Research Article

Bayesian Information Criterion Based Feature Filtering for the Fusion of Multiple Features in High-Spatial-Resolution Satellite Scene Classification

Da Lin,¹ Xin Xu,¹ and Fangling Pu²

¹Signal Processing Laboratory, School of Electronic Information, Wuhan University, Wuhan 430072, China

²Wireless Communication and Sensor Network Laboratory, School of Electronic Information, Wuhan University, Wuhan 430072, China

Correspondence should be addressed to Xin Xu; xinxu@whu.edu.cn

Received 12 November 2014; Accepted 18 February 2015

Academic Editor: Tianfu Wu

Copyright © 2015 Da Lin et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This paper presents a novel classification method for high-spatial-resolution satellite scene classification introducing Bayesian information criterion (BIC)-based feature filtering process to further eliminate opaque and redundant information between multiple features. Firstly, two diverse and complementary feature descriptors are extracted to characterize the satellite scene. Then, sparse canonical correlation analysis (SCCA) with penalty function is employed to fuse the extracted feature descriptors and remove the ambiguities and redundancies between them simultaneously. After that, a two-phase Bayesian information criterion (BIC)-based feature filtering process is designed to further filter out redundant information. In the first phase, we gradually impose a constraint via an iterative process to set a constraint on the loadings for averting sparse correlation descending below to a lower confidence limit of the approximated canonical correlation. In the second phase, Bayesian information criterion (BIC) is utilized to conduct the feature filtering which sets the smallest loading in absolute value to zero in each iteration for all features. Lastly, a support vector machine with pyramid match kernel is applied to obtain the final result. Experimental results on high-spatial-resolution satellite scenes demonstrate that the suggested approach achieves satisfactory performance in classification accuracy.

1. Introduction

Scene classification has aroused more and more attention in remote sensing domain. For satellite imagery of high-spatial-resolution, it evokes a great deal of challenging problems in scene classification due to high intraclass variability, low interclass disparity, and other external factors such as changes of viewpoint, illuminations and shadows, background clutter, partial occlusions, and multiple instances. In addition, with considerable increase of the spatial resolution of images, details of the targets become clearer, and a host of cues also become more distinctive, such as structure and colour. As a consequence, it is of great importance to appropriately combine and fuse them in various respects. In the past decade or so, many researchers and practitioners have made great efforts in exploiting different sources of information in

the high-spatial-resolution satellite imagery to enhance classification performance [1–5].

Unlike the case of low-spatial-resolution satellite images, where solitary type of feature descriptor has been proved to be effective and efficient for classification [6, 7], it is universally recognized that instead of adopting a solitary type of feature, it is more favorable to fuse and combine a set of diverse and complementary features such as features based on structure and colour information [8, 9]. Hence, how to fuse these pieces of diverse and complementary information and eliminate the ambiguities and redundancies between them becomes a critical problem. One widely praised approach is feature-level fusion, in which features from different channels are fused, developing a new pattern for scene classification. Many approaches have been reported in the literature [10–12].

The canonical correlation analysis (CCA) [13] method has been spotlighted, especially in feature fusion realm [14, 15], due to its capability in expressing inherent correlation between two sets of features. In order to extract canonical correlation features from two groups of features, the CCA method firstly constructs a correlation criterion function by extracting two diverse features from the identical samples. Then the CCA method creates efficient and effective discriminant features for classification.

However, when the dimensions of features are too large, such as in case of high-spatial-resolution satellite scene classification researches, traditional CCA methods are no longer proper. Besides, when the features are extracted from the identical image, sample covariance matrices turn into being undefined or unstable, inducing extra difficulty in parameter estimation. As a consequence, a dimension reduction approach is imperative for CCA to tackle this problem. In the past decade, abundant approaches have been proposed for feature shrinkage and selection, containing the non-negative garrote by Breiman [16], least absolute shrinkage and selection operator (Lasso) by Tibshirani [17], smoothly clipped absolute deviation (SCAD) by Fan and Li [18], and Elastic-net by Zou and Hastie [19]. Recently, these approaches have been employed to CCA for relationship assessment between two sets of high-dimensional remote sensing data. The feature selection strategies are utilized to the canonical feature loadings which set some of the coefficients to exact zeros for selecting the remaining features. The crucial features selected are then called the sparse set of features, and the canonical correlation analysis exploiting these features is often known as sparse canonical correlation analysis (SCCA). Initially, Waaijenborg et al. [20] suggested a penalized form of CCA adopting an iterative regression process with the Univariate Soft Threshold (UST) form of the Elastic-net penalty. Subsequently, Parkhomenko et al. [21] suggested the SCCA approach utilizing a form of regularization resembling UST Elastic-net [19]. Witten and Tibshirani [22] incorporated the Lasso penalty in their SCCA. These approaches are interesting. Yet, they do not control the sparsity in a direct way. At the same time, it has been revealed that sundry penalized likelihood approaches have the oracle property under certain conditions [18, 23–26]. However, in practical application, without an appropriate model of feature selection, the oracle property cannot be realized. Accordingly, the approaches may not necessarily generate sparse set of features. To address the issue, this paper suggests a two-phase process in implementing SCCA in which L1 penalty is utilized on the feature loadings during the first phase, and then a Bayesian Information Criterion (BIC) based feature filtering algorithm is conducted to further remove redundant and noisy information. To be specific, in the first phase, we gradually impose a constraint via an iterative process to set a constraint on the loadings for averting sparse correlation descending below to a lower confidence limit of the approximated canonical correlation. In the second phase, Bayesian information criterion (BIC) is utilized to conduct the feature filtering which sets the smallest loading in absolute value to zero in each iteration for all features.

