sensors

MDPI

Article

Scene-Level Geographic Image Classification Based on a Covariance Descriptor Using Supervised Collaborative Kernel Coding

Chunwei Yang [1,2,] *, Huaping Liu [2], Shicheng Wang [1] and Shouyi Liao [1]

[1] High-Tech Institute of Xi'an, Xi'an 710025, China; wangsching@163.com (S.W.); liaoshouyi123@163.com (S.L.)

[2] Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China; hpliu@tsinghua.edu.cn

* Correspondence: yangchunwei081129@163.com; Tel.: +86-136-9144-4997

Academic Editor: Assefa M. Melesse
Received: 26 October 2015; Accepted: 15 March 2016; Published: 18 March 2016

Abstract: Scene-level geographic image classification has been a very challenging problem and has become a research focus in recent years. This paper develops a supervised collaborative kernel coding method based on a covariance descriptor (covd) for scene-level geographic image classification. First, covd is introduced in the feature extraction process and, then, is transformed to a Euclidean feature by a supervised collaborative kernel coding model. Furthermore, we develop an iterative optimization framework to solve this model. Comprehensive evaluations on public high-resolution aerial image dataset and comparisons with state-of-the-art methods show the superiority and effectiveness of our approach.

Keywords: scene-level geographic image classification; covariance descriptor; collaborative kernel coding

1. Introduction

Nowadays, high spatial resolution remote sensing images are easily acquired thanks to the rapid development of satellite and remote sensing technology, which has endowed us with the opportunity to interpret, analyze and understand the image. As a fundamental research area of remote sensing image analysis, scene-level geographic image classification is of great importance for land use and land cover (LULC) image classification [1–3], semantic interpretations of images [4], geographic image retrieval [5–7] and forest type mapping [8], which has drawn increasing attention and scholars' study [1–3,5,9–13]. Figure 1 shows geographic images whose spatial resolution is 30 m, 1 m and 0.3 m, respectively.

However, finding an efficient representation of the scene-level image is a challenging problem. The bag of visual words (BOVW) model [14] is one of the most successful models. The works in [2,5] detailed the application of BOVW on the scene-level image classification task. As is illustrated in [2,5], BOVW can represent the image by compact representation through a visual word counts histogram and provides further invariance to the image transformations. However, the tradeoff between invariance and discriminability is controlled by the visual dictionary size. What is more, BOVW disregards the information about the spatial layout of the features, which is of great importance to scene-level image classification [2,15,16]. In order to overcome this shortcoming, one successful extension of BOVW is spatial pyramid matching (SPM) [16], which partitions the image into increasing finer sub-images and computes histograms of local features from each sub-image. Although SPM is a computationally-efficient extension of BOVW and shows superior performance, it does not consider the relative spatial arrangement and only characterizes the absolute location of the visual words in an image. From this point of view, SPM also limits the descriptive ability of the scene-level geographic image representation. Hence, two new image representation models, which are termed spatial co-occurrence kernel (SCK) [1] and spatial pyramid co-occurrence kernel (SPCK) [2], are proposed by Yang and Newsam. What is more, in order to capture the absolute and relative spatial relationships of BOVW, a pyramid of spatial relations (PSR) model is developed by Chen and Tian. The work in [17] points out that the computational complexities of SCK and SPCK are high because of the need to use nonlinear Mercer kernels and developed a linear form of the SCK. Besides, [10] proposed an unsupervised feature learning method, in which the new sparse representations of the feature descriptors are generated by the low-level feature descriptors.

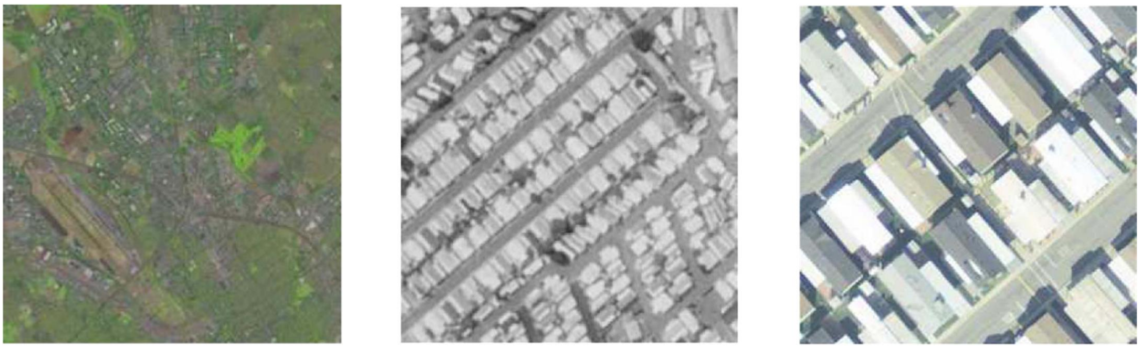

(a) (b) (d)

Figure 1. Images with a resolution of: (**a**) 30 m; (**b**) 1 m; (**c**) 0.3 m.

On the other hand, the covariance descriptor (covd) proposed by by Tuzel [18] can be used for feature representation of the image, which has been extensively adopted in vast computer vision tasks, e.g., texture discrimination [18], visual saliency estimation [19], object detection [18,20] and object tracking [21]. Covd is a covariance matrix of different features, e.g., color, gradient and spatial location, and it holds certain rotation and scale invariance. However, how to model and compute covd still remains a key problem. We all know that covd lies in the Riemannian manifold, which is a non-Euclidean space. As a result, traditional mathematical modeling and computation in Euclidean space cannot be directly utilized, which results in a great challenge. In [22], a discriminative learning method is developed to formulate the classification problem on Riemannian space by covd, which presents a kernel function and a log-Euclidean distance metric to solve Riemannian-Euclidean transformation. In [23], a coding strategy is introduced, and the descriptor can be transformed into a new feature; and then, extreme learning machine (ELM) can be used for dynamic texture video classification. However, such a method separately optimizes the reconstruction error of the coding and the classification error of ELM, and the design stage of coding and the classifier are totally independent. In order to solve this problem, a supervised collaborative kernel coding approach incorporating the linear classifier supervised term that can optimize both the reconstruction error and the linear classifier simultaneously is developed. There are three contributions as follows:

1. A supervised collaborative kernel coding model, illustrated in Figure 2, is proposed. This model can not only transform the covd to a discriminative feature representation, but also can obtain the corresponding linear classifier.
2. An iterative optimization framework is introduced to solve the supervised collaborative kernel coding model.
3. Experiments on public high-resolution aerial image dataset validate that the proposed supervised collaborative kernel coding model derives a satisfying performance on the scene-level geographic image classification.

The paper is organized as follows: After a review of our proposed methodology in Section 2, Section 3 shows the iterative optimization approach. In Sections 4 and 5, we give the experiments and conclusions.

