IEEE TRANSACTIONS ON
EMERGING TOPICS
IN COMPUTING

Received 13 October 2013; revised 22 May 2014; accepted 13 August 2014. Date of publication 11 September 2014;
date of current version 30 October 2014.

Digital Object Identifier 10.1109/TETC.2014.2356499

Estimating the Statistical Characteristics of Remote Sensing Big Data in the Wavelet Transform Domain

LIZHE WANG[1], (Senior Member, IEEE), HUI ZHONG[1,2], RAJIV RANJAN[3], ALBERT ZOMAYA[4], (Fellow, IEEE), AND PENG LIU[1]

[1]Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing 100864, China
[2]Department of Electrical Engineering and Information Sciences, University of Science and Technology of China, Hefei 230026, China
[3]Computational Informatics, Commonwealth Scientific and Industrial Research Organisation, Clayton South, VIC 3169, Australia
[4]School of Information Technologies, University of Sydney, Sydney, NSW 2006, Australia

CORRESPONDING AUTHOR: L. WANG (lizhe.wang@gmail.com)

This work was supported by the National Natural Science Foundation of China under Grant 61361120098.

ABSTRACT Since it is difficult to deal with big data using traditional models and algorithms, predicting and estimating the characteristics of big data is very important. Remote sensing big data consist of many large-scale images that are extremely complex in terms of their structural, spectral, and textual features. Based on multiresolution analysis theory, most of the natural images are sparse and have obvious clustering and persistence characters when they are transformed into another domain by a group of basic special functions. In this paper, we use a wavelet transform to represent remote sensing big data that are large scale in the space domain, correlated in the spectral domain, and continuous in the time domain. We decompose the big data set into approximate multiscale detail coefficients based on a wavelet transform. In order to determine whether the density function of wavelet coefficients in a big data set are peaky at zero and have a heavy tailed shape, a two-component Gaussian mixture model (GMM) is employed. For the first time, we use the expectation-maximization likelihood method to estimate the model parameters for the remote sensing big data set in the wavelet domain. The variance of the GMM with changing of bands, time, and scale are comprehensively analyzed. The statistical characteristics of different textures are also compared. We find that the cluster characteristics of the wavelet coefficients are still obvious in the remote sensing big data set for different bands and different scales. However, it is not always precise when we model the long-term sequence data set using the GMM. We also found that the scale features of different textures for the big data set are obviously reflected in the probability density function and GMM parameters of the wavelet coefficients.

INDEX TERMS Big data computing, remote sensing image processing, wavelet, parameters estimation.

I. INTRODUCTION

Big data is a collection of data sets so large and complex that it is difficult to deal with using traditional data processing algorithms and models. The challenges include acquisition, storage, searching, sharing, transferring, analysis and visualization. Scientists regularly encounter limitations due to large data sets in many areas, such as geosciences and remote sensing, complex physics simulations, and biological and environmental research. The three Vs (volume, variety, and velocity) are three defining properties or dimensions of big data. Volume refers to the amount of data, variety refers to the number of types of data, and velocity refers to the speed of data processing. According to the 3Vs model, the challenges of big data result from the expansion of all three properties, rather than just the volume alone.

With the fast increments of the volume or dimensions of data sets, researchers in different areas face different problems when they want to deal with big data using traditional methods. In the remote sensing application area, the size of data sets is growing, in part because they are increasingly being acquired and gathered by many different satellite sensors with different resolution and different spectral characteristics. More importantly, the data is often from a long-term sequence image and a large area of the earth's surface.

Assessing the statistical features of a remote sensing data set is a fundamental task in a lot of data analysis. For example, in the image clustering method, we often need to estimate the statistical similarity measure [1]; In many classification algorithms, we need spatial statistics-based expressions to create the decision boundary between various classes [2]; In researching of endmember detection, we also need to consider the spatial distribution of end members [3], [4] using the statistical characteristics of the data set. For big data, we often estimate a vector of model parameters given a training data set. Such statistical feature estimation provides us far more information than a simple inquiry and can be used to improve human interpretation of inferential outputs, do bias correction, perform hypothesis testing, make more efficient use of available resources, perform active learning, and optimize feature selection, among many more potential uses.

There are a large number of studies that focus on the statistical features of big data sets [5]–[11]. However, in many remote sensing applications related to big data sets, we often do not directly assess their statistical features. In order to manifest some of the statistical characteristics of the data sets, we often represent them based on some transforms. How to represent big data sets is one of the fundamental problems in researching big data, as most data processing tasks rely on an appropriate data representation. For many image processing tasks, the wavelet transform [12] of the data is the preferred transform. For remote sensing big data, multi-resolution representation by wavelet transform is more and more important for many algorithms such as image segmentation [13], image de-noising [14], image restoration [15], image fusion [16], change detection [17], feature extraction [18], and image interpretation. Therefore, the estimation of statistical features of big data in the wavelet transform domain is one of the most important problems.

Many researchers have worked on the statistical features of the wavelet domain of an image signal or a small data set. They have been applied in many image processing fields. Apart from the sparse characteristics, there are two other characteristics of the coefficients in the wavelet transform domain. The first is the inter-scale constraint known as the tree structure [19] and the second is the intra-scale constraints that are the statistical dependencies of the neighbourhood coefficients [20].

Some early research about the intra-scale statistical characteristics modelled the wavelet coefficients as Gaussian distributions. The theoretical formalization of filtering additive i.i.d Gaussian noise (with zero-mean and standard deviation) via thresholding wavelet coefficients was pioneered by Donoho [21]. However, a single Gaussian model is in conflict with some natural properties of the signal. Jointly, Gaussian models can efficiently represent the linear correlations between wavelet coefficients. A typical wavelet coefficient probability density is much more ''peaky'' at zero and heavy-tailed than the Gaussian distribution [19]. In reference [22], a framework for a near-optimal threshold is proposed. This approach can be formally described as Bayesian estimation, and it was pointed out that the wavelet coefficients in a subband of a natural image can be summarized adequately by a generalized Gaussian distribution (GGD). In [23], a bivariate probability density function is proposed to model the statistical dependence between a coefficient and its parent, and the corresponding bivariate shrinkage function is obtained. This method maintains the simplicity, efficiency, and intuition of soft thresholding. In reference [24], a de-noising method is proposed based on the statistical model of the coefficients of an overcomplete multiscale oriented basis. In this method, neighborhoods of coefficients at adjacent positions and scales are modelled as the product of two independent random variables: a Gaussian vector and a hidden positive scalar multiplier. Under this model, the Bayesian least squares estimate of each coefficient reduces to a weighted average of the local linear estimates over all possible values of the hidden multiplier variable. It obtained very good image reconstruction results, both visually and in terms of mean squared error. More recently, some propose using the product Bernoulli distributions (PBD) [25] for modeling coefficient histograms of wavelet subbands. For large filter outputs where the distribution of filter outputs peaked at zero, the PBD model has been shown to perform similarly to the GGD model. In particular, the bit-plane probability (BP) signature induced by the PBD model was applied successfully to supervised texture classifications [25]. In [26], the authors propose adopting three-parameter generalized Gamma density for modeling wavelet detail subband coefficient histograms. The advantage of generalized gamma density over the existing generalized Gaussian density is that it provides more flexibility to control the shape of the model, which is critical for practical histogram-based applications [26]. To measure the discrepancy between generalized Gamma densities, the symmetrized Kullback-Leibler distance (SKLD) is used to derive a closed form for the SKLD between generalized Gamma densities, which makes the proposed scheme particularly suitable for image retrieval systems with large image databases.

While modeling the wavelet coefficients in a dependent or joint Gaussian distribution without considering inter-scale constrains or tree structures, methods such as wavelet-based statistical signal processing techniques are unrealistic for many real world signals or data. To solve the problem, in reference [19], a wavelet-domain hidden Markov model (HMM) is developed as a framework for signal processing. It concisely models the statistical dependencies and non-Gaussian statistical characteristics using real signals or data. Wavelet-domain HMM well represent the intrinsic properties of the wavelet coefficients and provide a powerful and tractable model for signals. Therefore, it was quickly adopted by a wide range of applications, including de-noising [27], [28], detection, classification [29], segmentation [30], texturing [31], and compressive sensing [32], [33]. However, when using the expectation-maximization algorithm in reference [19], a potential drawback to the HMM framework is the computationally expensive iterative training. For the reduction of

the computation, in the reference [34], the HMM model is simplified by exploiting the inherent self-similarity of real-world images. Also introduced is a Bayesian universal HMM that fixes a few of the parameters no training is required. In reference [35], the authors propose a new "upward-downward" algorithm, in which a Viterbi-like algorithm for global restoration of the hidden state tree is introduced.

