Sparse Hierarchical Clustering for VHR Image Change Detection

Kun Ding, Chunlei Huo, Yuan Xu, Zisha Zhong, and Chunhong Pan

***Abstract*—The traditional clustering approaches are limited for the unsupervised change detection of very high resolution images due to the multimodal distribution of change features. To overcome this difficulty, a sparse hierarchical clustering approach is proposed. Discriminative change features are generated by stacking bitemporal multiscale center-symmetric local binary pattern features. In order to explore the multimodal and hierarchical distribution of the change features, a tree-structured dictionary is learned from the pseudotraining set and the unlabeled data. The sparse reconstruction error, a more robust distance compared to the Euclidean distance, is used to determine the label of each change feature. Comparative experiments demonstrate the effectiveness of the proposed method.**

***Index Terms*—Change detection, multimodal distribution, sparse hierarchical clustering (SHC).**

I. Introduction

CHANGE detection aims at detecting land-cover transitions from the coregistered remote sensing images taken over the same geographic area but at different times. It is important for practical applications such as disaster management, urban studies, etc. In general, the traditional change detection approaches consist of two steps: change feature extraction and change map generation by classification or clustering. The classification-based change detection methods [1], [2] require hand-labeled training samples, while the clustering-based methods [3], [4] can automatically split the change features into disjoint categories. However, the main difficulties in applying clustering to very high resolution (VHR) image change detection are the complex distribution of the change feature as explained by Fig. 1 and the limitation of the traditional clustering approaches for such a complex distribution.

In the literature, many change detection approaches construct change feature by differencing [3], [5] or stacking [1] the bitemporal features, e.g., spectral feature. Differencing is sensitive to the atmosphere, lighting, and seasonal variation. In addition, the low spectral resolution of VHR images results in high intraclass variability and low interclass variability [6]. The information loss caused by differencing will further aggravate this difficulty. Compared to differencing, stacking is more promising in improving interclass variability. However, the distribution of change feature by stacking is very complicated.

Manuscript received July 11, 2014; accepted August 10, 2014. Date of publication September 15, 2014; date of current version October 8, 2014. This work was supported by the Natural Science Foundation of China under Grants 61375024, 91338202, 61272331, and 61305049.

The authors are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: kding@nlpr.ia.ac.cn; clhuo@nlpr.ia.ac.cn; yxu@nlpr.ia.ac.cn; zszhong@nlpr.ia.ac.cn; chpan@nlpr.ia.ac.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.2014.2351807

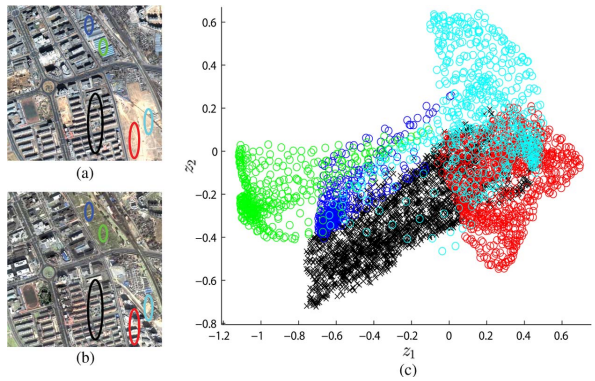

Fig. 1. Multimodal distribution of change feature. (a) Image $\mathbf{X}_1$. (b) Image $\mathbf{X}_2$. (c) Distribution of change feature. "o" and "x" stand for the changed and unchanged classes, respectively. Details are in Section II-A.

For illustration, Fig. 1(c) plots some change features generated by projecting the high-dimensional stacked center-symmetric local binary patterns (CS-LBP) [7] features to the 2-D principal component space. Apparently, the samples of the changed class (marked with "o") spread in wide range with multiple dense regions. Such a multimodal distribution makes the traditional clustering approaches, e.g., K-means [3], [4], difficult to cluster the change features. Note that these K-means-based methods represent each class by one center and assign a label for each change feature based on the Euclidean distance. However, one center is inadequate to capture the complex distribution, and the Euclidean distance is less robust to classify the change features.

