Learning Relationship for Very High Resolution Image Change Detection

Chunlei Huo, *Member, IEEE*, Keming Chen, Kun Ding, Zhixin Zhou, and Chunhong Pan

***Abstract*—The difficulty of very high resolution image change detection lies in the low interclass separability between the changed class and the unchanged class. According to experiments, we found that this separability can be improved by mining the relationship contained in the training samples. Based on this observation, a supervised change detection approach is proposed in this paper based on relationship learning. The proposed approach begins with enriching the training samples based on their neighborhood relationship and label coherence; this relationship is then learned simultaneously with the classifier, and, finally, the latter classification performance benefits from the learned relationship. Experiments demonstrate the effectiveness of the proposed approach.**

***Index Terms*—Change detection, distance tuning, interclass couple, intraclass couple, relationship learning, target neighborhood.**

I. Introduction

COMPARED with static object information (e.g., shape, texture, structure) contained in a single image, the dynamic change information between multitemporal images is more important for certain applications, such as disaster management, land-cover mapping, urban planning, etc. Generally, change information is extracted by the image change detection technique. With the development of very high resolution (VHR) satellites (e.g., QuickBird 2, WorldView, and GeoEye 1), the spatial resolution of remote sensing images has been improved significantly, and detecting changes from VHR multitemporal images is more attractive since not only the changed region, but also the changed type, can be recognized.

Remote sensing image change detection has been studied for a long time, but most early efforts were directed to low-to-medium resolution images. In the literature [1]–[6], many change detection approaches have been proposed for low-to-medium resolution images. Despite substantial efforts by various researchers on VHR image change detection, the existing techniques fall short of meeting practical requirements. There are many factors that make VHR image change detection difficult, such as misregistration [7], pan-sharpening impact [8], viewpoint variation, seasonal change, etc. In this paper, we focus on discussing this problem from the perspective of the discriminative capabilities of change features.

Change features are aimed at measuring the change confidence at each element (e.g., pixel, patch, or object). In this context, the change detection performance is determined mainly by the discriminative capabilities of change features. In other words, if the feature difference due to the real changes is larger than that in the unchanged region or the false changes, the changed class will be easily separated from the unchanged class. However, compared with the low-to-medium resolution images, the interclass separability of VHR images between the changed class and the unchanged one is very low, which forms the key difficulty in VHR image change detection. To clarify this point, we analyze the reasons for this difficulty in detail in the following paragraphs.

For illustration, two pairs of multitemporal images are shown in Fig. 1. One pair was acquired by Landsat 5, and the other by QuickBird 2. As shown by Fig. 1(a) and (b), the changes between the low-resolution images are mainly expressed by the radiance variation. As a consequence, the changed class and the unchanged class can be well separated automatically even by thresholding the spectral difference [cf. Fig. 1(d)]. However, for VHR images, the changes are difficult to represent. As illustrated by Fig. 1(e) and (f), the spectral difference [Fig. 1(h)] or the high-dimensional DAISY [9] feature difference [Fig. 1(i)] is incapable of being the qualified change feature. In other words, the discriminative capabilities of the change features based on the spectral difference or the high-dimensional local feature representation are low. It is worth noting that the powerful abilities of the local feature descriptors [such as HOG (histogram of gradient) or Dense SIFT [10], DAISY [9], etc.] for other applications such as object recognition and image classification, are beyond suspicion. Among the possible reasons, one important factor that causes the low interclass separability is the lack of mining of the "relationship"[1] between training samples.

First, accurately capturing the relationships (distances) between samples is essential to improving interclass separability. Considering the low interclass variation of change features of VHR images, the high-dimensional DAISY difference vector is expected to be classified reliably with the help of the training samples. The result obtained by applying a support vector machine (SVM) on the DAISY difference vector is shown in

Manuscript received February 4, 2016; revised April 20, 2016; accepted May 12, 2016. Date of publication April 23, 2015; date of current version August 24, 2016. This work was supported in part by Natural Science Foundation of China under Grant 91438105, Grant 61375024, Grant 61302170, and Grant 91338202.

C. Huo, K. Ding, and C. Pan are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences Beijing 100190, China (e-mail: clhuo@nlpr.ia.ac.cn; kding@nlpr.ia.ac.cn; chpan@nlpr.ia.ac.cn).

K. Chen is with the Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China (e-mail: kmchen.ie@gmail.com).

Z. Zhou is with the Beijing Institute of Remote Sensing, Beijing 100191, China (email: zhixin.zhou@mail.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/JSTARS.2016.2569598

[1]In this paper, relationship means the neighborhood relationship between training samples and their label coherence. Relationship learning is different from "relational learning [11]," where "relational" refers to the internal or external relational structure describing the examples. In detail, in this paper, relationship is captured by interclass and intraclass couples, and relationship learning is implemented by metric learning.

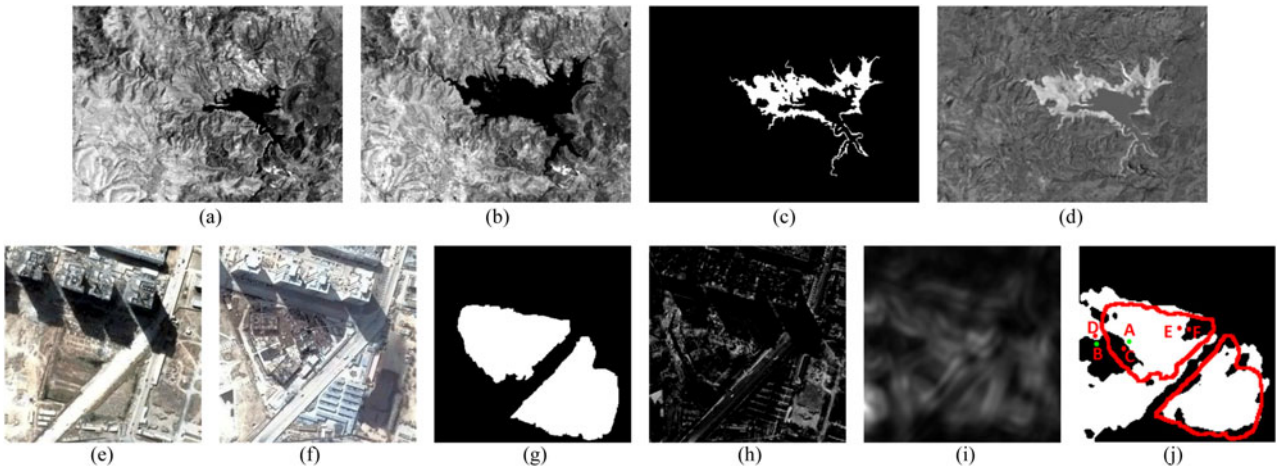


Fig. 1. Discriminative ability comparison of low resolution and VHR image change detection. (a)–(d) Low-resolution multitemporal images, reference ground truth, and spectral difference image. (e)–(h) VHR multitemporal images, reference ground truth, and spectral difference image. (i) Magnitude of DAISY difference. (j) Result achieved by SVM and DAISY-based change features. Training samples are chosen randomly from the ground truth. The red line in (j) is the border between the changed class and the unchanged class, and it is obtained based on (g).

Fig. 1(j), from which it is seen that the discriminative capabilities of the change features are improved effectively. This improvement demonstrates the importance of advanced feature representation and training samples, and this conclusion can be verified by comparing Fig. 1(h)–(j). Specifically, the uncertainties of the change features are diminished with the help of the training samples. However, as illustrated by pixels A–D in Fig. 1(j), it seems that the training samples lost their power near the border between the changed and unchanged regions. Although C and D are the nearest neighbors of A and B, respectively, they are being misclassified even with the help of the training samples. This implies that not only the closeness of change features in the feature space, but also that in the decision space, should be considered during the training procedure.

Second, accurately capturing the relationships (similarities) between samples is important for reducing intraclass difference. Pixels E and F in Fig. 1(j) are expected to be the changed class at the same time, but the computed change feature difference between them is large, which is harmful for the classification. In other words, the intraclass distance between E and F has to be diminished, guided by the training samples. Otherwise, the changed samples will be mixed up with the unchanged ones, and it is difficult to separate them.