The rest of this paper is organized as follows. The feature extraction process is presented in Section 2. Section 3 offers a concrete and detailed depiction of the methodology exploited in this paper. Section 4 displays experimental results and gives a performance assessment. Finally, Section 5 summarizes the work and points the directions for the future work.

2. Multiset Feature Extraction

2.1. Scale-Invariant Feature Transform (SIFT) Descriptor. In the detected region, SIFT descriptor extracts a gradient orientation histogram [27]. The gradient image is sampled over a 4×4 grid in each of eight orientation planes; thus resulting descriptor is of dimension 128. A weight to the magnitude of each sample point is gained by implementing a Gaussian window function. This puts more highlights on the gradients that are near the center of the region and renders the descriptor less sensitive to the small changes in the position of the detected region. The gradient magnitude is used to weigh the contribution to the orientation and location bins. The descriptor is immune to small errors in the region detection and small geometric distortions, largely due to the quantization of orientations and gradient locations. The square root of the sum of squared components is calculated to normalize the descriptor for acquiring illumination invariance.

2.2. Colour Histogram Descriptor. Colour histogram descriptor is three separated histograms for the R, G, and B channels [28]. A colour histogram represents the approximate distribution of the colours in an image, due to the fact that each histogram bin stands for a local colour range in the given colour space. Colour histograms are invariant to the translation and rotation of the image content; meanwhile they are unsophisticated to compute. In our experiments, the number of bins is 40, and the resulting descriptor is of dimension 120 through concatenating the three independent histograms.

3. Methodology

3.1. Sparse Canonical Correlation Analysis (SCCA). Canonical correlation analysis (CCA) is a multivariate statistical approach proposed to grope for the correlation between two sets of features [13]. Suppose that two feature sets $X = X_1, X_2, \dots, X_p$ and $Y = Y_1, Y_2, \dots, Y_q$ are of dimensions $n \times p$ and $n \times q$ ($p \leq n, q \leq n$), which extracted from the same image. Let the columns of X and Y be standardized to have standard deviation 1 and mean 0, let u and v be $p \times 1$ and $q \times 1$ vector of weights, and let $\xi = Xu$ and $\eta = Yv$ be linear combinations of the features of data sets X and Y , respectively. Note that ξ and η are $n \times 1$ vectors. Subsequently, (1) will be maximized to estimate coefficient vectors u and v :

$$\begin{aligned} \rho &= \frac{\text{Cov}(\xi, \eta)}{[\text{var}(\xi) \text{var}(\eta)]^{1/2}} \\ &= \frac{u^T X^T Y v}{[(u^T X^T X u)(v^T Y^T Y v)]^{1/2}}, \end{aligned} \quad (1)$$

where $X^T X$ and $Y^T Y$ are within data covariance matrices and $X^T Y$ is the between data covariance matrix. Equation (1) can be reformulated as follows for scaling u and v has very insignificant influence on the correlation coefficient:

$$\begin{aligned} \rho &= \text{corr}(\xi, \eta) = \max u^T X^T Y v \\ \text{subject to } u^T X^T X u &= v^T Y^T Y v = 1. \end{aligned} \quad (2)$$