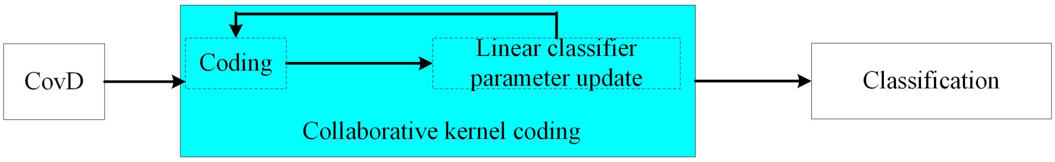

CovD
Coding
Linear classifier parameter update
Collaborative kernel coding
Classification

Figure 2. Illustration of the supervised collaborative kernel coding model.

2. Overview of the Methodology

Figure 3 shows the overview of the proposed method, which consists of 3 stages, the pre-processing stage, coding stage and classification stage. In the pre-processing stage, covd is extracted as the initial feature representation of the scene-level geographic image. Then, in the coding stage, the supervised collaborative kernel coding strategy involving dictionary coefficients, the coding representation phase and the linear classification phase is presented. Finally, in the classification stage, based on the dictionary coefficients and learned linear classifier, a label vector can be simply derived through the linear classifier, the index corresponding to the largest value of which is the label of a testing scene-level geographic image.

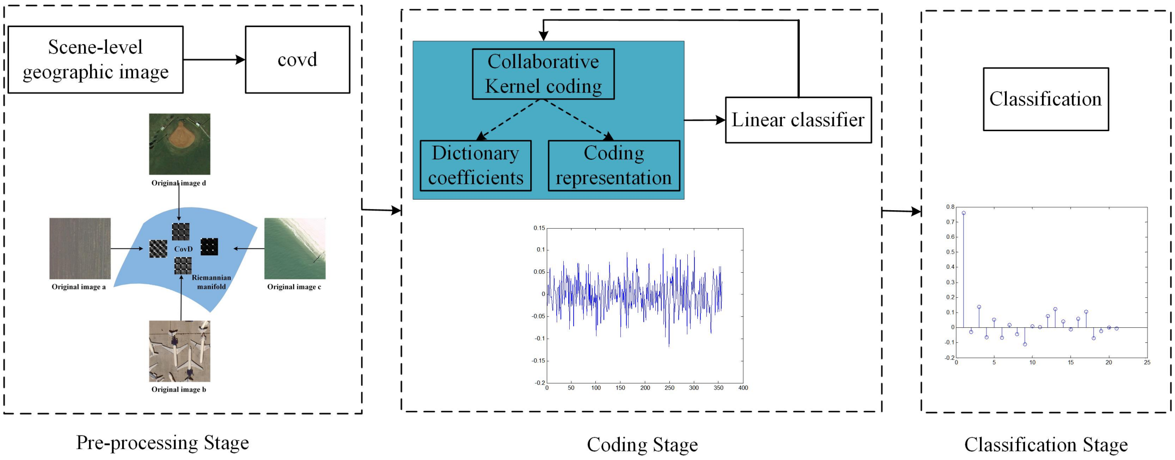


Figure 3. The overview of the proposed method. covd, covariance descriptor.

2.1. Covariance Descriptor

Covd was first proposed by Tuzel *et al.* [18] as a compact descriptor. Formally, let $\{\mathbf{f}_k\}_{k=1,\cdots,d}$ be a feature vector denoting the feature points of p-dimension as color, gradient filter response, *etc.* Then, a covd $\mathbf{C}$ of $s \times s$ dimensions of an image can be described as:

$$\mathbf{C} = \frac{1}{d-1}\sum_{k=1}^{d}(\mathbf{f}_k - \mathbf{v})(\mathbf{f}_k - \mathbf{v})^T \tag{1}$$

where d and $\mathbf{v}$ denote the pixel number and the mean value, respectively.

The feature vector $\mathbf{f}$ is established using the image intensity of each channel, the norm of the first and second derivatives of intensity in the x and y directions. As for a geographic image, a feature vector $\mathbf{f}_{x,y} = [\mathbf{c}_{R,x,y}^T, \mathbf{c}_{G,x,y}^T, \mathbf{c}_{B,x,y}^T]^T$ of 15 dimensions is computed at each pixel (x, y), and here,

$\mathbf{c}_{C,x,y} = [I_{C,x,y}, |\frac{\partial I_C}{\partial x}|, |\frac{\partial^2 I_C}{\partial^2 x}|, |\frac{\partial I_C}{\partial y}|, |\frac{\partial^2 I_C}{\partial^2 y}|]$, where I_C and $C \in \{R, G, B\}$ denote the the C channel intensity image and the channel of the color, respectively.

The work in [18] points out that covd has at least three characteristics: (1) it is enough to describe the image of different poses and views; (2) multiple features can be fused in a natural way through covd, the diagonal and non-diagonal elements of which describe the variance and correlations of different features, respectively; (3) comparing to other descriptors, such as raw values and the histogram, covd is low-dimensional, and it has only $\frac{s^2+s}{2}$ different values due to symmetry.

Nevertheless, covd is a symmetric positive definite matrix. The key issue for a symmetric positive definite matrix is how to model and compute it. As is illustrated in Figure 4, covd lies in a Riemannian manifold [24], which is not a Euclidean space.

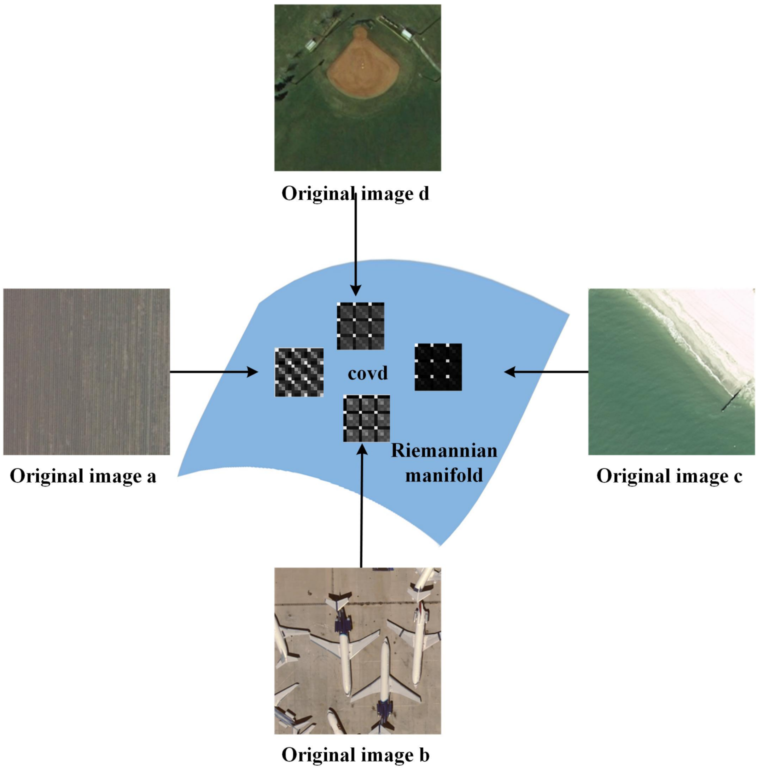


Figure 4. Sample geographic images and corresponding covariance descriptor (covd) features.