From the above analysis, it can be inferred that it is necessary to learn the relationship contained in the training samples for improving the interclass separability and reducing the intraclass discrepancy. Therefore, a supervised change detection approach is proposed in this paper. The power of this approach lies in learning the relationship hidden in the training samples and tuning the distance between change features, in order to enlarge the interclass distance and penalize the intraclass difference. Despite the great efforts made to improve the interclass separability, the relationship hidden in the training samples is, as far as we know, usually ignored. To the best of our knowledge, no approaches that address the difficulty by learning the relationship between the training samples have been reported in the literature.

The paper is organized as follows: in Section II we describe related work, and in Section III we discuss the proposed approach. Experiments are detailed in Section IV, and conclusions drawn in Section V.

II. Related Work

Unsupervised change detection is preferred for practical applications, but it is very difficult to achieve due to the user-specific interests in defining the changes [12] and the low interclass variation between change features. In the context of supervised or semisupervised change detection, the challenge is to build a classifier driven by the training samples and to classify the change feature at each element as the changed or unchanged class. Specifically, let $\boldsymbol{x_i}$ denote the change feature from the ith element, and $y_i = -1$ and $y_i = 1$ denote the unchanged and changed class, respectively. The key to change detection is to learn a decision function $f(\cdot)$ directed by the training samples $\{(\boldsymbol{x_i}, y_i)|i = 1, \ldots, N\}$, where N is the number of training samples. Based on the decision function $f(\cdot)$, the label of the test sample $\boldsymbol{x}$ can be determined. The difference between the supervised and semisupervised approach lies in the manner of obtaining the training samples. At present, they are usually obtained by one of the following two ways: manual labeling, the automatically initial pseudotraining samples selection followed by the subsequent updating.

Many existing approaches attempt to improve performance by extracting the representative features or classifying change features with the help of sophisticated classifiers. To better understand the proposed approach, this section describes the related work from the perspectives of feature extraction and classification.

A. Feature Extraction

Considering the limitation of pixel-based radiometric or spectral features for representing the complex changes in VHR images, advanced change features that take into account the spatial dependence among neighboring pixels are proposed,

i.e., object-specific change features [13], [14], change features based on morphological attribute profiles [15], the compensated change feature based on DAISY features [16], the normalized moment of inertia feature [17], multilevel features [18], [19], the morphological building index [20], HOG descriptor based on multiple support regions [21], etc.

Despite the creativity of the above approaches, as illustrated by Fig. 1, the performance improvement is limited if the changes are measured directly based on the elementwise difference and no distance tuning mechanism utilized. This topic will be discussed in detail in Section IV.

B. Feature Classification

The labels of change features are usually determined by a classifier, such as pulse-coupled neural networks [17], [22], SVM [14], [23]–[25], simultaneous feature learning and change feature classification [26], high-order distance-based multiview stochastic learning [27], etc. To clarify the difference between the proposed approach and existing techniques, we briefly introduce the following three classifiers.

1) *kNN:* kNN (k-nearest neighbors) algorithm is one of the most popular classification tools since it does not assume any particular statistical distribution of the training samples and requires only one free parameter. The basic idea of kNN is to assign the test sample $\boldsymbol{x}$ to the class that appears most frequently within the k-nearest training samples, i.e.,

$$k\text{NN}(\boldsymbol{x}) = \text{sgn}\left(\sum_{i=1}^{k} y_{\mathcal{N}(\boldsymbol{x})}(i)\right) \tag{1}$$

where $y_{\mathcal{N}(\boldsymbol{x})}(i)$ is the class label of the ith nearest training sample. Despite the fact that kNN converges to the optimal Bayes bound under certain conditions [28], it generally fails to exploit the potential discrimination because the geometric relationship within the k-nearest training samples is ignored in the decision strategy.

2) *SVM:* Compared with kNN, SVM is more promising due to its superior generalization capability provided by the maximal margin principle. SVM aims to maximize the class-separating margin in the transformed feature space spanned by the mapping function $\phi(\cdot)$, i.e.,

$$\min\left(\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_i \xi_i\right)$$
$$\text{s.t.}\ \forall_{i=1}^{n} : y_i\left(\boldsymbol{w}\cdot\phi(\boldsymbol{x_i}) + \boldsymbol{b}\right) \geq 1 - \xi_i, \xi_i \geq 0. \tag{2}$$

For the sample $\boldsymbol{x}$, the label is determined by

$$\text{SVM}(\boldsymbol{x}) = \text{sgn}\left(\sum_{i=1}^{N} \alpha_i y_i \boldsymbol{K}(\boldsymbol{x_i}, \boldsymbol{x}) + b\right) \tag{3}$$

where $\boldsymbol{K}(\boldsymbol{x_i}, \boldsymbol{x}) = \phi(\boldsymbol{x_i})\cdot\phi(\boldsymbol{x})$, and N denotes the number of support vectors. By comparing (1) and (3), it can be observed that there are significant differences between kNN and SVM with respect to the decision rule and decision domain. Specifically, the decision domain of SVM is the entire support vector set, while the decision domain of kNN is the nearest neighbors. In other words, SVM ignores the local coherence between the test samples and the training samples by its global decision domain, and kNN overlooks the geometric structure hidden in the training samples by the local decision domain.

3) *SVM-kNN:* The combination of kNN and SVM is expected to address the above limitations [29], [30]. In particular, SVM-kNN [30] aims to capture the local structure by making decision on the local decision domain, i.e.,

$$\text{SVM-}k\text{NN}(\boldsymbol{x}) = \text{sgn}\left(\sum_{i=1}^{k} \alpha_{\mathcal{N}(\boldsymbol{x})}(i) y_{\mathcal{N}(\boldsymbol{x})}(i) \boldsymbol{K}(\boldsymbol{x}_{\mathcal{N}(\boldsymbol{x})}(i), \boldsymbol{x}) + b\right) \tag{4}$$

where k is the number of nearest training samples. Differently than the global decision taken over all of the support vectors by SVM, SVM-kNN assigns each test sample based on its nearest support vectors.

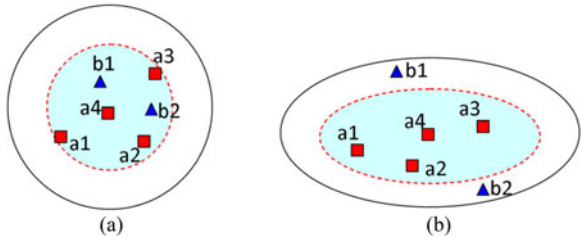


Fig. 2. Discriminative ability comparison of target neighbors and impostors before and after relationship learning. (a) Target neighbors and impostors before relationship learning. (b) Target neighbors and impostors after relationship learning.

III. Proposed Approach

As illustrated by Fig. 2, the rationale of the proposed approach is to improve the overall separability of change features by utilizing the neighborhood relationship between training samples and enlarging the interclass variation (especially near the border regions between the changed class and unchanged class). To this aim, the first step is to redefine the neighborhood relationship between change features, and the second step is to learn and tune the distance simultaneously with classifier learning. For convenience, the proposed approach is abbreviated rRL (robust Relation Learning).

A. Relationship Representation

Generally, the distance between two samples $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ is determined directly by the feature difference. For instance, in Fig. 2, the distance between b1 and a4 is smaller than that between a4 and a2, i.e., b1 is the nearest neighbor of a4. However, in the context of supervised learning, a2 is expected to be the nearest neighbor of a4 rather than b1. For this reason, the relationships between the training samples are redefined as follows:

Definition 1 (Target Neighbor): The sample $\boldsymbol{x_j}$ is the target neighbor of the sample $\boldsymbol{x_i}$ if $\boldsymbol{x_j}$ is one of the nearest neighbors of

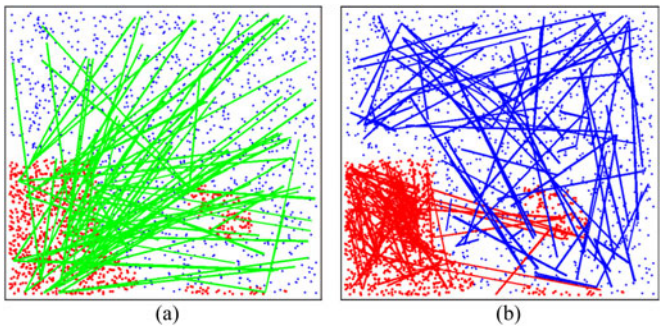


Fig. 3. Illustration of interclass and intraclass couples on dataset 1. (a) Interclass couples. (b) Intraclass couples. Original training samples are plotted with red dots and blue dots, respectively. In Fig. 3(a), each interclass couple is displayed by a green line, which connects two original training samples within nearest distances but with different labels. In Fig. 3(b), intraclass couples of changed and unchanged class are displayed by red lines and blue lines, respectively.