In this letter, a sparse-representation-based hierarchical clustering approach is proposed to address the aforementioned problems. Compared with the related change detection approaches [3], [4], the contributions of the proposed approach are twofold: 1) A tree-structured dictionary is learned from all of the change features to represent the multimodal distribution. This structure helps in capturing the multimodal nature of change feature. 2) Sparse reconstruction error (SRE) is utilized to measure the sample-to-class distance. The sparsity makes the error-based distance robust to false changes.

II. Proposed Approach

The proposed change detection approach consists of two steps: change feature extraction and sparse hierarchical clustering (SHC). These two steps are the basis of the proposed method, and they will be elaborated in the following sections.

A. Change Feature Extraction

The change feature should be discriminative to capture the salient structures of VHR images and robust to lighting condition variations, seasonal changes, and sensor noise. The simple

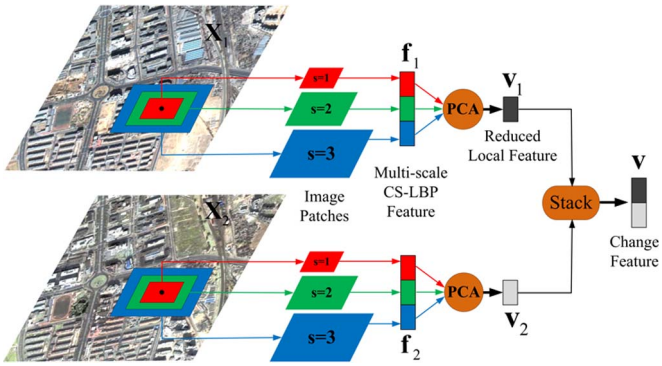

Fig. 2. Illustration of change feature extraction.

spectral feature has difficulty in meeting the aforementioned requirements simultaneously. This letter employs the CS-LBP feature [7], as it is robust to flat areas, tolerant to illumination changes, and efficient in computation.

Change is closely related to scale (i.e., the size of the observation window), and different changes can be detected at different scales. For this reason, the CS-LBP feature is extracted in a multiscale manner. Fig. 2 illustrates the feature extraction process using three different concentric patches.

For the coregistered multitemporal VHR images $\mathbf{X}_1$ and $\mathbf{X}_2$, CS-LBP features are computed at each pixel at S different scales. For pixel (i, j) of image $\mathbf{X}_t$ $(t = 1, 2)$, the multiscale CS-LBP feature $\mathbf{f}_t = [\mathbf{f}_{t1}^T, \ldots, \mathbf{f}_{tS}^T]^T$ is formed by stacking all of the single-scale CS-LBP features $\mathbf{f}_{ts}$ $(s = 1, \ldots, S)$. Principal component analysis (PCA) is then employed to reduce the dimension of multiscale CS-LBP features and merge the information from various scales (Fig. 2). Let us denote the reduced version of $\mathbf{f}_1$ and $\mathbf{f}_2$ as $\mathbf{v}_1$ and $\mathbf{v}_2$, respectively. By stacking $\mathbf{v}_1$ and $\mathbf{v}_2$, the final change feature $\mathbf{v} = [\mathbf{v}_1^T, \mathbf{v}_2^T]^T$ is formed. It is used in our SHC algorithm.

We observed that the change feature computed by the aforementioned strategy is of multimodal distribution,[1] which can be explained by Fig. 1. In Fig. 1(a) and (b), an unchanged region (ellipse in black) and some changed regions (ellipses in other colors) are labeled manually, and the change features $\mathbf{v}$ are extracted within these ellipses. To validate the multimodal distribution of the stacked change feature $(\mathbf{v})$, we use PCA to reduce the dimension of $\mathbf{v}_1$ and $\mathbf{v}_2$ to one, respectively, and get two scalars z_1 and z_2. Stacking them generates the 2-D change feature $\mathbf{z} = [z_1, z_2]^T$. Fig. 1(c) plots each $\mathbf{z}$ with the same color with the ellipses in Fig. 1(a) and (b). From Fig. 1(c), the black "x"s form an unchanged class modality, while other colored "o"s form a changed class modality. Hence, the distribution of $\mathbf{z}$ $[p(\mathbf{z})]$ is bimodal. As $\mathbf{z}$ is the linear projection of $\mathbf{v}$, the distribution of $\mathbf{v}$ $[p(\mathbf{v})]$ is also bimodal. Moreover, the changed class (marked with "o") has multiple modalities: the green modality, the red modality, and other modalities. In detail, the green modality indicates the change from the buildings to the grass, while the red modality is the change from the land to the buildings. In consequence, both the conditional distributions $p(\mathbf{z}|\text{change})$ and $p(\mathbf{v}|\text{change})$ are multimodal. In short, one center is not representative enough for each class (the changed and unchanged classes).