Accordingly, the mathematical modeling of covd is not the same as what we usually do in the Euclidean space. Here, we adopt the idea of Ruiping Wang [22] and compute the distance of two covds $\mathbf{C}_1$ and $\mathbf{C}_2$ using log-Euclidean distance [25,26]:

$$d(\mathbf{C}_1, \mathbf{C}_2) = ||\text{logm}(\mathbf{C}_1) - \text{logm}(\mathbf{C}_2)||_F \tag{2}$$

where logm is the logarithm computation of the matrix and $|| \cdot ||_F$ denotes the Frobenius norm.

Moreover, there is a tricky problem regarding how to use covd in the geographic image classification. It is a fact that covd lies in a non-Euclidean space; thus, the traditional linear classifier based on Euclidean space cannot be directly utilized. Therefore, in the following, how to solve this problem is the theme.

2.2. Supervised Collaborative Kernel Coding Model

As is shown in Figure 5, here, we propose a supervised collaborative kernel coding model, which consists of two jointly working components: (1) the dictionary learning and feature representation phase; and (2) the linear classification phase. First, the linear classifier is incorporated into the dictionary learning and feature representation phase, making the resulting coding vector **A** more discriminative. Then, based on the coding vector **A**, the linear classifier **W** is derived. In this way, the objectives function in each phase are combined into a unified optimization framework, through which a collaborative coding vector and the corresponding linear classifier can be simultaneously obtained. At last, based on the dictionary coefficients **V**, testing signal $\mathbf{s}_i$ is transformed into a feature vector, which is used for linear classification directly.

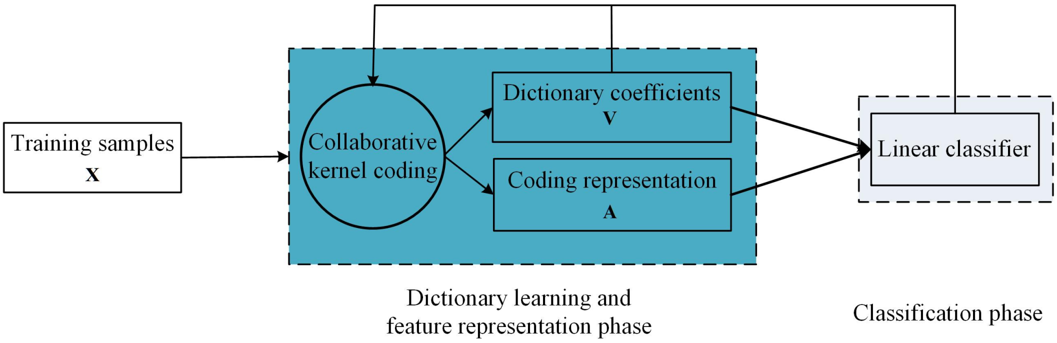


Figure 5. The illustration of the proposed model.

Denote $\{\mathbf{x}_i\}_{i=1}^N \in \mathbf{H}$ as the training samples, where **H** is a Riemannian manifold. Through the proper mapping function, $\{\mathbf{x}_i\}_{i=1}^N$ are mapped into a higher dimensional space. Namely, let $\Phi(\cdot) : \mathbf{H} \to \mathbf{P}$ be the nonlinear mapping process from the original space **H** into a high or infinite dimensional space **P**. For convenience, the dimension of **P** is denoted as $\widetilde{m}$. The mapping function here is associated with a kernel $\kappa(\mathbf{x}_i, \mathbf{x}_j) = < \Phi^T(\mathbf{x}_i), \Phi(\mathbf{x}_j) >$, where $\mathbf{x}_i, \mathbf{x}_j \in \mathbf{H}$. As for covd computation, the Gaussian kernel is chosen as the mapping function for its superior performance in vast computer vision tasks [27]:

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\beta||\text{logm}(\mathbf{x}_i) - \text{logm}(\mathbf{x}_j)||^2) \tag{3}$$

where the decay parameter β is empirically set as 0.02 and $\kappa(\mathbf{x}_i, \mathbf{x}_j)$ is the Gaussian kernel between two samples $\mathbf{x}_i$ and $\mathbf{x}_j$.

The aim of dictionary learning is to empirically learn a dictionary adapted to the training sample set; therefore, we need to determine some atoms $\mathbf{d}_1, \cdots, \mathbf{d}_K \in \mathbf{P}$ to represent the training samples, where K is the dictionary size and $K < N$. Let $\Phi(X) = [\Phi(\mathbf{x})_1, \cdots, \Phi(\mathbf{x})_N] \in \mathbf{R}^{\widetilde{m} \times K}$, and the kernel dictionary learning process can be formulated as:

$$\min_{\mathbf{D},\mathbf{A}} ||\Phi(X) - \Phi(D)\mathbf{A}||_2^2 + \lambda||\mathbf{A}||_2^2 \tag{4}$$

where $\mathbf{A} \in R^{K \times N}$ is the coding matrix and λ is the penalty parameter.

Thanks to the kernel trick [28,29], through the mapping function $\Phi(\cdot)$, the problem on the Riemannian manifold can be transformed to a collaborative coding problem in the Euclidean space. Nevertheless, since the number of dictionary atoms $\mathbf{d}_j$ may be infinite, there exists a new challenge to the dictionary learning process in such a formulation. Fortunately, [30,31] prove that the dictionary **D** can be represented as $\mathbf{D} = \Phi(X)V$, where $\mathbf{V} \in R^{N \times K}$ is a coefficient matrix. This indicates that the training samples can linearly represent the dictionary in the feature space. As a result, Equation (4) can be reformulated as:

$$\min_{\mathbf{V},\mathbf{A}} ||\Phi(X) - \Phi(X)\mathbf{VA}||_2^2 + \lambda||\mathbf{A}||_2^2 \tag{5}$$

Such a formulation provides two significant advantages: (1) the dictionary learning process becomes searching the matrix **V**; (2) for any kernel function, this formulation reduces the dictionary learning process to linear problems.

Now, we propose a novel objective function combining both the collaborative kernel coding phase and classification phase as:

$$\min_{\mathbf{V},\mathbf{A},\mathbf{W}} ||\Phi(X) - \Phi(X)\mathbf{VA}||_2^2 + \lambda||\mathbf{A}||_2^2 + \eta||\mathbf{L} - \mathbf{WA}||_2^2 + \rho||\mathbf{W}||_2^2 \tag{6}$$

where $||\Phi(X) - \Phi(X)\mathbf{VA}||_2^2$ and $||\mathbf{L} - \mathbf{WA}||_2^2$ denotes the reconstruction error and the linear classification error, respectively, and **W** represents the classifier parameters. η, λ and ρ are all penalty parameters.