$\boldsymbol{x_i}$ and shares the same class label with $\boldsymbol{x_i}$, i.e., $\boldsymbol{x_j} \in \mathcal{N}(\boldsymbol{x_i})$ and $y_i = y_j$, where y_i and y_j are the labels of $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$, respectively. $\mathcal{N}(\boldsymbol{x_i})$ means the neighborhood of $\boldsymbol{x_i}$. For convenience, $\boldsymbol{z} = (\boldsymbol{x}_i, \boldsymbol{x}_j)$ is called an intraclass couple.

Definition 2 (Impostor): The sample $\boldsymbol{x_j}$ is an impostor of the sample $\boldsymbol{x_i}$ if $\boldsymbol{x_j}$ is one of the nearest neighbors of $\boldsymbol{x_i}$ but has a different class label with $\boldsymbol{x_i}$, i.e., $\boldsymbol{x_i} \in \mathcal{N}(\boldsymbol{x_j})$ and $y_i \neq y_j$. For convenience, $\boldsymbol{z} = (\boldsymbol{x}_i, \boldsymbol{x}_j)$ is called an interclass couple.

The definitions of target neighbor and impostor imply the following facts:

1) A qualified training sample should be supported by the samples within its neighborhood, i.e., an isolated training sample is not stable for training the reliable classifier.
2) The distance between a training sample and its neighbors should be adjusted adaptively, i.e., the distances between intraclass samples should be smaller than the distances between interclass samples. Otherwise, it is difficult for the learned classifier to separate the samples that are close but have different labels.

For any training sample $(\boldsymbol{x_i}, y_i)$, we can find $k1 (k1 > 1)$ target neighbors $\boldsymbol{x}_{i,m}^T (m = 1, \ldots, k1)$ and impostors $\boldsymbol{x}_{i,n}^I (n = 1, \ldots, k2)$. The constraints $k1 > 1$ and $k2 > 1$ are aimed at making the changed and unchanged regions smoother. For high-dimensional change features and a large training set, the nearest-neighbor searching procedure can be accelerated by taking advantage of k-dimensional tree strategy, which is usually used for scale-invariant feature transform (SIFT) feature matching.

In addition to the mechanism of utilizing the training samples for distance learning and tuning, the other advantage of the above relationship representation approach is enrichment of the training information by introducing additional virtual training couples (i.e., interclass and intraclass couples). As illustrated by Fig. 3, when 1000 training samples are chosen, $1000 * (k1 + k2)$ interclass and intraclass training couples will be generated. For this reason, richer training information can be utilized: more training samples are generated, and the intraclass similarity and interclass variability are simultaneously captured.

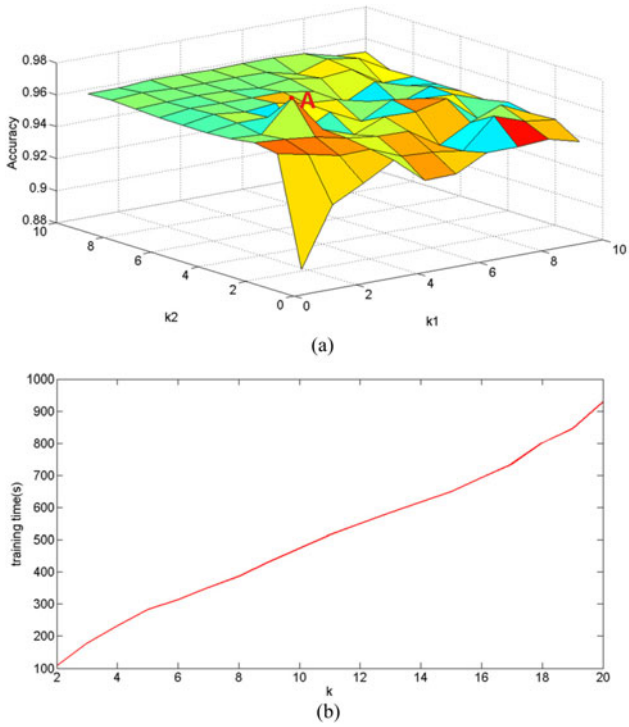


Fig. 4. Relation between the number of nearest neighbors k and classification performance as well as training time. (a) Classification accuracy versus $k1$ and $k2$. (b) Training time versus $k = k1 + k2$. Plot is based on dataset 3, and similar conclusions can be drawn from other datasets.

B. Relationship Learning

Given an interclass or intraclass couple $(\boldsymbol{x}_i, \boldsymbol{x}_j)$, relationship learning aims at learning and tuning the distance by $\boldsymbol{M}$ as follows:

$$d_{\boldsymbol{M}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i - \boldsymbol{x}_j)^T \boldsymbol{M} (\boldsymbol{x}_i - \boldsymbol{x}_j). \tag{5}$$

Specifically, $d_{\boldsymbol{M}}(\boldsymbol{x}_i, \boldsymbol{x}_j) \leq \tau_1$ is expected to hold if $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is an intraclass couple, and $d_{\boldsymbol{M}}(\boldsymbol{x}_i, \boldsymbol{x}_j) > \tau_2$ if $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is an interclass couple, where τ_1 is a relatively small value and τ_2 is a sufficiently large one. In other words, the distance between change features should not be computed directly based on the feature difference, but learned and tuned adaptively driven by the relationship hidden in the training samples.

In the literature, there are many novel approaches [31]–[37] for distance learning. For instance, Du [37] tried to increase the separability between anomalous pixels and other background pixels by exploiting a robust anomaly degree metric. Despite the great potential of metric learning for VHR image change detection, there are, to our knowledge, no related techniques in the literature.

Motivated by [34], a kernel-based metric learning approach is utilized for our relationship learning task. Other effective metric learning approaches can be used without any problem. However, since the focus of this paper is on improving the interclass separability of change features for VHR image change detection, the comparison of different approaches is beyond the scope of our discussion. More specifically, distance learning

and tuning functions are implemented by M, and it is obtained simultaneously with the classifier learning, i.e.,

$$\min_{\boldsymbol{M},\boldsymbol{b},\xi} \frac{1}{2}\|\boldsymbol{M}-\boldsymbol{I}\|_F^2 + C\sum_l \xi_l$$
$$\text{s.t. } h_l((\boldsymbol{x}_{l,1}-\boldsymbol{x}_{l,2})^T \boldsymbol{M}(\boldsymbol{x}_{l,1}-\boldsymbol{x}_{l,2})+b) \geq 1-\xi_l,$$
$$xi_l \geq 0 \quad \forall l \tag{6}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, and $\boldsymbol{I}$ is the identity matrix. $h_l = -1$ if $(\boldsymbol{x}_{l,1}, \boldsymbol{x}_{l,2})$ is an intraclass couple and $h_l = 1$ if $(\boldsymbol{x}_{l,1}, \boldsymbol{x}_{l,2})$ is an interclass couple. It is worth noting that $\|\boldsymbol{M}-\boldsymbol{I}\|_F^2$ is to be minimized in the objective function, which is different from [34], where $\|\boldsymbol{M}\|_F^2$ is used. The role of $\|\boldsymbol{M}-\boldsymbol{I}\|_F^2$ is to prevent the transformed change features from being distorted too much and to enhance the stability of the solution (distorted transformation means a high generalization error). The effectiveness of this modification will be verified in the next section.

Compared with the traditional training technique, the above matrix $\boldsymbol{M}$ is helpful for penalizing large distances between samples within the same class and small distances between samples from different classes. Considering the fact that $d^2_{\boldsymbol{M}}(\boldsymbol{x}_i,\boldsymbol{x}_j) = (\boldsymbol{x}_i-\boldsymbol{x}_j)^T\boldsymbol{M}(\boldsymbol{x}_i-\boldsymbol{x}_j) = (\boldsymbol{x}_i-\boldsymbol{x}_j)^T\boldsymbol{W}^T\boldsymbol{W}(\boldsymbol{x}_i-\boldsymbol{x}_j) = \|\boldsymbol{W}\boldsymbol{x}_i-\boldsymbol{W}\boldsymbol{x}_j\|_2^2$, the other role of $\boldsymbol{M}$ is to seek a transformation $\boldsymbol{W}$ that projects each sample $\boldsymbol{x}_i$ into a new subspace, under which the overall separability of the change features is improved.

