM dataset are shown in Table 3. Here, it can be seen that the FK-O and FK-S scene classification methods acquired a very competitive accuracy when compared to the state-of-the-art.

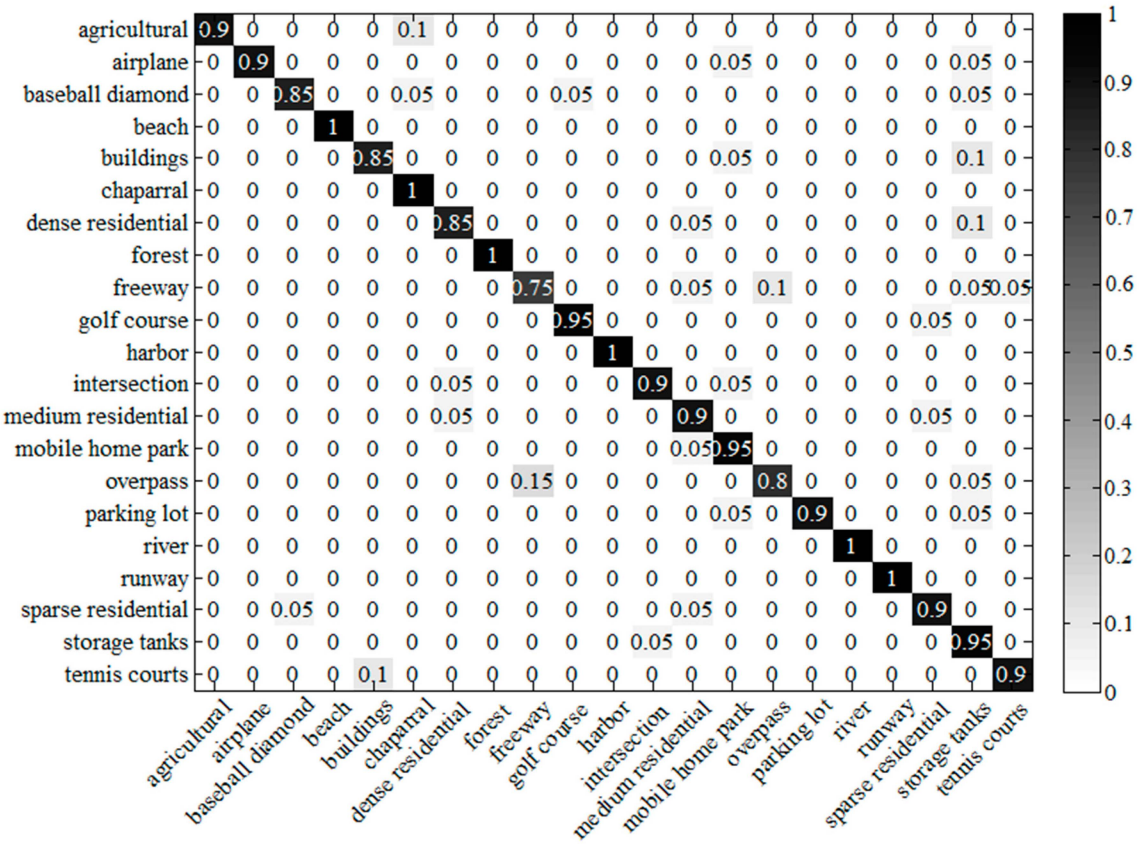
Table 3. Accuracy comparison for the UC Merced (UCM) dataset.

Methods	Accuracy (%)
SCK [33]	73.14
SCK++ [34]	77.38
CCM+BOVW [16]	86.64 $\pm$ 0.81
PSR [15]	89.61
UFL [35]	81.67 $\pm$ 1.23
SG-UFL [22]	82.72 $\pm$ 1.18
UFL-SC with LPP [17]	90.26 $\pm$ 1.51
Partlets-based method [19]	91.33 $\pm$ 0.11
FK-O	91.38 $\pm$ 1.54
FK-S	91.63 $\pm$ 1.49