The derived dictionary through this formulation can generate more discriminative codes **A**, which is of great importance to the performance of the classifier and also adaptive to the underlying structure of training samples. The resulting codes **A** are then directly used for classification.

For a testing sample $\mathbf{s}_i$, through Equation (7), the feature representation code $\mathbf{z}_i$ is firstly computed with dictionary coefficients **V**. Then, in order to derive the label vector, we can use $\mathbf{l}_i = \mathbf{W}\mathbf{z}_i$. The index corresponding to the largest value of $\mathbf{l}_i$ is the label of $\mathbf{s}_i$.

$$\min_{\mathbf{z}_i} ||\Phi(s) - \Phi(X)\mathbf{V}\mathbf{z}_i||_2^2 + \lambda||\mathbf{z}_i||_2^2 \tag{7}$$

3. Optimization Algorithm

There are three variables as **V**, **A** and **W** in the objective function Equation (6). Here, an iterative optimization algorithm for each variable by fixing the other two is introduced. (Equation (6) is denoted as $\mathbf{F}(\mathbf{V},\mathbf{A},\mathbf{W})$, and the obtained variables from the k-th and $(k+1)$-th iteration are denoted as the subscripts (k) and $(k+1)$, respectively, and $k = 0, \cdots, N-1$).

Step 1: Initialization. We randomly set coefficient matrix $\mathbf{V}_0 \in R^{N\times K}$. Next, we compute the corresponding coding coefficient **A** by taking the derivative of **A** of Equation (6):

$$\mathbf{A}_0 = (\mathbf{V}^T\mathbf{K}(\mathbf{X},\mathbf{X})\mathbf{V} + \lambda\mathbf{I})^{-1}\mathbf{V}^T\mathbf{K}^T(\mathbf{X},\mathbf{X})) \tag{8}$$

where $\mathbf{K}(\mathbf{X},\mathbf{X})$ is an $N \times N$ square matrix of which the (i,j)-th element is $\boldsymbol{\kappa}(\mathbf{x}_i,\mathbf{x}_j)$.

Step 2: Fixing **A**, taking the derivative of **V**:

$$\frac{\partial\mathbf{F}(\mathbf{V},\mathbf{A}_{(k)},\mathbf{W}_{(k)})}{\partial\mathbf{V}} = 0 \tag{9}$$

Additionally, the corresponding solution is:

$$\mathbf{V}_{(k+1)} = \mathbf{A}_{(k)}^T(\mathbf{A}_{(k)}\mathbf{A}_{(k)}^T)^{-1} \tag{10}$$

Step 3: Fixing **V** and **A**, taking the derivative of **W**, we can derive the optimal solution of **W**.

$$\frac{\partial\mathbf{F}(\mathbf{V}_{(k+1)},\mathbf{A}_{(k+1)},\mathbf{W})}{\partial\mathbf{W}} = 0 \tag{11}$$

$$\mathbf{W}_{(k+1)} = (\eta\mathbf{A}_{(k+1)}\mathbf{A}_{(k+1)}^T + \rho\mathbf{I})^{-1}\eta\mathbf{I}\mathbf{A}_{(k+1)}^T \tag{12}$$

Step 4: Fixing $\mathbf{V}$ and $\mathbf{W}$, and taking the derivative of $\mathbf{A}$:

$$\frac{\partial \mathbf{F}(\mathbf{V}_{(k+1)}, \mathbf{A}, \mathbf{W}_{(k)})}{\partial \mathbf{A}} = 0 \tag{13}$$

Then, the optimal solution of $\mathbf{A}$ is:

$$\mathbf{A}_{(k+1)} = (\mathbf{V}^T_{(k+1)}\mathbf{K}(\mathbf{X},\mathbf{X})\mathbf{V}_{(k+1)} + \lambda\mathbf{I} + \eta\mathbf{W}^T_{(k)}\mathbf{W}_{(k)})^{-1}(\mathbf{V}^T_{(k+1)}\mathbf{K}(\mathbf{X},\mathbf{X}) + \eta\mathbf{W}^T_{(k)}\mathbf{l}) \tag{14}$$

Step 5: Iteration from Step 2 to Step 4 until convergence.

A whole algorithm summary, which includes the above optimization procedures, is given in Algorithm 1, and the representative reconstruction error of the objective function is shown in Figure 6. In case of the optimal $\mathbf{A}$, we can derive the optimal solution of $\mathbf{z}$ based on Equation (7) as:

$$\mathbf{z}_i = (\mathbf{V}^T\mathbf{K}(\mathbf{X},\mathbf{X})\mathbf{V} + \lambda\mathbf{I})^{-1}\mathbf{V}^T\mathbf{K}(\mathbf{s}_i,\mathbf{X}) \tag{15}$$

where $\mathbf{K}(\mathbf{s}_i,\mathbf{X}) = [\boldsymbol{\kappa}(\mathbf{s}_i,\mathbf{x}_i),\cdots,\boldsymbol{\kappa}(\mathbf{s}_i,\mathbf{x}_N)]$.

Algorithm 1. The Iteration Optimization Procedure.

```
Input:
$\mathbf{K}(\mathbf{Y},\mathbf{Y}) \in R^{N\times N}$
Output:
$\mathbf{V} \in R^{N\times K}$, $\mathbf{A} \in R^{K\times N}$, $\mathbf{W} \in R^{L\times m}$
1. Initialization: Randomly set $\mathbf{V}_{(0)}$ with appropriate dimensions
and obtain initial $\mathbf{A}$ according to Equation (8).
2. while Not convergent
do
3. Fixing $\mathbf{A}_{(k)}$, update $\mathbf{V}_{(k+1)}$ according to Equation (10)
4. Fixing $\mathbf{V}_{(k+1)}$ and $\mathbf{A}_{(k)}$, update $\mathbf{W}_{(k+1)}$ according to Equation (12)
5. Fixing $\mathbf{V}_{(k+1)}$ and $\mathbf{W}_{(k+1)}$, update $\mathbf{A}_{(k+1)}$ according to Equation (14)
6. end while
```

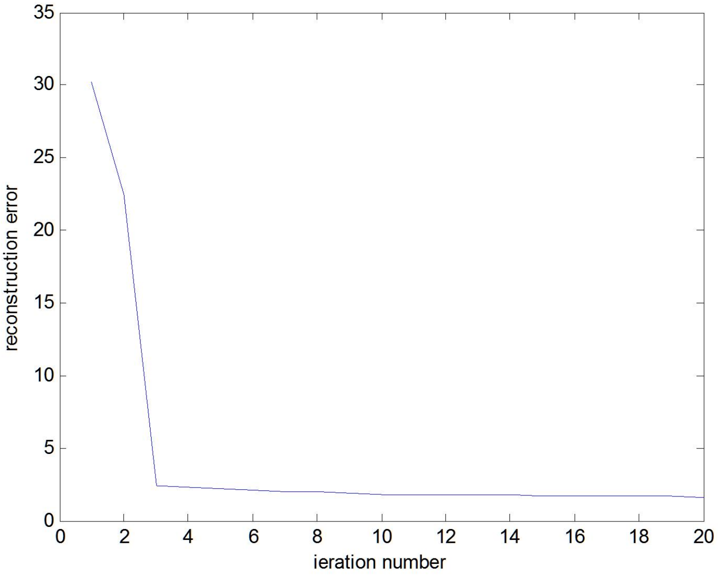


Figure 6. The representative reconstruction error of the objective function.

