remote sensing

MDPI

Article

Hierarchical Coding Vectors for Scene Level Land-Use Classification

Hang Wu [1], Baozhen Liu [1], Weihua Su [1], Wenchang Zhang [2] and Jinggong Sun [1,*]

[1] Institute of Medical Equipment, Academy of Military Medical Science, Tianjin 300161, China; 2008.wuhang@163.com (H.W.); liubaozhen91@126.com (B.L.); directorsu@126.com (W.S.)

[2] The State Key Laboratory of Intelligent Technology and System, Computer Science and Technology School, Tsinghua University, Beijing 100084, China; zwc0501@163.com

* Correspondence: sunjg@vip.sina.com; Tel.: +86-22-8465-6813

Academic Editors: Soe Myint and Prasad S. Thenkabail

Received: 26 March 2016; Accepted: 18 May 2016; Published: 23 May 2016

Abstract: Land-use classification from remote sensing images has become an important but challenging task. This paper proposes Hierarchical Coding Vectors (HCV), a novel representation based on hierarchically coding structures, for scene level land-use classification. We stack multiple Bag of Visual Words (BOVW) coding layers and one Fisher coding layer to develop the hierarchical feature learning structure. In BOVW coding layers, we extract local descriptors from a geographical image with densely sampled interest points, and encode them using soft assignment (SA). The Fisher coding layer encodes those semi-local features with Fisher vectors (FV) and aggregates them to develop a final global representation. The graphical semantic information is refined by feeding the output of one layer into the next computation layer. HCV describes the geographical images through a high-level representation of richer semantic information by using a hierarchical coding structure. The experimental results on the 21-Class Land Use (LU) and RSSCN7 image databases indicate the effectiveness of the proposed HCV. Combined with the standard FV, our method (FV + HCV) achieves superior performance compared to the state-of-the-art methods on the two databases, obtaining the average classification accuracy of 91.5% on the LU database and 86.4% on the RSSCN7 database.

Keywords: land use classification; Bag of Visual Word; Fisher Vectors; Hierarchical Coding Vectors

1. Introduction

Scene level land-use classification aims to assign a semantic label (e.g., building and river) to a remote sensing image according to its content. As remote sensing techniques continue to develop, overwhelming amounts of fine spatial resolution satellite images have become available. It is necessary to develop effective and efficient scene classification methods to annotate the massive remote sensing images.

By far, the Bag of Visual Words (BOVW) [1,2] framework and its variants [3,4] based on spatial relations have become promising remote sensing image representations for land-use classification. The pipeline for the BOVW framework consists of five main steps: feature extraction, codebook generation, feature coding, pooling, and normalization. For BOVW, we usually extract local features from the geographical images, learn a codebook in the training set by K-means or Gaussian mixture model (GMM), encode the local features and pool them to a vector, and normalize this vector as the final global representation. The representation is subsequently fed into a pre-trained classifier to obtain the annotation result for remote sensing images.

In a parallel development, *deep learning methods* have attracted continuous attention in the computer vision community in recent years. Deep neural networks (DNNs) [5] build and train deep architectures to capture graphical semantic information, achieving a large performance boost

compared to the previous hand-crafted system with mid-level features. Although their methods can describe the geographical images from low level features with a more abstract and semantic representation of deep structures, it is computationally expensive to directly train effective DNNs for scene level land-use classification. One important property of the DNNs is its hierarchical organization in layers of increasing processing complexity. We adopt a similar idea, and concentrate on a shallow but hierarchic layer framework based on *off-the-shelf* encoding methods [6,7].

Inspired by the success of DNNs in computer vision applications and encoding methods for remote sensing applications, we proposed Hierarchical Coding Vectors (HCV), a new representation based on hierarchically coding structures, for scene level land-use classification. We apply the traditional coding pipeline as corresponding to the layers of a standard DNN and stack multi-BOVW coding layers and one Fisher coding layer to develop the hierarchical feature learning structure. The complex graphical semantic information is refined by feeding the output of one layer into the next computation layer. Through hierarchical coding, the HCV contains richer semantic information and is more powerful to describe those remote sensing images. Our experimental results on the 21-Class Land Use (LU) and RSSCN7 geographical image databases demonstrate the excellent performance of our HCV for land-use classification. Furthermore, HCV provides complementary information to the traditional Fisher Vectors (FV). When combining traditional FV with our HCV, we obtain superior classification performance compared to the current state-of-the-art results on the LU and RSSCN7 databases.

There are two main contributions of our work:

- We devise the Hierarchical Coding Vectors (HCV) by organizing off-the-shelf coding methods into a hierarchical architecture and evaluate the parameters of HCV for land-use classification on the LU database.
- The HCV achieves excellent performance for land-use classification. Further, combining HCV with standard FV, our method (FV + HCV) outperforms the state-of-the-art performance reported on the LU and RSSCN7 databases.

The remainder of this paper is organized as follows. Section 2 discusses the related work on both computer vision and remote sensing applications. Section 3 describes the details of our proposed Hierarchical Coding Vectors (HCV). Section 4 presents the experimental results. Section 5 is the conclusion.

2. Related Work

In both the computer vision and remote sensing communities, the recent efforts in scene classification can be divided into three directions: (1) the development of more elaborate hand-crafted features (e.g., Scale Invariant Feature Transformation (SIFT) [8], Histogram of Oriented Gradient (HOG) [9], GIST [10], Local Binary Pattern (LBP) [11]); (2) more sophisticated encoding methods (e.g., Hard Assignment (HA) [12], Soft Assignment (SA) [6], Local Coordinate Coding (LCC) [13], Locality-constrained Linear Coding (LLC) [14], Vector of Locally Aggregated Descriptors (VLAD) [15], FV [7]), and (3) more complex classifiers (e.g., Support Vector Machine (SVM) [16], Extreme Learning Machine (ELM) [17]). Recently, the second direction (*i.e.*, encoding methods) has attracted more attention and become an effective representation for scene level land-use classification. Typical encoding methods are based on the BOVW framework. The traditional BOVW methods, including HA, SA, LCC, and LLC, are designed from the perspective of activation concept to obtain 0-order statistics of the distribution from descriptors space, and the core issue is to decide which visual word will be activated in the 'visual vocabulary' and to what extent they will be activated. Then, the Fisher Kernel introduced by Jaakkola [18] has been used to extend the BOVW framework. It describes the difference between the distribution of descriptors in an input image and that of the 'visual vocabulary', encoding multi-dimensional information (0th, 1st, 2nd) from the descriptors space. The typical Fisher

Kernel methods conclude Fisher Vector (FV) and Vector of Locally Aggregated Descriptors (VLAD). The VLAD can be viewed as a simplified nonprobabilistic version of the FV.

Some researchers have also attempted to use the multi-layers model to further improve the classification performance in the remote sensing community. Chen [3] stacks two BOVW layers with the HA coding method to represent the spatial relationship among local features. A two-layer sparse coding method is used in [19]. The authors apply two different optimum formulas to guarantee the image sparsity and category sparsity simultaneously, improving the discriminability of the output coding result. In the computer vision community, the hierarchical structure helps DNN [5] to achieve a large performance boost. However, it is difficult to be directly applied for the scene level land-use due to its huge computational cost. Xiaojiang Peng *et al.* [20] stacked multiple Fisher coding layers to build a hierarchical network for action recognition in video. The Fisher coding method causes increasing dimensions of the layer output. Thus, the dimensions of the final representation exponentially increase with the number of layers. A dimensionality reduction method has to be used between calculation layers. Inspired by the success of DNNs in computer vision applications and encoding methods for remote sensing applications, we use the off-the-shelf encoding methods to construct the hierarchical structure and stack multi-BOVW coding layers with only one Fisher coding layer to solve the dilemma in [20]. The overall framework and methods used in each layer of HCV are different from those in [20]. Generally speaking, our HCV develops the hierarchical feature learning structure by stacking N + 2 coding layers, which produces a much higher level representation of richer semantic information and achieves superior performance for scene level land-use classification.

