Unsupervised ensemble change detection using kernel PCA

Panu SRESTASATHIERN, Siam LAWAWIROJWONG, Narut SOONTRANON and Preesan RAKWATIN
Geo-informatics and Space Technology Development Agency (GISTDA)
120 The Government Complex, Ratthaprasasanabhakti Building
6th and 7th Floor, Chaeng Wattana Road, Lak Si, Bangkok, Thailand 10210
E-mail: panu@gistda.or.th; siam@gistda.or.th; narut@gistda.or.th; preesan@gistda.or.th

KEY WORDS: Change detection, kernel PCA, Bagging, Ensemble, Unsupervised

ABSTRACT: In this paper, we present a novel approach for unsupervised change detection on multi-spectral satellite images. The advantage of unsupervised approach over the supervised one is that the generation of an appropriated ground truth is not required. Especially, when the ground truth is not available, the unsupervised approach is the fundamental one. The unsupervised change detection method used in this paper is based on the concept of kernel Principal Component Analysis (PCA). The advantage of using kernel PCA over standard PCA is that it can handle non-linear relationships between data by projecting the data into higher dimension using kernel function. Assuming that relationship between data point after projecting data into higher dimensional space is linear, the principal basis vectors can then be extracted. A pixel is classified as unchanged if the absolute value of its second principal component is larger than a threshold. To improve the performance of kernel PCA, the ensemble approach is applied i.e., bootstrapped aggregating, which is also known as the bagging. The concept of bagging is to combine the change detection results from weak change detectors to improve overall performance. The proposed method begins with generating bootstrap samples. That is, the image pixels are random sampled with replacement. Each bootstrap sample is then used to construct a change detector using kernel PCA; a set of change detector is hence obtained. As a result, to classify a pixel as change or un-change, the change detection results from those change detectors are then fused by majority voting. By using random samples instead of using all image pixels at once for detecting change, the computation time is reduced. In other words, a large amount of computational resource is not required. To evaluated, the performance of the proposed methods, it was tested with multispectral satellite images i.e., Landsat 8. The experimental results will be illustrated as a change map. The analysis on principal basis vector extracted from the data using kernel PCA will be performed.

1. Introduction

In remote sensing, change detection is a fundamental task in multi-temporal analysis of satellite imagery and still on-going research agenda [8]. The purpose of change detection is to detect or label pixels or areas having surface component alternations. Such changes including land use or land cover changes are very useful in many applications such as deforestation, urban expansion monitoring.

The change detection algorithms can be roughly categorized into two approaches i.e., supervised and unsupervised change detections. The supervised approach exploits the scene knowledge or ground truth as training data which can be obtained manually. The training data is used to train or tune the parameters of change detectors. Examples of supervised change detection techniques are support vector machine [7] and artificial neural network [6]. The accuracy of the change detection is hence depend on the quality of the training data. The major drawback of this approach is that the training data must be known a priori and can be time consuming.

The later approach is the unsupervised one which is adopted in this work. The advantage of the unsupervised change detection is that the change can be automatically detected without having any ground truth to train the change detectors. Some popular unsupervised change detection approaches include, but no limited to, image differencing [2], change vector analysis [9], texture based analysis [12] and Principal Component Analysis (PCA) [4]. Image differencing technique uses *difference image*, which is calculated by pixel-wise subtracting image acquired from different time. The pixels having high positive or negative change are most likely change. Similar to image differencing technique, change vector analysis technique uses pixelwise vector subtraction. That is, a multi-spectral image pixel is represented as a vector. The magnitude and direction of change vector are used in the analysis. Therefore, the changes are not only detected, but also classified. The texture based change detection uses statistics of the spatial distribution of the image pixels for analysis. The drawback of this approach is the selection of appropriated window size for computing statistical parameters. The PCA technique is based on the principle of dimensionality reduction. By projection data onto new basis vectors, the principal components are then used for change detection. For readers interested in reviewing change detection techniques, [8] is recommended.

Motivated by the work of Celik [3], the proposed method adopts PCA technique to detect change on multi-temporal satellite images because of its simplicity. The limitation of [3], is that the multi-spectral satellite image must be converted to a single band image e.g., vegetation index; the standard PCA is then used to extract the change on the image. The proposed method exploits kernel PCA for change detection. Unlike standard PCA, the kernel PCA can handle non-linear relationship in the data [10]. In terms of computation, using all image pixels at once in PCA technique can be a problem because the size of a satellite image is usually large. That is, the size of the matrix used in PCA computation can be very large. Thus, the computational time can be very long and large computational resource is required. To fix this problem, the proposed method applied ensemble concept by using sampled image pixels to construct various change detectors. The final change detection result is then the combination of the results from those detectors.