Almost all the methods that model the statistical characteristics of wavelet coefficients are only applicable to a small data set. Most of them are related to de-noising, texture analysis, segmentation, classification, and retrieval algorithms. For big data sets, such as space-temporal remote sensing data sets, we also need to model their wavelet coefficients to find the changing trends, discover the intrinsic mechanisms, and represent the rules of their evolution process. In this paper, we use the GMM to denote the statistical properties of wavelet coefficients of a remote sensing big data set. Our contribution is to estimate the model parameters of a big data set using different aspects or dimensions such as time, spectral bands, scales, and textures.

II. MULTIRESOLUTION ANALYSIS AND THE WAVELET TRANSFORM

For the convenience of narration, before estimating the model parameters of wavelet coefficients of the remote sensing big data set, we first simply review the multiresolution analysis theory and the wavelet transform. A multiresoluiton analysis (MRA) consists of a collection of nested subspaces $\{V_j\}_{j\in Z}$, satisfying the following four properties [12]:

1) $\bigcap_{j\in Z} V_j = \{0\}$, $\bigcup_{j\in Z} V_j$ is dense in $L^2(R)$
2) $V_j \subset V_{j-1}$
3) $x(t)\in V_j \leftrightarrow x(2^j t)\in V_0$
4) There exists a function $\phi_0(t)$ in V_0, which is called the scaling function, such that the collection $\{\phi_0(t-k), k\in Z\}$ is an unconditional Riesz basis for V_0.

Similarly, there are scaled and shifted functions like

$$\{\phi_{j,k}(t) = 2^{-j/2}\phi_0(2^{-j}t - k), k\in Z\}, \tag{1}$$

where j is the dilation parameter about dilation, and k is the position parameter. Equation (1) constitutes the Riesz basis for the space V_j. Performing a multiresolution analysis of signal x means successively projecting it into each of the approximation subspaces V_j to get

$$approx\_x_j(t) = \sum a_x(j,k)\phi_{j,k}(t), \tag{2}$$

where $a_x(j,k)$ is the approximation coefficient for the basic function $\phi_{j,k}(t)$. Since $V_j \subset V_{j-1}$, $approx\_x_j(t)$ is a coarser approximation of $x(t)$ than $approx\_x_{j-1}(t)$; therefore, the main idea of the MRA consists of measuring the loss of information. They could be seen as the detail $detail\_x_j(t) = approx\_x_{j-1}(t) - approx\_x_j(t)$ when going from one approximation to the next coarser one. The MRA analysis shows that $detail\_x_j$ as the detail signals can be directly obtained from projections of x onto a collection of subspaces, the W_j, called the wavelet subspaces. Moreover, in MRA theory, there exists a function ψ_0, called the mother wavelet, derived from ϕ_0, such that its templates $\{\psi_{j,k}(t) = 2^{-j/2}\psi_0(2^{-j}t - k), k\in Z\}$ constitute a Riesz basis for W_j. We then get

$$detail\_x_j(t) = \sum_k d_x(j,k)\psi_{j,k}(t). \tag{3}$$

In practice, we may want to see all the data with the "desired" resolution, that is for some resolution j. We define the subspace

$$Vj = Span\{\phi_{j,k}(t)\} \quad \text{and} \tag{4}$$

$$Wj = Span\{\psi_{j,k}(t)\}. \tag{5}$$

Once we combine infinite wavelet sets, the sets are equal to the set $L^2(R) = \{f(x) | \int |f(x)|^2 dx < \infty\}$. In mathematical terms,

$$L^2(R) = V_0 \oplus W_0 \oplus W_1 \cdots \tag{6}$$

Based on MRA theory, we represent the signal x as a collection of details at different resolutions and a low-resolution approximation.

$$\begin{aligned} x(t) &= approx\_x_j(t) + \sum_{j=1}^{j=J} detail\_x_j(t) \\ &= \sum_k a_x(J,k)\phi_{J,k} + \sum_{j=1}^{J}\sum_k d_x(j,k)\psi_{j,k}(t) \end{aligned} \tag{7}$$

The $approx\_x_j$ is essentially coarser and a coarser approximation of x means that ϕ_0 is a low-pass function. The $detail\_x_j$, being an information "differential," indicates rather that ψ_0 is a bandpass function and therefore a small wave called a wavelet. MRA theory also shows that the mother wavelet function must satisfy $\int \psi_0(t)dt = 0$, and its Fourier transform obeys $|\Psi_0(\nu)| \sim \nu_N$, $\nu \to 0$, where N is a positive integer called the number of vanishing moments of the wavelet.

Given a scaling funtion ϕ_0 and mother wavelet ψ_0, the discrete (or non-redundant) wavelet transform (DWT) is a mapping from $L^2(R) \to l^2(Z)$ given by

$$x(t) \to \{\{a_x(J,k), k\in Z\}, \{d_x(j,k), j=1,\ldots,J, k\in Z\}\}. \tag{8}$$

These coefficients are defined using the inner products of x with two sets of basis functions:

$$\begin{cases} a_x(j,k) = \langle x, \phi_{j,k}\rangle \\ d_x(j,k) = \langle x, \psi_{j,k}\rangle, \end{cases} \tag{9}$$

where $\psi_{j,k}$ (respectively, $\phi_{j,k}$) are shifted and dilated templates of ψ_0 (respectively, ϕ_0), called the dual mother wavelet (respectively, the dual scaling function). Their definitions depend on whether one chooses to use an orthogonal, semi-orthogonal, or bi-orthogonal DWT [36]. Using a fast recursive filterbank-based pyramidal algorithm, they can be computed with extremely low computational cost [36].

What we next deal with is 2-D remote sensing image data. We define $x(t_1, t_2) \in L^2(R^2)$, which is a 2-D signal or image, where t_1 and t_2 are the coordinates of two directions.

The transform coefficient becomes two variable functions as well as the 2D wavelet transform. In reference [37], the scaling and wavelet functions are two variable functions, denoted as $\phi_{j,k_1,k_2}(t_1,t_2)$ and $\psi_{j,k_1,k_2}(t_1,t_2)$. The scaled and translated basis functions for the j^{th} level are defined as

$$\phi^i_{j,k_1,k_2}(t_1,t_2) = 2^{-j/2}\phi^i_0(2^{-j}t_1-k_1, 2^{-j}t_2-k_2) \quad \text{and} \tag{10}$$

$$\psi^i_{j,k_1,k_2}(t_1,t_2) = 2^{-j/2}\psi^i_0(2^{-j}t_1-k_1, 2^{-j}t_2-k_2), \tag{11}$$

where $k_1, k_2 \in Z$, and $i = \{H, V, D\}$ in which H is horizontal, V is vertical, and D is diagonal.

Similar to the representation of a 1-D signal, based on the MRA, the information in $x(t_1,t_2)$ is written as a collection of details at different resolutions and a low-resolution approximation. For each level, there are three different wavelet functions, $\psi^H(t_1,t_2)$, $\psi^V(t_1,t_2)$, and $\psi^D(t_1,t_2)$. Conceptually, the scaling function is the relatively low-frequency component. Therefore, there is one 2D scaling function. However, the wavelet function is related to the order to apply the decompositions. If the wavelet function is separable, that is, $f(x,y) = f_1(x)f_2(y)$, these functions can be easily rewritten as

$$\phi(t_1,t_2) = \phi(t_1)\phi(t_2), \tag{12}$$

$$\psi^H(t_1,t_2) = \psi(t_1)\phi(t_2), \tag{13}$$

$$\psi^V(t_1,t_2) = \phi(t_1)\psi(t_2), \quad \text{and} \tag{14}$$

$$\psi^D(t_1,t_2) = \psi(t_1)\psi(t_2). \tag{15}$$

Based on the definition of a 2-D basis function, performing a multiresolution analysis of 2-D image x means successively projecting it into each of the approximation subspaces V_j

$$approx\_x_j(t_1,t_2) = \sum_{k_1,k_2} a_x(j,k_1,k_2)\phi_{j,k_1,k_2}(t_1,t_2). \tag{16}$$

If we define the functions as separable functions, it is easier to analyze the 2D function. The analysis and synthesis equations are modified to

$$\begin{aligned} x(t_1,t_2) &= approx\_x_j(t_1,t_2) + \sum_{j=1}^{J} detail\_x_j(t_1,t_2) \\ &= \sum_k a_x(J,k_1,k_2)\phi_{J,k_1,k_2} \\ &\quad + \sum_{i=H,V,D}\sum_{j=1}^{J}\sum_{k_1,k_2} d_x(j,k_1,k_2)\psi^i_{j,k_1,k_2}(t_1,t_2). \end{aligned} \tag{17}$$

Based on the research on small image or data sets, the coefficients of the scale function in the wavelet transform domain are not easy to characterize using an analytic density function. However, as previously mentioned in the introduction, there have been much research on how to model the wavelet detail coefficients of $detail\_x_j(t_1,t_2)$. For big data sets, the coefficients of a wavelet function, as well as the detail coefficients, could also be subject to an obvious distribution, such as a Gaussian mixture model a generalized Gaussian model, or generalized gamma density. In this paper, we model the detail coefficients using a Gaussian mixture model and employ the expectation-maximization likelihood method (EM) to estimate the set of parameters in the next section.