3. Hierarchical Coding Vector

The conventional coding methods effectively encode each local feature in an image into a high-dimensional space and aggregate these codes into a single vector by a pooling method over the entire image (followed by normalization). The representation describes the geographical image in terms of the local patch features, which cannot capture more global and complex structures. Deep neural networks [5] can model complex graphical semantic structures by passing an output of one feature computation layer as the input to the next and by hierarchical refining of the semantic information. Along the line of a similar idea, we devised a hierarchical structure by stacking multi-BOVW coding layers and one Fisher coding layer, which we call the *Hierarchical Coding Vector*. The architecture of the Hierarchical Coding Vector (HCV) is depicted in Figure 1.

We devised the HCV to describe the whole geographical image with higher level representation of richer semantic information by a hierarchical coding structure. As shown in Figure 1, the HCV framework contains N + 2 coding layers (N + 1 BOVW coding layers and one Fisher coding layer). The coding result of one coding layer is fed into the next as the input. These coding layers are then stacked into a hierarchical network. We used BOVW coding layers to describe the local patches. Multi-BOVW coding layer superposition does not trigger dimension disaster because of the stable coding dimension of BOVW methods. The BOVW coding layers refine the local semantic information layer-by-layer and then feed the information into the Fisher coding layer to produce global deep representation. Multi-BOVW coding layers provide a better coding 'material' for the Fisher coding layer, giving the global representation (*i.e.*, HCV) stronger discriminability for scene classification.

Theoretically, a HCV with more coding layers can learn more complicated abstract features, but this may significantly increase the complexity of the model. Considering the effectiveness and efficiency, in this paper, we consider a HCV with two coding layers (*i.e.*, one BOVW coding layer and one Fisher coding layer), because it has already provided compelling quality. The HCV can be generalized to more layers without difficulty. The BOVW coding layer uses a Soft Assignment (SA) [6] coding method to map the low-level descriptors $\mathbf{X} = (x_1, x_2, \ldots, x_k, \ldots, x_K) \in \mathbb{R}^{E \times K}$ from the geographical image to the coding space $\mathbf{D} = (d_1, d_2, \ldots, d_k \ldots, d_K) \in \mathbb{R}^{M \times K}$ using the K-means codebook $\mathbf{B_1} = (b_1, b_2, \ldots, b_m, \ldots, b_M) \in \mathbb{R}^{E \times M}$. After local pooling and normalization, the semi-local features $\mathbf{F} = (f_1, f_2, \ldots f_t \ldots, f_T) \in \mathbb{R}^{M \times T}$ are fed into the Fisher coding layer. With the Gaussian Mixture

Model (GMM) codebook $\mathbf{B_2}$= (b_1,b_2,...b_n,..,b_N)$\in \mathbb{R}^{M\times N}$, the *Hierarchical coding Vector* **HCV**$\in \mathbb{R}^{M\times 2N}$ is produced by Fisher vector (FV) coding. Finally, HCV is input into a classifier such as a Support Vector Machine (SVM) for scene-level land use classification. The detailed description of each layer is as follows. The parameters used in this paper are summarized in Table 1.

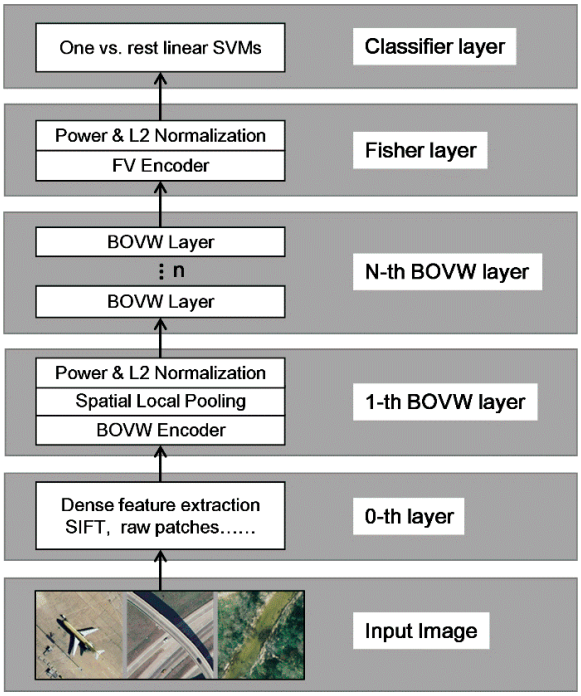

Figure 1. The architecture of the proposed Hierarchical Coding Vector (HCV). The representation of HCV is deeper with richer semantic information by constructing a hierarchical coding structure. SVMs, Support Vector Machines; FV, Fisher Vectors; BOVW, Bag of Visual Words; SIFT, Scale Invariant Feature Transformation.

Table 1. The definitions of parameters used in this paper.

Parameter	Dim.	Definition
X	$E \times K$	Low-level descriptors
B_1	$E \times M$	K-means codebook
D	$M \times K$	Coding result of BOVW coding layer
F	$M \times T$	Semi-local features
B_2	$M \times N$	Gaussian mixture model (GMM) codebook
G	$M \times 2N$	Hierarchical coding Vector
x_k	E	The k-th low-level descriptor
d_k	M	The k-th coding result in D
b_m	E	The m-th codeword in B_1
b_n	M	The n-th codeword in B_2
f_t	M	The t-th semi-local feature
$g_n^{(1)}$	M	Gaussian mean difference
$g_n^{(2)}$	M	Gaussian variance difference
E	1	Dimension of low-level descriptors
T	1	Number of semi-local features
M	1	Size of K-means codebook
N	1	Size of GMM codebook
K	1	Number of low-level descriptors
P	-	Local pooling region
$\hat{e}(x_k, b_m)$	1	Euclidean distance between x_k and b_m
β	1	Smoothing factor in SA coding
α	1	Smoothing factor in Power-normalization
$\alpha_t(n)$	1	Soft assignment weight of f_t to b_n
w_n	1	Mixture weights of b_n
μ_n	1	Means of b_n
σ_n	1	Diagonal covariance of b_n

3.1. The BOVW Coding Layer

The BOVW coding layer maps the input descriptors $\mathbf{X} \in \mathbb{R}^{E \times K}$ to the semi-local features $\mathbf{F} \in \mathbb{R}^{M \times T}$. The pipeline of the BOVW coding layer is shown in Figure 2. Let $\mathbf{X}$ be a set of D-dimensional local descriptors extracted from a geographical image $\mathbf{X} \in \mathbb{R}^{E \times K}$ with densely sampled interest points. Through clustering, a codebook is formed with M entries $\mathbf{B_1} \in \mathbb{R}^{E \times M}$. The codebook is used to express each descriptor and to develop the coding result $\mathbf{D} \in \mathbb{R}^{M \times K}$. Then, pooling and normalization methods are used to produce the local patch coding representation (*i.e.*, a semi-local features $\mathbf{F} \in \mathbb{R}^{M \times T}$). Finally, the features, **F,** are fed into the next Fisher coding layer as the input.

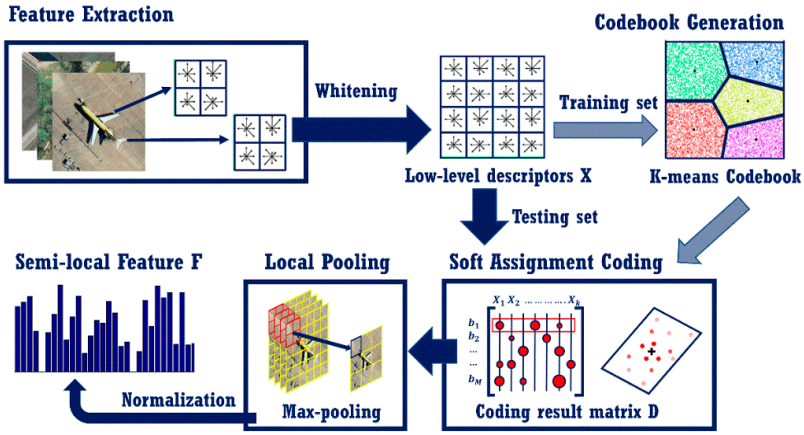

Figure 2. The pipeline of the BOVW coding layer.