The aforementioned CCA approach is not applicable when the quantity of features is excessive. Latent multicollinearity between predictor features further complicates the computation for the covariance matrices can turn into undefined or unstable. As a consequence, some critical features should be selected by standard model selection criteria. In the subsequent process, the selected set of features is utilized to compute the canonical correlation for making the results understandable, which is called sparse canonical correlation analysis (SCCA). Theoretically, SCCA is implemented by maximizing the penalized objective function below:

$$\begin{aligned} \rho &= \text{corr}(\xi, \eta) = \max_{u,v} u^T X^T Y v \\ \text{subject to } u^T X^T X u &= v^T Y^T Y v = 1. \end{aligned} \quad (3)$$

In order to tackle the multicollinearity problem, a host of approaches have been introduced. Vinod suggested incorporating penalty terms to the diagonal elements of the covariance matrix, which appears to resemble the ridge regression thought in regression analysis [29]. This needs to estimate additional ridge parameters. Other regularization forms have been proposed where the variance matrices are substituted with their corresponding identity matrices [22] or diagonal matrices [21]. In our work, the matrices $X^T X$ and $Y^T Y$ are substituted with their corresponding diagonal matrices.

3.2. Shrinkage Methods. Penalized linear regression mechanisms have been widely applied to analyze high-dimensional data, and it has incorporated feature selection and shrinkage techniques. Assume that y is an $n \times 1$ vector and x is an $n \times p$ matrix. Then the estimation of penalized regression coefficient β can be yielded by using following penalized regression model:

$$\hat{\beta}^{\text{penalized}} = \arg \min |y - x\beta|^2 + p_\lambda(\beta), \quad (4)$$

where $p_\lambda(\beta)$ is the penalty term and λ is a tuning parameter which is estimated utilizing permutation approaches or cross validation (CV). As the first penalized regression approach, ridge regression was proposed to temper the multicollinearity among the predictors in which a quadratic penalty term is embedded to the regular least square estimating equations [30]. Ridge regression implements a penalty on the coefficients to shrink them towards zero. Yet the shrunken coefficients are never equal to zero. As a consequence, ridge regression fails to conduct feature selection. The least absolute shrinkage and selection operator (Lasso), Elastic-net, and

smoothly clipped absolute deviation (SCAD) are different from ridge regression which solve the multicollinearity problem (i.e., shrinkage) and set some of the coefficient to exact zero, creating sparse set of features (i.e., feature selection). In this paper, we apply different penalty functions to SCCA utilizing the algorithm elaborated by Parkhomenko et al. [21]. The tuning parameters for all penalty functions are estimated through cross validation (CV).

3.2.1. Least Absolute Shrinkage and Selection Operator (Lasso) Penalty. The least absolute shrinkage and selection operator (Lasso) penalty is a shrinkage approach and has the competence of selecting discriminant features by shrinking some coefficients and setting others to zero [31]. The penalty term of Lasso is defined as follows:

$$p_\lambda^{\text{lasso}}(\beta) = \lambda \sum_{j=1}^p |\beta_j|, \quad (5)$$

where λ is a tuning parameter. The solution of Lasso is given as

$$\hat{\beta}_j^{\text{lasso}} = (|\hat{\beta}_j| - \lambda)_+ \text{sign}(\hat{\beta}_j). \quad (6)$$

This is similar to the soft thresholding rule introduced by Donoho et al. [32] and Donoho and Johnstone [33], which was utilized to estimate wavelet coefficients.

3.2.2. Elastic-Net Penalty. Elastic-net is a regularization mechanism that carries out continuous shrinkage and feature selection simultaneously [34]. To be specific, this approach utilizes both L_1 penalty of Lasso and L_2 quadratic penalty of ridge regression to create a convex combination. Consequently, this approach preserves the abilities of feature selection and coefficients shrinkage. The definition of the Elastic-net penalty can be formulated as follows:

$$p_\lambda^{\text{elastic-net}}(\beta) = \lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2. \quad (7)$$

Nevertheless, thanks to two tuning parameters needed to be estimated, the computational cost of Elastic-net is somewhat higher.

As a substitute to Elastic-net, Zou and Hastie proposed a predigested version of the Elastic-net called univariate soft thresholding (UST) [34], the solution of which is shown as

$$\hat{\beta}_j^{\text{elastic-net}} = (|\hat{\beta}_j| - 0.5\lambda)_+ \text{sign}(\hat{\beta}_j). \quad (8)$$

In this paper, the Elastic-net based on the univariate soft threshold is adopted to implement feature loadings u and v as follows:

$$\begin{aligned} \hat{u}_j &= (|\hat{u}_j| - 0.5\lambda_1)_+ \text{sign}(\hat{u}_j), \\ \hat{v}_j &= (|\hat{v}_j| - 0.5\lambda_2)_+ \text{sign}(\hat{v}_j). \end{aligned} \quad (9)$$