[1] In this letter, the item "multimodal" is to specify the multimodal distribution (i.e., a statistical distribution of values with multiple peaks) rather than the different acquisition modalities (e.g., radar and optical).

B. SHC

In this letter, SHC is proposed to cluster the aforementioned change features. In this method, SRE is adopted to measure the sample-to-class distance. The sparsity in SRE makes it more robust than the Euclidean distance used in K-means clustering [3], [4]. The robustness helps in reducing the false changes caused by registration error and lighting variation. On the other hand, the hierarchical dictionary learning is employed to capture the complex distribution of change features. This dictionary learning method has three advantages over the traditional one [8]: 1) Each dictionary atom is a cluster center that has the explicit meaning; 2) the hierarchical structure of the dictionary is helpful in reducing the false alarms; and 3) the dictionary updating strategy is simple to implement. In detail, SHC consists of three steps: 1) dictionary initialization; 2) label assignment; and 3) dictionary updating. Steps 2) and 3) are iterated alternatively until the convergence is reached.

1) Dictionary Initialization: This step is to learn an initial structured dictionary that roughly models the distribution of change feature. An appropriate initialization is helpful for the stable clustering.

A pseudotraining set is required to produce the initial dictionary. An intuitive idea to obtain the pseudotraining set is picking the most reliable samples that are much easier to be identified as change or no change in an unsupervised manner. Given the bitemporal CS-LBP features $\mathbf{f}_{13}$ and $\mathbf{f}_{23}$ at the scale $s = 3$, the magnitude of the CS-LBP change vector is $a = \|\mathbf{f}_{13} - \mathbf{f}_{23}\|_2$, where $\|\cdot\|_2$ denotes the Euclidean norm of a vector. As [2], the magnitude is assumed to be the bimodal Gaussian mixture distribution (GMD), where one modal near the origin stands for the unchanged class and the other corresponds to the changed class. Therefore, the reliability of a training sample can be measured by the magnitude a. Formally, the pseudotraining set is defined as

$$\mathcal{L} = \{\mathbf{v} | a \geq \theta T + (1-\theta)\mu_c \; or \; a \leq (1-\theta)\mu_u + \theta T\}. \quad (1)$$

T is the Bayesian optimal threshold that separates the changed class from the unchanged class, which can be obtained by maximizing *a posteriori* (MAP). θ controls the number of selected pseudotraining samples. μ_c and μ_u are the mean values of the GMD, and they are estimated by the expectation maximization (EM) algorithm.

For convenience, we denote the set of unlabeled data as $\mathcal{U}$. In addition, a new set $\mathcal{F}$ is defined to denote the samples whose labels need assigning or updating. It is related to $\mathcal{U}$ and $\mathcal{L}$ and has different formulations at the different stages of SHC. In the dictionary initialization step, $\mathcal{F} = \mathcal{L}$ since only the labeled pseudotraining set is needed. At the first iteration ($\tau = 1$, where τ represents the number of iterations) of the label assignment step, only the labels of unlabeled data need being assigned, and thus, $\mathcal{F} = \mathcal{U}$. At the remaining iterations ($\tau \geq 2$), $\mathcal{F} = \mathcal{L} \cup \mathcal{U}$, since the labels of all of the samples need to be updated.