III. PARAMETER ESTIMATION OF WAVELET COEFFICIENTS

To analyze the intra-scale correlation feature of wavelet coefficients of remote sensing big data, we use a GMM to denote its detail coefficients distribution. For arbitrary scales or for several scales, drawing samples $y_1, \cdots, y_n$ from details of x (such as $d_x(j,k_1,k_2)$), we assume that they are subject to a GMM with k components. Without loss of generality, we assume $y_1, \cdots, y_n$ are vectors, but actually, in our experiments, when we focus on a certain scale, they are scalars. Our goal is to estimate the parameter set $\theta = \{(\omega_j, \mu_j, \Sigma_j)\}_{j=1}^k$ using the EM method. For any sample y_i and parameters μ_j and Σ_j, we denote the Gaussian distribution as

$$p(y_i|\mu_j, \Sigma_j) = \frac{1}{2\pi^{d/2}|\Sigma_j|^{1/2}} exp(-\frac{1}{2}(y-\mu_j)^T \Sigma_j^{-1}(y-\mu_j)). \tag{18}$$

In the GMM, we have the probability density

$$p(Y_i = y_i|\theta) = \sum_{j=1}^{k} \omega_j p(y_i|\mu_j, \Sigma_j), \tag{19}$$

where $\omega_j > 0$, $\sum_{j=1}^k \omega_j = 1$, and $\theta = \{(\omega_j, \mu_j, \Sigma_j)\}_{j=1}^k$.

Let γ_{ij}^m be the estimate at the *mth* iteration of the probability that the *ith* sample was generated by the *jth* Gaussian component, that is,

$$\gamma_{ij}^m = P(Z_i = j|y_i, \theta^{(m)}) = \frac{\omega_j^{(m)} p(y_i|\mu_j^{(m)}, \Sigma_j^{(m)})}{\sum_{l=1}^k \omega_l^{(m)} p(y_i|\mu_l^{(m)}, \Sigma_l^{(m)})}, \tag{20}$$

which satisfies $\sum_{j=1}^k \gamma_{ij}^{(m)} = 1$.

Because we assume the samples are i.i.d., we can apply

$$\begin{aligned} Q_i(\theta|\theta^{(m)}) &= E_{Z_i|y_i,\theta_{(m)}}[\log p(y_i, Z_i|\theta)] \\ &= \sum_{j=1}^{k} P(Z_i|y_i, \theta^{(m)}) \log p(y_i, Z_i|\theta) \\ &= \sum_{j=1}^{k} \gamma_{ij}^{(m)} \log(\omega_j p(y_i|\mu_j, \Sigma_j)) \\ &= \sum_{j=1}^{k} \gamma_{ij}^{(m)} (\log\gamma_j - \frac{1}{2}\log|\Sigma_j| \\ &\quad -\frac{1}{2}(y_i-\mu_j)^T \Sigma_j^{-1}(y_i-\mu_j)) + C, \end{aligned} \tag{21}$$

where C is a constant that does not depend on θ and that can thus be dropped without affecting the M-step. Then

$$\begin{aligned} Q(\theta|\theta^{(m)}) &= \sum_{i=1}^{n}\sum_{j=1}^{k} \gamma_{ij}^{(m)} (\log\gamma_j - \frac{1}{2}\log|\Sigma_j| \\ &\quad -\frac{1}{2}(y_i-\mu_j)^T \Sigma_j^{-1}(y_i-\mu_j)), \end{aligned} \tag{22}$$

which completes the E-step. For notational simplicity we denote the total membership weight of the *jth* Gaussian as

$$n_j^{(m)} = \sum_{i=1}^{n} \gamma_{ij}^{(m)}.$$

Then $Q(\theta|\theta^{(m)})$ can be rewritten as

$$Q(\theta|\theta^{(m)}) = \sum_{j=1}^{k} n_j^{(m)}((\log\gamma_j - \frac{1}{2}\log|\Sigma_j|) - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{k}(y_i-\mu_j)^T {\Sigma_j}^{-1}(y_i-\mu_j)). \quad (23)$$

The M-step is to solve

$$\begin{aligned} &\text{maximize} \quad Q(\theta|\theta^{(m)}) \\ &\text{subject to} \quad \sum_{j=1}^{k}\omega_j = 1, \omega_j > 0 \\ &\Sigma_j > 0, \quad j = 1, \cdots, k, \end{aligned} \quad (24)$$

where $\Sigma_j > 0$ means that Σ_j is positive definite.

From equation (23), we can independently maximize the Q-function with respect to the weights, which requires maximizing the term $\sum_{j=1}^{k} n_j^{(m)}\log\gamma_j$. The solution is:

$$\omega_j^{m+1} = \frac{n_j^m}{\sum_{j=1}^{k} n_j^m} = \frac{n_j^m}{n}, \quad j = 1, \cdots, k. \quad (25)$$

The optimal μ_j and Σ_j can be found by setting the corresponding derivatives to zero. To solve for the means μ_j, we let

$$0 = \frac{\partial Q(\theta|\theta^{(m)})}{\partial \mu_j} = {\Sigma_j}^{-1}, \quad (26)$$

which yields

$$\mu_j^{(m+1)} = \frac{1}{n_j^{(m)}}\sum_{i=1}^{n}\gamma_{ij}^m y_i, j = 1, \cdots, k. \quad (27)$$

To solve for covariance matrices Σ_j, we let

$$\begin{aligned} 0 &= \frac{\partial Q(\theta|\theta^{(m)})}{\partial \Sigma_j} \\ &= \frac{1}{2}n_j^{(m)}\frac{\partial\log|\Sigma_j|}{\partial\Sigma_j} - \frac{1}{2}\sum_{i=1}^{n}\gamma_{ij}^{(m)}\frac{\partial((y_i-\mu_j)^T{\Sigma_j}^{-1}(y_i-\mu_j))}{\partial\Sigma_j} \\ &= \frac{1}{2}n_j^{(m)}{\Sigma_j}^{-1} - \frac{1}{2}\sum_{i=1}^{n}\gamma_{ij}^{(m)}{\Sigma_j}^{-1}((y_i-\mu_j)^T(y_i-\mu_j)){\Sigma_j}^{-1}, \end{aligned} \quad (28)$$

and thus

$$\Sigma_j^{(m+1)} = \frac{1}{n_j^{(m)}}\sum_{i=1}^{n}\gamma_{ij}^{(m)}(y_i-\mu_j^{(m+1)})(y_i-\mu_j^{(m+1)})^T. \quad (29)$$

Now, for a GMM with k components, we could estimate its model parameters $\theta = \{(\omega_j, \mu_j, \Sigma_j)\}_{j=1}^{k}$ using equations (25), (27), and (29). In the next section, we will model the wavelet coefficients of big data sets for different bands, different scales, long-term sequences and different textures.

IV. EXPERIMENTS AND RESULTS

In the experiments, we use the image data set from the Landsat satellite, which represents the world's longest continuously acquired collection of space-based moderate-resolution land remote sensing data. Over the past four decades, the imagery data set has provided a unique and extremely rich resource for research on agriculture, geology, forestry, regional planning, education, mapping, and global change.

Since July 23, 1972, there has been a series of Landsat satellites missions, from Landsat1 to Landsat8. In our experiments, different data sets from different Landsat satellites are included. The first is the data set from Landsat1 through Landsat5. The Landsat Multispectral Scanner (MSS) sensor was onboard Landsat1 through Landsat5 and acquired images of the earth nearly continuously from July 1972 to October 1992. Of the Landsats, Landsats 1, 2, and 3, which had 4–7 bands, had an 18-day revisit cycle. However, Landsats 4 and 5, which had 1–4 bands, maintained a 16-day revisit cycle. Their resolution is 60 m, and they have different spectral characteristics. The second is the image data set acquired by the Landsat Thematic Mapper (TM) sensor carried on Landsats4 and 5 with a 16-day repeat cycle. The TM images from Landsats 4 and 5 consist of seven spectral bands. The resolution is 30 m for bands 1–7 (thermal infrared band 6 was collected at 120 m, but was re-sampled at 30 m). The third is the image data set acquired by the Landsat Enhanced Thematic Mapper Plus (ETM+) sensor carried by the Landsat 7 satellite with a 16-day revisit cycle. Landsat 7 ETM+ images consist of eight spectral bands with a spatial resolution of 30 m for bands 1–7. The panchromatic band 8 has a resolution of 15 m. All the images use the map projection of UTM-WGS84 with polar stereographic for the continent of Antarctica.