3.2.3. Smoothly Clipped Absolute Deviation (SCAD) Penalty. Fan and Li proposed a nonconvex penalty function named smoothly clipped absolute deviation (SCAD) [18]. They suggested three criteria for determining an excellent penalty function, namely, (i) sparsity, (ii) continuity, and (iii) unbiasedness. They made further efforts to claim that the SCAD penalty possesses these properties. The SCAD penalty is shown as follows:

$$p_{\lambda}^{\text{SCAD}}(\beta) = \begin{cases} \lambda |\beta_j|, & \text{if } |\beta_j| \leq \lambda, \\ -\left(\frac{|\beta_j|^2 - 2a\lambda |\beta_j| + \lambda^2}{2(a-1)} \right), & \text{if } \lambda < |\beta_j| \leq a\lambda, \\ \frac{(a+1)\lambda^2}{2}, & \text{if } |\beta_j| > a\lambda. \end{cases} \quad (10)$$

When value within the range of λ and $a\lambda$, SCAD penalty function coincides with a quadratic spline function. The function is continuous, and when $a > 2$ and $\beta > 0$ the first derivative can be formulated as follows:

$$p'_{\lambda}(\beta) = \lambda \left\{ I(\beta \leq \lambda) + \frac{(a\lambda - \beta)_+}{(a-1)\lambda} I(\beta > \lambda) \right\}. \quad (11)$$

The SCAD penalty is continuously differentiable on $(-\infty, 0) \cup (0, \infty)$, but singular at 0, with its derivatives zero outside the range $[-a\lambda, a\lambda]$. This penalty function sets small coefficients to zero, shrinks mid-size coefficients towards zero, and keeps large coefficients untouched. Consequently, the SCAD penalty generates almost unbiased coefficients and a sparse solution for large coefficients. The solution of SCAD penalty is shown as follows:

$$\hat{\beta}_j^{\text{SCAD}} = \begin{cases} (|\hat{\beta}_j| - \lambda)_+ \text{sign}(\hat{\beta}_j), & \text{if } |\hat{\beta}_j| \leq 2\lambda, \\ \frac{(a-1)\hat{\beta}_j - \text{sign}(\hat{\beta}_j)a\lambda}{a-2}, & \text{if } 2\lambda < |\hat{\beta}_j| \leq a\lambda, \\ \hat{\beta}_j, & \text{if } |\hat{\beta}_j| > a\lambda. \end{cases} \quad (12)$$

This thresholding rule has two unknown parameters: λ and a . In an ideal situation, the optimal results (λ, a) can be acquired utilizing a scheme involving a two-dimensional grid-search with criteria resembling cross validation approaches. However, such an execution is computationally intensive. In the Bayesian perspective, Fan and Li advised $a = 3.7$ is a wise option for many issues [18]. They have further highlighted that the performance of feature selection issues does not boost tremendously when data-driven approaches are adopted. In this paper, we set a to 3.7 and λ was selected by cross validation. Meanwhile, the thresholding rule (12) was adopted to load vectors u and v .

3.2.4. Hard-Threshold Penalty. Hard-thresholding directly sets several coefficients to zero [35, 36]. However, this penalty function does not tackle the issue of multicollinearity among the predictors, for it does not shrink any coefficients toward zero. Nevertheless, the results obtained by this penalty are unbiased estimators with large effects. The solution of the hard-thresholding rule is revealed as follows:

$$\hat{\beta}_j^{\text{hard}} = \beta_j I(|\beta_j| > \lambda). \quad (13)$$

3.3. The Suggested BIC Based Feature Filtering Algorithm. The major drawback of the current SCCA approaches is that they do not control sparsity directly. Hence, it is hard to complete efficient and effective elimination of noisy and redundant information. There is a trade-off between the sparsity of the features and the maximum correlation. In this paper, we suggest a two-phase process to establish an equilibrium between the sparsity of the features and the maximum correlation. In the first phase, we gradually impose a constraint via an iterative process to set a constraint on the loadings for averting sparse correlation descending below to a lower confidence limit of the approximated canonical correlation. In the second phase, Bayesian information criterion (BIC) is utilized to conduct the feature filtering which sets the smallest loading in absolute value to zero in each iteration for all features.

The proposed feature filtering process is iterative and simple. One more coefficient of $\hat{\beta}_j$, at each iteration, is set to be 0 according to the magnitudes of the coefficients in absolute value. Let $\hat{\beta}_i$ and $\hat{\beta}_j$ be the constrained effective dimension reduction direction; the proposed feature filtering process is shown as follows.

