Satellite Image Classification via Two-Layer Sparse Coding With Biased Image Representation

Dengxin Dai and Wen Yang, *Member, IEEE*

***Abstract*—This letter presents a method for satellite image classification aiming at the following two objectives: 1) involving visual attention into the satellite image classification; biologically inspired saliency information is exploited in the phase of the image representation, making our method more concentrated on the interesting objects and structures, and 2) handling the satellite image classification without the learning phase. A two-layer sparse coding (TSC) model is designed to discover the "true" neighbors of the images and bypass the intensive learning phase of the satellite image classification. The underlying philosophy of the TSC is that an image can be more sparsely reconstructed via the images (sparse I) belonging to the same category (sparse II). The images are classified according to a newly defined "image-to-category" similarity based on the coding coefficients. Requiring no training phase, our method achieves very promising results. The experimental comparisons are shown on a real satellite image database.**

***Index Terms*—Satellite image classification, two-layer sparse coding (TSC), visual attention.**

I. Introduction

OVER the recent decades, overwhelming amounts of high-resolution satellite images have become available, enabling accurate Earth observations and topographic measurements. But the problem of how we can effectively use these images with a computer is still far from being addressed. In the literature, there mainly exist two challenges in the high-resolution satellite image classification.

On the one hand, with the increasing spatial resolution, more details (structures and objects) on the Earth's surface emerge in satellite images, invalidating the features proposed for the low-resolution satellite image classification. The intensity and texture cues of entire scenes have proven sufficient for the low-resolution satellite image classification [1], [2]. However, for high-resolution satellite images, the structures and objects often dominate their categories, and the whole image representation may sometimes lead to mistakes. Fig. 1 shows that the examples in different categories may have similar global responses while those in the same category may be different. This observation entails the image representation of focusing on the salient structures for high-resolution satellite images.

Manuscript received January 25, 2010; revised March 4, 2010, April 10, 2010, and May 14, 2010; accepted May 26, 2010. Date of publication August 9, 2010; date of current version December 27, 2010. This work was supported in part by the National Natural Science Foundation of China under Grants 40801183 and 60890074 and in part by the Laboratory for Information Engineering in Surveying, Mapping, and Remote Sensing (LIESMARS) Special Research Funding.

The authors are with the School of Electronic Information and State Key LIESMARS, Wuhan University, Wuhan 430079, China (e-mail: dxdai@mail.whu.edu.cn; yangwen@whu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LGRS.2010.2055033

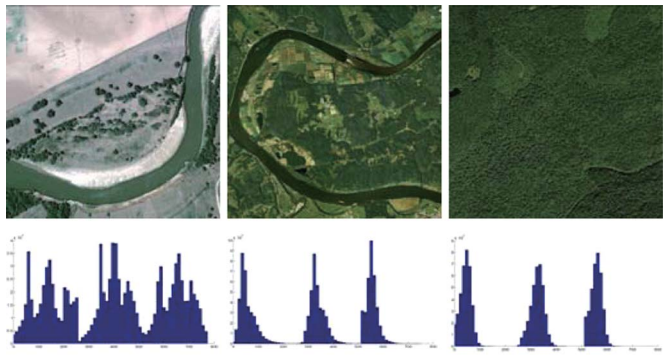

Fig. 1. Illustration of structures and objects dominating the image category. The whole image-based representation may sometimes lead to mistakes for the image classification.

On the other hand, tremendously acquired satellite images and increasingly demanding category requirements make most classification methods incompetent. State of the art as it is, modern machine learning algorithms (e.g., support vector machine (SVM), boosting, etc.) [3], [4] are incapable of scaling up to thousands of categories as they require a sophisticated learned model for each category. Recently, nonparametric nearest-neighbor schemes have attracted great attention in the image classification community [5], [6]. Despite its popularity, it is still unclear how to define the "true" neighbors of the optical satellite images.

Motivated by the two challenges, in this letter, we present a method for the high-resolution satellite image classification, which puts more resources on the salient structures and relaxes the learning phase. The main idea is that our image representation, hereinafter referred to as the biased image representation, is guided by a bottom-up biologically inspired saliency measure. Then, we recognize the image categories based on the coding coefficients of a novel two-layer sparse decomposition model.