Given the pseudotraining set $\mathcal{F} = \mathcal{L}$, we compute the initial dictionary via hierarchical K-means as illustrated in Fig. 3. For the unchanged class $\mathcal{W}^u$, all of the pseudotraining samples belonging to this class are clustered into N subclasses $\{\mathcal{C}_i^u\}_{i=1}^N$. Hence, the dictionary of this class is

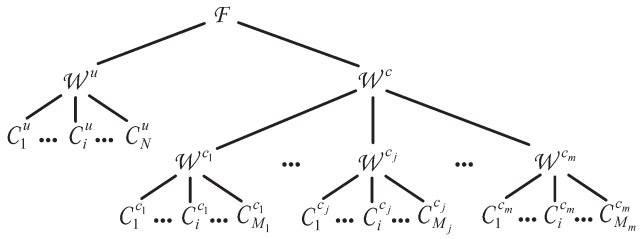

Fig. 3. Tree structure of clustering. Each node is a subset of the change features, and each child node is a subset of its father node.

$\mathbf{D}^u = [\mathbf{d}_1^u, \ldots, \mathbf{d}_i^u, \ldots, \mathbf{d}_N^u]$, where $\mathbf{d}_i^u$ is the cluster center of class $\mathcal{C}_i^u$. For the changed class $\mathcal{W}^c$, considering the multimodal distribution, all of the samples having the pseudolabel "change" are grouped into m $(m \geq 1)$ subclasses $\{\mathcal{W}^{c_j}\}_{j=1}^m$; each class $\mathcal{W}^{c_j}$ is further clustered into M_j clusters $\{\mathcal{C}_i^{c_j}\}_{i=1}^{M_j}$. The dictionary of class $\mathcal{W}^c$ is denoted as $\mathbf{D}^c = [\mathbf{D}^{c_1}, \ldots, \mathbf{D}^{c_j}, \ldots, \mathbf{D}^{c_m}]$, where $\mathbf{D}^{c_j} = [\mathbf{d}_1^{c_j}, \ldots, \mathbf{d}_i^{c_j}, \ldots, \mathbf{d}_{M_j}^{c_j}]$ $(j = 1, \ldots, m)$, and $\mathbf{d}_i^{c_j}$ is the cluster center of class $\mathcal{C}_i^{c_j}$. Concatenating $\mathbf{D}^u$ and $\mathbf{D}^c$ produces a tree-structured dictionary $\mathbf{D} = [\mathbf{D}^u, \mathbf{D}^c]$. The parameters satisfy the constraints: $M = \sum_{j=1}^m M_j$ and $M_j/M = |\mathcal{W}^{c_j}|/|\mathcal{W}^c|$ $(j = 1, 2, \ldots, m)$, where $|\cdot|$ denotes the size of a set. N and M are the dictionary sizes of the unchanged and changed classes, respectively. Note that N and M are related to the number of clustering center; the excessive unbalance between them would produce biased classification results. Therefore, we set them to be equal. After the initialization step, a tree-shaped clustering structure (Fig. 3) is built, which captures the hierarchical distribution of the change feature in Fig. 1(c).

Since the pseudotraining set is a part of all of the samples, the initial dictionary is difficult to model the statistics of the data completely. In consequence, it is necessary to take the unlabeled change features into account and refine the dictionary. The refinement process is implemented by alternatively iterating the following two steps.

2) Label Assignment: Given the dictionary $\mathbf{D}$ learned at the $(\tau - 1)$th iteration, only the labels of the unlabeled change features (i.e., the samples in $\mathcal{U}$) are assigned $(\tau = 1)$, or the labels of all of the change features (i.e., the samples in $\mathcal{L} \cup \mathcal{U}$) are assigned $(\tau > 1)$. An appropriate distance is needed for assigning the labels. Since the SRE is more robust than the Euclidean distance [9], it is employed at the nodes $\mathcal{F}$ and $\mathcal{W}^c$, while at the nodes $\mathcal{W}^u$ and $\mathcal{W}^{c_j}$ $(j = 1, \ldots, m)$, the Euclidean distance is adopted for its sufficiency in measuring the distance locally and its computational efficiency.