3.1.1. BOVW Coding

The BOVW coding step was based on the idea of using overcomplete basis vectors to map the local descriptors $\mathbf{X} \in \mathbb{R}^{E \times K}$ to the coding result $\mathbf{D} \in \mathbb{R}^{M \times K}$.

Given a geographical image, we first extracted the D-dimensional local descriptors X with densely sampled interest points. The raw input local descriptors **X** were usually strongly correlated, which created significant challenges in the subsequent codebook generation [12]. The feature pre-processing approach, *Whitening*, was used to realize the decorrelation. The overcomplete basis vectors (*i.e.*, *codebook* $\mathbf{B_1} \in \mathbb{R}^{E \times M}$) were computed on the training set using the K-means clustering method [21]. To retain spatial information, the dense local descriptors (e.g., Scale Invariant Feature Transformation (SIFT) [8]) were augmented with their normalized x, y location before codebook clustering.

We chose the SA coding method rather than another BOVW coding methods such as HA [12], LCC [13], and LLC [14], which led to strong sparsity in the semi-local features **F**. The strong sparsity caused great challenges in the next Fisher coding layer. SA chose to activate the entire codebook and used the kernel function of distance as the coding representation:

$$\mathrm{d}_k = \frac{\exp(-\beta \hat{e}(x_k, b_m))}{\sum_{m=1}^{M} \exp(-\beta \hat{e}(x_k, b_m))} \tag{1}$$

$$\mathbf{SA}: \ \hat{e}(x_k, b_m) = \| x_k - b_m \|^2 \tag{2}$$

where β is the smoothing factor that controls the softness of the assignment, and the Euclidean distance $\hat{e}$ is used. Smoothing factor β, the sole parameter in SA coding, determines the sensitivity of likelihood to the distance $\hat{e}$ and is critical to the coding and classification performance.

3.1.2. Spatial Local Pooling

Spatial local pooling aggregates the coding result $\mathbf{D} \in \mathbb{R}^{M \times K}$ into the semi-local features $\mathbf{F} \in \mathbb{R}^{M \times T}$, thus achieving greater invariance to image transformations and better robustness to noise and clutter.

Compared to the regions used in the traditional global pooling, the regions are much smaller and sampled much more densely in our HCV framework. The semi-local feature representation captures more complex image statistics with the spatial local pooling.

In the HCV, we performed the spatial local pooling in adjacent scales and spaces. The 2×2 pooling region is illustrated in Figure 2. The optimal spatial structure for local pooling will be evaluated in the following experiment. We used the Max-pooling method in this step, which avoids the semi-local features being strongly influenced by frequent yet often uninformative descriptors [22].

$$\textbf{Max}: \ \mathrm{f}_t = \max(\{\mathrm{d}_k\}_{k \in \mathrm{P}}) \tag{3}$$

where $\mathrm{f_t}$ is the tth element in the semi-local features **F** and the $\mathrm{d_k}$ is the coding result. P refers to the local pooling region. The Max-pooling method has demonstrated its effectiveness in many studies [6,13,14,23].

3.1.3. Normalization

Normalization is used to make the semi-local features have the same scale. Unlike the traditional BOVW coding pipeline, we injected power normalization before the $\mathrm{L_2}$ normalization method as a pre-processing step.

$$\mathbf{L}_2: \ \mathrm{f}_t = \mathrm{f}_t / ||\mathrm{f}_t||_2 \tag{4}$$

$$\textbf{Power}: \ \mathrm{f}_t = sign(\mathrm{f}_t)\,|\mathrm{f}_t|^{\alpha} \tag{5}$$

where $0 \leqslant \alpha \leqslant 1$ is a smoothing factor of normalization (we set $\alpha = 0.5$ the same as [24]). Power normalization is usually used in the Fisher coding method to further improve the classification performance [7]. Meanwhile, BOVW coding methods generally do not apply due to the minimal effect on the performance. However, in our proposed HCV framework, the output of the BOVW coding layer is not used for classification but as the input for the Fisher coding layer. The Fisher vector captures the Gaussian mean and variance differences between the input features and the codebook, and it is very sensitive to the sparsity of the input features. Power normalization deceases the sparsity of the semi-local features **F** and make their distribution smoother, improving the classification performance of HCV (with the experiment on the LU database, we found that the power-normalization can improve the classification accuracy 3%~5%).

To retain the spatial information, the semi-local features **F** were also augmented with their normalized x, y location before they were fed into the next layer.

3.2. The Fisher Coding Layer

The Fisher coding layer maps the input semi-local features $\mathbf{F} \in \mathbb{R}^{M \times \mathrm{T}}$ into the final global representation *Hierarchical coding vector* $\mathbf{HCV} \in \mathbb{R}^{M \times 2N}$ using the Fisher vector (FV) coding method. The pipeline of the Fisher coding layer is shown in Figure 3. All the semi-local features were decorrelated using Whitening technology before being fed into the Fisher coding layer.

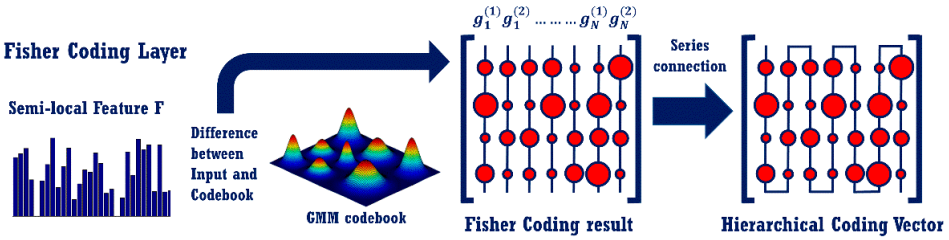

Figure 3. The pipeline of the Fisher coding layer.

The FV coding method is based on fitting a parametric generative model (e.g., GMM) to the input semi-local features **F** and then encoding the derivatives of the log-likelihood of the model with respect to its parameters [25]. The GMMs with diagonal covariance are used in our HCV framework, leading to a HCV representation that captures the Gaussian mean (1st) and variance (2nd) differences between the input semi-local features **F** and each of the GMM centers.

$$g_n^{(1)} = \frac{1}{T\sqrt{w_n}} \sum_{t=1}^{T} \alpha_t(n) \left(\frac{\mathrm{f}_t - \mu_n}{\sigma_n} \right) \tag{6}$$

$$g_n^{(2)} = \frac{1}{T\sqrt{2w_n}} \sum_{t=1}^{T} \alpha_t(n) \left(\frac{(\mathrm{f}_t - \mu_n)^2}{\sigma_n^2} - 1 \right) \tag{7}$$

where $\{w_n, \mu_n, \sigma_n\}_n$ are the respective mixture weights, means, and diagonal covariance of the GMM codebook $\mathbf{B_2}$ = (b_1,b_2,...b_n,...,b_N)$\in \mathbb{R}^{M \times N}$. f_t is one semi-local feature fed into the Fisher coding layer and T is the number of the semi-local features. $\alpha_t(n)$ is the soft assignment weight of the t-th semi-local features f_t to the n-th Gaussian.

$$\alpha_t(n) = \frac{w_n N(\mathrm{f}_t; \mu_n, \sigma_n)}{\sum_{n=1}^{N} w_n N(\mathrm{f}_t; \mu_n, \sigma_n)} \tag{8}$$

where $N(\mathrm{f}_\mathrm{t}; \mu_n, \sigma_n)$ is a M-dimensional Gaussian distribution and N is the size of GMM codebook. Finally, global representation **HCV**$\in \mathbb{R}^{M \times 2N}$ is obtained by stacking the first and second differences:

$$\mathbf{HCV}: \; G = [g_1^{(1)}, g_1^{(2)}, g_2^{(1)}, g_2^{(2)}, \cdots g_n^{(1)}, g_n^{(2)}, \cdots, g_N^{(1)}, g_N^{(2)}] \tag{9}$$

The output vector is subsequently normalized using the power + L_2 scheme, and serves as the final scene representation of HCV.