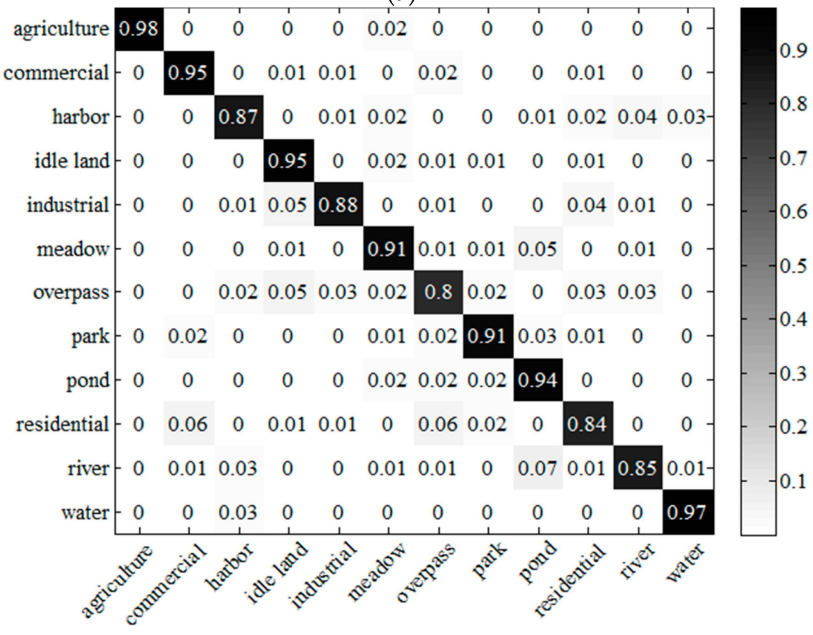
(6) Evaluation of the performance of the scene classification and annotation methods developed under the FK coding framework for each scene class. Taking FK-S as an example, the classification confusion matrices of the three datasets are shown in Figure 11, and the annotated image for the large Wuhan IKONOS image is shown in Figure 9b.

From the confusion matrix of the UCM dataset (Figure 11a), it can be seen that the accuracies of all the scenes, except for the freeway class, are more than 80%, and the relatively low accuracy of the freeway scene is mainly caused by the confusion with the overpass scene. In addition, the confusion levels of the following pairs of scenes exceed 10%: agricultural/chaparral, buildings/storage tanks, and dense residential/storage tanks. For the Google dataset (Figure 11b), the accuracies of all the scenes are more than 80%, and the main confusion occurs in the pairs of scenes of residential/commercial, river/pond, and residential/overpass. For the Wuhan IKONOS dataset (Figure 11c), the accuracies of all the scenes are higher than 80%, except for the commercial scene, and the main confusion occurs between the commercial scene and the medium residential scene. One of the main reasons for the confusion is that some images in these pairs of scenes are very similar in spectral value, and the mean and standard deviation statistics of the spectral values have a limited ability to describe the difference. Therefore, finding a proper feature extractor for the HSR scene classification, or combining different feature extractors with different characteristics, are potential ways to improve the performance. For the annotation experiment, although there is some confusion between industrial, parking lot, commercial,

dense residential, and medium residential, the annotated large image is still satisfactory, based on our remote sensing image analysis expertise.



(**a**)



(**b**)

Figure 11. *Cont.*

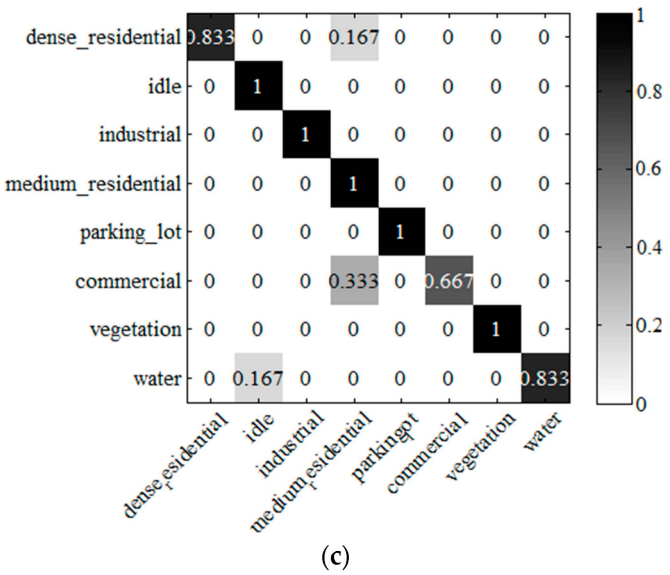


Figure 11. Confusion matrices obtained by the FK-S scene classification method for the three datasets. (**a**) UCM dataset; (**b**) Google dataset; (**c**) Wuhan IKONOS dataset.