For label assignment, two SREs should be computed for each sample $\mathbf{v} \in \mathcal{F}$

$$e^l = \left\| \mathbf{v} - \mathbf{D}^l \hat{\boldsymbol{\alpha}}^l \right\|_2^2, \ l \in \{u, c\} \tag{2}$$

where the column vector $\hat{\boldsymbol{\alpha}} = [(\hat{\boldsymbol{\alpha}}^u)^T, (\hat{\boldsymbol{\alpha}}^c)^T]^T$ is the best representing coefficient under the structured dictionary $\mathbf{D} = [\mathbf{D}^u, \mathbf{D}^c]$, and it can be obtained by solving

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \|\mathbf{v} - \mathbf{D}\boldsymbol{\alpha}\|_2^2, \text{ s.t. } \|\boldsymbol{\alpha}\|_0 \leq L \tag{3}$$

where $\|\cdot\|_0$ is the l_0-norm, which counts the number of nonzero entries in a vector. L controls the sparsity of $\hat{\boldsymbol{\alpha}}$. The aforementioned problem can be solved by the orthogonal matching pursuit algorithm [10]. Given e^u and e^c, the label of $\mathbf{v}$ will be $\mathcal{W}^{\hat{l}}$, where

$$\hat{l} = \arg\min_{l \in \{u, c\}} \ e^l. \tag{4}$$

The reason that the category of $\mathbf{v}$ can be determined by (4) lies in the following fact: if $\mathbf{v}$ belongs to the class $\mathcal{W}^u$, then the nonzero elements of $\hat{\boldsymbol{\alpha}}$ will concentrate in $\hat{\boldsymbol{\alpha}}^u$, and the representation error e^u will be much smaller than e^c and vice versa. The aforementioned class-determination process is consistent with the sparse-representation-based classification proposed in [9].

As illustrated by Fig. 3, the sublabels of $\mathbf{v}$ can be determined sequentially based on the label of its father node $\mathcal{W}^u$ or $\mathcal{W}^c$. In detail, if the label of $\mathbf{v}$ is $\mathcal{W}^u$ and $\hat{i} = \arg\min_{i \in \{1, \ldots, N\}} \|\mathbf{v} - \mathbf{d}_i^u\|_2$, $\mathbf{v}$ will be assigned to the class $\mathcal{C}_{\hat{i}}^u$. Otherwise, the subclass label $\mathcal{W}^{c_{\hat{j}}}$ $(\hat{j} \in \{1, 2, \ldots, m\})$ can be assigned to $\mathbf{v}$ by a multiclass extension of (2)–(4). The extension is naive by replacing the dictionary $\mathbf{D} = [\mathbf{D}^u, \mathbf{D}^c]$ with $\mathbf{D}^c = [\mathbf{D}^{c_1}, \ldots, \mathbf{D}^{c_j}, \ldots, \mathbf{D}^{c_m}]$ and by replacing the set $\{u, c\}$ with $\{c_1, \ldots, c_j, \ldots, c_m\}$ in these equations. Furthermore, the subclass label $\mathcal{C}_{\hat{i}}^{c_{\hat{j}}}$ $(\hat{i} \in \{1, 2, \ldots, M_{\hat{j}}\})$ is assigned to $\mathbf{v}$, where $\hat{i} = \arg\min_{i \in \{1, \ldots, M_{\hat{j}}\}} \|\mathbf{v} - \mathbf{d}_i^{c_{\hat{j}}}\|_2$.

3) Dictionary Updating: Dictionary updating is to approximate the real distribution of change feature progressively by achieving a new dictionary (i.e., a collection of cluster centers) based on the updated labels. More specifically, for a class $\mathcal{C}_i^j$ $(j \in \{c_1, \ldots, c_m\}, i \in \{1, \ldots, M_j\})$ or $\mathcal{C}_i^u$ $(i \in \{1, \ldots, N\})$, the new cluster center is the mean vector of all of the labeled and unlabeled data that belong to it, and the new dictionaries are $\mathbf{D}^{c_j} = [\mathbf{d}_1^{c_j}, \ldots, \mathbf{d}_i^{c_j}, \ldots, \mathbf{d}_{M_j}^{c_j}]$ $(j = 1, \ldots, m)$, $\mathbf{D}^c = [\mathbf{D}^{c_1}, \ldots, \mathbf{D}^{c_j}, \ldots, \mathbf{D}^{c_m}]$, $\mathbf{D}^u = [\mathbf{d}_1^u, \ldots, \mathbf{d}_i^u, \ldots, \mathbf{d}_N^u]$, and $\mathbf{D} = [\mathbf{D}^u, \mathbf{D}^c]$.