4. Experiment

We now evaluate the effectiveness of the proposed HCV framework and traditional FV for remote sensing land-use scene classification using two standard public databases, the 21-class Land Use (LU) database and the RSSCN7 [26] database. The classification performances of the proposed method are compared with several state-of-the-art methods.

4.1. Experimental Data and Setup

The 21-Class Land Use (LU) database [1] is one of the first publicly available geographical image databases (http://vision.ucmerced.edu/datasets.html) with ground truth, which is collected by University of California at Merced Computer Vision Lab (UCMCVL). The database consists of 21 land-use classes, and each class contains 100 images of the same size (*i.e.*, 256 pixels × 256 pixels). The pixel resolutions of all images are 30 cm per pixel. Sample images of each land-use class are shown in Figure 4. To be consistent with other researchers' experimental settings on the LU database [1,27–29], the database was randomly partitioned into five equal subsets. Each subset contained 20 images from each land-use category. Four subsets were used for training, and the remaining subset was used for testing.

The RSSCN7 database [26] is the recently public remote sensing database (https://sites.google.com/site/qinzoucn/documents) and was released in 2015. It contains 2800 remote sensing scene images that are from seven typical scene categories. There are 400 images with sizes of 400 × 400 pixels for each class. Each scene category is of four different scales with 100 images per scale. Sample images from RSSCN7 are shown in Figure 5. The same experimental setup in [26] is used. Half of the images in each category were fixed for training and the rest for testing.

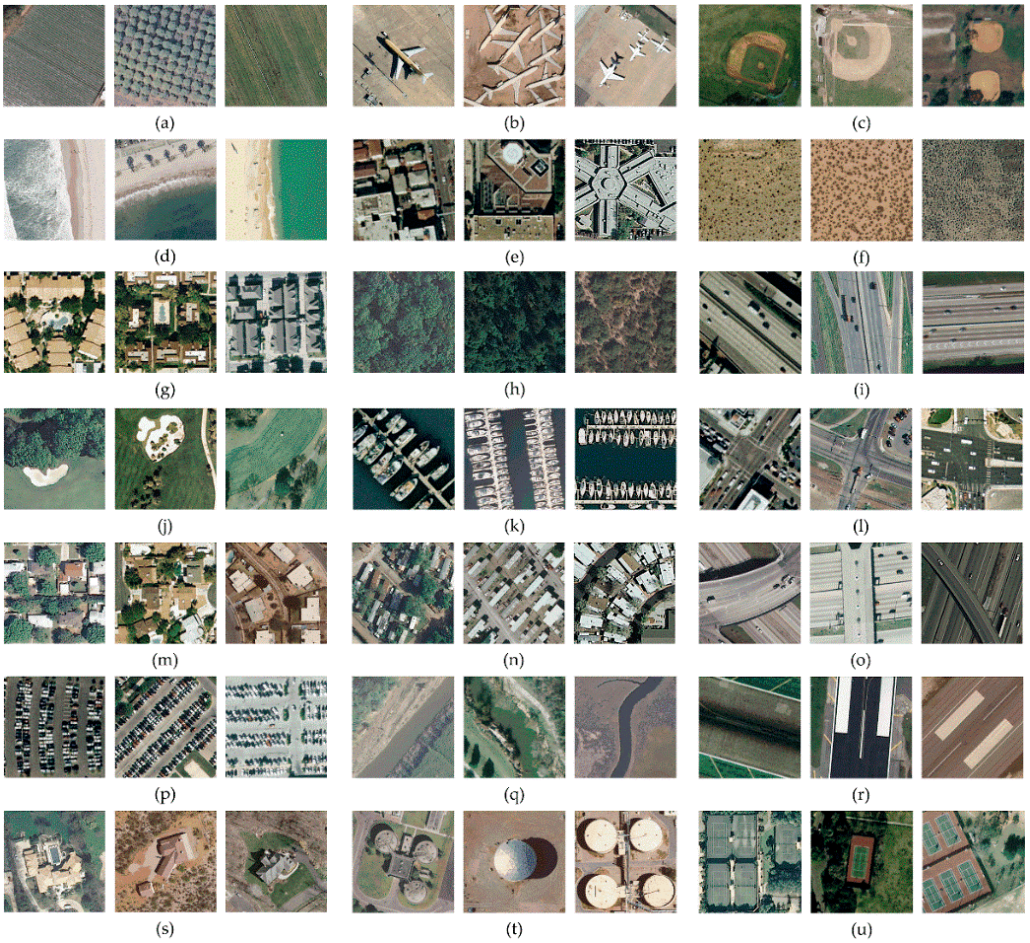

Figure 4. Sample images from each of the 21 categories in the Land Use (LU) database: (**a**) agricultural; (**b**) airplane; (**c**) baseball diamond; (**d**) beach; (**e**) buildings; (**f**) chaparral; (**g**) dense residential; (**h**) forest; (**i**) freeway; (**j**) golf course; (**k**) harbor; (**l**) intersection; (**m**) medium density residential; (**n**) mobile home park; (**o**) overpass; (**p**) parking lot; (**q**) river; (**r**) runway; (**s**) sparse residential; (**t**) storage tanks; (**u**) tennis courts.

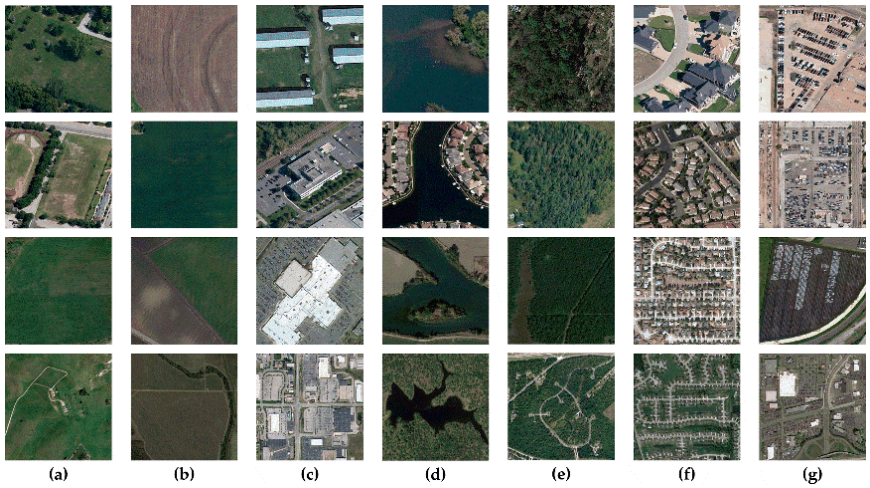

Figure 5. Sample images from the RSSCN7 database: (**a**) grassland; (**b**) farmland; (**c**) industrial and commercial regions; (**d**) river and lake; (**e**) forest field; (**f**) residential region; (**g**) parking lot. There are four scales, from top to bottom (in rows): 1:700, 1:1300, 1:2600, and 1:5200.

In the paper, we adopted Scale Invariant Feature Transformation (SIFT) as the local feature and the SIFT features were extracted from the interest point every six pixels in both the x and y directions under four scales (16, 24, 32, 48). The one *vs.* rest linear SVM classifier was employed in our experiments. The experiments were repeated ten times by randomly selecting the training and testing data with the experimental settings above. The average classification accuracy was set as the evaluation index.