To solve the above problem, the Lagrangian version is derived as follows:

$$\begin{aligned} L(\boldsymbol{M},\boldsymbol{b},\xi,\alpha,\beta) = {} & \frac{1}{2}\|\boldsymbol{M}-\boldsymbol{I}\|_F^2 + C\sum_l \xi_l \\ & -\sum_l \alpha_l \left[h_l((\boldsymbol{x}_{l,1}-\boldsymbol{x}_{l,2})^T\boldsymbol{M}(\boldsymbol{x}_{l,1}-\boldsymbol{x}_{l,2})\right. \\ & \left. +b)-1+\xi_l\right] - \sum_l \beta_l\xi_l \end{aligned} \tag{7}$$

where α and β are the Lagrange multipliers that satisfy $\alpha \geq 0$ and $\beta \geq 0 \quad \forall l$. To convert the original problem to its dual version, we set the derivative of the Lagrangian version with respect to $\boldsymbol{M}$, b, and ξ to be 0

$$\frac{\partial L(\boldsymbol{M},b,\xi,\alpha,\beta)}{\partial \boldsymbol{M}} = \boldsymbol{0} \tag{8}$$
$$\Rightarrow (\boldsymbol{M}-\boldsymbol{I}) - \sum_l \alpha_l h_l(\boldsymbol{x}_{l,1}-\boldsymbol{x}_{l,2})(\boldsymbol{x}_{l,1}-\boldsymbol{x}_{l,2})^T = \boldsymbol{0} \tag{9}$$
$$\frac{\partial L(\boldsymbol{M},b,\xi,\alpha,\beta)}{\partial b} = 0 \tag{10}$$
$$\Rightarrow \sum_l \alpha_l h_l = 0 \tag{11}$$
$$\frac{\partial L(\boldsymbol{M},b,\xi,\alpha,\beta)}{\partial \xi_l} = 0 \tag{12}$$
$$\Rightarrow C-\alpha_l-\beta_l = 0 \tag{13}$$
$$\Rightarrow 0 < \alpha_l < C \quad \forall l. \tag{14}$$

Equation (8) implies that the relationship between $\boldsymbol{M}$ and α can be represented as follows:

$$\boldsymbol{M} = \boldsymbol{I} + \sum_l \alpha_l h_l(\boldsymbol{x}_{l,1}-\boldsymbol{x}_{l,2})(\boldsymbol{x}_{l,1}-\boldsymbol{x}_{l,2})^T. \tag{15}$$

Substituting (8)–(12) back into $L(\boldsymbol{M},b,\xi,\alpha,\beta)$, we obtain the Lagrange dual problem as follows:

$$\max_\alpha -\frac{1}{2}\sum_{i,j}\alpha_i\alpha_j h_i h_j \boldsymbol{K}_{\boldsymbol{D}}(\boldsymbol{z}_i,\boldsymbol{z}_j) + \sum_i \alpha_i(1-h_i\boldsymbol{z}_i^T\boldsymbol{z}_i)$$
$$\text{s.t. } 0 \leq \alpha_l \leq C \quad \forall l,$$
$$\sum_l \alpha_l h_l = 0 \tag{16}$$

where $\boldsymbol{K}_{\boldsymbol{D}}(\boldsymbol{z}_i,\boldsymbol{z}_j) = [\boldsymbol{z}_i^T\boldsymbol{z}_j]^2$, and $\boldsymbol{z}_i = \boldsymbol{x}_{i,1}-\boldsymbol{x}_{i,2}$. The above problem is a standard quadratic program, and it can be solved by a variety of approaches, such as the interior point method, active set method, etc., [38]. Considering the similarity between (16) and the Lagrange dual problem of (2), the above problem is solved using LibSVM [39].

For each test couple $\boldsymbol{z}$, the label is determined by

$$\text{rRL}(\boldsymbol{z}) = \text{sgn}\left(\sum_l \alpha_l h_l \boldsymbol{K}_{\boldsymbol{D}}(\boldsymbol{z}_l,\boldsymbol{z}) + b\right). \tag{17}$$

For each change feature $\boldsymbol{x}$ to be classified, $k1+k2$ virtual couples $\boldsymbol{z}\boldsymbol{x}_i = (\boldsymbol{x},\boldsymbol{x}_i)$ can be constructed, where $\boldsymbol{x}_i (i = 1,\ldots,k1+k2)$ are the nearest neighbors of $\boldsymbol{x}$ within the original training set $\{(\boldsymbol{x_i}, y_i)|i = 1,\ldots,N\}$. $\text{rRL}(\boldsymbol{z}\boldsymbol{x}_i) = -1$ means that the label of $\boldsymbol{x}$ is the same as $\boldsymbol{x}_i$, and $\text{rRL}(\boldsymbol{z}\boldsymbol{x}_i) = 1$ means that the label of $\boldsymbol{x}$ is different from $\boldsymbol{x}_i$. The final label of $\boldsymbol{x}$ is determined by voting on the above $k1+k2$ decisions. Noting that the above decision procedure can be implemented on the global decision domain or on the local decision domain.

IV. Experiments

A. Experiments Description

To validate the effectiveness of the proposed approach, many experiments were conducted on different datasets. Owing to space limitations, only the results for three datasets are illustrated in this paper. The images, taken over Beijing (China), were acquired by the QuickBird 2 and Pleiades 1A satellites. The details of the images are described in Table I. The panchromatic image and multispectral image of the same time are registered by multilevel SIFT matching [40] and merged based on sparse matrix-vector multiplication [41]. For change detection, the multitemporal pan-sharpened images are registered by Ref. [40]. Pan-sharpened RGB images are shown in Figs. 5(a), (b), 6(a), (b), 7(a), and (b). As stated in Section II, for VHR images, not all the changes are of the user's interests. In this paper, we focus on the addition and removal of buildings. For this reason, the vegetation changes caused by seasonal variations are not considered.

The main advantages of the proposed approach lie in the exploitation of relationship learning and distance tuning for improving the ability to discriminate between the change features.

TABLE I
DATASET DESCRIPTION

Dataset	Image	Sensor	GSD resolution	Date	Resolution and size of coregistered pan-sharpened images
1	Fig. 5(a)	QuickBird 2	Pan: 0.61 m/pixel, mul: 2.44 m/pixel	Feb. 12, 2004	0.63 m/pixel, 1167×1260
	Fig. 5(b)	QuickBird 2	Pan: 0.63 m/pixel, mul: 2.52 m/pixel	Oct. 18, 2005	0.63 m/pixel, 1167×1260
2	Fig. 6(a)	QuickBird 2	Pan: 0.61 m/pixel, mul: 2.44 m/pixel	Feb. 12, 2004	0.63 m/pixel, 1120×1120
	Fig. 6(b)	QuickBird 2	Pan: 0.63 m/pixel, mul: 2.52 m/pixel	Oct. 18, 2005	0.63 m/pixel, 1120×1120
3	Fig. 7(a)	QuickBird 2	Pan: 0.61 m/pixel, mul: 2.44 m/pixel	Feb. 12, 2004	0.61 m/pixel, 1190×1404
	Fig. 7(b)	Pleiades 1A	Pan: 0.50 m/pixel, mul: 2.00 m/pixel	Feb. 8, 2012	0.61 m/pixel, 1190×1404

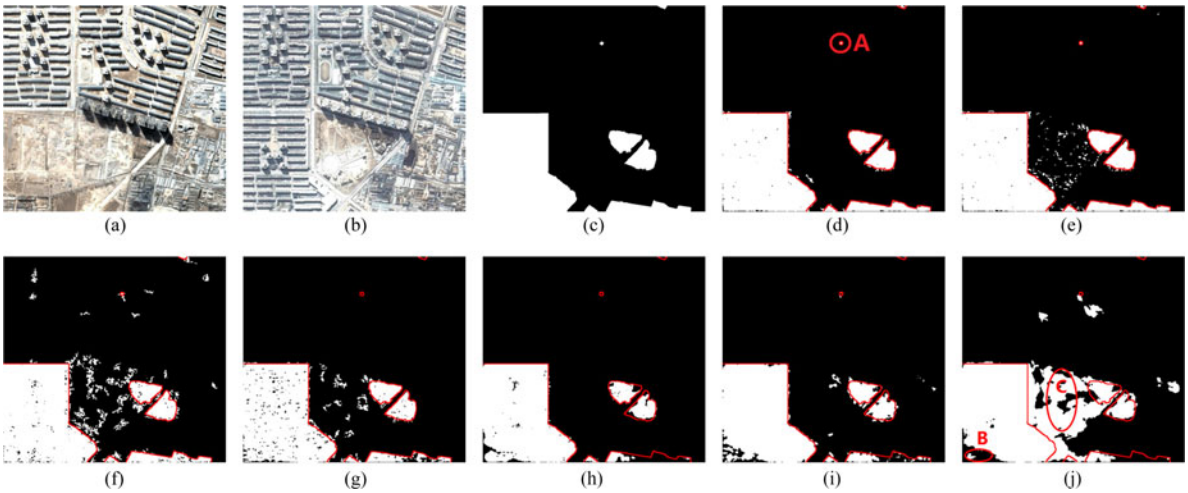


Fig. 5. Results comparison for dataset 1. (a) and (b) Multitemporal images, 1167×1260; (c) ground truth; (d) rRL; (e) RL-SVM; (f) SVM; (g) SVM-*k*NN; (h) CF; (i) SHC; and (j) *k*NN.