The rest of this paper is organized as follow. In Section 2, the principles of PCA and kernel PCA are briefly discussed. The proposed change detection method will be presented in Section 3. The experimental result will be illustrated in Section 3. In Section 4, the conclusion and future work will be discussed.

2. Principal component analysis

In machine learning and data mining, the major problem is the high-dimensional data. PCA is a popular technique for reducing the dimension of the data. The concept is to extract the intrinsic dimension of data which is relatively small to the original dimension. In this section, the standard PCA is first reviewed to provide the basic concept of principal component. A more generic analysis i.e., kernel PCA, which can handle non-linearity in the data, is then presented.

2.1. Standard PCA

The principal component analysis is a data analysis technique, especially for dimensionality reduction and feature extraction. The goal of PCA is to find a transformation into linear sub-space having dimentionality lower than that of original data space such that the data in the linear sub-space has the largest variance.

Let there be N data points $\mathbf{x}_i \in \mathbb{R}^P, i = 1 \ldots, N$. For example, in the satellite remote sensing applications, sampled image pixels $\mathbf{x}_i$ is the set of sampled image pixels and P is the number of bands which can be used to form a matrix X with P rows and N columns:

$$\mathbf{X} = [\mathbf{x}_1 \quad \mathbf{x}_2 \quad \ldots \quad \mathbf{x}_N] \in \mathbb{R}^{P \times N} \tag{1}$$

Moreover, the observed data is centered such that $\sum_{i=1}^{N} \mathbf{x}_i = \mathbf{0}$, which can be done by translating the origin of the data coordinate system to the mean of the data. The covariance matrix of the data can be computed as:

$$\mathtt{C} = \frac{1}{N}\mathbf{X}\mathbf{X}^\top = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^\top \tag{2}$$

As it is already mentioned that, the goal of PCA is to project the data into a lower-subspace. Let the dimensionality of the linear subspace is M i.e., $M < P$. We therefore attempts to find the transformation $\mathtt{V}$ to transform the data into the subspace:

$$\widehat{\mathbf{x}}_i = \mathtt{V}\mathbf{x}_i, \tag{3}$$

where $\mathtt{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_M]^\top$ and $\mathbf{v}_i^\top \mathbf{v}_i = 1$ for $i = 1, \ldots, M$. Moreover, it is assumed that the covariance matric of the data in the subspace is diagonal. The transformation matrix $\mathtt{V}$ hence maximize the trance of the covariance matrix in the subspace:

$$\widehat{\mathtt{V}} = \arg\max_{\mathtt{V}} \ \text{trace}\left(\frac{1}{N}\widehat{\mathbf{x}}^\top \widehat{\mathbf{x}}\right). \tag{4}$$

According to (3), one obtains:

$$\widehat{\mathtt{V}} = \arg\max_{\mathtt{V}} \ \text{trace}\left(\mathtt{V}\mathtt{C}\mathtt{V}^\top\right). \tag{5}$$

The transformation $\mathtt{V}$ can be computed using Lagrange multiplier technique. By setting the derivative of the Lagrangian function to be zero, one obtains:

$$\mathtt{C}\mathtt{V} = \boldsymbol{\lambda}\mathtt{V}. \tag{6}$$

That is, the transformation matrix $\mathtt{V}$ can be computed by solving the eigenvalue problem, where $\mathbf{v}_l \in \mathbb{R}^P \setminus \{\mathbf{0}\}$ is the eigenvector and $\boldsymbol{\lambda}$ eigenvalue.

The new bases of the observed data are then the eigenvectors having first M largest eigenvalues. Moreover, eigenvectors are the directions of maximum variance in the data. Particularly, the principal components can be obtained by projecting the data to the new basis vectors. The new data coordinates in the new bases are called principal components.

2.2. Kernel PCA

A major drawback of the ordinary PCA is that it cannot handle non-linear relationship in the data. In other word, the standard PCA performs poorly when the structure of the data is more complicated and cannot be represented by a linear subspace.