- (i) Let $d = p + q$.
- (ii) Define a new direction $\hat{\beta}'_j(d)$ by maintaining the largest d coefficients of $\hat{\beta}_j$ in absolute value and assigning the other $p+q-d$ coefficients to 0. Searching $\hat{\beta}^p_j(d)$ as projection of $\hat{\beta}'_j(d)$ into the space B_i , the set of all $\hat{\beta}^p_j(d)$ should satisfy the following.
 - (1) The set of zero coefficients in $\hat{\beta}^p_j(d)$ is the same as that in $\hat{\beta}'_j(d)$.
 - (2) Consider $\hat{\beta}^p_j(d)^T X^T X \hat{\beta}_i = 0, i \leq j-1$.
 - (3) Consider $\hat{\beta}^p_j(d)^T X^T X \hat{\beta}^p_j(d) = 1$.

(iii) Compute the correlation

$$r_d = \text{Corr}(\hat{\beta}'_j(d), \hat{\beta}^p_j(d)) = \frac{\text{Cov}(\hat{\beta}'_j(d), \hat{\beta}^p_j(d))}{[\text{var}(\hat{\beta}'_j(d)) \text{var}(\hat{\beta}^p_j(d))]^{1/2}}, \quad (14)$$

and the BIC-type criterion $\text{BIC}(d) = n \log(1 - r_d^2) + d \log(n)$, where n is the sample size.

- (iv) Let $d = d - 1$. Repeat Steps (ii)–(iv) until $d = 0$.

After the above feature filtering process is implemented, we obtain a sequence of $BIC(d)$ as d descends from $p + q$ to 0. Let d_0 be the integer at which $BIC(d)$ is minimized. Then, the $p + q - d_0$ smallest coefficients of $\hat{\beta}_j$ in absolute value are assigned to 0. This proposed feature filtering process is a streamlined feature selection process. In feature filtering, at most $p + q$ possibilities are taken into account, which make it viable to conduct even when $p + q$ are large. Lastly, the features corresponding to the minimum BIC value are the final selected features.

3.4. Support Vector Machine (SVM) with Pyramid Match Kernel (PMK-SVM) Classifier. Establishment of kernel-based learning algorithms is based on the notion of mapping data into a Euclidean space and then discovering linear relations within the mapped data. Taking a typical issue as an example, the SVM unearths the optimal separating hyperplane between two classes in a feature space. The assistance provided by a kernel function is to map pairs of data points in an input space to their inner product in the feature space, thereby estimating the similarities between all points and deciding their relative positions. Linear relations are discovered in the feature space, although a decision boundary may still be nonlinear in the input space, depending on the method of a feature mapping function. The support vector machine (SVM) with pyramid match kernel (PMK-SVM) [37] furnishes an accurate and time-saving solution for classification and the pyramid match kernel function is formulated as follows:

$$K(\Psi(y), \Psi(z)) = \sum_{i=0}^L \frac{1}{2^i} (\Gamma(H_i(y), H_i(z)) - \Gamma(H_{i-1}(y), H_{i-1}(z))), \quad (15)$$

where y, z are the input sets, $L = \lceil \log_2 D \rceil$ in which D is a sphere of diameter, Ψ is the feature extraction function, $H_i(x)$ is the i th histogram in $\Psi(x)$, and Γ is a histogram intersection function which measures the overlap between two histograms' bins:

$$\Gamma(A, B) = \sum_{j=1}^r \min(A^{(j)}, B^{(j)}), \quad (16)$$

where A and B are histograms with r bins, and $A^{(j)}$ denotes the count of the j th bin of A . Since in the construction of the pyramid phase $\Gamma(H_L(y), H_L(z)) = \min(|y|, |z|)$, and $\Gamma(H_{-1}(y), H_{-1}(z)) = 0$, (15) is equivalent to

$$K(\Psi(y), \Psi(z)) = \frac{\min(|y|, |z|)}{2^L} + \sum_{i=0}^{L-1} \frac{1}{2^{i+1}} \Gamma(H_i(y), H_i(z)). \quad (17)$$

In order to preserve generality and obtain promising and satisfactory classification results. Here, we employ support vector machine (SVM) with pyramid match kernel (PMK-SVM) as classifier. Under the multiclass circumstances, a set

of binary classifiers and majority vote technique are utilized to perform multiclass categorization. Figure 1 illustrates our classification scheme based on the suggested two-phase BIC filtering process.