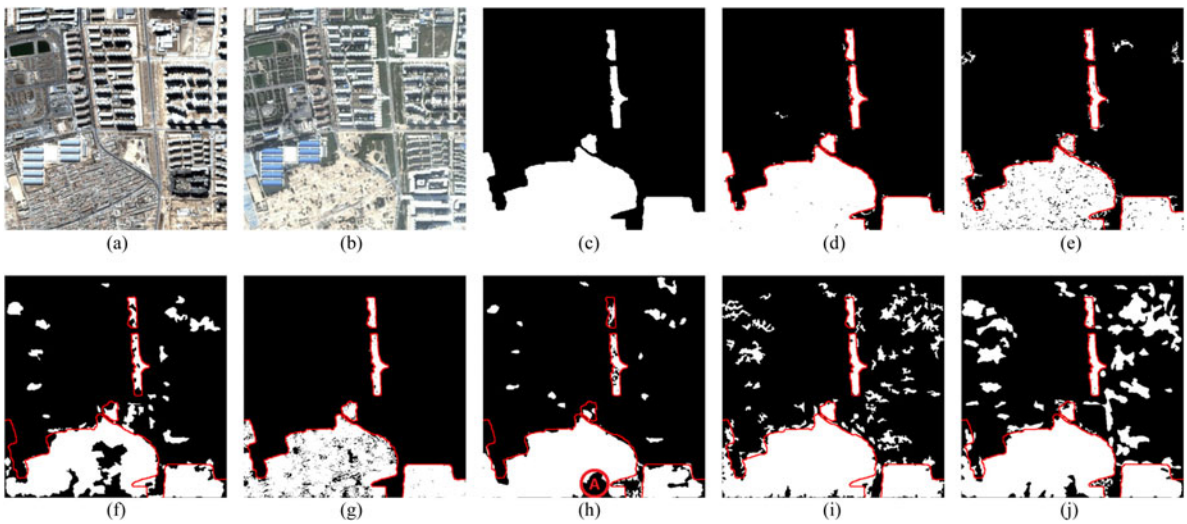


Fig. 6. Results comparison for dataset 2. (a) and (b) Multitemporal images, 1120×1120; (c) ground truth; (d) rRL; (e) RL-SVM; (f) SVM; (g) SVM-*k*NN; (h) CF; (i) SHC; and (j) *k*NN.

To validate the effectiveness of the proposed approach, it is compared to six related approaches.

1) *kNN:* Different from the unsupervised *k*NN approach for the traditional clustering task, *k*NN is used as a supervised classifier in this paper, i.e., each unlabeled change feature is assigned to the label based on the majority voting rule and the k closest labeled training samples.
2) *SVM:* The SVM classifier is used directly on the original training samples(not interclass and intraclass couples), and the unlabeled change features are classified by the learned model. In this paper, the radial basis function kernel is used.
3) *SVM-kNN:* The change features are classified by SVM-*k*NN [30], which considers the local relationship between

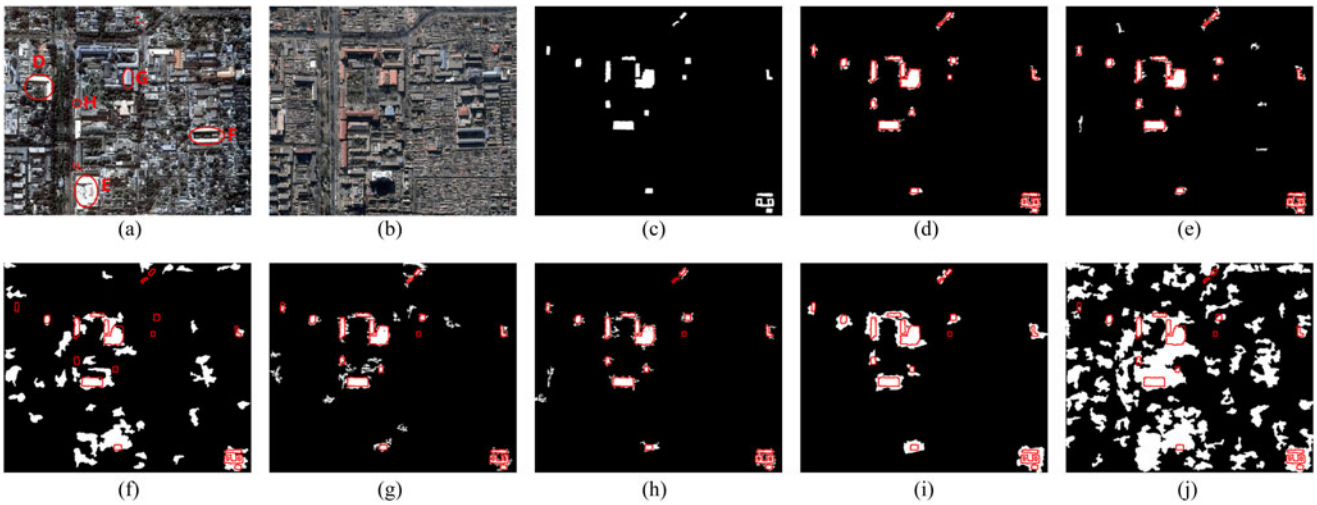


Fig. 7. Results comparison for dataset 3. (a) and (b) Multitemporal images, 1190×1404; (c) ground truth; (d) rRL; (e) RL-SVM; (f) SVM; (g) SVM-*k*NN; (h) CF; (i) SHC; and (j) *k*NN.

the test sample and the nearest training samples. When k is the number of support vectors, SVM-*k*NN will be equivalent to SVM.

4) *RL-SVM (Relation Learning With SVM):* In RL-SVM, the change features are classified by the doublet-SVM [34]. The difference between rRL and RL-SVM lies in the constraint over $\boldsymbol{M}$, i.e., the objective function of RL-SVM is $\frac{1}{2}\|\boldsymbol{M}\|_F^2 + C\sum_i \xi_l$.
5) *Change Field (CF):* Change features are extracted under the CF framework [16], i.e., change features consist of the compensated DAISY difference magnitude and the changes with respect to the feature shift along the X and Y directions. For fair comparison, the change features are classified by SVM.
6) *Sparse Hierarchical Clustering (SHC):* The change features are classified by SHC [42], where the clustering procedure is directed by the same training samples as the proposed approach.

In this paper, the change feature is extracted based on DAISY [9] difference vector. For fair comparison, all the above seven approaches used the same training samples. Specifically, 30% of the changed pixels were selected randomly from the ground truth, and they were viewed as the positive training samples. Similarly, 30% of the unchanged pixels were selected and taken as the negative training samples. The entire change feature set comprised the test sample. A fivefold cross-validation strategy was utilized to tune the parameters for SVM, RL-SVM, SVM-*k*NN, and CF. The above supervised procedures were repeated five times, and the averaged performance over the five runs used as the final indicator of performance.

B. Parameter Setting

For RL-SVM and rRL, the numbers of target neighbors and impostors, $k1$ and $k2$, are important parameters that affect the classification performance and training time. On one hand, the size of the kernel matrix $\boldsymbol{K_D}$ is increased proportionally to the increase of the number of interclass and intraclass couples, $k = k1 + k2$. As can be observed from Fig. 4(b), the training time is approximately linear with k. On the other hand, with larger k, more intraclass couples and interclass couples are added into the training set, and the similarities between the intraclass samples and the differences between the interclass samples are enhanced, which is helpful for improving the discrimination of the trained classifier. However, the ambiguities will be imported into the training samples if k is too large, where the margin between interclass couples and intraclass couples will be narrowed and the classification hyperplane will be impacted. The influences of $k1$ and $k2$ on the classification accuracy are shown in Fig. 4(a), from which it can be observed that the peak A is achieved at $k1 = 3$ and $k2 = 4$. In the following experiments, this configuration is set for RL-SVM and rRL. By experiments, we found that the best performance is achieved at $k = 1$ for *k*NN and $k = 200$ for SVM-*k*NN.