To extend the standard PCA to handle non-linear data distribution, kernel PCA technique projects the data in P-dimensional space into higher dimensional space. Particularly, the data is project into the higher dimensional space via a nonlinear function:

$$\boldsymbol{\phi} : \mathbb{R}^P \to \mathbb{F}, \tag{7}$$

The space $\mathbb{F}$ is referred to as the feature space which is an inner product space. Hence, $\boldsymbol{\phi}(\mathbf{x}_i)$ is the image of the data point $\mathbf{x}_i$ in the feature space. Moreover, the dimension of the feature space $\mathbb{F}$ is much larger than that of data space. It is hypothesize that, the distribution of the data after projection into the feature space i.e., $\mathbf{x} \to \boldsymbol{\phi}(\mathbf{x}_i)$ is linear. The standard PCA can thus be performed. Given N data points, which is a set of sampled image pixels, the observation in the feature space is:

$$\Phi = [\boldsymbol{\phi}(x_1) \quad \boldsymbol{\phi}(x_2) \quad \ldots \boldsymbol{\phi}(x_N)], \tag{8}$$

Assuming that the data after projection into the feature space is centered i.e., $\sum_{i=1}^{N} \boldsymbol{\phi}(\mathbf{x}_i) = \mathbf{0}$. Similar to the ordinary PCA, the covariance matrix of the data after projection into higher dimensional space is:

$$\mathtt{C} = \frac{1}{N} \Phi\Phi^\top. \tag{9}$$

Similar to the standard PCA, its eigenvalue problem is stated as:

$$\mathtt{C}\mathbf{v}_l = \lambda_l \mathbf{v}_l \tag{10}$$

Since the eigenvector lies within the span of $\{\boldsymbol{\phi}(\mathbf{x}_1), \boldsymbol{\phi}(\mathbf{x}_2), \ldots, \boldsymbol{\phi}(\mathbf{x}_N)\}$ i.e., $\mathbf{v}_l \in \text{span}\{\boldsymbol{\phi}(\mathbf{x}_1), \boldsymbol{\phi}(\mathbf{x}_2), \ldots, \boldsymbol{\phi}(\mathbf{x}_N)\}$. That is, the eigenvector can be represented as a linear combination:

$$\mathbf{v}_l = \sum_{i=1}^{N} \alpha_{li} \boldsymbol{\phi}(\mathbf{x}_i). \tag{11}$$

We therefore come up with the eigenvalue problem [11]:

$$N\lambda_l \boldsymbol{\alpha}_l = \mathtt{K}\boldsymbol{\alpha}_l, \tag{12}$$

where $\boldsymbol{\alpha}_l = [\alpha_{l1}, \ldots, \alpha_{lN}]^\top$.

The matrix $\mathtt{K}$ is an $N \times N$ matrix. Its element is the inner product of points $\boldsymbol{\phi}(\mathbf{x}_i)$ and $\boldsymbol{\phi}(\mathbf{x}_j)$ in the feature space:

$$\mathtt{K}_{ij} = \langle \boldsymbol{\phi}(\mathbf{x}_i), \boldsymbol{\phi}(\mathbf{x})_j \rangle, \tag{13}$$

which is symmetric and positive definite. The matrix $\mathtt{K}$ is also known as the *Gramm matrix*. In [11], it is shown that, with a proper non-linear function Φ, the inner product in the feature space can be computed without knowing the values of $\boldsymbol{\phi}(\mathbf{x}_i)$ and $\boldsymbol{\phi}(\mathbf{x}_j)$. Namely, the elements of $\mathtt{K}$ can be computed in the original space without mapping the data point into the feature space which possibly has infinite dimensionality. Therefore, the pairwise relation of points in the original space can be used to construct the matrix $\mathtt{K}$:

$$k : \mathbb{R}^P \times \mathbb{R}^P \to \mathbb{R}, \quad \mathtt{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j), \tag{14}$$

where the function k is called *kernel function*. Some popular kernel function are:

- Gaussian: $k(\mathbf{x}_i, \mathbf{x}_j) = e^{\frac{\|\mathbf{x}-\mathbf{x}_j\|^2}{2\sigma^2}}$,
- Power: $k(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^\top \mathbf{x}_j)^d$,
- Polynomial: $k(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^\top \mathbf{x}_j + h_0)^d$.