Because the series of Landsat satellites have continuously acquired image data for four decades, the image data volume is large enough to be big data. Furthermore, the remote sensing big data set exhibits different characteristics in different dimensions, such as time, spectral, space, and textual dimensions. In general, it is hard to precisely model the statistical characteristics of remote sensing big data. In the following experiments, we estimate their statistical characteristics by randomly selecting some subsets of the big data and transforming them into a wavelet domain. In all experiments, the db1 [36] wavelet basis function is employed. Based on some research results for small image data [19], we assume that there are two components in the GMM model for wavelet coefficients. Therefore, the parameters to be estimated are $\theta = \{(\omega_j, \mu_j, \Sigma_j)\}_{j=1}^{2}$. Σ_j will be the variance not covariance variance because it is more intuitive in visuals to show the statistical characteristics by sampling in low dimension data or in a certain scale or one band.

A. EXPERIMENTS 1

In this experiment, we separately decompose the different bands of the remote sensing image data set using a wavelet

transform and make some comparisons. A subset of the Landsat global image data set is selected for the validation. The subset data used in this experiment is from the 2008 whole year data, which cover all China, with more than 9,000,000 square kilometers. In this data subset, only one multispectral image set is selected for every location on earth's surface. Some image data that are defective or that have too many clouds are not included in the test data set. The volume of the data set is about 70 GB.

The data set are transformed into six scales, which have horizontal, vertical, and diagonal directions. We selected results from the first, third, fifth, and seventh bands of the data subset and show them in Fig. 1. The statistical features of the fourth and fifth scales of this multiresolution analysis are shown in Fig. 1. Figure. 1 (a) shows the distribution density of wavelet coefficients in three directions and in two scales, which come from the first band of the image data set. Figure. 1 (b) shows the distribution density of wavelet coefficients in three directions and in two scales, which come from the third band of the image data set. Figure. 1 (c) and (d) have a similar definition and meaning as Fig. 1 (a) and (b) except for the fifth and seventh bands. In each curve graph of Fig. 1 (a–d), the GMM model estimated using the EM method are shown on the left, while the real distribution density of wavelet coefficients calculated and counted by random sampling are show on the right.

When we compare the results of the EM estimation with the real statistical features of the wavelet coefficients, we find that for every band data subset of the big data the GMM could well represent the distribution in most cases. Therefore, it is still relatively reasonable to believe that the wavelet coefficients follow the GMM for different bands of a remote sensing big data set.

In Fig. 1 (a–d), we can observe that in a certain scale, the statistical characteristics of the three directions such as horizontal, vertical, and diagonal details, do not show very obvious differences. It is because that in a large data set, all the directions almost have similar possibility in global. However, there are indeed some small differences in the statistical features of the wavelet coefficients for different directions; their degrees of concentration or clusters are different. The diagonal wavelet coefficients concentrate more on the center part when compared to the other two directions, which means that there are more small or near zero coefficients in diagonal wavelet coefficients. This phenomenon is more obvious in the fifth or larger scale of each band.

In Fig. 2, we compare the changing trend of the model parameters of the wavelet coefficients with the different scales in decomposition. Figure. 2(a) shows the two variances of the two components in the GMM and their changing scales for the first band. The variance of the first component is on the left and the variance of the second component is on the right. Similarly, Fig. 2 (c–d) shows the variance of the wavelet coefficients from the third, fifth, and seventh bands. We do not show all the variances of all the directions, but only the variance of the diagonal direction. We find that for the Landsat image big data set, the model parameters of the first and seventh bands in the multispectral image vary obviously, but the model parameters of the third and fifth band data are relatively similar. It is hard to explain this based only on the characteristics or properties of the wavelet transform, but it may be related to the spectral response function of the multi-spectral sensor on Landsat satellites and the spectral characteristics of terrestrial objects.

B. EXPERIMENTS 2

In the experiment, we mainly focus on the statistical characteristics of long-term sequence data set of remote sensing images. Remote sensing big data from Landsat satellites contain many long time sequence data sets for many locations on the earth's surface. The data subset for the Beijing area in northern China, which covers 16411 square kilometers, was selected for the tests. The time period is from 1983 to 2013, and some of the data with too much cloud coverage was removed from the data set. There are many forests, cities, and mountains in this area, which makes the textual information very rich. The high degree of climatic seasonality is another characteristics of the data subset. The volume of the tested data subset is about 110 GB. Figure. 3 shows the GMM estimation and distribution density by real sampling of long-term sequence data set for different scales and different subband directions. Figure. 4 shows the changing trends of the GMM parameters with the changing of time.

Figure. 3 shows graphs of the distribution density of wavelet coefficients in the fourth scale, fifth scale, and sixth scale. In each scale, the probability density distribution estimated by EM on the left and the histogram counted by sampling on the right are shown on every graph. Furthermore, the statistical characteristics of subband wavelet coefficients in three directions are sampled, estimated, and shown separately for every scale. We can observe that there are also cluster characteristics similar to the those of the different band image data set in Fig. 1. However, the concentration or cluster characteristics of the wavelet coefficients are not as regular as in Fig. 1. There are obvious estimation errors in Fig. 3 (a) regarding the horizontal direction of the fourth scale in Fig. 3 (b) regarding the vertical direction of the fourth scale Fig. 3 (d) regarding the horizontal direction of the fifth scale, and in Fig. 3 (g) regarding the horizontal direction of the sixth scale. In the long-term sequence data set, there are often highly redundant image features. Therefore, in some scales and in some directions, there could be too many similar wavelet coefficient values that are not zeros but concentrate together and become a new cluster center. These wavelet coefficients near the new cluster centers make the GMM inapplicable to denote the distribution density of wavelet coefficients of the long-term sequence data set. We can observe that in Fig. 3 (a–i), these errors are more obvious in the horizontal and vertical directions of small scales than in the diagonal direction of large scales.

In Fig. 4, the x-axis is the time and the y-axis is the value of the variance in the GMM. There are two variances in

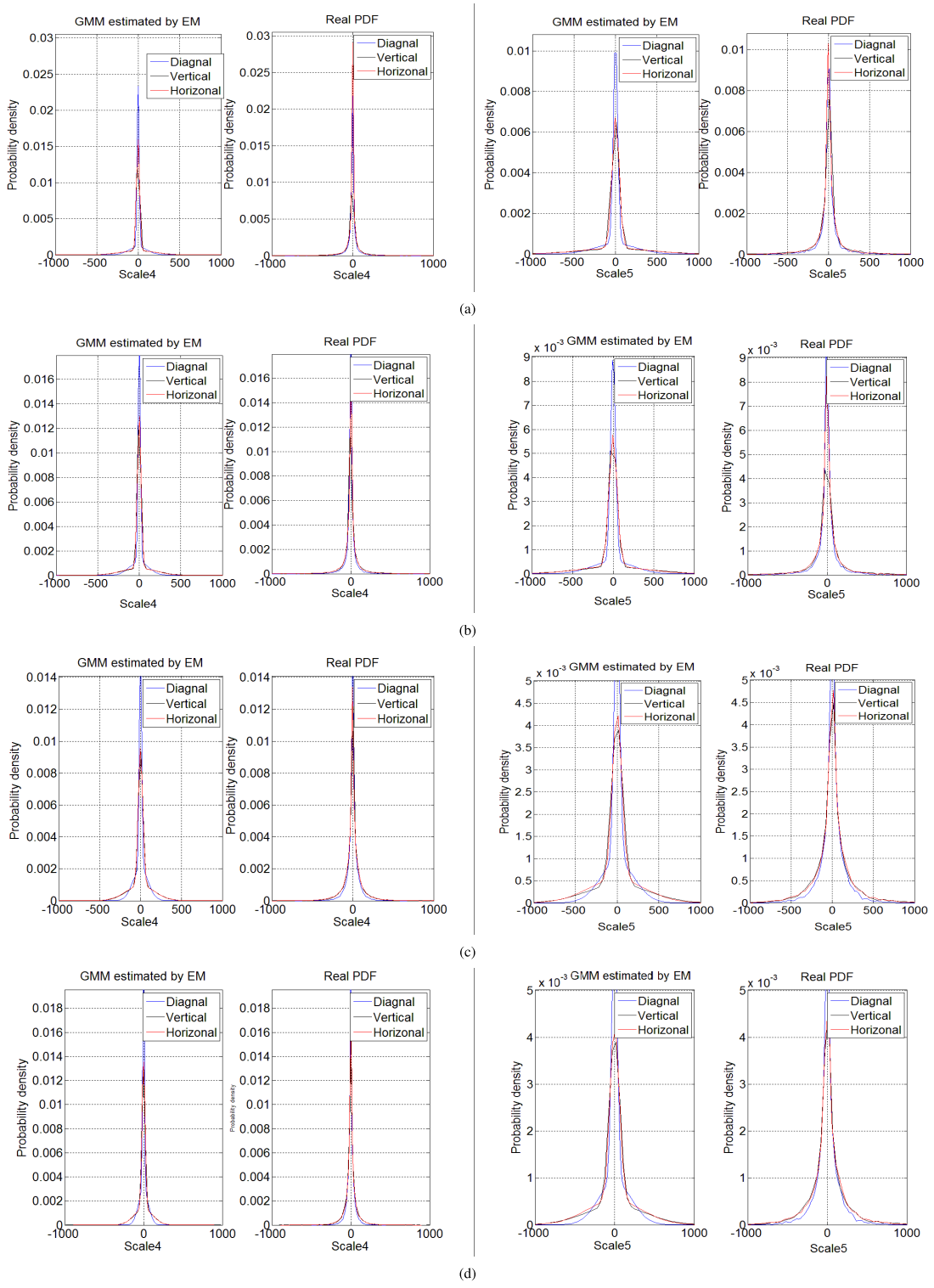


FIGURE 1. Wavelet coefficients distribution in different scales and in different bands. (a) Band-1. (b) Band-3. (c) Band-5. (d) Band-7.