4.2. Experimental Results

We evaluated the classification performance by the default parameters on the two databases. On the LU database, the classification accuracy of our proposed HCV was 90.5%. We also evaluated the traditional FV [7] with the same size of the GMM codebook in HCV. The classification accuracy of the traditional FV was 88.2%. On the RSSCN7 database, the results were similar (*i.e.*, HCV: 84.7% and FV 82.6%). On the two databases, the HCV achieved better performance than the traditional FV, which has shown great success in computer vision [7,20,24,25,30].

Furthermore, the proposed HCV also provided complementary information to the traditional FV. We used the multiple kernel learning [31] method with the average kernel to combine HCV with FV. When combining FV and HCV, we achieved a mean classification accuracy of 91.8% on the LU database and 86.4% on the RSSCN7 database.

To further investigate the performance of HCV, FV, and the combination of the two, we illustrate the per-class accuracies of the LU database in Figure 6.

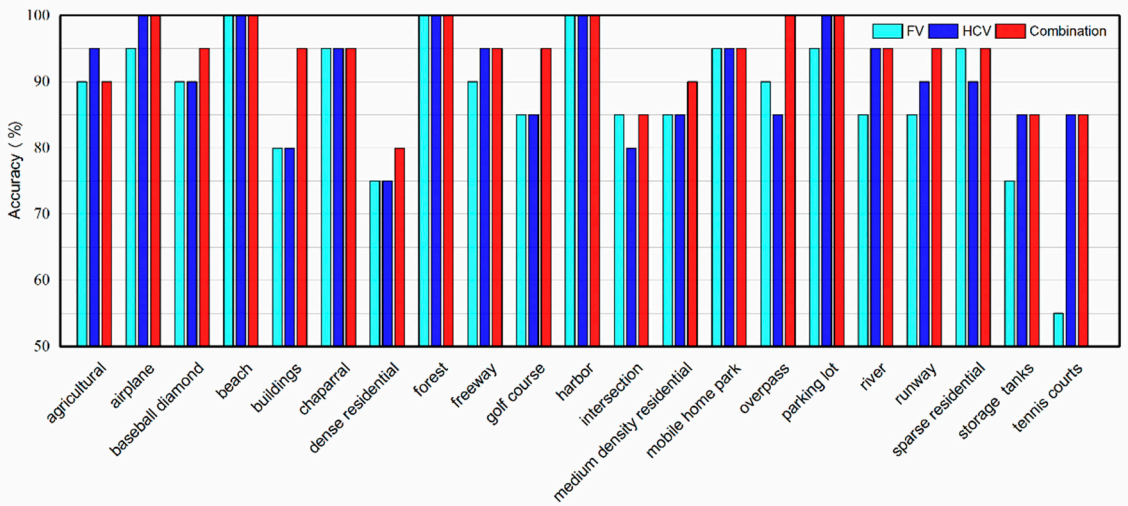

Figure 6. Comparison of the pre-class accuracies of Hierarchical Coding Vector (HCV) with the Fisher Vector (FV) and the combination of the two on the LU database.

From Figure 6, we observe that the proposed HCV is effective for almost all geographical classes on the LU database. Except for the intersection, overpass, and sparse residential categories, the HCV has better or comparable performance to FV in all other categories. The performance improvement is especially profound over the Tennis Courts category, which is approximately 30%, as shown in Figure 6.

Figure 7 shows some geographical images from three categories on the LU database that were predicted correctly by HCV, but not by the traditional FV. The traditional FV misclassified the two images in Figure 7a as buildings and the two images in Figure 7b as runways. The rivers in the Figure 7b do not have any curves and can easily be misclassified as runways, even by a human observer. The two images in Figure 7a are similar to buildings, and the storage tanks are not in a conspicuous position. The four images in Figure 7c were misclassified as other classes (e.g., parking lot, river, and sparse residential) by the traditional FV. Those images contain visually deceptive information, which makes

the recognition challenging. The correct classification requires sufficient semantic information. HCV described those geographical images correctly through higher level representation of richer semantic information by hierarchical coding structure.

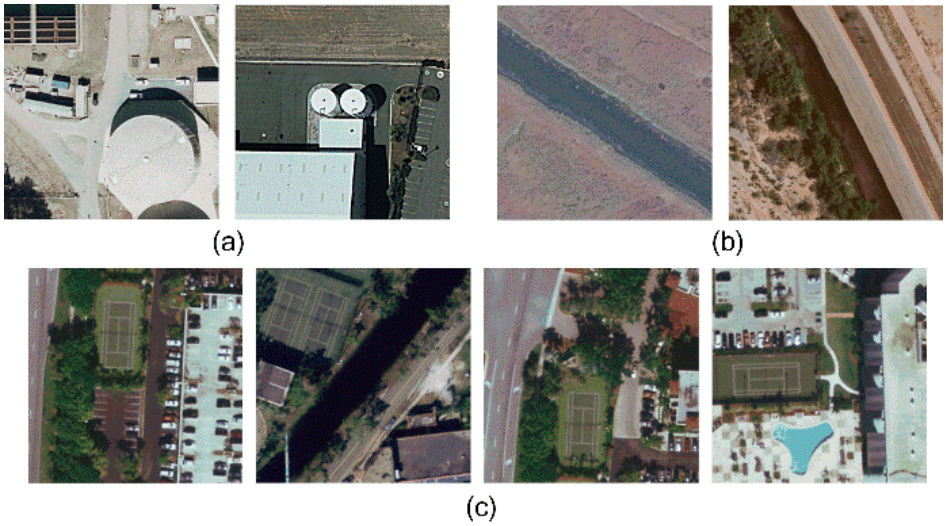

Figure 7. Some images are predicted correctly by the HCV, but not by the FV on the LU database: (**a**) storage tanks images; (**b**) river images; (**c**) tennis courts images.

Moreover, the classification performance was improved by the combination for almost all geographical classes, as shown in Figure 6, due to the complementarity between FV and HCV. By using HCV to capture the deep visual semantic information and combining FV with HCV, our method (FV + HCV) achieved very good classification performance.

4.3. Evaluation of the Parameters in HCV

In the proposed Hierarchical Coding Vector (HCV) framework, the dictionary size of each of the coding layers, the key parameter β in the SA coding method, and the different spatial structures in local pooling are the important parameters. We investigated these parameters on the LU database and chose the optimum HCV parameters for scene level land-use classification. The evaluation was carried out for one parameter at a time and the other ones were fixed to the default. The most important parameter (*i.e.*, the codebook size of each coding layers) was investigated first and then we studied the key parameter β. In the end, the different spatial structures in local pooling were evaluated. Furthermore, we also evaluated the effect of the number of coding layers.

4.3.1. The Effect of Different Codebook Size

First, we estimated the optimum codebook sizes for each coding layer. The BOVW coding layer used the K-means codebook. The FV coding layer used the GMM codebook. We set $\beta = 0.01$ and the spatial structure as 2×2. The classification results of HCV with varying K-means/GMM codebook size on the LU database are listed in Table 2.

The sizes of the K-means and GMM codebooks are critical to the classification performance of HCV. Too small of a codebook cannot capture enough graphical statistics. Meanwhile, too large of a codebook can cause over-partitioning in the descriptor space. As shown in Table 2, the classification performance increased with the larger codebooks and reached a plateau (even decreased) when the codebooks' size exceeded a threshold for both K-means and GMM codebooks. Based on the experimental results, we chose the codebook size of K-means/GMM as 1000/8 in terms of the classification accuracy and computational complexity.

Table 2. Classification accuracy (%) of HCV with varying K-means/GMM codebook size on the LU database.