To compute the l^{th}-principal component by projecting the data onto the l^{th} eigenvector, with the kernel function, the projection can be computed easily:

$$\text{PC}_l(\mathbf{x}) = \langle \mathbf{v}_l, \boldsymbol{\phi}(\mathbf{x}) \rangle = \sum_{i=1}^{N} \alpha_{li} k(\mathbf{x}, \mathbf{x}_i). \tag{15}$$

The principal component will be used to classify a pixel as “change” or “unchange”.

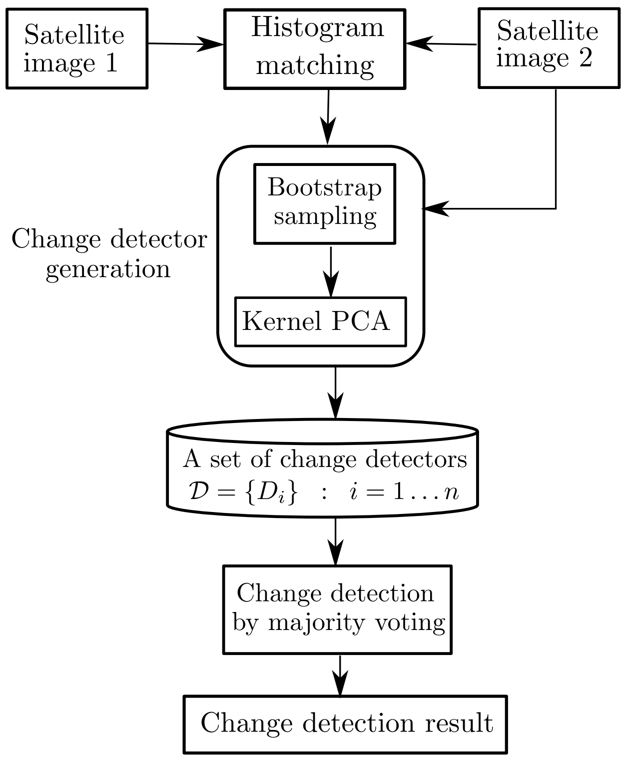


Figure 1: Workflow of the proposed change detection method begins with the histogram matching between image taken at two different times. The image pixels are then bootstrap sampled to generate a set of change detectors. The final change detection result is then the majority vote of the result from all change detectors.

3. The proposed change detection method

In this section, the proposed unsupervised change detection method will be explained. The process begins with the histogram matching between two image taken in different dates. The purpose of the histogram matching is to reduce the effect of sensor response that can change overtime. The proposed change detection is based on the ensemble concept [5]. That is, the results from many (weak) change detectors are combined to make a final change detection result. Therefore, image pixels are sampled with replacement. Each sample set is then used to generate a change detector using kernel PCA. The reason of using ensemble approach is that using all image pixels at once to generate a single change detector required large amount of computation resource, especially, for solving the eigenvalue problem (13) and computing the principal component (15) because the amount of satellite image's pixels is huge. In summary, the work flow of the proposed change detection algorithm is illustrated in Figure 1.

3.1. Histogram matching

The purpose of histogram matching is to adjust the digital numbers of an image such that its histogram is match with that of the reference one. The process begins with the calculation of cumulative distribution function of two images' histograms. Let the cumulative distribution functions of reference and the input image be F_{ref} and F_{input}, respectively. The histograms of those images can be matched under the relation:

$$F_{\text{ref}}(G_{\text{ref}}) = F_{\text{input}}(G_{\text{input}}), \tag{16}$$

where G_{input} and G_{ref} are the grey values of input and reference images, respectively. Therefore, the color of the input image can be adjusted to be matched with the reference one by:

$$\text{adjusted grey value} = F_{\text{ref}}^{-1}\left(F_{\text{input}}\left(G_{\text{input}}\right)\right) = F_{\text{ref}}^{-1} \circ F_{\text{input}}\left(G_{\text{input}}\right) \tag{17}$$

The image after color adjustment is used in the change detector generation.

3.2. Kernel PCA for change detection

The core process of the proposed change detection algorithm is the kernel PCA. As it is already mention, the proposed method uses the several change detectors generated from sampled image pixels. Without loss of generality, the construction of a single change detector is explained.