III. Experimental Results

A. Data Set Description

For space limitation, two data sets (DS1 and DS2) are used to assess the effectiveness of the proposed change detection method. The data sets and the corresponding reference change maps are shown in Fig. 4(a)–(c). Each data set is composed of two coregistered and pansharpened VHR images taken by QuickBird 2 satellite over Beijing in 2002 and 2003. The resolution of these pansharpened images is 0.7 m/pixel. The image sizes of DS1 and DS2 are 1024×1024 and 600×519 pixels, respectively.

B. Experimental Settings

To investigate SHC in detail, it is compared with the following five related methods.

1) EM-based method (EM) [11]. For the EM-based method, the magnitude $\|\mathbf{f}_1 - \mathbf{f}_2\|_2$ is utilized and assumed to obey a GMD. The optimal threshold for generating the binary change map is computed by MAP.
2) K-means-based method (K-means). K-means clusters the stacked change feature $\mathbf{v}$ into two classes. To obtain better performances, we make two modifications: a) The two cluster centers are initialized on the pseudotraining set that is used in SHC, and b) the labels of the initial labeled

Fig. 4. Data sets and change maps. From up to down: DS1 and DS2. From left to right: images in 2002, images in 2003, ground truth (GT), EM-based method, K-means-based method, parcel-based method, IR-MAD method, SVM-based method, SHC with $m = 1$, SHC with $m = 2$, and SHC with $m = 3$. In (c)–(i), the changed class is in red. In (j), the changed class is in red and cyan, while in (k), the changed class is in red, green, and blue. Different colors mean different change types (best viewed in color). (a) 2002. (b) 2003. (c) GT. (d) EM. (e) K-means. (f) Parcel. (g) IR-MAD. (h) SVM. (i) SHC-1. (j) SHC-2. (k) SHC-3.

samples keep fixed in the label updating step. Note that updating these labels will make K-means fail due to the complex change feature distribution.

3) Parcel-based method (parcel) [5]. In this method, the parameters in hierarchical segmentation are tuned to obtain the best performances.
4) Regularized iteratively reweighted multivariate alteration detection method (IR-MAD) [12].
5) Support vector machine based method (SVM) [1]. SVM directly classifies the stacked features $\mathbf{v}$'s. The same pseudotraining samples as SHC are used to train SVM. RBF kernel SVM is adopted, and its parameters are selected by fivefold cross-validation.

For a fair comparison, K-means, SVM, and SHC use the same pseudotraining sets. In the parcel-based method and IR-MAD method, we use the multiscale CS-LBP feature ($\mathbf{f}_1$ and $\mathbf{f}_2$) instead of spectral feature as we found that CS-LBP can get better performances. The thresholds for getting the change maps in these two methods are manually set to obtain the lowest total error rates (TERs). Comparison is made qualitatively by checking the final change maps and quantitatively by computing false alarm rate (FAR), missed alarm rate (MAR), TER, and kappa coefficient.

For our method, m determines the structure of clustering. As the change types contained in DS1 and DS2 are not very rich, we change m from 1 to 3 and denote the approach as SHC-m ($m \in \{1, 2, 3\}$).

The influence of θ, L, and N on FAR, MAR, and TER is shown in Fig. 5, where the curves are presented for DS1 when $m = 2$, and similar results could be obtained on the other data set when $m \geq 1$. First, the increase of θ will increase the lower threshold and decrease the higher threshold in (1), which enlarges the number of pseudotraining samples and mislabeled samples simultaneously. Therefore, FAR increases slightly with θ [Fig. 5(a)]. Second, the increased L will result in the increased FAR and the reduced MAR [Fig. 5(b)]. The underlying reason is the dropped discriminative ability of SRE. In consequence, L could be set to be a small value for different data sets, e.g., five. Finally, the dictionary is required to be overcompleted; a too small dictionary size will produce high FAR and MAR. When the dictionary size is large enough, its effects on TER can be ignored [Fig. 5(c)]. We recommend to set the dictionary size according to the image size and the change richness, and the balance between the performance and the computational complexity. Considering these, we use the settings $\theta = 0.15$, $L = 5$, and $N = M = 1200$ for both data sets.