The underlying mechanism of our method accords with the vision mechanism of human beings strictly and is described as follows. The higher visual processes of human beings appear to select a subset of the available sensory information, such as intensity, color, texture, structure, and objects, before further processing [7]. These selections are implemented by the so-called "focus of attention," which scans the scene in a fast saliency-driven[1] manner [7]. Considering the neurons in the focus of attention to form an overcomplete dictionary for basic visual elements, the stimulation of these neurons by a given

[1]Saliency is the component in visual scenes which more easily stimulates the human visual system.

image is highly sparse. These discoveries inspire us that the combination of the biased image representation and the sparse coding is similar to the visual mechanism of human beings, thus suitable for the image classification. The motivation of adding another sparsity to the conventional sparse coding is that an image can be more sparsely reconstructed via the images belonging to the same category.

The remainder of this letter is organized as follows. We describe the biased image representation in Section II and formulate the two-layer sparse coding (TSC) and the image classification in Section III. Then, we devote Section IV to the comparison experiments. Finally, we conclude this letter in Section V.

II. Biased Image Representation

Whatever strategies are employed, the goal of the image representation is to reserve enough discrimination power while reducing the amount of information. In this letter, we attempt to investigate whether saliency-based attention model is useful in the high-resolution satellite image representation.

A. Biologically Inspired Saliency Map

There exist many ways to learn a saliency map. In this letter, we employ the method in [8] to obtain the saliency map. The principle of the method is summarized in the following paragraphs.

An image is first interpolated to its finer scale and then subsampled into a Gaussian pyramid, where several independent channels, red (R), green (G), blue (B), yellow (Y), intensity (I), and local orientation (O_θ), are obtained. R, G, B, and Y are simple transforms of the red (r), green (g), and blue (b) values of the image: $R = r - (g+b)/2$, $G = g - (r+b)/2$, $B = b - (r+g)/2$, and $Y = r + g - 2(|r-g| + b)$. O_θ is extracted by using Gabor filters to the intensity of each pyramid level. By these simple operations, a set of "center-surround" features are constructed and normalized

$$\begin{aligned} I(c,s) &= \mathcal{N}\left(|I(c) \ominus I(s)|\right) \\ RG(c,s) &= \mathcal{N}\left(|(R(c)-G(c)) \ominus (G(s)-R(s))|\right) \\ BY(c,s) &= \mathcal{N}\left(|(B(c)-Y(c)) \ominus (Y(s)-B(s))|\right) \\ O(c,s,\theta) &= \mathcal{N}\left(|O(c,\theta) \ominus O(s,\theta)|\right) \end{aligned} \tag{1}$$

where $\ominus$ indicates the difference between the neighboring pyramid levels at the center (c) and the surround (s) and $\mathcal{N}(\bullet)$ is a normalization operator. Then, the feature maps are summed and normalized into three "saliency channels": $\bar{\mathcal{I}}$ for the intensity [obtained from I in (1)], $\bar{\mathcal{C}}$ for the colors [obtained from RG and BY in (1)], and $\bar{\mathcal{O}}$ for the orientation [obtained from O in (1)]. Finally, the contributions of the these independent features are linearly combined and normalized to yield the final saliency map

$$H = \frac{1}{3}\left(\mathcal{N}(\bar{\mathcal{I}}) + \mathcal{N}(\bar{\mathcal{C}}) + \mathcal{N}(\bar{\mathcal{O}})\right). \tag{2}$$

B. Biased Image Representation

After computing the saliency map, interesting objects would likely stand out. To reduce the solution complexity, in this letter, we use superpixels to represent the satellite images. For feature type t, $t \in \{1, 2, \ldots, T\}$, all extracted descriptors are clustered into K^t clusters (for example, words), where T is the number of feature types. The words of the regions are weighted by the extracted saliency map, which makes our representation more focused on the interesting structures and objects. Then, we summarize an image using a word-frequency histogram formed by counting all the weighted words in the image.

Specifically, given the ith image I_i, we denote its jth superpixel as $S_{i,j}$ and indicate the tth word-based representation for $S_{i,j}$ as $f^t_{i,j}$. Then, the weighted word histogram can be constructed as follows:

$$F^t_i(k) = \sum_{j=1}^{n_i} \bar{H}_{i,j}\delta\left[f^t_{i,j} = k\right], \qquad \forall k \in \{1, \ldots, K^t\} \tag{3}$$

where $\bar{H}_{i,j}$ is the average saliency map in $S_{i,j}$, n_i is the number of superpixels in image i, and the function $\delta[\bullet]$ is one if its argument is true and zero otherwise.