C. Experiments Analysis

Comparison is made qualitatively by visually checking the final change maps and quantitatively by computing false positive (FP), false negative (FN), total error (TE), CPU time, and Kappa coefficient (Kappa) and its variance. The core algorithms are implemented in C++, and they are evaluated on a desktop computer (Intel Core i7-2720QM CPU, 2.20GHz, 8-GB DDR RAM). The performances are listed in Table II, and the results using different approaches are shown in Figs. 5–7. To check how the results obtained by each approach coincide with the ground truth, the edges of the ground truth are shown in red.

Table II indicates that the proposed approach outperforms the other techniques. For instance, for the dataset 1, Kappa of *k*NN is 0.67, which is the worst among the seven approaches. As illustrated by the regions B and C in Fig. 5(j), due to the low interclass variability between the changed class and unchanged class, many FNs and FPs are difficult to avoid by the simple decision strategy. By taking advantage of the maximal margin principle, SVM can achieve the better performance. For example, TE is reduced from $173\,224$ to $53\,332$, and Kappa

TABLE II
PERFORMANCE COMPARISON

Dataset	Approach	TE	FN	FP	Kappa $\pm$ variance	CPU (s)
	rRL	12 792	6516	6276	0.97 ± 0.0031	435
	RL-SVM	19 964	6524	13 440	0.95 ± 0.0039	439
1	SVM	53 332	43 352	9980	0.88 ± 0.0037	354
	SVM-*k*NN	36 264	20 052	16 212	0.92 ± 0.0029	366
	CF	33 788	30 916	2872	0.92 ± 0.0035	268
	SHC	37 984	28 336	9648	0.92 ± 0.0025	298
	*k*NN	173 224	38 004	135 220	0.67 ± 0.0191	61
	rRL	11 248	1624	9624	0.98 ± 0.0034	440
	RL-SVM	34 604	15 596	19 008	0.92 ± 0.0042	437
2	SVM	118 480	71 168	47 312	0.73 ± 0.0056	341
	SVM-*k*NN	34 328	32 792	1536	0.91 ± 0.0051	362
	CF	72 920	48 868	24 052	0.83 ± 0.0039	259
	SHC	108 700	17 196	91 504	0.78 ± 0.0024	292
	*k*NN	250 480	42 452	208 028	0.54 ± 0.0274	59
	rRL	22 556	304	22 252	0.76 ± 0.0056	496
	RL-SVM	24 808	380	24 428	0.74 ± 0.0073	490
3	SVM	147 064	4696	142 368	0.27 ± 0.0062	377
	SVM-*k*NN	36 316	2584	33 732	0.64 ± 0.0059	379
	CF	27 936	840	27 096	0.71 ± 0.0046	269
	SHC	62 216	304	61 912	0.52 ± 0.0039	301
	*k*NN	110 516	13 404	97 112	0.27 ± 0.0629	63

is improved from 0.67 to 0.88. This demonstrates the importance of the relationship for improving the overall separability. However, due to ignoring the intraclass similarity in training the classifier and the local coherence in the decision procedure, the performance improved by SVM is limited. SHC is superior to SVM in that it takes the complex structures of the change features into account (i.e., SHC considers not only the ability to discriminate between the changed class and unchanged class, but also the separability between the subclasses of the changed class), which is another important factor for improving the overall separability between change features. FN is decreased from 43 352 by SVM to 28 336 by SHC. CF outperforms SVM in that it compensates for the misregistration impact, and FP is reduced from 9980 by SVM to 2872 by CF. SVM-*k*NN is superior to SVM. The underlying reason lies in the neighborhood relationship between the test sample and training samples being considered. With the help of the simple local decision strategy, Kappa is improved from 0.88 to 0.92. Moreover, by taking advantage of the adaptive distance tuning mechanism, RL-SVM outperforms SVM-*k*NN, and RL-SVM achieved the lower TE ($19\,964$) and higher Kappa (0.95). Furthermore, by utilizing the robust Mahalanobis matrix estimation, the Kappa of rRL is improved to 0.97. The advantage of rRL over RL-SVM can be validated by the higher Kappa and lower variances.

Similar conclusions can be drawn from dataset 2. For example, *k*NN is still worse than other approaches, and Kappa is 0.54. SVM, whose Kappa is 0.73, is slightly better than *k*NN. CF and SHC are superior to SVM, and Kappa coefficients are increased to 0.83 and 0.78, respectively. CF is robust for the misregistration impact since the feature difference is computed based on the compensated correspondence. As a consequence, CF achieves lower FP (24 052) than that of SHC (91 504). However, CF is limited in enhancing the weak change features, and it is prone to obtaining a higher number of FNs. For instance, the changes within region A of Fig. 6(a) are undetected. In contrast, SVM-*k*NN reduces FP and FN significantly by capturing the local relationship in the decision procedure. TE is reduced from 118 480 by SVM to 34 328 by SVM-*k*NN. On the other hand, RL-SVM and rRL capture the relationship in the classifier training procedure, and thus obtain better performances. RL-SVM obtained a lower FN (15 596) and higher Kappa (0.92), and rRL achieved the lowest TE ($11\,248$) and the highest Kappa (0.98). The above comparisons illustrate the importance and effectiveness of relationship representation and relationship learning. In short, rRL addresses overall discriminative capacity from two perspectives: the neighborhood relationships of change features with respect to the feature difference and the label, and the adaptive distance learning and tuning driven by the training samples. As expected, these two factors are promising in processing the complex change features. The advantages of the proposed approach can also be validated by the visual comparison of the results obtained by the different approaches shown in Figs. 5–7.

Compared with the first two datasets, change features extracted from dataset 3 are more difficult to separate since the ratio of the changed areas is very small (about $1/45$). Moreover, the false changes caused by different sensors and long-term intervals are overwhelming. For instance, as illustrated by Fig. 7(a), the appearance differences within regions D–G are even stronger than that in region H. In fact, from the semantic perspective, regions D–G are "unchanged." In this case, the performances of three creative approaches (CF, SHC, and SVM-*k*NN) are not as promising as before. For instance, the Kappa of SVM-*k*NN is 0.64, and it is even worse than that of CF. In contrast, the performance of rRL is still the best. To investigate how the proposed approach improves performance, we compare different approaches from the following two aspects.

1) *Relationship Representation:* The purpose of a relationship representation is to capture the intraclass similarity and interclass variability. To verify the effectiveness of the relationship representation, Fig. 8 illustrates the difference in overall separability before and after relation learning. *k*NN and SVM-*k*NN measure the relationship between samples based on the Euclidean distance (noting that the Gaussian kernel leads to an ordering function that is equivalent to using the Euclidean metric). For the test sample u[e.g., point u in Fig. 7(a)], two of three nearest training samples chosen by *k*NN are from the changed class since the Euclidean metric ignores the latent discrimination information hidden in the training labels. At the point c, only one of three nearest neighbors is from the changed class. For this reason, improper nearest neighbors are obtained for SVM-*k*NN, and they will be propagated in the subsequent decision procedure. The naive Mahalanobis distance driven by the training samples is more robust than the Euclidean distance, and three nearest neighbors of *u* are found correctly; however, it is limited to capturing the complex relations. For instance, two of three nearest neighbors of c are found incorrectly. In contrast, under the Mahalanobis metric learned by rRL and RL-SVM, no incorrect nearest neighbors are found at *c* or *u*. The above

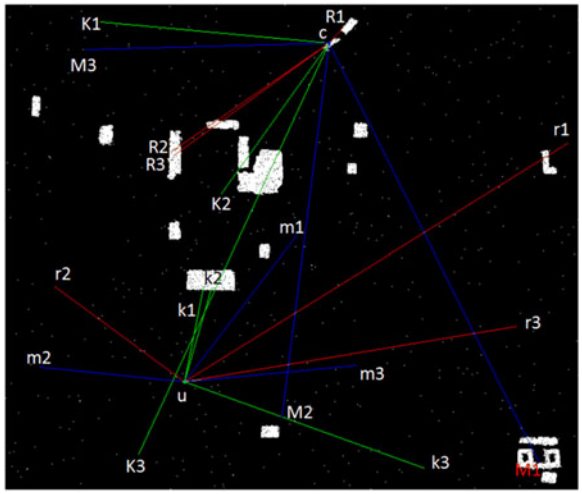


Fig. 8. Difference in overall separability before and after relationship learning. For illustration, three nearest neighbors within the training samples are obtained for the unchanged change feature u and the changed feature c. Under the Euclidean distance metric, three nearest neighbors of u are $k1$, $k2$, and $k3$, and three nearest neighbors of c are $K1$, $K2$, and $K3$. Under the Mahalanobis distance metric, where M is computed based on the training samples, three nearest neighbors of u are $m1$, $m2$, and $m3$, and three nearest neighbors of c are $M1$, $M2$, and $M3$. Under the Mahalanobis distance metric where M is learned by relationship learning driven by the training samples, three nearest neighbors of u are $k1$, $k2$, and $k3$, and three nearest neighbors of c are $R1$, $R2$, and $R3$.