our GMM. Figure. 4 (a) is the variance vs. time of the fourth scale, Fig. 4 (b) is the variance vs. time of the fifth scale, and Fig. 4 (c) is the variance vs. time of the sixth scale. We can observe that the curve of variances in different scales is not smooth. It illustrates that the changing of the texture in the long-term sequence data set is relatively obvious. For Landsat

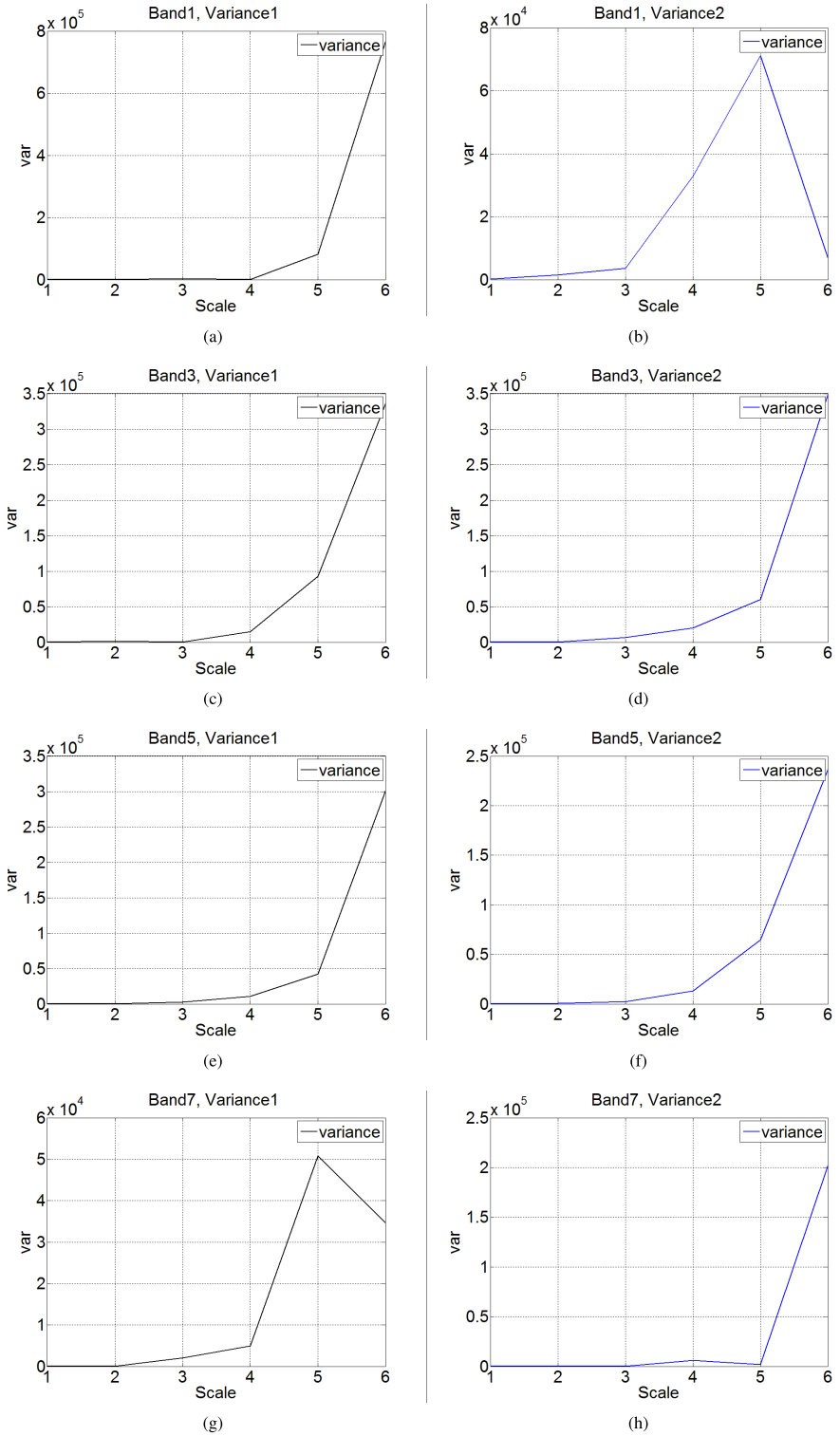


FIGURE 2. The variation of wavelet coefficients with different bands and different scales. (a) Band-1. (b) Band-1. (c) Band-3. (d) Band-3. (e) Band-5. (f) Band-5. (g) Band-7. (h) Band-7.

satellites, their revisit cycle is more than half a month, so there are many differences between the multitemporal image data sets. This is why there is no obvious continuity, as in the scale dimension of Fig. 2.

C. EXPERIMENTS 3

In this experiment, we decompose the different texture data sets, such as cities, mountains and vegetables into different scales using a wavelet transform. The subsets of data for city

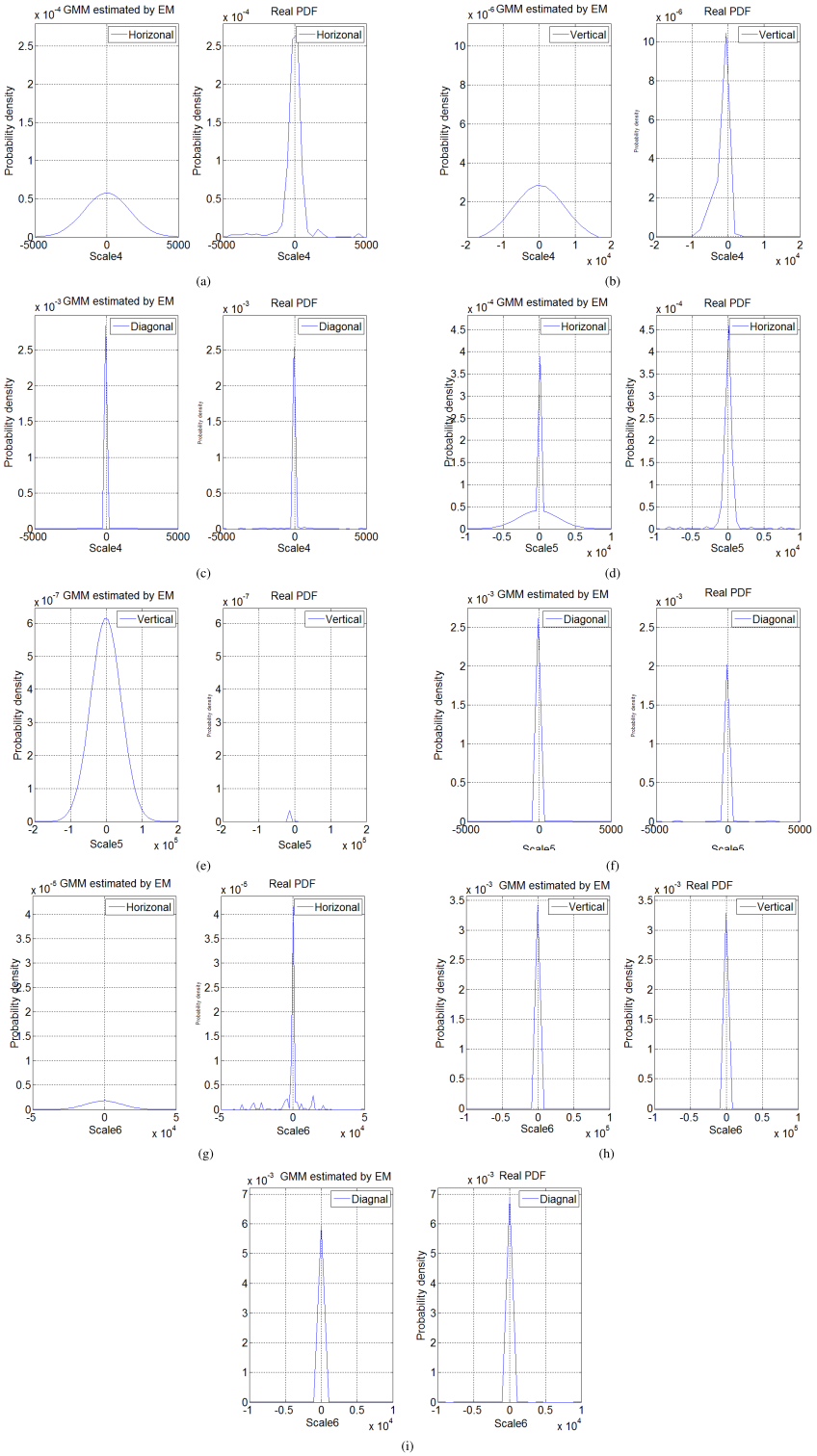


FIGURE 3. The wavelet coefficients distribution in different scales and in a long time sequence (a) The fourth scale, Horizontal. (b) The fourth scale, Vertical. (c) The fourth scale, Diagonal. (d) The fifth scale, Horizontal. (e) The fifth scale, Vertical. (f) The fifth scale, Diagonal. (g) The sixth scale, Horizontal. (h) The sixth scale, Vertical. (i) The sixth scale, Diagonal.