4. Experiments

4.1. Dataset and Experiment Setup

In this section we demonstrate the application of our method in the classification experiments using a publicly available dataset [1], which includes twenty one scene categories with one hundred images of each class. This dataset corresponds to various land LULC types, which is shown in Figure 7.

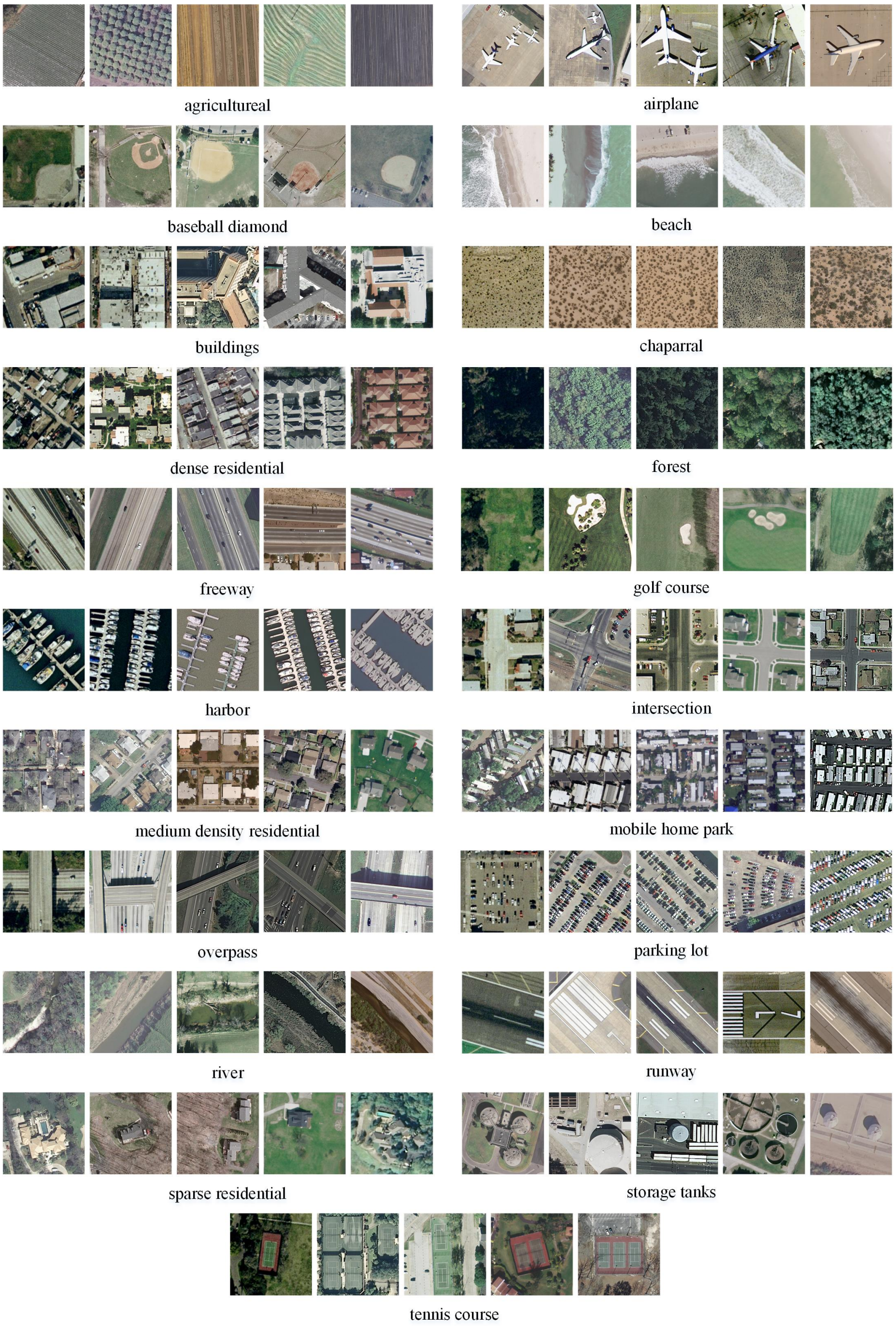


Figure 7. Samples from UCMERCED. Example geographic images associated with 21 categories are shown here.

For each category, it is randomly partitioned into five subsets, and each subset contains twenty samples. During the experiments, one subset is used for testing, and the remaining four subsets are used for training. Finally, we report the average classification accuracy.

4.2. Parameter Analysis

Equation (6) has four parameters, λ, η, ρ and dictionary size K, which need to be tuned. In order to determine their values, n-fold cross-validation is adopted. Each parameter is investigated by fixing the other parameters. It is noted that the initialization of K is 210.

Figure 8 shows the classification accuracy of each tuned parameter. It is easy to find that our approach obtains the best performance (83.81%) when $\lambda = 0.001$, $\eta = 1$, $\rho = 0.1$ (or 1).

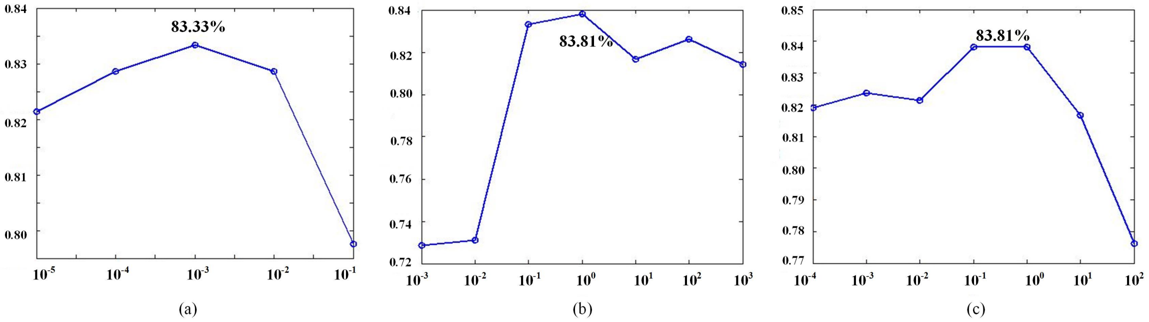


Figure 8. Evaluation of the effect on the classification accuracy for parameters: (**a**) λ; (**b**) η; and (**c**) ρ.