K-means/GMM	2	4	8	16	32
50	71.55	76.98	81.62	84.33	87.62
100	77.05	82.02	85.79	85.98	87.93
200	83.00	84.74	87.31	88.10	88.21
600	86.86	88.69	89.50	89.45	88.81
1000	88.36	89.29	90.00	88.57	88.40
1400	88.26	89.76	89.17	88.49	88.36

4.3.2. The Key Parameter β in the SA Coding Method

To show the effect of β on the HCV more clearly, we selected five images from five different land-use classes and visualized those coding results under different values of β. The visualization result is illustrated in Figure 8. Each vertical column represents the coding result with a different value of β for the same image. Each horizontal row represents the coding result with the same value of β for the different images. The left-most column is the visualization of the semi-local feature $\mathbf{f_t} \in \mathbb{R}^M$ output by the BOVW coding layer, and the remaining part is the visualization of HCV. The visualizations of the semi-local feature $\mathbf{f_t}$ (output of the BOVW coding layer) for the five different images are quite similar, so we have only displayed one representative of the feature $\mathbf{f_t}$ for each value of β in Figure 8.

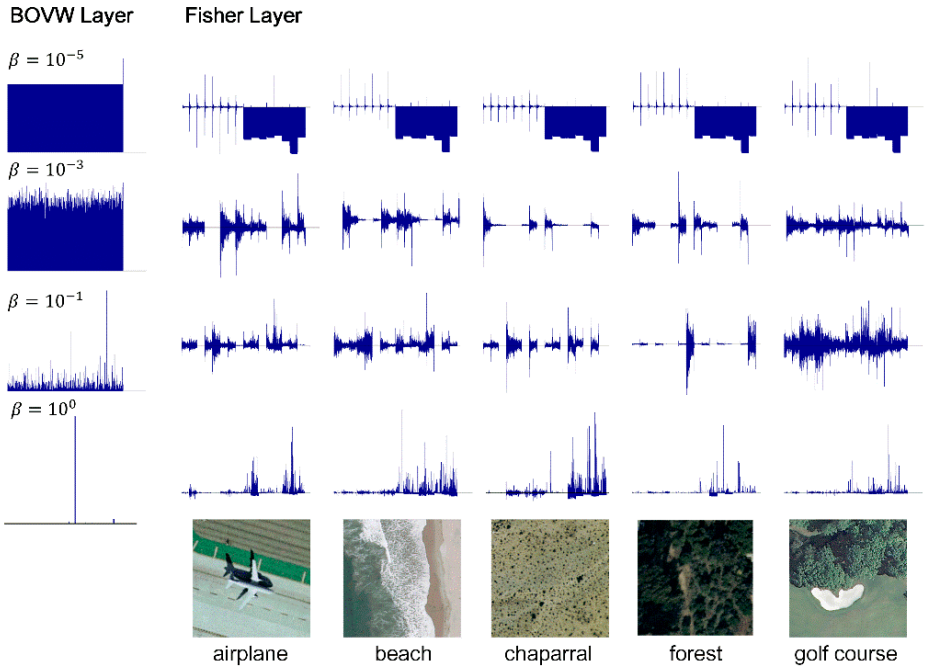

Figure 8. Visual coding result of the Hierarchical Coding Vector (HCV) of different parameters on the LU database. Each vertical column represents the coding result of a different β for the same image. Each horizontal row represents the coding result of same β for different images.

When β is too small (e.g., $\beta = 10^{-5}$), SA coding is not sensitive to the distance $\hat{e}$ between descriptors x_k and codeword b_m. The codebook is almost activated in the same intensity. The BOVW coding layer cannot capture enough discriminable image information, and the HCV is not able to represent the complex semantic structure. We can observe that the BOVW layer output seems to be meaningless and the HCV of the five images are very similar in this situation, as shown in Figure 8. It is easy to cause misclassification. With the increase of β, the SA coding method can express the distance information $\hat{e}$ appropriately and the BOVW layer output appears to be undulating. The HCV

output by the Fisher coding layer of different images shows the obvious difference and increasing classification performance is expected. When β becomes too large, the SA coding response decreases rapidly with the increasing distance $\hat{e}$. Figure 8 shows that the sparsity of the BOVW layer output increases and the HCV of the five images becomes similar. The increasing sparsity is a challenge for the Fisher vector coding and weakens the discriminability of the HCV.

With the visualization result, we found the value of parameter β is critical to the classification performance of HCV. We evaluated the effect of different values of β on the classification performance of HCV and determined the optimal value. Sizes of 1000 and 8 were our choices for the K-means codebook and GMM codebook, respectively. The spatial structure is 2×2. The classification accuracy of HCV for the different parameters β on the LU database is shown in Figure 9.

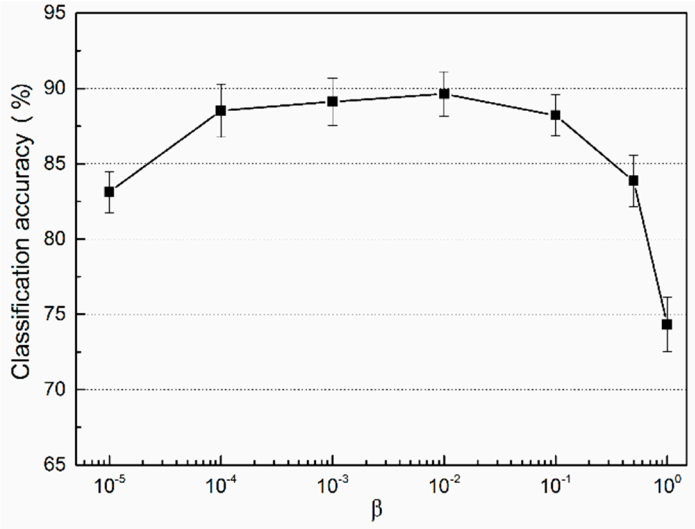

Figure 9. Evaluation of the effect on the classification accuracy of HCV of the parameter β on the LU database.

The experiment results confirm our previous analysis. The parameter β is a key factor for HCV. Too small or too large value of a β weakens the classification performance by a large margin. Based on the results in Figure 9, we chose $\beta = 0.01$.

4.3.3. The Effect of Different Spatial Structures in Local Pooling

Local pooling aggregates the coding results of SIFT features under four scales inside the spatial structure. We evaluated the effect of different spatial structures on the classification performance of HCV in this section. The five different spatial structures (1×1, 2×2, 3×3, 4×4, and 5×5) were evaluated on the LU database. The Max-pooling method was applied. We set $\beta = 0.01$ and the size of K-means/GMM codebook was 1000/8. The classification performance of different spatial structures for HCV is illustrated in Figure 10.

As seen from Figure 10, the classification performance of HCV gradually decreases with the larger spatial structure, which can be explained by two factors: (1) the increasing spatial structure leads to the repeated expression of some mutation points, creating a new challenge in the FV coding; and (2) the number of the input points of the Fisher coding layer proportionately decreases with the larger spatial structure, weakening the discriminability of the HCV.

Based on the experiment results, the spatial structure 1×1 was applied in our HCV framework. Inside the 1×1 spatial structure, the coding results $d_k \in \mathbb{R}^M$ of the SIFT features $x_k \in \mathbb{R}^D$ under four scales were aggregated to semi-local feature $f_t \in \mathbb{R}^M$ using the Max-pooling methods.