5. Discussion

In the FK-O and FK-S scene classification methods, the number of Gaussians K is an important parameter, which is discussed in this section (Figure 12). In addition, the effect of the number of regions M for the FK-S scene classification method is also analyzed (Figure 13).

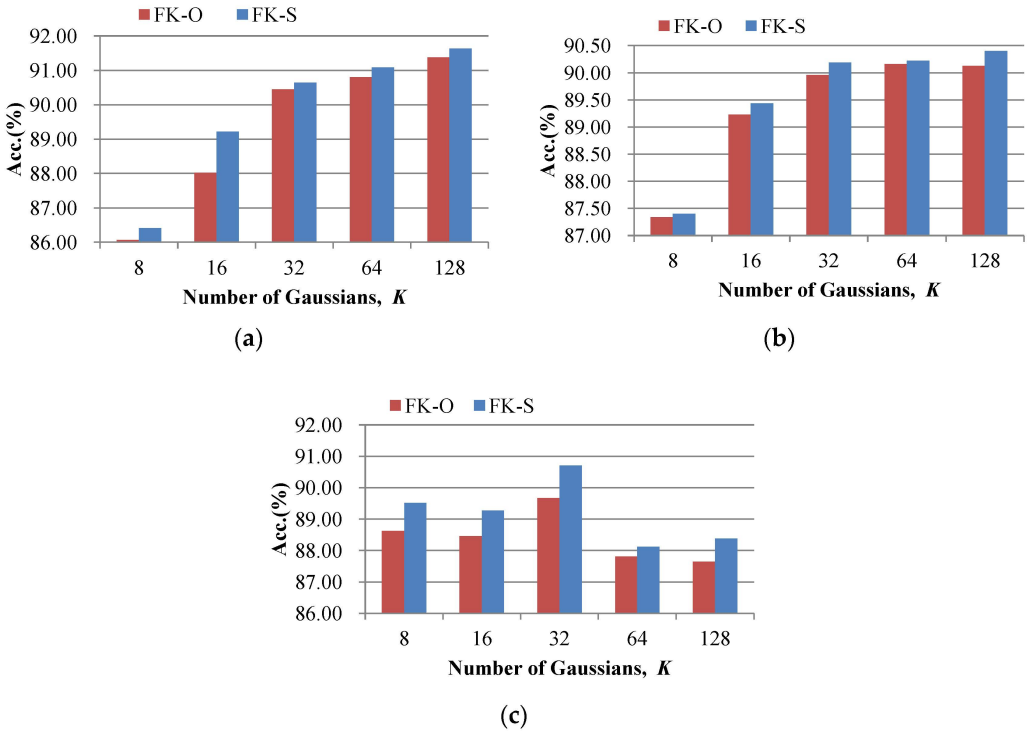


Figure 12. Accuracies of the FK-O and FK-S scene classification methods with different numbers of Gaussians. (**a**) UCM dataset; (**b**) Google dataset; (**c**) Wuhan IKONOS dataset.

(1) The effect of the number of Gaussians K. In the experiments, K was varied between 8, 16, 32, 64, and 128. The accuracies of the FK-O and FK-S scene classification methods with different K values

are shown in Figure 12, where the number of regions was set to nine for FK-S. From Figure 12, it can be seen that the classification accuracies of the FK-O and FK-S scene classification methods increased rapidly with the increase in K from eight to 32, but the magnitude of the increase was small when K was increased from 32 to 128 for the UCM dataset and the Google dataset. For the Wuhan IKONOS dataset, the best performances for the FK-O and FK-S scene classification methods were acquired when K was set to 32, and a smaller or bigger K caused a decrease in the classification accuracy. This is because a small codebook lacks the descriptive ability for the low-level features, while a large codebook contains redundant visual words, which leads to the high dimension of the coding vector ($2KD+M(K-1)$) and high correlation between the components. When compared to the FK-O scene classification method, the FK-S scene classification method obtained higher accuracies.

(2) The effect of the number of regions M for the FK-S scene classification method. In the experiments, M was varied between 4, 9, 16, and 25. The accuracies of the FK-O and FK-S scene classification methods with different M values are shown in Figure 13. In Figure 13, the best accuracies for the FK-S scene classification method were acquired when M was set to nine for all three datasets. A larger number of regions, e.g., M = 16, led to a decrease in the classification accuracy, because there were too many components in the LFK coding vector describing the spatial information. Meanwhile, a smaller number of regions led to a smaller number of spatial components, which resulted in less use of the spatial information during the scene classification.