4.3. Experiment Results and Comparison

The following three baseline methods are designed for comparison:

1. This method isolates the feature representation and classification process, which means that $\mathbf{A}_0$ is used as the feature representation and that $\mathbf{W}_0$ is used as the linear classifier.
2. This method is the same as the proposed method, except that the covd is established based on image intensities and the magnitude of the first and second gradients. Namely, $\mathbf{f}_{x,y} = [\mathbf{c}_{R,x,y}^T, \mathbf{c}_{G,x,y}^T, \mathbf{c}_{B,x,y}^T]^T$ and $\mathbf{c}_{C,x,y} = [I_{C,x,y}, \sqrt{(\frac{\partial I_C}{\partial x})^2 + (\frac{\partial I_C}{\partial y})^2}, \sqrt{(\frac{\partial^2 I_C}{\partial^2 x})^2 + (\frac{\partial^2 I_C}{\partial^2 y})^2}]$.
3. This method is the same as baseline Method 1, except that the covd is a 9 × 9 matrix, which is the same as baseline Method 2.

Figure 9 shows the classification accuracy *versus* dictionary size K. From this figure, we can find some interesting results:

1. Our approach is always better than the three baseline methods, and when $K = 357$, our approach obtains the best performance (87.14%).
2. Comparing to baseline Method 1, our proposed method obtains a higher classification accuracy, which indicates the effectiveness of the optimization algorithm.
3. Comparing the proposed method to baseline Method 2, the only difference is the covd. The former uses a 15 × 15 matrix, which is a covariance format of intensity of each channel and the norms of the first and second gradients of intensities, while the covd of the latter is a 9 × 9 covariance format of the intensity of each channel and the magnitude of the first and second gradients. It is clear that both covds are not rotationally invariant, especially that the former covd is not direction invariant. However, the proposed method obtains a higher classification accuracy. This may indicate that the covariance format offsets the rotations to some extent.

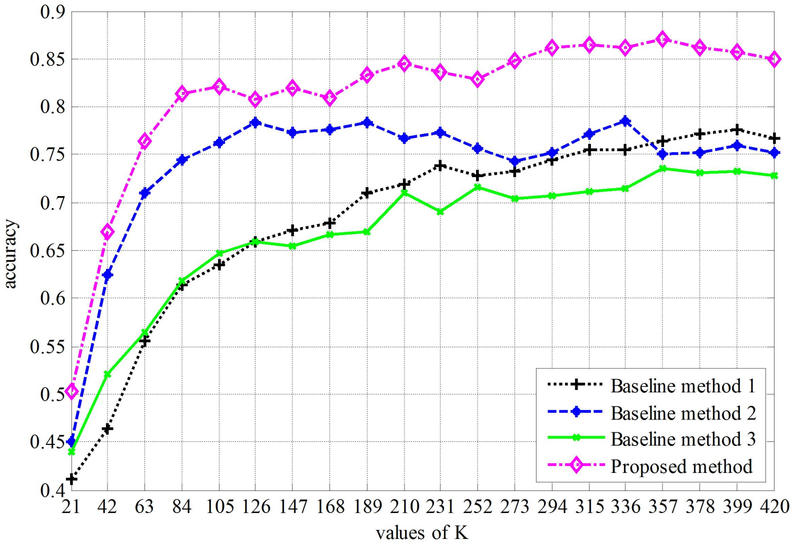
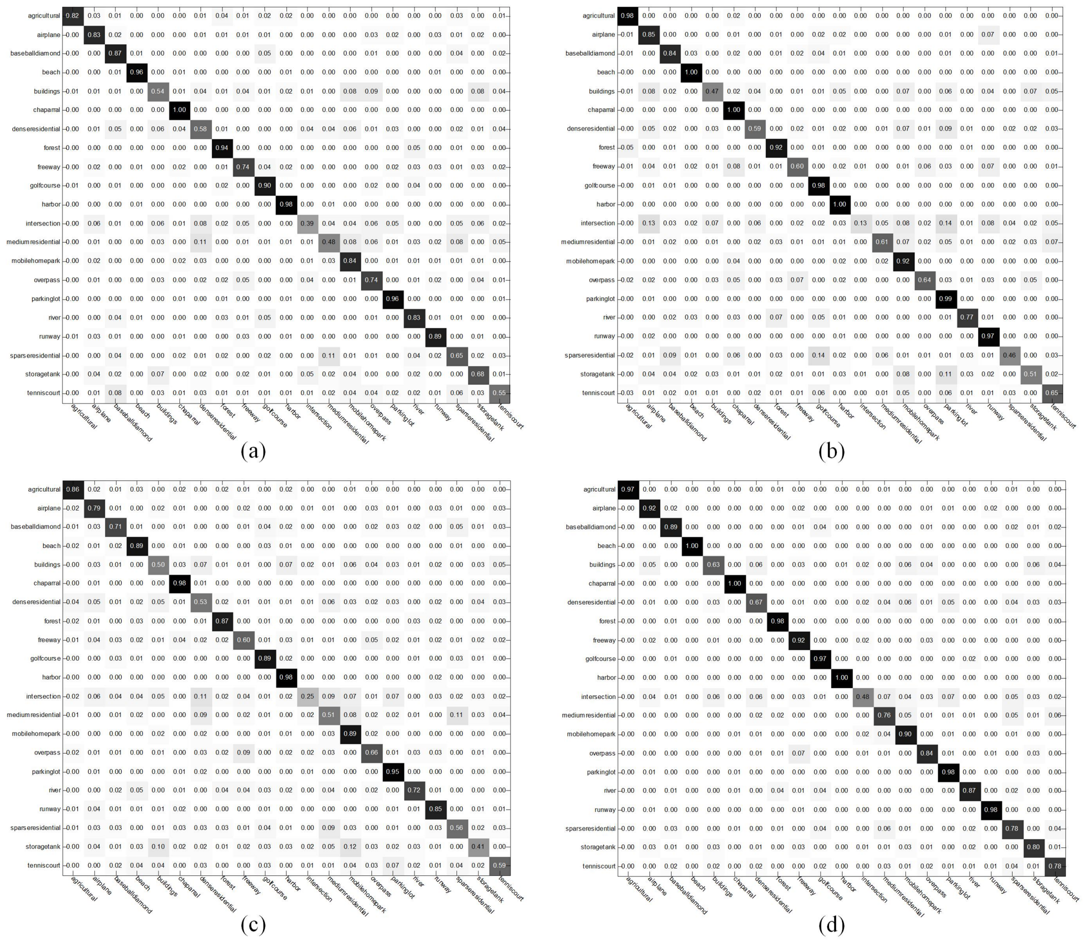


Figure 9. Comparison of different methods.