and mountain textures were randomly selected from the area north of the Yellow River in northern China, which is about 2,000,000 square kilometres. The images in the subset data for vegetable texture are mostly from the area of the Great Lakes on the US-Canada border. The total volume of the three types of textures is about 5 GB.

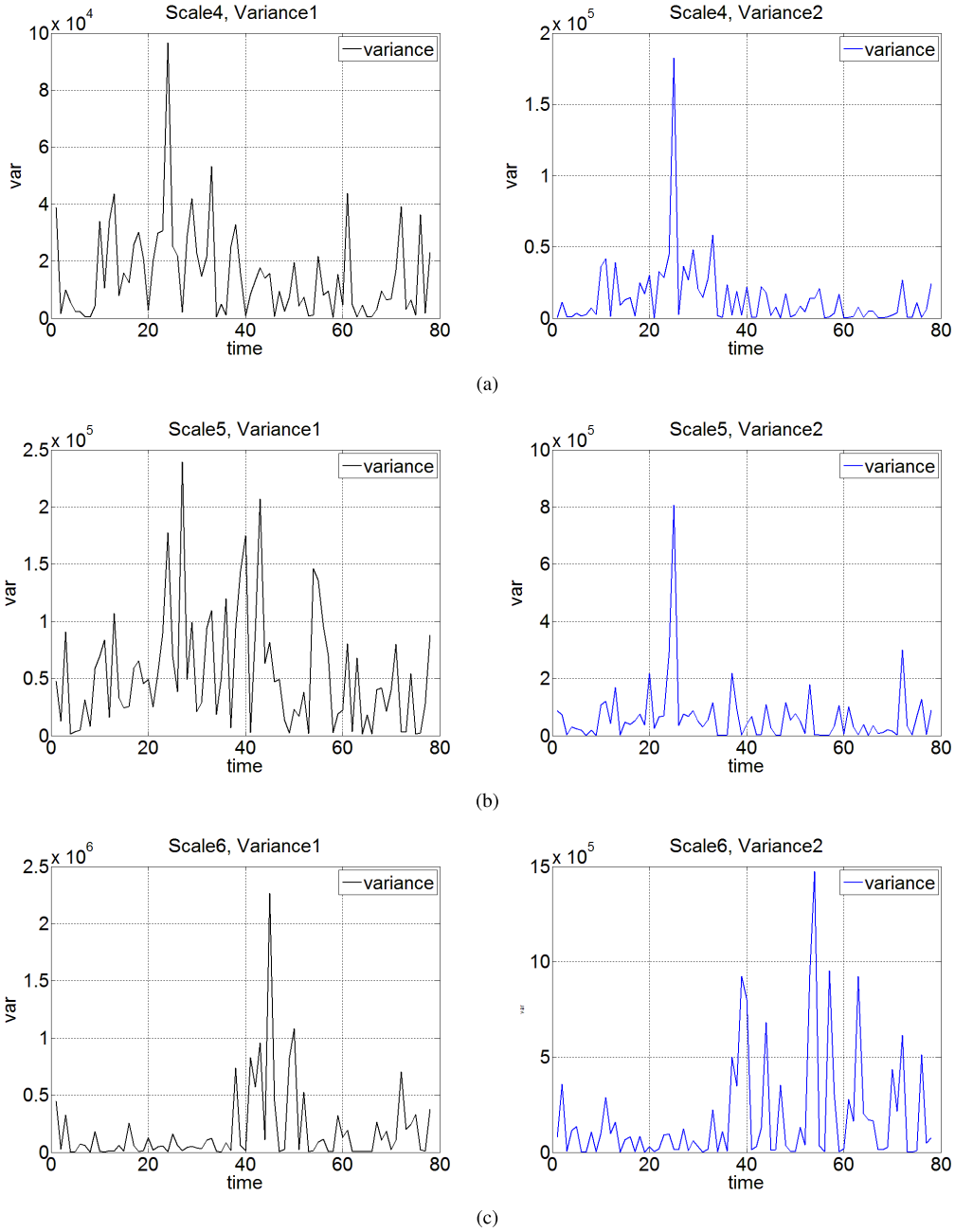


FIGURE 4. The variation of wavelet coefficients in different scales and in a long-term sequence. (a) The fourth scale. (b) The fifth scale. (c) The sixth scale.

Figure. 5(a-c) shows the statistical characteristics of wavelet coefficients of city textures in different scales, different directions. Figure. 5(d-f) shows the statistical characteristics of wavelet coefficients of mountain textures in different scales and different directions. Figure. 5(g-i) shows the statistical characteristics of the wavelet coefficients of vegetable textures in different scales and different directions. We find that these coefficients of texture more likely concentrate to zeros in small scales than in large scales. The values of coefficients of mountain texture in the fourth scale mostly range from -380 to $+380$, the values of coefficients of city texture in the fourth scale mostly range from -200 to $+200$, and the values of coefficients of vegetable texture in the fourth scale mostly range from -100 to $+100$. These scale features agree well with vegetable textures that have smaller image features than the city texture, while buildings in cities have smaller image features than mountains. Therefore, it is reasonable to infer some natural characteristics of the textures using the statistical characteristics of wavelet coefficients for big data sets.