Let $\mathbf{x}^1$ and $\mathbf{x}^2$ be image pixels sampled from the first and second images, respectively. The kernel PCA can be performed using bi-temporal data:

$$\mathtt{X} = \begin{bmatrix} \mathbf{x}_1^1 & \dots & \mathbf{x}_n^1 \\ \\ \mathbf{x}_1^2 & \dots & \mathbf{x}_n^2 \end{bmatrix}_{2P \times N}, \tag{18}$$

where N is the number of sampled image pixel. The image of $\mathtt{X}$ in the feature space is then:

$$\Phi = \left[\phi\left(\left[\mathbf{x}_1^{1\top}\ \mathbf{x}_1^{2\top}\right]^\top\right) \ \dots\ \phi\left(\left[\mathbf{x}_n^{1\top}\ \mathbf{x}_n^{2\top}\right]^\top\right)\right]. \tag{19}$$

The kernel PCA is then performed to extract the principal component from the data.

To detect change, it is hypothesized that the relation between unchange pixels in feature space is linear. That is, the unchange pixels in the feature space lie in a narrow elongated cluster along the first principal axis. The change pixels therefore lie far from the first principal axis. As a consequence, the second principal component can be used for quantifying the amount of change such that a pixel is classified as "change" if its absolute second principal component is larger than a threshold:

$$D(\mathbf{x}) = \begin{cases} 1 & \mathrm{PC}_2(\mathbf{x}) > \tau \ \text{(change)} \\ 0 & \text{otherwise} \ \text{(unchange)} \end{cases} \tag{20}$$

The threshold is selected using k-means algorithm to find clusters in the absolute values of the second principal components. The largest cluster center is then used as the threshold.

3.3. Ensemble change detection

In this work, the ensemble approach is employ. The concept of the ensemble approach is to use many classifiers, which is change detector in this work, for classification. The final result is then the combination of results from all classifiers. The statistical rational behind the concept of ensemble approach is the that the chance of selecting a poorly performing classifier can be decreased by combining the results from several classifiers. Moreover, the benefit of using small data subset is the small computation complexity.

The ensemble method used in this work is the Bootstrap Aggregating method also known as Bagging [1]. Bagging creates bootstrap sample by randomly drawing the data with replacement. Each sample is then used to train a classifier. Finally, a majority or weighted vote is used to classify an instance.

Using the bagging concept, the proposed change detection method generate a set of change detectors, which is illustrated in Algorithm 1. The algorithm uses bootstrap method to sample the image pixels with replacement. The sampled pixels is then used to extract eigenvectors from kernel PCA. The threshold is then determined from the second principal components. The final change detection result is determined from all change detectors using majority voting:

$$\text{Final change detection result} = \text{Majority}(\{D_t(\mathbf{x})\}). \tag{21}$$

Algorithm 1: Change detector generation

Input : S : image pixels; T : number of change detectors; n : Bootstrap size(number of sampled pixel); $\mathbf{v}$: eigenvector; τ : threshold

Output: A set of change detectors $\mathcal{D} = \{D_t\}$

```
for t ← 1 to T do
    S_t ← BootstrapSample(n, S);
    v ← ComputeKernelPCA(S_t);
    τ ← ComputeThreshold(v, S_t);
    D_t ← {v, τ}
end
```

3.4. Experimental result

To test the performance of the proposed change detection method, Landsat-8 images were used, see Figure 7. There sizes are 500pixels $\times$ 400pixels. The process begins with histogram matching of all image bands. In this

Figure 2: Landsat-8 images for experiment.