Besides the aforementioned parameters, some other parameters also need to be set properly. The number of scales S is set to be three. A higher image resolution enables the use of larger scales. The window size parameters h_s and w_s determine the spatial range of the extracted feature. A larger window would strengthen the robustness of change features to the false changes caused by registration error and viewpoint change but may lose some detail information. For DS1, the windows with the sizes $h_s = w_s = 32, 48, 64$ are used, since it contains mainly large-scale changes. Considering that there are some subtle changes, $h_s = w_s = 24, 32, 40$ are adopted in DS2. The dimension of the multiscale local features ($\mathbf{f}_1$ and $\mathbf{f}_2$) after PCA is 200, which retains about 95% of the total energy and generates a 400-dimensional change feature ($\mathbf{v}$).

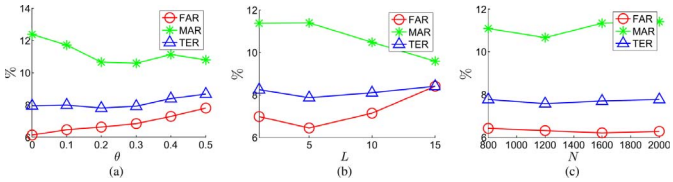

Fig. 5. Influence of θ, L, and N on FAR, MAR, and TER by SHC-2 on DS1.

C. Results and Analysis

The results of different approaches are shown in Fig. 4. The EM-based method [Fig. 4(d)] performs poor on both DS1 and DS2 because of the unreasonable assumption of the bimodal GMD and the pure usage of change vector magnitude. K-means [Fig. 4(e)] has many missed alarms and false alarms. This poor performance mainly lies in the multimodal distribution of the change feature and the limitation of utilizing K-means of only two clusters and the Euclidean distance to deal with the multimodal distribution problem. The parcel-based method is prone to detecting some fragment-like false changed regions [e.g., the areas marked by the green rectangles in Fig. 4(f)]. IR-MAD is robust to the illumination changes for the low-to-medium resolution images, but it is less effective for VHR images. It has missed some real changed areas on DS1 and regarded some false changes caused by the illumination variation as real changes on DS2 [e.g., the regions marked with the blue rectangles in Fig. 4(g)]. The results of SVM are comparable to that of SHC on DS1, since the kernel technique makes the multimodal change feature more separable in the high-dimensional space. However, due to the less representative training samples, some changed regions [e.g., some small regions in the yellow squares of Fig. 4(h)] are missed by SVM. By taking the advantages of SHC, they are correctly detected by the proposed approach.

The performances of the different approaches are listed in Table I. From the table, we can conclude that SHC is superior to other methods in terms of TER and kappa coefficient. The advantages of the proposed change detection approach are

TABLE I
PERFORMANCE COMPARISON

Dataset	Method	FAR (%)	MAR (%)	TER (%)	Kappa
DS1	EM	10.56	12.26	11.05	0.7420
	K-means	25.38	14.08	22.10	0.5300
	Parcel	7.39	22.19	11.69	0.7130
	IR-MAD	10.25	35.22	17.51	0.5623
	SVM	4.84	16.84	8.32	0.7949
	SHC-1	6.38	11.39	7.83	0.8121
	SHC-2	6.22	11.19	7.66	0.8161
	SHC-3	5.90	11.93	7.65	0.8156
DS2	EM	1.58	44.97	12.39	0.6174
	K-means	17.09	20.60	17.96	0.5648
	Parcel	5.57	46.42	15.89	0.5328
	IR-MAD	34.80	6.73	27.71	0.4434
	SVM	2.04	27.68	8.43	0.7573
	SHC-1	4.57	10.37	6.02	0.8411
	SHC-2	4.05	9.93	5.52	0.8537
	SHC-3	1.98	5.45	2.85	0.9241

mainly taken from the multiscale change feature representation, hierarchical dictionary, and SRE. The stacked multiscale CS-LBP feature improves the discriminative ability and the robustness of the change feature. As the hierarchical dictionary is learned from all features, it captures the multimodal distribution effectively. This dictionary makes both FAR and MAR low. Furthermore, SRE is robust to false changes, which further reduces FAR.