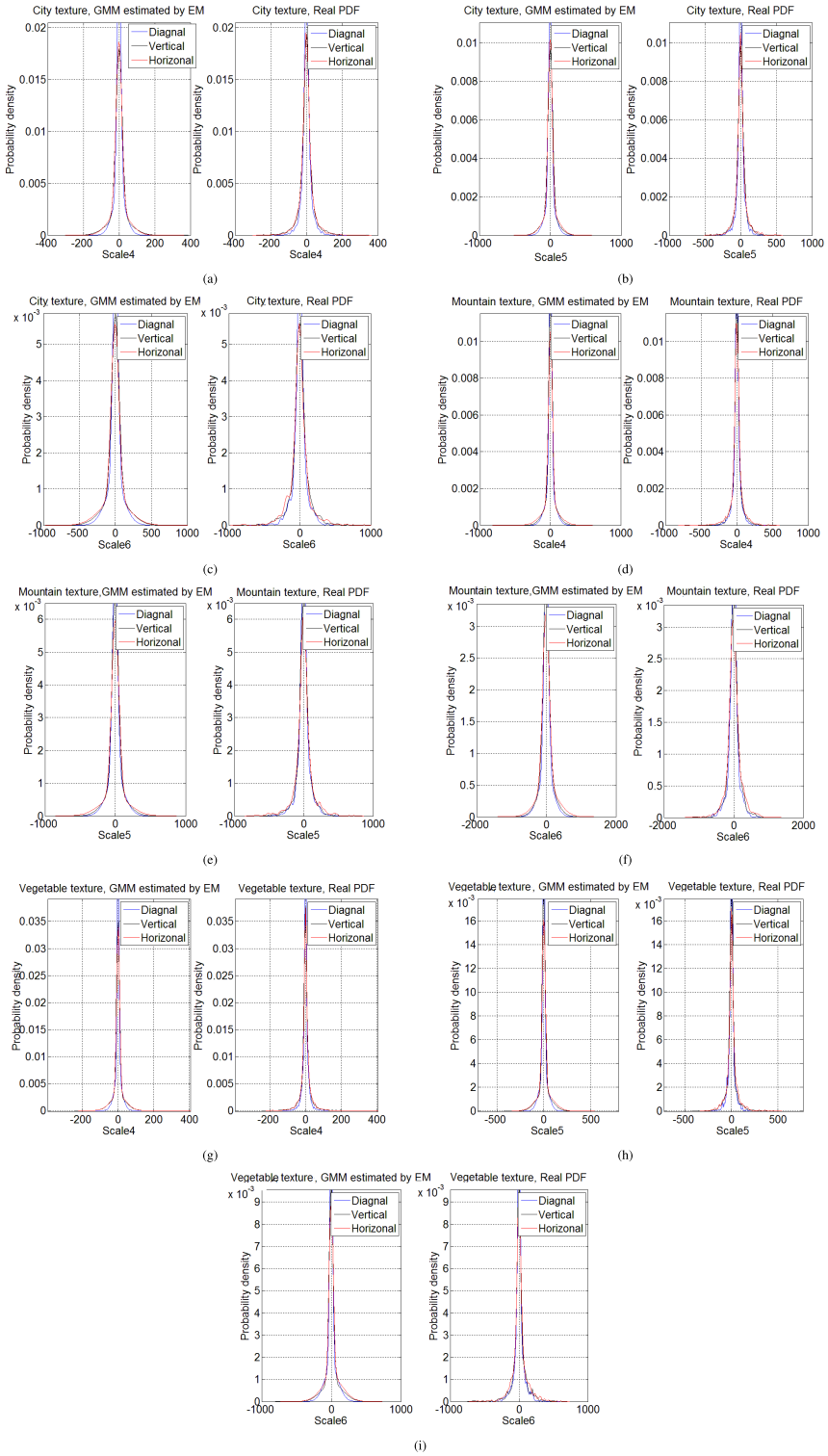


FIGURE 5. The distribution of wavelet coefficients in different scales and in different textures. (a) City, the fourth scale. (b) City, the fifth scale. (c) City, the sixth scale. (d) Mountain, the fourth scale. (e) Mountain, the fifth scale. (f) Mountain, the sixth scale. (g) Vegetable, the fourth scale. (h) Vegetable, the fifth scale. (i) Vegetable, the sixth scale.

Figure. 6 shows the variances in the GMM and their changing scales. We can observe that in Fig. 6 (a), for city textures, the two parameters change smoothly and synchronously. In Fig. 6 (b), for mountain textures, the two parameters do not change regularly. However, due to the complexity of the big data, we cannot conclude they are the rules of the city

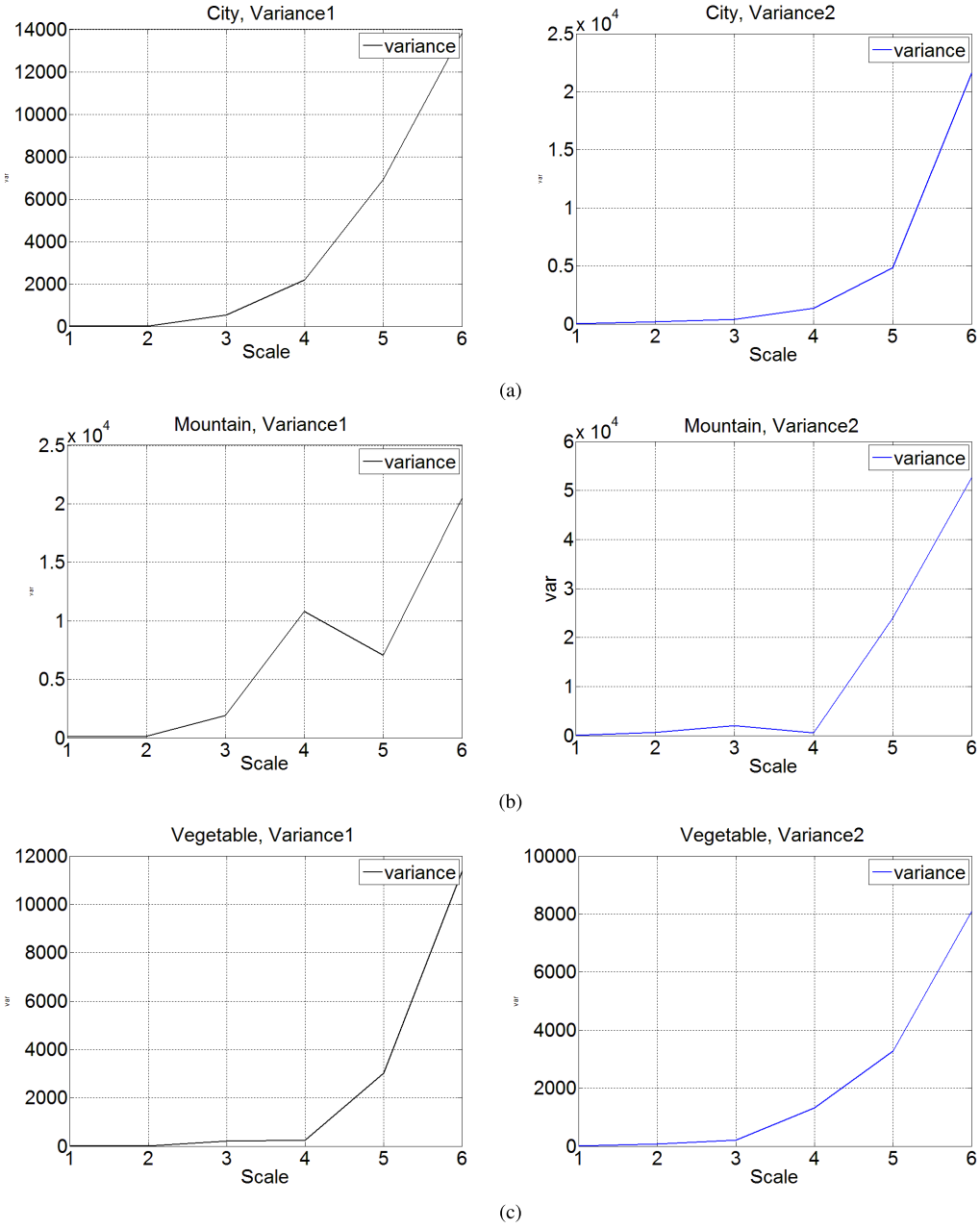


FIGURE 6. The variation of wavelet coefficients in different scales and in different textures. (a) City. (b) Moutain. (c) Vegetable.

textures and mountain texture in the wavelet domain because of the uncertainly of the two components from the EM estimation. What we can conclude is that, for big data, there are more large wavelet coefficients in large scales and the cluster characteristics are more obvious in small scales.

V. CONCLUSION

In this paper, we sampled and transformed the remote sensing big data set into a wavelet domain. The statistical characteristics of wavelet coefficients in terms of the scale, time, and band of the data set were comprehensively analyzed and compared. The data set of different textures was decomposed into different scales, and the parameters of the GMM of the wavelet coefficients were estimated. The statistical characteristics of different textures were also compared. We found that the cluster characteristics of the wavelet coefficients are still obvious in the remote sensing big data set for different bands and different scales. However, it is not always well estimated when we modeled the long-term sequence big data set using a GMM. We also found that the features of different textures for the big data set are obviously reflected in the probability density function and model parameters of wavelet coefficients.

REFERENCES

[1] C. Yang, L. Bruzzone, F. Sun, L. Lu, R. Guan, and Y. Liang, "A fuzzy-statistics-based affinity propagation technique for clustering in multispectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 6, pp. 2647–2659, Jun. 2010.

[2] P. Mishra and D. Singh, "A statistical-measure-based adaptive land cover classification algorithm by efficient utilization of polarimetric SAR observables," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 2889–2900, May 2013.

[3] A. Plaza, P. Martinez, R. Perez, and J. Plaza, "A quantitative and comparative analysis of endmember extraction algorithms from hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 3, pp. 650–663, Mar. 2004.

[4] D. R. Thompson, L. Mandrake, M. S. Gilmore, and R. Castano, "Superpixel endmember detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 11, pp. 4023–4033, Nov. 2010.

[5] J. Lu and D. Li, "Bias correction in a small sample from big data," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 11, pp. 2658–2663, Nov. 2013.

[6] A. Kleiner, A. Talwalkar, P. Sarkar, and M. I. Jordan, "The big data bootstrap," in *Proc. 29th Int. Conf. Mach. Learn. (ICML)*, 2012, pp. 1–8.

[7] R. Li, D. K. J. Lin, and B. Li, "Statistical inference in massive data sets," *Appl. Stochastic Models Bus. Ind.*, vol. 29, no. 5, pp. 399–409, Sep./Oct. 2012.

[8] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 1, pp. 97–107, Jan. 2014.

[9] G. Cormode and M. Garofalakis, "Histograms and wavelets on probabilistic data," in *Proc. IEEE 25th Int. Conf. Data Eng. (ICDE)*, Mar./Apr. 2009, pp. 293–304.

[10] Q. Yang, Y. Chen, G.-R. Xue, W. Dai, and Y. Yu, "Heterogeneous transfer learning for image clustering via the social web," in *Proc. Joint Conf. 47th Annu. Meeting ACL 4th Int. Joint Conf. Natural Lang. Process. AFNLP*, Stroudsburg, PA, USA, 2009, pp. 1–9.

[11] J. Portilla and E. P. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," *Int. J. Comput. Vis.*, vol. 40, no. 1, pp. 49–70, 2000.