4. Experiments and Results

Experiments were performed on a high-spatial-resolution satellite image with a size of $4000 \text{ pixels} \times 4000 \text{ pixels}$, as presented in Figure 2. The view was captured by the GeoEye-1 satellite on 21 November 2009 at Majuqiao town, which locates in the southwest of Tongzhou, southeast of Beijing, where the latitudes and longitudes at the lower right and upper left corners are $39^\circ 43' \text{N}$, $116^\circ 32' \text{E}$ and $39^\circ 44' \text{N}$, $116^\circ 30' \text{E}$ respectively. The ground sampling distance is approximately 0.5 m and the band assignment is red for band 3, green for band 2, and blue for band 1. The image mainly contains eight-class satellite scenes: factories, roads, water, farm land, high buildings, low buildings, bare land, and green land. With associated geographic information, the reference data presented in Figure 6(a) was manually labeled. Moreover, some classes have not been taken into account, which can be neglected, such as clearance between the green land and its neighbours. Figure 3 exhibits one example of each class from the eight-class satellite scene.

In this paper, the k -fold cross validation approach is used to optimize the penalty parameters for each canonical feature pair, where k in k -fold cross validation is set to 10 by empirical. Specifically, the obtained feature data are split into two parts: $1/k$ proportion of feature data for validation (testing) and the remaining $k - 1/k$ proportion of feature data for training. The loading vectors are yielded in the training procedure and used in the testing procedure. We maximize the correlation of testing feature data to select the sparseness parameters λ_u and λ_v using cross validation approach. The weighted vectors u and v are obtained by the preset values of λ_u and λ_v . Subsequently, the correlation is calculated by

$$\Delta_{\text{cor}} = \frac{1}{k} \sum_{j=1}^k |\text{cor}(X_j u^{-j}, Y_j v^{-j})|, \quad (18)$$

where u^{-j} and v^{-j} are the weight vectors in the training sets X_j and Y_j , respectively, and k is the implementing times of cross validation. Here, X_j and Y_j are the test sets. Lastly, λ_u and λ_v values are determined as optimum sparseness parameters according to the maximum value of Δ_{cor} .

To achieve reliable results and see the convergence tendency of the classification accuracy, we conducted the entire experiments with different proportions of training-testing samples by using linear increasing number of training samples and learnt that classification accuracy is gradually boosted with an increasing number of training samples. To avert training and testing samples are too spatially correlated, we trained on 5, 10, 15, 20, 25 and 30 images in each class and tested on the remainder. At the same time, we chose training data and testing data from the different high-spatial-resolution satellite scene for independent data set and reconducted the classification experiments with these

FIGURE 1: Flow chart of the suggested classification method.

FIGURE 2: High-spatial-resolution satellite scene.

selected data; the acquired results were coincident as well. The classifiers were not conspicuously biased within this particular image for the moderate selected proportion of training samples and hence classification results would not be extremely exaggerated. We divided training\testing data at random and reiterated the experiments ten times. Meanwhile, we kept an account of the average of each-class classification accuracies for every run. The mean and standard deviation of the results from each individual experiment were used to quantify the final result.

For the sake of boost in classification performance to some extent, these two features should be fused in an apt tactic. For comparison purposes, SCCA with different penalty functions (Lasso, Elastic-net, SCAD, and Hard-threshold) was introduced to fuse the features yielded by feature extraction methods adopted in this paper. It is all obvious from Figure 4 and easy to observe that SCCA with SCAD penalty function acquires a relatively outstanding result over the other penalty functions. The reasons are sketched as follows. On one side, in order to grip the complete source of image information, we utilized two typical and representative features to depict structure and colour properties of the image, and these features contain the essential and intrinsic information of the image. On the other side, SCAD penalty function possesses the competence to manipulate multicollinearity between features as compare to the other penalty functions. However, these methods do not have immediate control over the sparsity. Consequently, an extra two-phase feature filtering process was indispensable to conduct for further removing redundant and opaque information between and within features. Figure 5 compares the classification performance of four different penalty functions with the suggested two-phase BIC based filtering process. It is evident that classification accuracy enhances after utilizing the suggested BIC filter, which can be explained by the possibility that without the suggested BIC filter the SCCA may not necessarily engender sparse set of features. In other words, after using the suggested BIC filter the noisy and redundant information among features have been further eliminated.

Note that the SCAD penalty function with the suggested BIC filter yielded the best classification accuracy. Figures

FIGURE 3: Examples of each class in the eight-class satellite scene.

FIGURE 4: Performance comparison using different penalty functions on SCCA without the suggested BIC filter.

FIGURE 5: Performance comparison using different penalty functions on SCCA with the suggested BIC filter.