III. TSC and Image Classification

In this section, we first review the conventional sparse coding, hereinafter referred to as the one-layer sparse coding (OSC), and then demonstrate the principle of the TSC. Finally, the image classification based on the TSC is illustrated.

A. OSC

Sparse coding was heightened to a spotlight position in the statistical signal processing community, particularly after an exciting declaration that, when the solution is sparse enough, it can be efficiently solved by convex $\ell 1$-norm minimization [9].

Suppose to solve a linear equation: $x = D\alpha$, where $x \in \mathbb{R}^m$ is the descriptor to be reconstructed, $\alpha \in \mathbb{R}^n$ is the vector of the coding coefficients, and $D \in \mathbb{R}^{m \times n} (m < n)$ is the n bases in the dictionary. When the assumption that α is sparse enough holds, the equation can be solved by the following convex $\ell 1$-norm optimization

$$\arg\min_{\alpha} \|\alpha\|_1, \qquad \text{s.t. } x = D\alpha. \tag{4}$$

Specifically, given a set of labeled images $\mathcal{D} = \{I_1, I_2, \ldots, I_N\}$ and suppose that there are L categories, the corresponding category labels are $\mathcal{C} = \{c_1, c_2, \ldots, c_N\}$, where $c_i \in \{1, \ldots, L\}$. For the tth type of feature, we use x^t to represent the test image and embody the base dictionary D_t as $D^t = [F^t_1, F^t_2, \ldots, F^t_N] \in \mathbb{R}^{K^t \times N}$. Then, (4) is substantiated as follows:

$$\arg\min_{\alpha^t} \|\alpha^t\|_1, \qquad \text{s.t. } x^t = D^t\alpha^t. \tag{5}$$

The minimization of (5) guarantees the sparsity of the selected images, hereinafter referred to as the image sparsity, and the coefficients α^t can then be used for the image classification.

B. TSC

Since one test image only belongs to one category, it is natural to enforce the selected bases (labeled images) into as few categories as possible, which motivates the second layer of the sparsity referred to as the category sparsity.

Fig. 2. Some samples of the test high-resolution satellite image database. For each class, there are 50 samples, four of which are shown.

To introduce the category sparsity, we bring a set of coefficients, $\zeta^t = [\zeta_1^t, \zeta_2^t, \ldots, \zeta_L^t]^T$, to measure the weights of the selected categories

$$\zeta_l^t = \lambda \sum_{i=1}^{N} \delta[c_i = l]\alpha_i^t \tag{6}$$

where λ is a tuning parameter.

By defining a matrix

$$B^t = \begin{bmatrix} \delta_1^1 & \cdots & \delta_i^1 & \cdots & \delta_N^1 \\ \delta_1^2 & \cdots & \delta_i^2 & \cdots & \delta_N^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \delta_1^L & \cdots & \delta_i^L & \cdots & \delta_N^L \end{bmatrix}$$

where $\delta_i^l = \delta[c_i = l]$, we can reformulate (6) as

$$\zeta^t = \lambda B^t \alpha^t. \tag{7}$$

Then, we obtain the following optimization problem:

$$\arg\min_{\alpha^t, \zeta^t} \|\alpha^t\|_1 + \|\zeta^t\|_1, \qquad \text{s.t. } x^t = D^t\alpha^t,\ \zeta^t = \lambda B^t \alpha^t. \tag{8}$$

By letting

$$\tilde{x}^t = \begin{bmatrix} x^t \\ 0_{L\times 1} \end{bmatrix} \quad \tilde{\alpha}^t = \begin{bmatrix} \alpha^t \\ \zeta^t \end{bmatrix} \quad \tilde{D}^t = \begin{bmatrix} D^t, & 0_{K^t\times L} \\ \lambda B^t, & -I_{L\times L} \end{bmatrix}$$

we can reformulate (8) as a standard $\ell 1$-norm minimization problem

$$\arg\min_{\tilde{\alpha}^t} \|\tilde{\alpha}^t\|_1, \qquad \text{s.t. } \tilde{x}^t = \tilde{D}^t\tilde{\alpha}^t. \tag{9}$$

The minimization of (9) guarantees the image sparsity and the category sparsity simultaneously. We refer to this model as the TSC.