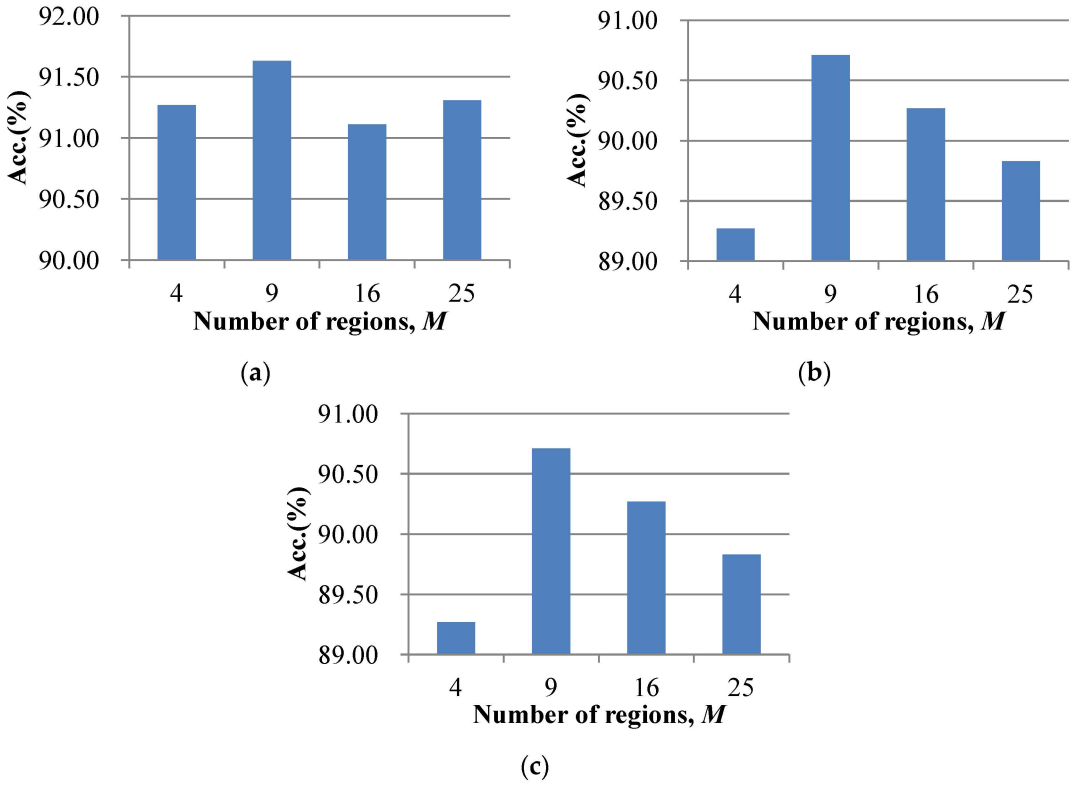


Figure 13. Accuracies of the FK-S scene classification method with different numbers of regions. (**a**) UCM dataset; (**b**) Google dataset; (**c**) Wuhan IKONOS dataset.

6. Conclusions

In order to bridge the semantic gap between the low-level features and high-level semantic concepts for high spatial resolution (HSR) imagery, we introduce a compact representation for HSR scenes under the Fisher kernel (FK) coding framework by coding the low-level features with a gradient vector instead of the count statistics in the BOVW model. Meanwhile, a scene classification method is proposed under the FK coding framework to incorporate the spatial information, where the local

Gaussian mixture model (LGMM) is used to consider the spatial arrangement by estimating the different sets of priors of the Gaussians for the low-level features in different regions, and a local FK (LFK) coding method is developed to deliver the spatial information into the coding vectors. The scene classification methods developed under the FK coding framework, with and without the incorporation of the spatial information, are called FK-S and FK-O, respectively. The experimental results with the UCM dataset, a Google dataset, and an IKONOS dataset infer that the scene classification methods developed under the FK coding framework are able to generate a compact representation for the HSR scenes, and can decrease the size of the codebook. In addition, the experimental results show that the scene classification method incorporating the spatial information, FK-S, can acquire a slightly better performance than the scene classification method that does not consider the spatial information, FK-O. When compared to the published accuracies of the state-of-the art for the UCM dataset, the scene classification methods under the FK coding framework can obtain a very competitive accuracy.