(a)

(b)

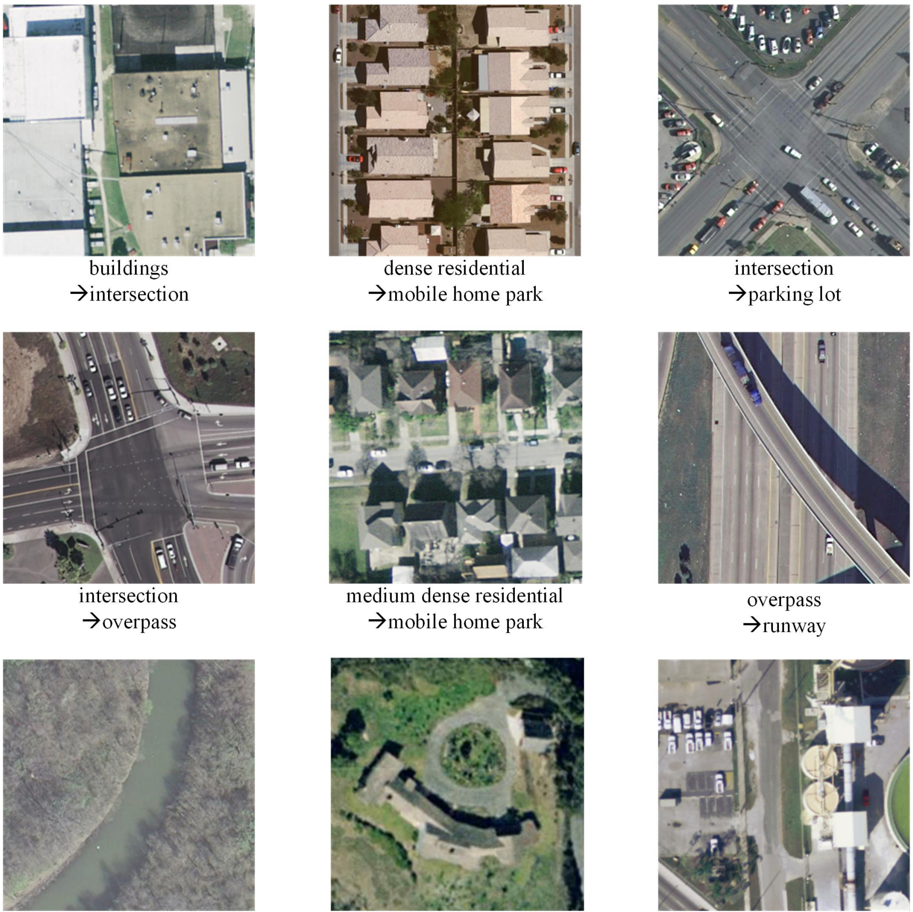
(c)

(d)

Figure 10. The average confusion matrices of: (**a**) baseline Method 1; (**b**) baseline Method 2; (**c**) baseline Method 3; and (**d**) the proposed method.

Figure 10 shows the confusion matrices of the baseline methods and our approach, respectively. The classification accuracy of fifteen categories is more than 80%, and eleven categories are more than 90%. Nevertheless, the classification accuracy of three categories, buildings, dense residential and intersection, is less than 70%.

In order to analyze the proposed method, Figure 11 lists some representative misclassification samples of the proposed method. Some misclassification couples, such as intersection/overpass, overpass/runway and river/forest, shown in Figure 11 are hard to identify, even with our own eyes.



buildings
→intersection

dense residential
→mobile home park

intersection
→parking lot

intersection
→overpass

medium dense residential
→mobile home park

overpass
→runway

river
→forest

sparse residential
→golf course

storage tank
→mobile home park

Figure 11. The representative misclassification samples.

Besides, we report the classification accuracies of both baseline methods and our method over groups rotated five times in Table 1. Then, the comparison with the classical approaches [1,2], BOVW, SPM, SCK, BOVW + SCK, color histogram, such as RGB, HLS and CIE Lab, texture, SPCK, BOVW + SPCK and SPCK + SPM, is shown in Figure 12.

Table 1. classification accuracies over all five groups of our method.

Subset Number	1	2	3	4	5	Average
baseline Method 1	78.10%	79.29%	76.90%	78.81%	71.90%	77.00%
baseline Method 2	78.33%	75.48%	74.29%	75.71%	74.29%	75.62%
baseline Method 3	73.10%	71.43%	70.00%	73.33%	69.05%	71.38%
proposed Method	**87.14%**	**84.52%**	**88.10%**	**87.14%**	**84.52%**	**86.28%**

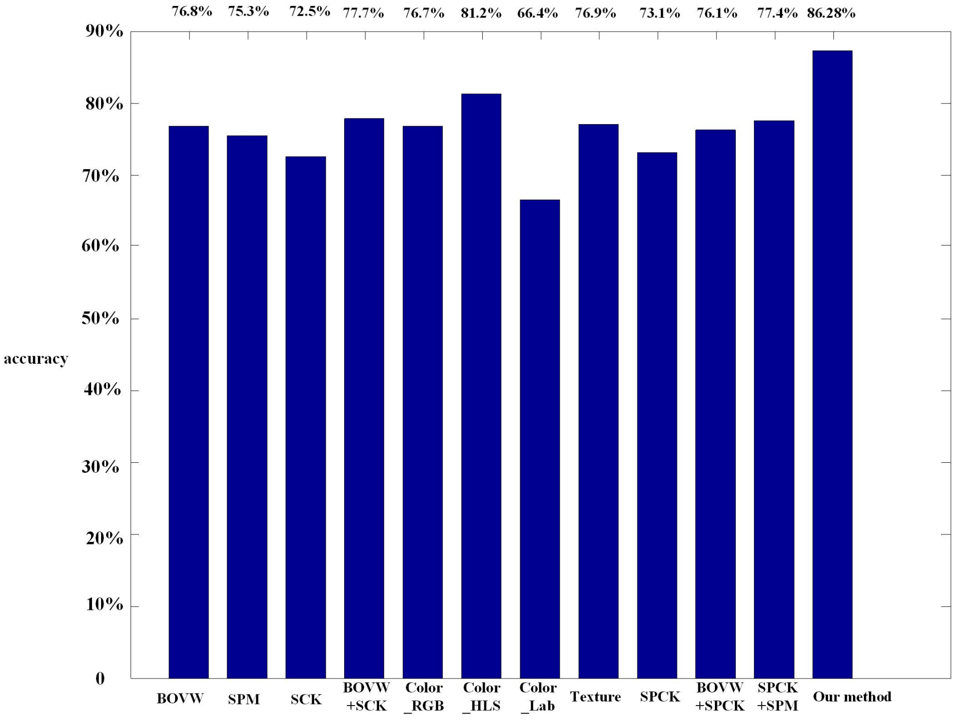


Figure 12. The comparison of our method with the state-of-the-art performance reported in the literature on the dataset UCMERCED. BOVW, bag of visual words; SPM, spatial pyramid matching; SCK, spatial co-occurrence kernel; SPCK, spatial pyramid co-occurrence kernel.