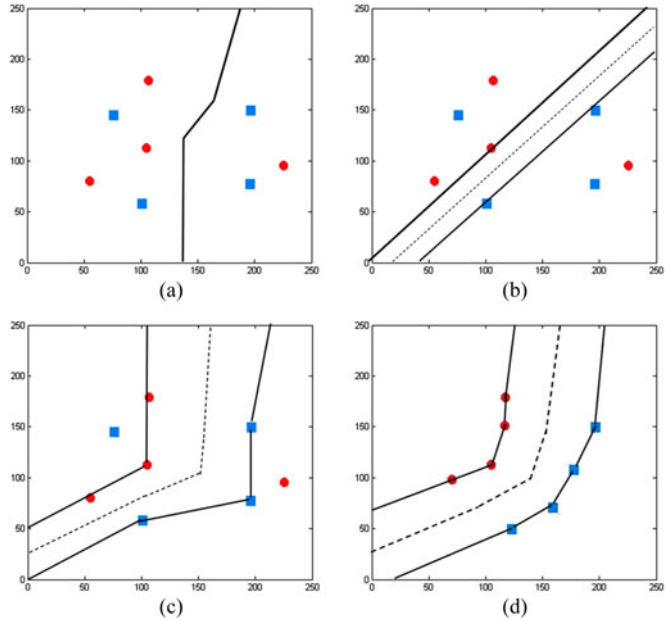


Fig. 9. Differences between different approaches in decision strategy. (a) kNN, (b) SVM, (c) SVM-kNN, (d) rRL. Training samples of different classes are plotted by different colors. SVM-kNN is superior to SVM due to the robust local decision domain. However, this improvement is limited for the high overlap of different classes. rRL achieves the robust decision hyperplane by dynamically tuning the distances between the interclass and intraclass couples.

difference indirectly demonstrates the effectiveness of the relationship representation.

2) *Decision Strategy:* On one hand, the impacts of improper decision domain will be propagated to the decision function. For instance, u is wrongly classified as the changed class by kNN. On the other hand, a robust decision strategy is helpful for reducing the impacts. For instance, SVM-kNN has the same nearest neighbors with kNN, but it makes the correct decision on u by taking advantage of the local decision strategy. However, this decision strategy is limited for the high overlap between the changed and unchanged class. For instance, it makes the wrong decision at c guided by the improper decision domain. Noting that rRL can also make decisions on local domain as SVM-kNN; however, through experiments we found that the performance difference between global decision domain and local decision domain can be ignored. The possible reason lies in the fact that the local decision hyperplane nearly coincides with the global one after rRL adjustment of the distances between the training couples. For comparison, the decision strategies of different approaches are illustrated in Fig. 9.

Owing to the above two important factors, the other advantage of the proposed approach is the high fidelity in preserving the region shape, especially near the borders between the changed and unchanged regions. It is worth noting that even for small regions, such as region A in Fig. 5(d), the proposed approach, rRL, can detect the changes correctly and preserve the shape with high confidence. In contrast, such small regions are missed by other approaches due to their ignoring of the local relationship. From Fig. 7(d), it can be observed that the shape-preserving ability of rRL on dataset 3 is not as good as the first two datasets. The underlying reason is that the interclass variability is impacted by the imbalanced training samples of different classes, i.e., some pixels in the unchanged regions are "forcibly dragged" across the border by the overwhelming number of unchanged training samples. Even so, compared with other approaches, rRL is still promising for such a difficult task. This fact implies that the overall discriminative ability of the changed features is indeed improved by the proposed approach.

Regarding computation time, rRL and RL-SVM are inferior to other approaches. Specifically, rRL, RL-SVM, CF, and SVM-kNN are essentially SVMs. However, CF is computationally faster than SVM due to its classification of the low-dimensional features (for CF, the dimension of change features is 3), and rRL and RL-SVM are computationally slower than SVM due to the extra computation time for training on the increased interclass and intraclass training couples, which are multiple times greater in number than the original training samples. There are no differences between SVM-kNN and SVM in the training procedure, and the complexity difference in the decision procedure can be ignored. For this reason, SVM-kNN is comparable to SVM. The core operation of SHC is sparse clustering, and the efficiency is improved by the hierarchical strategy. kNN is fastest since there are only two simple operators involved in: distance computation and distance comparison. Considering the overall performance, rRL and RL-SVM are more promising for practical applications.

V. Conclusion

The low interclass separability between change features has been an obstacle for VHR image change detection. However,

traditional efforts have focused on change feature extraction and/or feature classification, and the latent relationships hidden in training samples are usually neglected. Through experimentation, we found that the overall separability can be improved significantly by learning the relationships between change features. Training information can be enriched by utilizing the compensated neighborhood relationships between the considered training samples and their neighbors, and the distance between change features can be tuned automatically driven by the learned projections. Experiments demonstrate the effectiveness of the proposed approach. In light of the creative direction embodied by relationship learning, more efforts should be directed to further strengthening interclass separability. Our future work will focus on relationship learning in the context of unsupervised or semisupervised change detection, or on enhancing the interclass separability by other advanced techniques such as semisupervised distance learning [43] and multiple kernel learning [44], [45].

ACKNOWLEDGMENT

The authors would like to thank Satellite Imaging Corporation for providing the Pleiades-1A image, and the anonymous reviewers for their suggestions.

REFERENCES

[1] A. Singh, "Digital change detection techniques using remotely-sensed data," *Int. J. Remote Sens.*, vol. 10, no. 8, pp. 989–1003, 1989.

[2] D. Lu, P. Mausel, E. Brondizio, and E. Moran, "Change detection techniques," *Int. J. Remote Sens.*, vol. 25, no. 12, pp. 2365–2407, 2004.

[3] R. J. Radke, S. Andra, O. Al-Kofahi, and B. Roysam, "Image change detection algorithms: A systematic survey," *IEEE Trans. Image Process.*, vol. 14, no. 3, pp. 294–307, Mar. 2005.

[4] P. Coppin, E. Lambin, and I. Jonckheere, "Digital change detection methods in natural ecosystem monitoring: A review," *Int. J. Remote Sens.*, vol. 25, no. 9, pp. 1565–1596, 2004.

[5] J. Mas, "Monitoring land cover changes: A comparison of change detection techniques," *Int. J. Remote Sens.*, vol. 20, no. 1, pp. 139–152, 1999.

[6] J. Chan, K. Chan, and A. Yeh, "Detecting the nature of change in an urban environment: A comparison of machine learning algorithms," *Photogramm. Eng. Remote Sens.*, vol. 67, no. 2, pp. 213–226, 2001.

[7] H. Wang and E. C. Ellis, "Image misregistration error in change measurements," *Photogramm. Eng. Remote Sens.*, vol. 71, no. 9, pp. 1037–1044, 2005.

[8] F. Bovolo, L. Bruzzone, L. Capobianco, S. Marchesi, F. Nencini, and A. Garzelli, "Analysis of the effects of pansharpening in change detection on VHR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 1, pp. 53–57, Jan. 2010.

[9] E. Tola, V. Lepetit, and P. Fua, "A fast local descriptor for dense matching," presented at the IEEE Conf. Computer Vision Pattern Recognition, Anchorage, AK, USA, 2008.

[10] C. Liu, J. Yuen, and A. Torralba, "SIFT flow: Dense correspondence across scenes and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 978–994, May 2011.