[12] S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 7, pp. 674–693, Jul. 1989.

[13] V. P. Shah, N. H. Younan, S. S. Durbha, and R. L. King, "Feature identification via a combined ICA–wavelet method for image information mining," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 1, pp. 18–22, Jan. 2010.

[14] P. Liu, F. Huang, G. Li, and Z. Liu, "Remote-sensing image denoising using partial differential equations and auxiliary images as priors," *IEEE Geosci. Remote Sens. Lett.*, vol. 9, no. 3, pp. 358–362, May 2012.

[15] P. Liu and K. B. Eom, "Restoration of multispectral images by total variation with auxiliary image," *Opt. Lasers Eng.*, vol. 51, no. 7, pp. 873–882, 2013.

[16] M. Gonzalez-Audicana, J. L. Saleta, R. G. Catalan, and R. Garcia, "Fusion of multispectral and panchromatic images using improved IHS and PCA mergers based on wavelet decomposition," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 6, pp. 1291–1299, Jun. 2004.

[17] T. Celik and K.-K. Ma, "Unsupervised change detection for satellite images using dual-tree complex wavelet transform," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 3, pp. 1199–1210, Mar. 2010.

[18] L. M. Bruce, C. H. Koger, and J. Li, "Dimensionality reduction of hyperspectral data using discrete wavelet transform feature extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 10, pp. 2331–2338, Oct. 2002.

[19] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Trans. Signal Process.*, vol. 46, no. 4, pp. 886–902, Apr. 1998.

[20] J. Liu and P. Moulin, "Information-theoretic analysis of interscale and intrascale dependencies between image wavelet coefficients," *IEEE Trans. Image Process.*, vol. 10, no. 11, pp. 1647–1658, Nov. 2001.

[21] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 613–627, May 1995.

[22] S. G. Chang, B. Yu, and M. Vetterli, "Adaptive wavelet thresholding for image denoising and compression," *IEEE Trans. Image Process.*, vol. 9, no. 9, pp. 1532–1546, Sep. 2000.

[23] L. Sendur and I. W. Selesnick, "Bivariate shrinkage with local variance estimation," *IEEE Signal Process. Lett.*, vol. 9, no. 12, pp. 438–441, Dec. 2002.

[24] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli, "Image denoising using scale mixtures of Gaussians in the wavelet domain," *IEEE Trans. Image Process.*, vol. 12, no. 11, pp. 1338–1351, Nov. 2003.

[25] H. Pi, C. S. Tong, S. K. Choy, and H. Zhang, "A fast and effective model for wavelet subband histograms and its application in texture image retrieval," *IEEE Trans. Image Process.*, vol. 15, no. 10, pp. 3078–3088, Oct. 2006.

[26] S. K. Choy and C. S. Tong, "Statistical wavelet subband characterization based on generalized gamma density and its application in texture retrieval," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 281–289, Feb. 2010.

[27] G. Fan and X.-G. Xia, "Image denoising using a local contextual hidden Markov model in the wavelet domain," *IEEE Signal Process. Lett.*, vol. 8, no. 5, pp. 125–128, May 2001.

[28] H. Xie, L. E. Pierce, and F. T. Ulaby, "SAR speckle reduction using wavelet denoising and Markov random field modeling," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 10, pp. 2196–2212, Oct. 2002.

[29] J. Li, R. M. Gray, and R. A. Olshen, "Multiresolution image classification by hierarchical modeling with two-dimensional hidden Markov models," *IEEE Trans. Inf. Theory*, vol. 46, no. 5, pp. 1826–1841, Aug. 2000.

[30] H. Choi and R. G. Baraniuk, "Multiscale image segmentation using wavelet-domain hidden Markov models," *IEEE Trans. Image Process.*, vol. 10, no. 9, pp. 1309–1321, Sep. 2001.

[31] G. Fan and X.-G. Xia, "Wavelet-based texture analysis and synthesis using hidden Markov models," *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. 50, no. 1, pp. 106–120, Jan. 2003.

[32] L. He and L. Carin, "Exploiting structure in wavelet-based Bayesian compressive sensing," *IEEE Trans. Signal Process.*, vol. 57, no. 9, pp. 3488–3497, Sep. 2009.

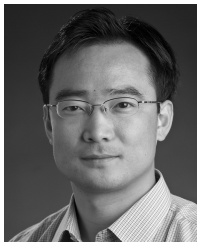
[33] S. Som and P. Schniter, "Compressive imaging using approximate message passing and a Markov-tree prior," *IEEE Trans. Signal Process.*, vol. 60, no. 7, pp. 3439–3448, Jul. 2012.

[34] J. K. Romberg, H. Choi, and R. G. Baraniuk, "Bayesian tree-structured image modeling using wavelet-domain hidden Markov models," *IEEE Trans. Image Process.*, vol. 10, no. 7, pp. 1056–1068, Jul. 2001.

[35] J.-B. Durand, P. Goncalves, and Y. Guedon, "Computational methods for hidden Markov tree models-an application to wavelet trees," *IEEE Trans. Signal Process.*, vol. 52, no. 9, pp. 2551–2560, Sep. 2004.

[36] I. Daubechies, *Ten Lectures on Wavelets*. Philadelphia, PA, USA: SIAM, 1992.

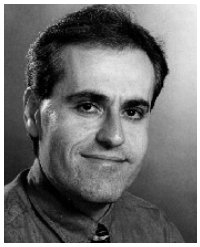
[37] R. C. Gonzalez and R. E. Woods, *Digital Image Processing (2nd Edition)*. Englewood Cliffs, NJ, USA: Prentice-Hall, Jan. 2002.



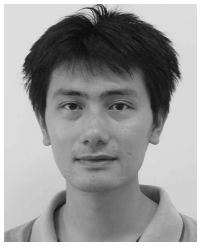
LIZHE WANG (SM'09) received the B.Eng. (Hons.) and M.Eng. degrees from Tsinghua University, Beijing, China, and the Dr.Ing. (*magna cum laude*) degree from the Karlsruhe Institute of Technology, Karlsruhe, Germany. He is currently a 100-Talent Program Professor with the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing, and a ChuTian Chair Professor with the School of Computer Science, China University of Geosciences, Wuhan, China. He is a fellow of the Institution of Engineering and Technology and the British Computer Society. His research interests include cluster/grid/cloud computing, high-performance geocomputing, and spatial information processing.

HUI ZHONG received the B.E. degree in communication engineering from Anhui University, Hefei, China, in 2012. She is currently pursuing the master's degree with the University of Science and Technology of China, Hefei. Since 2013, she has been an Intern with the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing, China. Her research interests include sparse coding and correlation analysis.

RAJIV RANJAN is currently a Research Scientist and Julius Fellow with the CSIRO Computational Informatics Division, Brisbane, QLD, Australia. His expertise is in datacenter cloud computing, application provisioning, and performance optimization. He received the Ph.D. degree in engineering from the University of Melbourne, Melbourne, VIC, Australia, in 2009. He has authored 62 scientific and peer-reviewed papers (seven books, 25 journals, 25 conferences, and five book chapters). His H-index is 20, with a life-time citation count of over 1660 (Google Scholar). His papers have also received over 140 ISI citations. His 70% of journal papers and 60% of conference papers have been A*/A ranked ERA publications. He has been invited to serve as the Guest Editor for leading distributed systems journals, including the *IEEE Transactions on Cloud Computing, Future Generation Computing Systems, and Software Practice and Experience*. One of his papers was in 2011s top computer science journal, *IEEE Communication Surveys and Tutorials*.



ALBERT ZOMAYA (F'04) is currently the Chair Professor of High-Performance Computing and Networking, and an Australian Research Council Professorial Fellow with the School of Information Technologies, University of Sydney, Sydney, NSW, Australia. He is also the Director of the Centre for Distributed and High Performance Computing, University of Sydney, which was established in 2009. He was the CISCO Systems Chair Professor of Internetworking, from 2002 to 2007, and the Head of the School of Information Technologies from 2006 to 2007. He was a Full Professor with the School of Electrical, Electronic and Computer Engineering, University of Western Australia, Crawley, WA, USA, where he led the Parallel Computing Research Laboratory from 1990 to 2002. He served as the Associate, Deputy, and Acting Head of the Department of Electrical, Electronic and Computer Engineering, held numerous visiting positions, and has extensive industry involvement. He received the Ph.D. degree from the Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield, U.K.



PENG LIU received the M.S. and Ph.D. degrees in signal processing from the Chinese Academy of Sciences, Beijing, China, in 2004 and 2009, respectively. Since 2009, he has been an Assistant Professor with the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences. From 2012 to 2013, he was with the Department of Electrical and Computer Engineering, George Washington University, Washington, DC, USA, as a Visiting Scholar. His research is focused on sparse representation, compressive sensing, image processing, and their applications to remote sensing.