C. Image Classification

The satellite image classification is directly based on the coding coefficients α^t, $t \in \{1, \ldots, T\}$, obtained during the coding process. According to the tth type of feature, the probability of classifying image I into category l is defined as

$$P_l^t(I) = \frac{\sum_{i=1}^{N} \max\{\alpha_i^t \delta[c_i = l], 0\}}{\sum_{l'=1}^{L}\sum_{i=1}^{N} \max\{\alpha_i^t \delta[c_i = l'], 0\}}. \tag{10}$$

Prior to computing (10), all the feature vectors should be normalized to alleviate the coefficient sensitivities to the magnitudes of the feature vectors. Then, the test images can be classified by the maximum likelihood mechanism.

In order to boost further the classification performance, multiple (T) types of features are combined: The decision rule linearly combines the similarity, defined in (10), of each feature. Thus, the final classification rule is formulated as

$$\hat{l}(I) = \arg\max_{l} \sum_{t=1}^{T} \omega_t P_l^t(I) \tag{11}$$

where ω_t is determined by the variance of the Parzen Gaussian kernel $\mathbf{K}_t$ corresponding to the tth type of feature.

IV. Experiments

A. Data Set and Experiment Setting

To validate the proposed method, we collected a set of satellite images (Google Inc.) from Google Earth.[2] Some typical samples are displayed in Fig. 2. It contains the 12 categories of the physical scenes in the satellite imagery, including *Airport*, *Bridge*, *River*, *Forest*, *Meadow*, *Pond*, *Parking*, *Port*, *Viaduct*, *Residential area*, *Industrial area*, and *Commercial area*. For each class, there are 50 samples.

Three types of features, color, texture, and shape are used to characterize the properties of the segmented regions. For the color, we use the hue descriptor [10] and quantify all the hue descriptors into $K^c = 100$ clusters (words). For the texture, we compute the scale-invariant feature transform (SIFT) [11] at each pixel over a patch of 16 pixels $\times$ 16 pixels. The SIFT transforms a patch into a 128-dimension vector, which is invariant to image translation, scaling, and rotation. All the SIFT descriptors are quantized into $K^t = 200$ clusters. The

[2]http://earth.google.com/.

TABLE I
PERFORMANCE COMPARISONS ON THE 12-CATEGORY SATELLITE IMAGE DATA SET(%). IKNN IS IMAGE-BASED KNN AND CKNN MEANS CATEGORY-BASED KNN. U DENOTES UNBIASED REPRESENTATION, AND B INDICATES BIASED REPRESENTATION

	IKNN		CKNN		SVM		Sparse Coding	
	KL	HI	KL	HI	Linear kernel	RBF kernel	One-layer	Two-layer
U	72.4	73.0	74.0	74.1	79.2	81.4	75.2	80.1
B	75.5	76.6	77.2	78.6	83.3	84.7	78.9	84.2

moment invariants [12] are employed to measure the geometric cues where a set of region properties (shape and part cues) are captured. For each superpixel, we extract a 7-dimension vector to describe its shape, and all these vectors are clustered into $K^s = 100$ clusters. The K-means algorithm is used for our feature clustering. The formula of the TSC is resolved by the publically available ℓ^1-MAGIC package.[3]

In the experiments, 25 samples are randomly chosen from each category as the labeled images, leaving the others for the testing. We test the OSC and the TSC on 50 random training-test partitions and compare them to the K-nearest neighbor (KNN) classifier (both image-based and category-based) with Kullback–Leibler and histogram-intersection distances and to the SVM with both the linear and radial-basis-function kernels. The feature sets and the classification decision rule used for the KNN classifier are the same as that of our method, and the input of the SVM is the concatenation of all the descriptors. In our implementation, the library for SVMs[4] is used. We test the K of the KNN classifier from one to ten, search the "optimal" parameters for the SVM over a grid, and list the the best results of these classifiers.

B. Results

Table I lists the classification accuracies yielded by different methods using several input configurations. The accuracy values in the table are computed as the percentage of images classified to the correct categories.