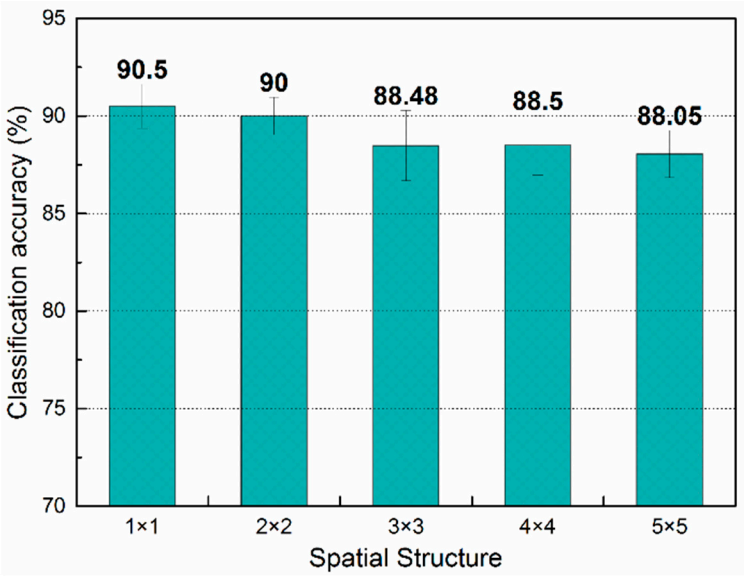

Figure 10. Classification performance of different spatial structures for HCV on the LU database.

4.3.4. The Effect of the Number of Coding Layers

We also evaluated the effect of the number of coding layers. The classification accuracy over different number of coding layers in the HCV framework is shown in Figure 11. One coding layer represents only the Fisher coding layer used in the HCV. Two coding layers contain one BOVW coding layer and one Fisher coding layer. Similarly, the three coding layers consist of two BOVW coding layers and one Fisher coding layer.

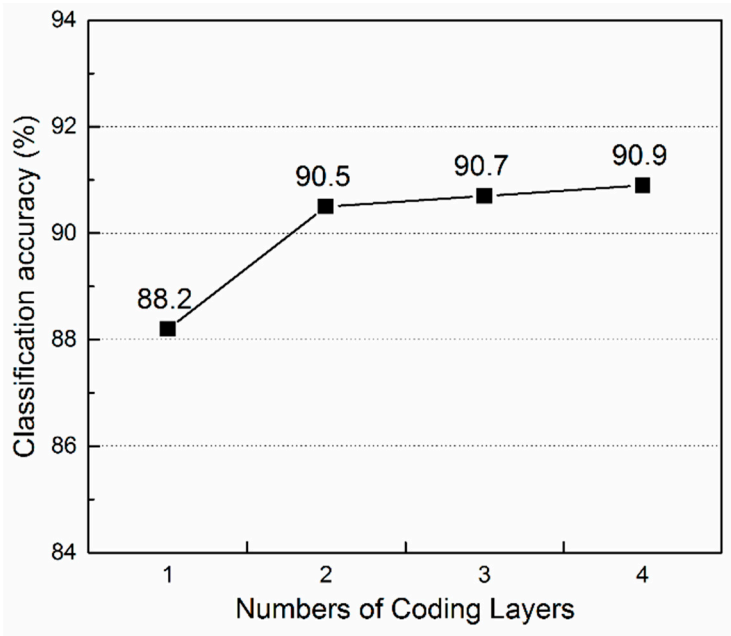

Figure 11. Evaluation of the effect on the classification accuracy of the number of coding layers.

From Figure 11, we can observe that the performance has been improved significantly from one layer (88.2) to two layers (90.5) due to the hierarchical structure. However, as the layer number continued to increase, there was no further substantial improvement in the classification performance due to the parameter tuning. With an increasing number of layers, the number of parameters to tune

grows exponentially. The lack of the good parameter tuning for the larger models (*i.e.*, three layers and four layers) prevented the optimal performance of HCV. This is a problem that needs to be solved in the future.

For a good tradeoff between effectiveness and efficiency, we only used two coding layers (*i.e.*, one BOVW coding layer and one Fisher coding layer) to perform scene level land-use classification in this paper.

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

To prove the effectiveness of our proposed method, a comparison of its performance with the state-of-the-art performance reported in the literature was performed on the two public databases under the same experimental setup. The comparison results of LU database are reported in Table 3.

Although the MS-CLBP described in [27] achieves comparable performance with HCV, the Extreme Learning Machine (ELM) and Radial Basis Function (RBF) nonlinear kernel were used in their approach. The nonlinear classifier has to bear additional complexity and bear the poor scalability, which is important for real application. Our proposed method relies on the one *vs.* rest linear SVM classifier. The linear classifier makes the framework simpler and more conducive to practical application. The classification performance of our method should be improved further with a sophisticated classifier.

As shown in Table 3, our method (FV + HCV) outperformed the current state-of-the-art results on the LU database, which demonstrates the effectiveness of our method (FV + HCV) for remotely sensed land use classification. Furthermore, the statistical z-test was used to test whether the performance improvement is meaningful. The z-test is a hypothesis test based on the Z-statistic, which follows the standard normal distribution under the null hypothesis [32]. It is often used to determine whether the difference between two means is significant. When the $Z \geqslant 1.96$, the difference is significant ($p \leqslant 0.05$). On the contrary, when the $Z < 1.96$, the difference is not significant ($p > 0.05$). A comparison of our method to other methods is provided in Table 3; $p \leqslant 0.05$ for our method (FV + HCV). The minimum value of Z is 1.99 when compared to MS-CLBP and the p is still less than 0.05. The performance boost of our method is statistically significant.

Table 3. Comparison of our approach (FV + HCV) with the state-of-the-art performance reported in the literature on the LU database under the same experimental setup: 80% of images from each class are used for training and the remaining images are used for testing. The average classification accuracy (mean $\pm$ SD) is set as the evaluation index.

Method	Accuracy (%)
BOVW [1]	76.8
SPM [1]	75.3
BOVW + spatial co-occurrence kernel [1]	77.7
Color Gabor [1]	80.5
Color histogram [1]	81.2
SPCK [4]	73.1
SPCK + BOW [4]	76.1
SPCK + SPM [4]	77.4
Structural texture similarity [33]	86.0
Wavelet BOVW [29]	87.4 $\pm$ 1.3
Unsupervised feature learning [34]	81.1 $\pm$ 1.2
Saliency-guided feature learning [35]	82.7 $\pm$ 1.2
Concentric circle-structured BOVW [2]	86.6 $\pm$ 0.8
Multifeature concatenation [36]	89.5 $\pm$ 0.8
Pyramid-of-spatial-relations [3]	89.1
CLBP [27]	85.5 $\pm$ 1.9
MS-CLBP [27]	90.6 $\pm$ 1.4
HCV	90.5 $\pm$ 1.1
Our method	91.8 $\pm$ 1.3

The comparison results for RSSCN7 database are listed in Table 4. It was observed that our method improved the performance significantly with a noticeable margin on the RSSCN7 database. We also used the statistical z-test and the result showed that the performance boost is statistically significant. It should be noted that our method here directly used the parameters tuning results on the LU database, thereby showing that this parameters set has some reasonable applicability to other datasets. The classification performance on the RSSCN7 database should be further improved by integral fine parameter tuning.

Table 4. Comparison of our approach (FV + HCV) with the state-of-the-art performance reported in the literature on the RSSCN7 database under the same experimental setup: half of images from each class are used for training and the rest are used for testing. The average classification accuracy (mean $\pm$ SD) is set as the evaluation index. DBN: Deep Belief Networks.

Method	Accuracy (%)
GIST *	69.5 ± 0.9
Color histogram *	70.9 ± 0.8
BOVW *	73.1 ± 1.1
LBP *	75.3 ± 1.0
DBN based feature selection [26]	77.0
HCV	84.7 ± 0.7
Our method	86.4 ± 0.7

* Our own implementation.

4.5. Computational Complexity

Many approaches with a nonlinear classifier have to pay a computational complexity $O(n^2)$ or $O(n^3)$ in the train phase and $O(n)$ in the testing phase, where n is the training size. It implies a poor scalability for the real application. Our method, using a simple linear SVM, reduces the training complexity to $O(n)$, and obtains a constant complexity in testing, while still achieving a superior performance. In the end, we evaluated the computation complexity of our method (HCV + FV) and used the 21-class land-use (LU) database to obtain the processing time. Our codes are all implemented in MATLAB 2014a and were run on a computer with an Inter (R) Xeon (R) CPU E5-2620 v2 @ 2.1GHZ and 32G RAM in a 64-bit Win7 operation system. As observed from our experiment, the train phase takes about 27 min and the average processing time for a test remote sensing image (size of 256×256 pixels) is 0.55 ± 0.02 second (including dense local descriptors extraction, HCV, and FV coding to get the final representation).