5. Conclusions

This paper proposes a novel supervised collaborative kernel coding model based on covd for scene-level geographic image classification. Since covd lies in non-Euclidean space, the linear classifier, which is based on Euclidean distance, cannot be utilized. Additionally, our main contribution is explicitly integrating the discriminative feature coding and a linear classifier into the objective function. Moreover, the solution to the new objective function is efficiently achieved by simply employing the optimization algorithm. Experiments implemented on the UCMERCED dataset show the effectiveness of our approach.

Acknowledgments: This study was jointly supported by National Key Project for Basic Research of China under Grant 2013CB329403, National Natural Science Foundationof China under Grant 61327809, and National High-tech Research and Development Plan under Grant 2015AA042306.

Author Contributions: Chunwei Yang initiated the research and designed the experiments; Huaping Liu performed the experiments; Shicheng Wang analyzed the data; Shouyi Liao wrote the paper.

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

References

1. Yang, Y.; Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification. In Proceedings of the 18th SIGSPATIAL International Conference Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010; pp. 270–279.
2. Yang, Y.; Newsam, S. Spatial pyramid co-occurrence for image classification. In Proceedings of the 7th IEEE International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 1465–1472.
3. Xu, S.; Fang, T.; Wang, S. Object classification of aerial images with bag-of-visual words. *IEEE Geosci. Remote Sens. Lett.* **2010**, *7*, 366–370.
4. Aksoy, S.; Koperski, K.; Tusk, C.; Marchisio, G.; Tilton, J.C. Learning bayesian classifiers for scene classification with a visual grammar. *IEEE Trans. Geosci. Remote Sens.* **2005**, *43*, 581–589.
5. Yang, Y.; Newsam, S. Geographic image retrieval using local invariant features. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 818–832.
6. Schroder, M.; Rehrauer, H.; Seidel, K.; Datcu, M. Interactive learning and probabilistic retrieval in remote sensing image archives. *IEEE Trans. Geosci. Remote Sens.* **2000**, *38*, 2288–2298.
7. Shyu, C.; Klaric, M.; Scott, G.J.; Barb, A.S.; Davis, C.H.; Palaniappan, K. GeoIRIS: Geospatial information retrieval and indexing system-content mining, semantics modeling and complex queries. *IEEE Trans. Geosci. Remote Sens.* **2000**, *45*, 839–852.
8. Kim, M.; Madden, M.; Warner, T.A. Forest type mapping using object-specific texture measures from multispectral Ikonos imagery: Segmentation quality and image classification issues. *Photogramm. Eng. Remote Sens.* **2000**, *75*, 819–829.
9. Zhang, Y.; Wu, L.; Neggaz, N.; Wang, S.; Wei, G. Remote-sensing image classification based on an improved probabilistic neural network. *Sensors* **2009**, *9*, 7516–7539.
10. Cheriyadat, A. Unsupervised feature learning for aerial scene classification. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 439–451.
11. Du, P.; Xia, J.; Zhang, W.; Tan, K.; Liu, Y.; Liu, S. Multiple classifier system for remote sensing image classification: A review. *Sensors* **2012**, *12*, 4764–4792.
12. Cheng, G.; Han, J.; Zhou, P.; Guo, L. Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *ISPRC J. Photogramm. Remote Sens.* **2014**, *98*, 119–132.
13. Li, J.; Du, Q.; Li, W.; Li, Y. Optimizing extreme learning machine for hyperspectral image classification. *J. Appl. Remote Sens.* **2015**, *8*, 097296:1–097296:13.
14. Csurka, G.; Dance, C.; Fan, L.; Willamowski, J.; Bray, C. Visual categorization with bags of keypoints. In Proceedings of the ECCV International Workshop on Statistical Learning in Computer Vision, Prague, Czech Republic, 11–14 May, 2004.
15. Cao, Y.; Wang, C.; Li, Z.; Zhang, L.Q.; Zhang, L. Spatial-bag-of-features. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 3352–3359.
16. Lazebnik, S.; Schmid, C.; Ponce, J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, NY, USA , 17–22 June 2006; pp. 2169–2178.
17. Yang, J.; Yu, K.; Gong, Y.; Huang, T. Linear spatial pyramid matching suing sparse coding for image classification. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 1794–1801.
18. Tuzel, O.; Porikli, F.; Meer, P. Region covariance: A fast descriptor for detection and classification. In Proceedings of the European Conference on Computer Vision, Graz, Austria, 7–13 May 2006.
19. Erdem, E.; Erdem, A. Visual saliency estimation by nonlinearly integrating features using region covariances. *J. Vis.* **2013**, *13*, 1–20.
20. Tuzel, O.; Porikli, F.; Meer, P. Pedestrian detection via classification on riemannian manifolds. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 1713–1727.
21. Porikli, F.; Tuzel, O.; Meer, P. Covariance tracking using model update based on lie algebra. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, NY, USA, 17–22 June 2006; pp. 728–735.

22. Wang, R.; Guo, H.; Davis, L.; Dai, Q. Covariance discriminative learning: A natural and efficient approach to image set classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2496–2503.
23. Wang, L.; Liu, H.; Sun, F. Dynamic texture video classification using extreme learning machine. *Neurocomputing* **2016**, *174*, 278–285.
24. Arsigny, V.; Fillard, P.; Pennec, X.; Ayache, N. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM J. Matrix Anal. Appl.* **2006**, *29*, 328–347.
25. Arsigny, V.; Fillard, P.; Pennec, X.; Ayache, N. Log-euclidean metrics for fast and simple calculus on diffusion tensors. *Magn. Reson. Med.* **2006**, *56*, 411–421.
26. Li, P.; Wang, Q.; Zhang, L. Log-euclidean kernels for sparse representation and dictionary learning. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 8–15 December 2013.
27. Bo, L.; Sminchisescu, C. Efficient match kernels between sets of features for visual recognition. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, B.C., Canada, 7–12 December 2009, pp. 135–143.
28. Gao, S.; Tsing, I.; Chia, L. Sparse representation with kernels. *IEEE Trans. Image Process.* **2013**, *22*, 423–434.
29. Harandi, M.; Salzmann, M. Riemannian coding and dictionary learning: Kernels to the rescue. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3926–3935.
30. Van Nguyen, H.; Patel, V.M.; Nasrabadi, N.M.; Chellappa, R. Design of non-linear kernel dictionaries for object recognition. *IEEE Trans. Image Process.* **2013**, *22*, 5123–5135.
31. Kim, M. Efficient kernel sparse coding via first-order smooth optimization. *IEEE Trans. Neural Netw. Learn. Syst.* **2014**, *25*, 1447–1459.