6(a) and 6(b) show the comparison between the reference image in the same satellite scene and the best classified image acquired by our method (SCAD penalty function SCCA with the suggested BIC filter), respectively, and the confusion matrices obtained via SCAD penalty function SCCA without and with the suggested BIC filter are presented in Figures 7 and 8, respectively.

It is as expected that misclassification was more inclinable to emerge between short buildings and factories. This is due to the fact that factories often include dense houses and level and perpendicular lines, which resemble short buildings. Meantime green land was misclassified as farm land or bare land; the most contributing factor to this

misclassification is that these classes possess similar elements and patterns. However, confusions between some classes were tough to understand. Namely, some short buildings, roads, and factories were classified as bare land.

5. Conclusions

In this paper, a BIC based feature filtering approach was presented for high-spatial-resolution satellite scene classification. The SCCA with a two-phase BIC feature filtering process acts as a crucial component in satellite scene classification and can strikingly meliorate the classification accuracy by

FIGURE 6: Classification results for the eight-class satellite image. (a) The hand-labelled reference data. (b) The best classified image using our method (SCAD penalty function SCCA with the suggested BIC filter).

FIGURE 7: Confusion matrix for eight-class high-spatial-resolution satellite scene data using SCAD penalty function SCCA without the suggested BIC filter. The entry in the i th row and j th column is the percentage of images from class i that were misclassified as class j .

efficaciously filtering and selecting two distinct and complementary features, SIFT and colour histogram. However, in the absence of a prior original features processing step before using the suggested SCCA feature fusion method, current method is greatly restricted in the experiments of Majuqiao town satellite scene classification. To solve the aforementioned issue, we intend to implement a sparse coding step to the original features before utilizing the feature fusion

FIGURE 8: Confusion matrix for eight-class high-spatial-resolution satellite scene data using SCAD penalty function SCCA with the suggested BIC filter. The entry in the i th row and j th column is the percentage of images from class i that were misclassified as class j .

method for reducing the computational cost and enhancing classification accuracy.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported in part by the National Basic Research Program of China (973 Program) under Grant no. 2011CB707102 and the National High Technology Research and Development Program of China (863 Program) under Grant no. 2013AA122301.