When comparing SHC-1, SHC-2, and SHC-3, the TER of SHC-2 (SHC-3) is lower than that of SHC-1 (SHC-2) on all data sets. The improvements could be attributed to the exploration of the distribution structure of the changed class. Besides the advantages in improving the performances, the achieved change structure partition is helpful in understanding the change types (e.g., from the inhomogeneous structure to the homogeneous structure). In contrast, EM, K-means, parcel, IR-MAD, and SVM can only provide the information on where the changes happened. As shown in the upmost figure of Fig. 4(j), the cyan regions denote the changes from the inhomogeneous structures (complex buildings and complex wasteland) to the homogeneous structures (grass and very simple buildings), and the red regions mean the changes from the homogeneous structures (wasteland) to the inhomogeneous structures (buildings) and other weak changes. Similarly, the change types can be inferred from other figures in the rightmost two columns of Fig. 4. The capability of recognizing change types is mainly attributed to the discriminant information contained in the CS-LBP features, i.e., similar local structures (e.g., buildings with different dimensions) are encoded by close features, and similar changes (e.g., changes from land to buildings with different dimensions) are grouped together by SHC.

IV. CONCLUSION

In this letter, a novel SHC approach has been presented for VHR image change detection. The promising performances on real data sets validate the effectiveness of the tree-structured dictionary learning and the sparse reconstruction error based distance, which are helpful in dealing with the complex distribution of the change feature. Future work will focus on combining the object level strategy with the proposed approach to further improve the performance.

REFERENCES

[1] M. Volpi, D. Tuia, F. Bovolo, M. Kanevski, and L. Bruzzone, "Supervised change detection in VHR images using contextual information and support vector machines," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 20, pp. 77–85, Feb. 2013.

[2] F. Bovolo, L. Bruzzone, and M. Marconcini, "A novel approach to unsupervised change detection based on a semisupervised SVM and a similarity measure," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 7, pp. 2070–2082, Jul. 2008.

[3] T. Celik, "Unsupervised change detection in satellite images using principal component analysis and K-means clustering," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 4, pp. 772–776, Oct. 2009.

[4] M. Volpi, D. Tuia, G. Camps-Valls, and M. Kanevski, "Unsupervised change detection with kernels," *IEEE Trans. Geosci. Remote Sens. Lett.*, vol. 9, no. 6, pp. 1026–1030, Nov. 2012.

[5] F. Bovolo, "A multilevel parcel-based approach to change detection in very high resolution multitemporal images," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 1, pp. 33–37, Jan. 2009.

[6] C. Huo, B. Fan, C. Pan, and Z. Zhou, "Combining local features and progressive support vector machine for urban change detection of VHR images," in *Proc. ISPRS Ann.*, 2012, pp. 221–226.

[7] M. Heikkilä, M. Pietikäinen, and C. Schmid, "Description of interest regions with center-symmetric local binary patterns," in *Proc. ICVGIP*, 2006, pp. 58–69.

[8] M. Aharon, M. Elad, and A. M. Bruckstein, "The K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representations," *IEEE Trans. Image Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[9] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[10] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Conf. Rec. 27th Asilomar Conf. Signals, Syst. Comput.*, 1993, pp. 40–44.

[11] L. Bruzzone and D. F. Prieto, "Automatic analysis of the difference image for unsupervised change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 38, no. 3, pp. 1171–1182, May 2000.

[12] A. A. Nielsen, "The regularized iteratively reweighted MAD method for change detection in multi- and hyperspectral data," *IEEE Trans. Image Process.*, vol. 16, no. 2, pp. 463–478, Feb. 2007.