Acknowledgments: The authors would like to thank the editor and the anonymous reviewers for their comments and suggestions. This work was supported by National Natural Science Foundation of China under Grant Nos. 41371344 and 41431175, State Key Laboratory of Earth Surface Processes and Resource Ecology under Grant No. 2015-KF-02, Program for Changjiang Scholars and Innovative Research Team in University under Grant No. IRT1278, Natural Science Foundation of Hubei Province under Grant No. 2015CFA002, and the Hong Kong Research Grants Council under GRF No. 14606315.

Author Contributions: All the authors made significant contributions to the work. Bei Zhao and Yanfei Zhong designed the research and analyzed the results. Liangpei Zhang and Bo Huang provided advice for the preparation and revision of the paper.

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

References

1. Blaschke, T.; Hay, G.J.; Kelly, M.; Lang, S.; Hofmann, P.; Addink, E.; Queiroz Feitosa, R.; van der Meer, F.; van der Werff, H.; van Coillie, F.; *et al*. Geographic object-based image analysis—Towards a new paradigm. *ISPRS J. Photogramm. Remote Sens.* **2014**, *87*, 180–191. [CrossRef] [PubMed]
2. Moser, G.; Serpico, S.B.; Benediktsson, J.A. Land-cover mapping by markov modeling of spatial-contextual information in very-high-resolution remote sensing images. *Proc. IEEE* **2013**, *101*, 631–651. [CrossRef]
3. Zhong, Y.; Zhao, B.; Zhang, L. Multiagent object-based classifier for high spatial resolution imagery. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 841–857. [CrossRef]
4. Zhong, Y.; Zhao, J.; Zhang, L. A hybrid object-oriented conditional random field classification framework for high spatial resolution remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 7023–7037. [CrossRef]
5. Huang, X.; Liu, X.; Zhang, L. A multichannel gray level co-occurrence matrix for multi/hyperspectral image texture representation. *Remote Sens.* **2014**, *6*, 8424–8445. [CrossRef]
6. Li, J.; Bioucas-Dias, J.M.; Plaza, A. Hyperspectral image segmentation using a new bayesian approach with active learning. *IEEE Trans. Geosci. Remote Sens.* **2011**, *49*, 3947–3960. [CrossRef]
7. Bouziani, M.; Goita, K.; Dong-Chen, H. Rule-based classification of a very high resolution image in an urban environment using multispectral segmentation guided by cartographic data. *IEEE Trans. Geosci. Remote Sens.* **2010**, *48*, 3198–3211. [CrossRef]
8. Kim, M.; Warner, T.A.; Madden, M.; Atkinson, D.S. Multi-scale geobia with very high spatial resolution digital aerial imagery: Scale, texture and image objects. *Int. J. Remote Sens.* **2011**, *32*, 2825–2850. [CrossRef]
9. Lienou, M.; Maitre, H.; Datcu, M. Semantic annotation of satellite images using latent dirichlet allocation. *IEEE Geosci. Remote Sens. Lett.* **2010**, *7*, 28–32. [CrossRef]
10. Bahmanyar, R.; Shiyong, C.; Datcu, M. A comparative study of bag-of-words and bag-of-topics models of eo image patches. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 1357–1361. [CrossRef]
11. Zhang, X.; Du, S. A linear dirichlet mixture model for decomposing scenes: Application to analyzing urban functional zonings. *Remote Sens. Environ.* **2015**, *169*, 37–49. [CrossRef]
12. Zhao, B.; Zhong, Y.; Xia, G.S.; Zhang, L. Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2015**. [CrossRef]