5. Conclusions

In this paper, we proposed using Hierarchical Coding Vectors (HCV), a novel representation based on hierarchically coding structures, for scene level land-use classification. We have shown that the traditional coding pipelines are amenable to stacking in multiple layers. Building a hierarchical coding structure is sufficient to significantly boost the performance of these shallow encoding methods. The experimental results on the LU and RSSCN7 databases demonstrate the effectiveness of our HCV representation. By combining HCV with the traditional Fisher vectors, our method (FV + HCV) outperforms the current state-of-the-art methods on the LU and RSSCN7 databases.

Acknowledgments: This work is supported by the National Science and Technology Major Project of China (2012ZX10004801) and the National Biological Cross Pre-research Foundation of China (9140A26020314JB94409).

Author Contributions: Hang Wu and Jinggong Sun conceived and designed the experiments; Hang Wu and Baozhen Liu performed the experiments; Wenchang Zhang and Baozhen Liu analyzed the data; Hang Wu and Weihua Su wrote the paper.

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

Abbreviations

The following abbreviations are used in this manuscript:

HCV	Hierarchical Coding Vector
BOVW	Bag of Visual Words
HOG	Histogram of Oriented Gradient
LBP	Local Binary Pattern
SA	Soft Assignment
FV	Fisher Vectors
VLAD	Vector of Locally Aggregated Descriptors
LU	21-Class Land Use
GMM	Gaussian Mixture Model
DNN	Deep Neural Network
SIFT	Scale Invariant Feature Transformation
SPCK	Spatial Pyramid Co-occurrence Kernel
CLBP	Completed Local Binary Pattern
HA	Hard Assignment
LCC	Local Coordinate Coding
LLC	Locality-constrained Linear Coding
SVC	Super Vector Coding
UCMCVL	University of California at Merced Computer Vision Lab
SVM	Support Vector Machine
ELM	Extreme Learning Machine
RBF	Radial Basis Function
DBN	Deep Belief Networks

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 on Advances in Geographic Information Systems, San Jose, CA, USA, 3–5 November 2010.
2. Zhao, L.-J.; Tang, P.; Huo, L.-Z. Land-use scene classification using a concentric circle-structured multiscale bag-of-visual-words model. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 4620–4631. [CrossRef]
3. Chen, S.; Tian, Y. Pyramid of spatial relatons for scene-level land use classification. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 1947–1957. [CrossRef]
4. Yang, Y.; Newsam, S. Spatial pyramid co-occurrence for image classification. In Proceedings of the 2011 IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011.
5. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–8 December 2012; pp. 1097–1105.
6. Liu, L.; Wang, L.; Liu, X. Defense of soft-assignment coding. In Proceedings of the 2011 IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011.
7. Sánchez, J.; Perronnin, F.; Mensink, T.; Verbeek, J. Image classification with the fisher vector: Theory and practice. *Int. J. Comput. Vis.* **2013**, *105*, 222–245. [CrossRef]
8. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
9. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–25 June 2005.
10. Oliva, A.; Torralba, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vis.* **2001**, *42*, 145–175. [CrossRef]
11. Ojala, T.; Pietikäinen, M.; Harwood, D. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognit.* **1996**, *29*, 51–59. [CrossRef]
12. Peng, X.; Wang, L.; Wang, X.; Qiao, Y. Bag of Visual Words and Fusion Methods for Action Recognition: Comprehensive Study and Good Practice. Available online: http://arxiv.org/abs/1405.4506 (accessed on 18 May 2016).
13. Yu, K.; Zhang, T.; Gong, Y. Nonlinear learning using local coordinate coding. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 6–10 December 2009; pp. 2223–2231.

14. Wang, J.; Yang, J.; Yu, K.; Lv, F.; Huang, T.; Gong, Y. Locality-constrained linear coding for image classification. In Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010.
15. Jégou, H.; Perronnin, F.; Douze, M.; Sanchez, J.; Perez, P.; Schmid, C. Aggregating local image descriptors into compact codes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 1704–1716. [CrossRef] [PubMed]
16. Lin, C.J.; Hsu, C.-W.; Chang, C.-C. A Practical Guide to Support Vector Classification. Available online: https://www.cs.sfu.ca/people/Faculty/teaching/726/spring11/svmguide.pdf (accessed on 18 May 2016).
17. Huang, G.-B.; Zhu, Q.-Y.; Siew, C.-K. Extreme learning machine: Theory and applications. *Neurocomputing* **2006**, *70*, 489–501. [CrossRef]
18. Jaakkola, T.S.; Haussler, D. Exploiting generative models in discriminative classifiers. *Adv. Neural Inf. Process. Syst.* **1999**, 487–493.
19. 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]
20. Peng, X.; Zou, C.; Qiao, Y.; Peng, Q. Action recognition with stacked fisher vectors. *Comput. Vis.* **2014**, *8693*, 581–595.
21. Arthur, D.; Vassilvitskii, S. K-means++: The advantages of careful seeding. In Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, New Orleans, LA, USA, 7–9 January 2007.
22. Murray, N.; Perronnin, F. Generalized max pooling. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 January 2014.
23. Chatfield, K.; Lempitsky, V.S.; Vedaldi, A.; Zisserman, A. The devil is in the details: An evaluation of recent feature encoding methods. In Proceedings of the BMVC, Dundee, UK, 29 August–2 September 2011; p. 8.
24. Perronnin, F.; Sánchez, J.; Mensink, T. Improving the fisher kernel for large-scale image classification. *Comput. Vis.* **2010**, *6314*, 143–156.
25. Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep fisher networks for large-scale image classification. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–10 December 2013; pp. 163–171.
26. Zou, Q.; Ni, L.; Zhang, T.; Wang, Q. Deep learning based feature selection for remote sensing scene classification. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 2321–2325. [CrossRef]
27. Chen, C.; Zhang, B.; Su, H.; Li, W.; Wang, L. Land-use scene classification using multi-scale completed local binary patterns. *Signal Image Video Process.* **2016**, *4*, 745–752. [CrossRef]
28. Mekhalfi, M.L.; Melgani, F.; Bazi, Y.; Alajlan, N. Land-use classification with compressive sensing multifeature fusion. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 2155–2159. [CrossRef]
29. Zhao, L.; Tang, P.; Huo, L. A 2-D wavelet decomposition-based bag-of-visual-words model for land-use scene classification. *Int. J. Remote Sens.* **2014**, *35*, 2296–2310.
30. Simonyan, K.; Parkhi, O.M.; Vedaldi, A.; Zisserman, A. Fisher vector faces in the wild. In Proceedings of the BMVC, Bristol, UK, 9–13 September 2013.
31. Gehler, P.; Nowozin, S. On feature combination for multiclass object classification. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 27 September–4 October 2009.
32. Mandel, J. *The Statistical Analysis of Experimental Data*; Courier Corporation: New York, NY, USA, 2012.
33. Risojević, V.; Babić, Z. Aerial image classification using structural texture similarity. In Proceedings of the 2011 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), 2011, 14–17 December 2011; IEEE: Bilbao, Spain.
34. Cheriyadat, A.M. Unsupervised feature learning for aerial scene classification. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 439–451. [CrossRef]
35. Zhang, F.; Du, B.; Zhang, L. Saliency-guided unsupervised feature learning for scene classification. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 2175–2184. [CrossRef]
36. Shao, W.; Yang, W.; Xia, G.-S.; Liu, G. A hierarchical scheme of multiple feature fusion for high-resolution satellite scene categorization. *Comput. Vis. Syst.* **2013**, *7963*, 324–333.