[11] J. Struyf and H. Blockeel, *Encyclopedia of Machine Learning*. Boston, MA, USA: Springer, 2010, ch. Relational Learning, pp. 851–857. [Online]. Available: http://dx.doi.org/10.1007/978-0-387-30164-8_719

[12] C. Paul and A. Alessandro, "Toward specification-driven change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 38, no. 3, pp. 1484–1488, May 2000.

[13] G. Chen, G. J. Hay, L. M. T. Carvalho, and M. A. Wulder, "Object-based change detection," *Int. J. Remote Sens.*, vol. 33, no. 14, pp. 4434–4457, 2012.

[14] C. Huo, Z. Zhou, H. Lu, and C. Pan, "Fast object-level change detection for VHR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 1, pp. 118–122, Jan. 2010.

[15] N. Falco, M. D. Mura, F. Bovolo, J. A. Benediktsson, and L. Bruzzone, "Change detection in VHR images based on morphological attribute profiles," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 3, pp. 636–640, May 2013.

[16] L. Huo, X. Feng, C. Huo, Z. Zhou, and C. Pan, "Change field: A new change measure for VHR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 1, pp. 1812–1816, Oct. 2014.

[17] Y. Zhong, W. Liu, J. Zhao, and L. Zhang, "Change detection based on pulse-coupled neural networks and the NMI feature for high spatial resolution remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 3, pp. 537–541, Mar. 2015.

[18] 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.

[19] Z. Wang *et al.*, "A multiscale and hierarchical feature extraction method for terrestrial laser scanning point cloud classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2409–2425, May 2015.

[20] X. Huang, L. Zhang, and T. Zhu, "Building change detection from multitemporal high-resolution remotely sensed images based on a morphological building index," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 1, pp. 105–115, Jan. 2014.

[21] B. Fan, C. Huo, C. Pan, and Q. Kong, "Registration of optical and SAR satellite images by exploring the spatial relationship of the improved sift," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 4, pp. 657–661, Jul. 2013.

[22] F. Pacifici and F. Del Frate, "Automatic change detection in very high resolution images with pulse-coupled neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 1, pp. 58–62, Jan. 2010.

[23] E. Pagot and M. Pesaresi, "Systematic study of the urban postconflict change classification performance using spectral and structural features in a support vector machine," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 1, no. 2, pp. 120–128, Jun. 2008.

[24] M. Volpi, D. Tuia, F. Bovolo, M. F. 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, no. 2, pp. 77–85, 2013.

[25] M. Fauvel, "Kernel matrix approximation for learning the kernel hyperparameters," presented at the IEEE Int. Symp. Geoscience Remote Sensing, Munich, Germany, 2012.

[26] K. Ding, C. Huo, Y. Xu, and C. Pan, "VHR image change detection based on discriminative dictionary learning," presented at the IEEE Conf. Acoustics, Speech, Signal Processing, Vancouver, BC, Canada, 2013.

[27] J. Yu, Y. Rui, Y. Y. Tang, and D. Tao, "High-order distance-based multiview stochastic learning in image classification," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2431–2442, Dec. 2014.

[28] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967.

[29] P. Ghamisi, M. Couceiro, M. Fauvel, and J. Benediktsson, "Integration of segmentation techniques for classification of hyperspectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 1, pp. 342–346, Jan. 2014.

[30] E. Blanzieri and F. Melgani, "Nearest neighbor classification of remote sensing images with the maximal margin principle," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 6, pp. 1804–1811, Jun. 2008.

[31] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10, pp. 207–244, 2009.

[32] K. Q. Weinberger, F. Sha, and L. K. Saul, "Convex optimizations for distance metric learning and pattern classification," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 146–158, May 2010.

[33] J. Davis, B. Kulis, P. Jain, S. Sra, and I. Dhillon, "Information-theoretic metric learning," in *Proc. Int. Conf. Mach. Learn.*, 2007, pp. 209–216.

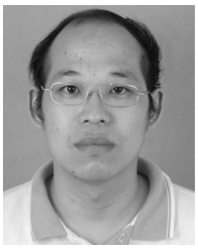
[34] F. Wang, W. Zuo, L. Zhang, D. Meng, and D. Zhang, "A kernel classification framework for metric learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 9, pp. 1950–1962, Sep. 2015.

[35] M. Fauvel, J. Chanussot, J. A. Benediktsson, and A. Villa, "Parsimonious Mahalanobis kernel for the classification of high dimensional data," *Pattern Recog.*, vol. 46, no. 3, pp. 845–854, Mar. 2013.

[36] B. Du, L. Zhang, L. Zhang, T. Chen, and K. Wu, "A discriminative manifold learning based dimension reduction method for hyperspectral classification," *Int. J. Fuzzy Syst.*, vol. 14, no. 2, pp. 272–277, 2012.

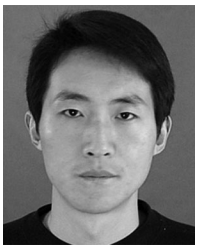
[37] B. Du and L. Zhang, "A discriminative metric learning based anomaly detection method," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 11, pp. 6844–6857, Nov. 2014.

[38] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed. Berlin, Germany: Springer-Verlag, 2006.
[39] C. Chang and C. Lin, "LibSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, 2011.
[40] C. Huo, C. Pan, L. Huo, and Z. Zhou, "Multilevel SIFT matching for large-size VHR image registration," *IEEE Geosci. Remote Sens. Lett.*, vol. 9, no. 2, pp. 171–175, Mar. 2012.
[41] Y. Wang, H. Yan, C. Pan, and S. Xiang, "Image editing based on Sparse Matrix-Vector multiplication," in *Proc. IEEE Conf. Acoustics, Speech, Signal Process.*, 2011, pp. 1317–1320.
[42] K. Ding, C. Huo, Y. Xu, Z. Zhong, and C. Pan, "Sparse hierarchical clustering for VHR image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 3, pp. 577–581, Mar. 2015.
[43] Y. Gu and K. Feng, "Optimized laplacian SVM with distance metric learning for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 6, no. 3, pp. 1109–1117, Jun. 2013.
[44] Y. Gu, G. Gao, D. Zuo, and D. You, "Model selection and classification with multiple kernel learning for hyperspectral images via sparsity," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2119–2130, Jun. 2014.
[45] Q. Wang, Y. Gu, and D. Tuia, "Discriminative multiple kernel learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 7, pp. 3912–3927, Jul. 2016.



Chunlei Huo (M'11) received the B.S. degree in applied mathematics from Hebei Normal University, Shijiazhuang, China, in 1999, the M.S. degree in applied mathematics from Xidian University, Xi'an, China, in 2002, and the Ph.D. degree in pattern recognition and intelligent system from the Institute of Automation, Chinese Academy of Sciences, Beijing, in 2009.

He is currently an Associate Professor at the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His current research interests include remote sensing image processing, computer vision and pattern recognition, etc.



Keming Chen received the M.S. degree in automatic control from Wuhan University of Technology, Wuhan, China, in 2006 and the Ph.D. degree in pattern recognition and intelligent system from the Institute of Automation, Chinese Academy of Sciences, Beijing, in 2011.

He is currently an Associate Professor at the Institute of Electronics, Chinese Academy of Sciences. His current research interests include remote sensing image processing and pattern recognition.

Kun Ding received the B.S. degree in automatic control from Tianjin University of Science and Technology, Tianjin, China, in 2011 and the M.S. degree in pattern recognition and intelligent system from the Institute of Automation, Chinese Academy of Sciences, Beijing in 2014, where he is currently working toward the Ph.D. degree.

His current research interests include pattern recognition and machine learning.

Zhixin Zhou received the B.S. degree in communication engineering from Electronic Engineering Institute of PLA, Hefei, China, in 1986, the M.S. degree and the Ph.D. degree in electronic engineering from Harbin Institute of Technology, Harbin, China, in 1989 and 1997.

He is currently a Professor at Beijing Institute of Remote Sensing, Beijing, China. His research interests include computer vision, pattern recognition, and remote sensing.

Chunhong Pan received the B.S. degree in automatic control from Tsinghua University, Beijing, China, in 1987, the M.S. degree from Shanghai Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, China, in 1990, and the Ph.D. degree in pattern recognition and intelligent system from the Institute of Automation, Chinese Academy of Sciences, Beijing, in 2000.

He is currently a Professor at National Laboratory of Pattern Recognition of Institute of Automation, Chinese Academy of Sciences. His research interests include computer vision, image processing, computer graphics, and remote sensing.