13. Weizman, L.; Goldberger, J. Urban-area segmentation using visual words. *IEEE Geosci. Remote Sens. Lett.* **2009**, *6*, 388–392. [CrossRef]
14. Xu, S.; Fang, T.; Li, D.; Wang, S. Object classification of aerial images with bag-of-visual words. *IEEE Geosci. Remote Sens. Lett.* **2010**, *7*, 366–370.
15. Chen, S.; Tian, Y. Pyramid of spatial relatons for scene-level land use classification. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 1947–1957. [CrossRef]
16. Zhao, L.; Tang, P.; Huo, L. Land-use scene classification using a concentric circle-structured multiscale bag-of-visual-words model. *IEEE J. Sel. Topics Appl. Earth Observ.* **2014**, *7*, 4620–4631. [CrossRef]
17. Hu, F.; Xia, G.; Wang, Z.; Huang, X.; Zhang, L.; Sun, H. Unsupervised feature learning via spectral clustering of multidimensional patches for remotely sensed scene classification. *IEEE J. Sel. Topics Appl. Earth Observ.* **2015**, *8*, 2015–2030. [CrossRef]
18. Zhang, Y.; Sun, X.; Wang, H.; Fu, K. High-resolution remote-sensing image classification via an approximate earth mover's distance-based bag-of-features model. *IEEE Geosci. Remote Sens. Lett.* **2013**, *10*, 1055–1059. [CrossRef]
19. Cheng, G.; Han, J.; Guo, L.; Liu, Z.; Bu, S.; Ren, J. Effective and efficient midlevel visual elements-oriented land-use classification using vhr remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 4238–4249. [CrossRef]
20. Cheng, G.; Han, J.; Zhou, P.; Guo, L. Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *ISPRS J. Photogramm. Remote Sens.* **2014**, *98*, 119–132. [CrossRef]
21. Han, J.; Zhang, D.; Cheng, G.; Guo, L.; Ren, J. Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 3325–3337. [CrossRef]
22. Zhang, F.; Du, B.; Zhang, L. Saliency-guided unsupervised feature learning for scene classification. *IEEE Trans. Geosci. Remote Sens.* **2014**, *53*, 2175–2184. [CrossRef]
23. Cheng, G.; Zhou, P.; Han, J.; Guo, L.; Han, J. Auto-encoder-based shared mid-level visual dictionary learning for scene classification using very high resolution remote sensing images. *Comput. Vis. IET* **2015**, *9*, 639–647. [CrossRef]
24. Haralick, R.M.; Shanmugam, K.; Dinstein, I.H. Textural features for image classification. *IEEE Trans. Syst. Man Cybern.* **1973**, *SMC-3*, 610–621. [CrossRef]
25. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
26. Csurka, G.; Dance, C.; Fan, L.; Willamowski, J.; Bray, C. Visual categorization with bags of keypoints. In Proceedings of the Workshop on Statistical Learning in Computer Vision, ECCV, Prague, Czech Republic, 10–16 May 2004; pp. 1–2.
27. Luo, W.; Li, H.; Liu, G.; Zeng, L. Semantic annotation of satellite images using author-genre-topic model. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 1356–1368. [CrossRef]
28. Blei, D. Probabilistic topic models. *Commun. ACM* **2012**, *55*, 77–84. [CrossRef]
29. Zhong, Y.; Zhu, Q.; Zhang, L. Scene classification based on the multifeature fusion probabilistic topic model for high spatial resolution remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 6207–6222. [CrossRef]
30. Zhao, B.; Zhong, Y.; Zhang, L. Scene classification via latent dirichlet allocation using a hybrid generative/discriminative strategy for high spatial resolution remote sensing imagery. *Remote Sens. Lett.* **2013**, *4*, 1204–1213. [CrossRef]
31. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
32. 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.
33. Yang, Y.; Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification. In Proceedings of the ACM SIGSPATIAL Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010; pp. 270–279.
34. Yang, Y.; Newsam, S. Spatial pyramid co-occurrence for image classification. In Proceedings of the IEEE International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 1465–1472.

35. Cheriyadat, A.M. Unsupervised feature learning for aerial scene classification. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 439–451. [CrossRef]
36. 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]
37. 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]
38. 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]
39. Huang, Y.; Wu, Z.; Wang, L.; Tan, T. Feature coding in image classification: A comprehensive study. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *36*, 493–506. [CrossRef] [PubMed]
40. Han, J.; Zhou, P.; Zhang, D.; Cheng, G.; Guo, L.; Liu, Z.; Bu, S.; Wu, J. Efficient, simultaneous detection of multi-class geospatial targets based on visual saliency modeling and discriminative learning of sparse coding. *ISPRS J. Photogramm. Remote Sens.* **2014**, *89*, 37–48. [CrossRef]
41. Perronnin, F.; Sánchez, J.; Mensink, T. Improving the fisher kernel for large-scale image classification. In Proceedings of the European Conference on Computer Vision (ECCV), Crete, Greece, 5–11 November 2010; Daniilidis, K., Maragos, P., Paragios, N., Eds.; Springer: Berlin, Germany; Heidelberg, Germany, 2010; Volume 6314, pp. 143–156.
42. 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.
43. Chang, C.-C.; Lin, C.-J. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 1–27. [CrossRef]
44. Barla, A.; Odone, F.; Verri, A. Histogram intersection kernel for image classification. In Proceedings of the IEEE International Conference on Image Processing, Barcelona, Spain, 14–17 September 2003; pp. 513–516.