References

- [1] A. Bekkari, S. Idbraim, D. Mammas, M. El Yassa, and D. Ducrot, "SVM classification of high resolution urban satellites images using composite kernels and haralick features," *Journal of Emerging Technologies in Web Intelligence*, vol. 6, no. 1, pp. 69–74, 2014.
- [2] G.-S. Xia, W. Yang, J. Delon et al., "Structural high-resolution satellite image indexing," in *Proceedings of the ISPRS TC VII Symposium—100 Years ISPRS*, pp. 298–303, Vienna, Austria, July 2010.
- [3] D. Dai and W. Yang, "Satellite image classification via two-layer sparse coding with biased image representation," *IEEE Geoscience and Remote Sensing Letters*, vol. 8, no. 1, pp. 173–176, 2011.
- [4] B. A. Johnson, "High-resolution urban land-cover classification using a competitive multi-scale object-based approach," *Remote Sensing Letters*, vol. 4, no. 2, pp. 131–140, 2013.
- [5] J. J. Mitchell, R. Shrestha, C. A. Moore-Ellison, and N. F. Glenn, "Single and multi-date landsat classifications of basalt to support soil survey efforts," *Remote Sensing*, vol. 5, no. 10, pp. 4857–4876, 2013.
- [6] L. Ruiz, A. Fdez-Sarria, and J. Recio, "Texture feature extraction for classification of remote sensing data using wavelet decomposition: a comparative study," in *Proceedings of the 20th ISPRS Congress*, 2004.
- [7] C.-S. Li and V. Castelli, "Deriving texture feature set for content-based retrieval of satellite image database," in *Proceedings of the International Conference on Image Processing*, pp. 576–579, IEEE, October 1997.
- [8] M. Shimoni, D. Borghys, R. Heremans, C. Perneel, and M. Acheroy, "Fusion of PolSAR and PolInSAR data for land cover classification," *International Journal of Applied Earth Observation and Geoinformation*, vol. 11, no. 3, pp. 169–180, 2009.
- [9] J. Muñoz-Marí, G. Camps-Valls, L. Gómez-Chova, and J. Calpe-Maravilla, "Combination of one-class remote sensing image classifiers," in *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS '07)*, pp. 1509–1512, IEEE, Barcelona, Spain, June 2007.
- [10] B. Bigdeli, F. Samadzadegan, and P. Reinartz, "A multiple svm system for classification of hyperspectral remote sensing data," *Journal of the Indian Society of Remote Sensing*, vol. 41, no. 4, pp. 763–776, 2013.
- [11] P. Du, Y. Chen, J. Xia, and K. Tan, "A novel remote sensing image classification scheme based on data fusion, multiple features and ensemble learning," *Journal of the Indian Society of Remote Sensing*, vol. 41, no. 2, pp. 213–222, 2013.
- [12] E. Pasolli, F. Melgani, D. Tuia, F. Pacifici, and W. J. Emery, "SVM active learning approach for image classification using spatial information," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 4, pp. 2217–2223, 2014.
- [13] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3–4, pp. 321–377, 1936.
- [14] Q.-S. Sun, Z.-D. Liu, P.-A. Heng, and D.-S. Xia, "A theorem on the generalized canonical projective vectors," *Pattern Recognition*, vol. 38, no. 3, pp. 449–452, 2005.
- [15] Q.-S. Sun, S.-G. Zeng, Y. Liu, P.-A. Heng, and D.-S. Xia, "A new method of feature fusion and its application in image recognition," *Pattern Recognition*, vol. 38, no. 12, pp. 2437–2448, 2005.
- [16] L. Breiman, "Better subset regression using the nonnegative garrote," *Technometrics*, vol. 37, no. 4, pp. 373–384, 1995.
- [17] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Series B: Methodological*, vol. 58, no. 1, pp. 267–288, 1996.
- [18] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.
- [19] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, vol. 67, no. 2, pp. 301–320, 2005.
- [20] S. Waaijenborg, P. C. Verselewe De Witt Hamer, and A. H. Zwinderman, "Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis," *Statistical Applications in Genetics and Molecular Biology*, vol. 7, no. 1, article 3, 2008.
- [21] E. Parkhomenko, D. Tritchler, and J. Beyene, "Sparse canonical correlation analysis with application to genomic data integration," *Statistical Applications in Genetics and Molecular Biology*, vol. 8, no. 1, article 1, 2009.
- [22] D. M. Witten and R. J. Tibshirani, "Extensions of sparse canonical correlation analysis with applications to genomic data," *Statistical Applications in Genetics and Molecular Biology*, vol. 8, no. 1, article 28, 2009.
- [23] J. Fan and H. Peng, "Nonconcave penalized likelihood with a diverging number of parameters," *The Annals of Statistics*, vol. 32, no. 3, pp. 928–961, 2004.
- [24] Y. Kim, H. Choi, and H.-S. Oh, "Smoothly clipped absolute deviation on high dimensions," *Journal of the American Statistical Association*, vol. 103, no. 484, pp. 1665–1673, 2008.
- [25] N. Meinshausen and B. Yu, "Lasso-type recovery of sparse representations for high-dimensional data," *The Annals of Statistics*, vol. 37, no. 1, pp. 246–270, 2009.
- [26] P. Zhao and B. Yu, "On model selection consistency of Lasso," *The Journal of Machine Learning Research*, vol. 7, pp. 2541–2563, 2006.
- [27] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [28] B. S. Manjunath, J.-R. Ohm, V. V. Vasudevan, and A. Yamada, "Color and texture descriptors," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 703–715, 2001.
- [29] H. D. Vinod, "Canonical ridge and econometrics of joint production," *Journal of Econometrics*, vol. 4, no. 2, pp. 147–166, 1976.
- [30] A. E. Hoerl and R. W. Kennard, "Ridge regression: biased estimation for nonorthogonal problems," *Technometrics*, vol. 42, no. 1, pp. 80–86, 2000.
- [31] R. Tibshirani, "Regression shrinkage and selection via the lasso: a retrospective," *Journal of the Royal Statistical Society, Series B. Statistical Methodology*, vol. 73, no. 3, pp. 273–282, 2011.
- [32] D. L. Donoho, I. M. Johnstone, G. Kerkycharian, and D. Picard, "Wavelet shrinkage: Asymptopia," *Journal of the Royal Statistical Society: Series B*, vol. 57, no. 2, pp. 301–369, 1995.

- [33] D. L. Donoho and J. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.
- [34] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, vol. 67, no. 2, pp. 301–320, 2005.
- [35] A. Antoniadis, "Wavelets in statistics: a review," *Journal of the Italian Statistical Society*, vol. 6, no. 2, pp. 97–130, 1997.
- [36] J. Fan, "Comments on 'wavelets in statistics: a review' by A. Antoniadis," *Journal of the Italian Statistical Society*, vol. 6, no. 2, pp. 131–138, 1997.
- [37] K. Grauman and T. Darrell, "The pyramid match kernel: discriminative classification with sets of image features," in *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV '05)*, vol. 2, pp. 1458–1465, October 2005.