Comparison Between Biased and Unbiased Representations: The first conclusion arrived at Table I is that the vision attention really plays an important role in the high-resolution satellite image classification. Compared to the unbiased representation, the image representation guided by the biological saliency extraction achieves superior performance, particularly on the object-dominated categories such as the river, bridge, etc.

Comparison Between One-Layer and Two-Layer Sparsities: The second conclusion drawn from Table I is that the involvement of the category sparsity is critical to the satellite image classification. Fig. 3 shows an exemplary comparison of the OSC and the TSC, where we can obviously observe that the category sparsity enforces the selected bases (labeled images) into very few categories which, to a great extent, benefits the image classification.

Comparison With Other Methods: The third conclusion reached from Table I is that, requiring no learning phase, our method significantly outperforms the KNN classifier and achieves comparable results to that of the SVM, a leading classifier. The validation of our method can be attributed to the combination of the biased image representation and the TSC, which is similar to the mechanism of the human visual system and can discover the true neighbors of the images.

[3]http://www.acm.caltech.edu/l1magic.

[4]Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

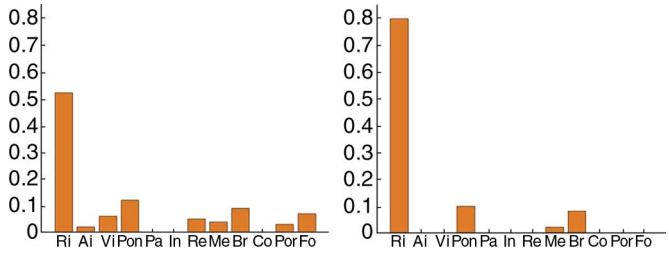

Fig. 3. Comparison between the OSC and the TSC for the river category. The left subfigure shows the results of the OSC, and the right one shows the results of the TSC. The results are obtained by summing the coding coefficients of all the test images belonging to the river category. All negative coefficients are set to zero.

V. CONCLUSION

The saliency-based attention model plays an important role in the high-resolution satellite image classification, and the proposed TSC model is an effective tool for the satellite image classification. The experiments on the real satellite data set show that our method can achieve very promising results without the learning phase.

REFERENCES

[1] L. A. Ruiz, A. Fdez-Sarría, and J. A. Recio, "Texture feature extraction for classification of remote sensing data using wavelet decomposition: A comparative study," in *Proc. Int. Archives Photogramm., Remote Sens. Spatial Inf. Sci.*, Istanbul, Turkey, Jul. 2004, pp. 1109–1115.

[2] C.-S. Li and V. Castelli, "Deriving texture feature set for content-based retrieval of satellite image database," in *Proc. Int. Conf. Image Process.*, Washington, DC, Oct. 1997.

[3] L. Bruzzone and L. Carlin, "A multilevel context-based system for classification of very high spatial resolution images," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 9, pp. 2587–2600, Sep. 2006.

[4] G. Camps-Valls and A. Rodrigo-González, "Classification of satellite images with regularized adaboosting of rbf neural networks," in *Proc. Speech, Audio, Image Biomed. Signal Process. Using Neural Netw.*, 2008, vol. 83, pp. 307–326.

[5] O. Boiman, E. Shechtman, and M. Irani, "In defense of nearest-neighbor based image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Anchorage, AK, Jun. 2008, pp. 1–8.

[6] A. Torralba, R. Fergus, and W. T. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1958–1970, Nov. 2008.

[7] J. K. Tsotsos, S. Culhane, W. Winky, Y. Lai, N. Davis, and F. Nuflo, "Modeling visual attention via selective tuning," *Artif. Intell.*, vol. 78, no. 1/2, pp. 507–545, Oct. 1995.

[8] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[9] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution," *Comm. Pure Appl. Math.*, vol. 59, no. 6, pp. 797–892, 2006.

[10] J. van de Weijer and C. Schmid, "Coloring local feature extraction," in *Proc. Eur. Conf. Comput. Vis.*, Graz, Austria, May 2006, pp. 334–348, Part II.

[11] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[12] M. K. Hu, "Visual pattern recognition by moment invariants," *IRE Trans. Inf. Theory*, vol. 8, no. 2, pp. 179–187, 1962.