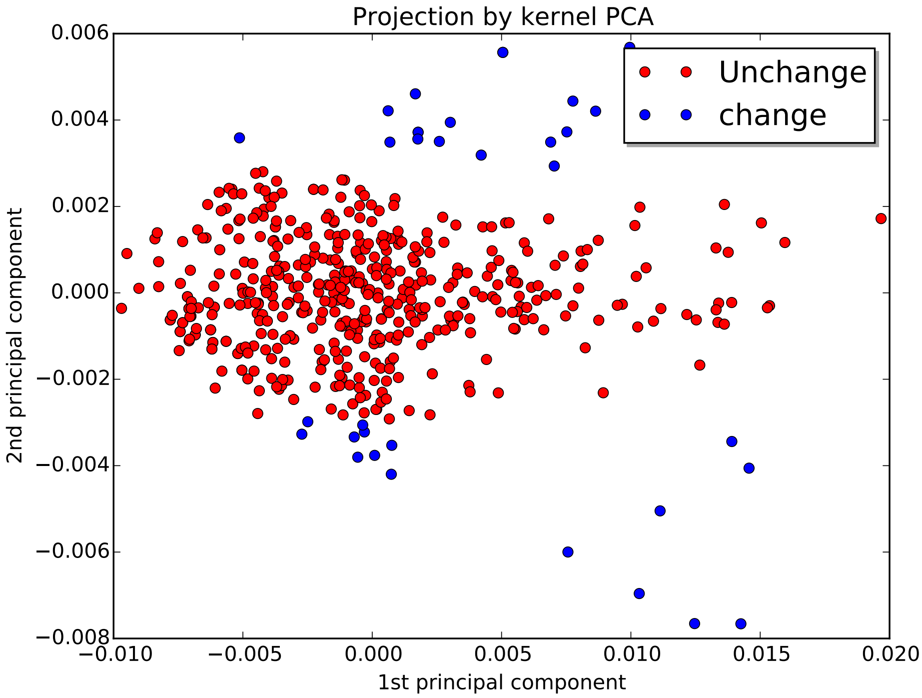


Figure 3: The first and second principal components using kernel PCA.

experiment, there were 150 change detectors; each detector was constructed using 500 pixels which were randomly sampling with replacement. The kernel function used in the experiment was Gaussian kernel.

In Figure 3, the scatter plot of the first and second principal components from the generation of a single change detection is presented. The threshold value is the cluster center of the absolute values of the second principal component. The change pixels of which the absolute values of the second principal component are greater than the threshold are plotted with blue color. It can be observed that the unchange pixels are in the cluster elongated in the direction of the firs principal axis.

According to the concept the of ensemble approach i.e., bagging method, the final change detection result is majority vote of the detection results from all detectors, where the detection from a single change detector is shown in (20). The cumulative vote (normalized to 1) from all 150 change detectors is illustrated in Figure 4. Therefore, a pixel is classified as change if its cumulative vote is greater than 0.5. The final change detection result is shown in Figure 5. The change detection result on the subarea of the test images is illustrated in Figure 6.

To show the advantage of the ensemble approach, the results from single detector using the same amount of sampled image pixels are illustrated in Figure 7. The change detection result shown in Figure 7.a is over-detection.

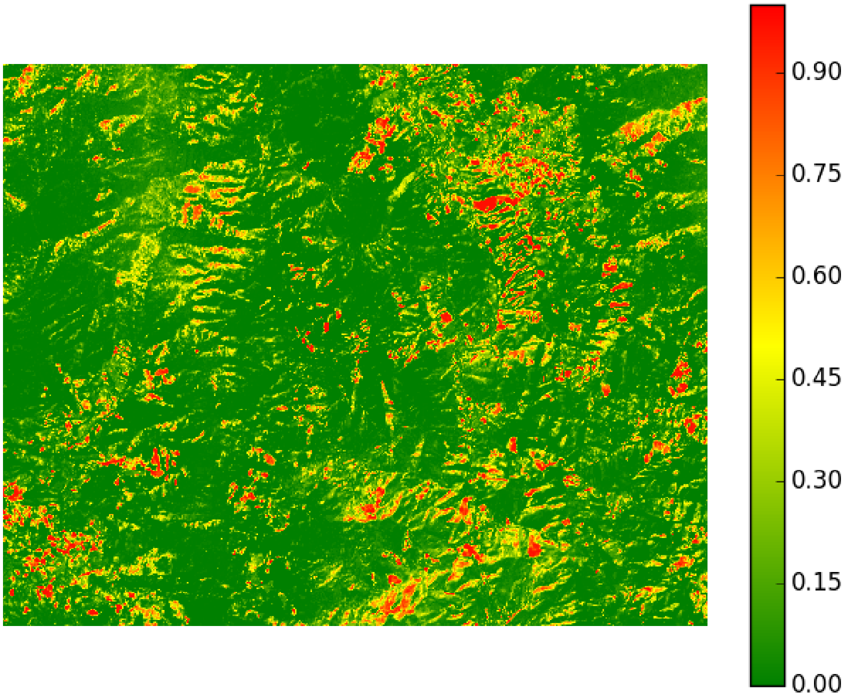


Figure 4: Cumulative voting, where the number of votes is normalized to 1. A pixel most likely change if its cumulative voting core close to 1.

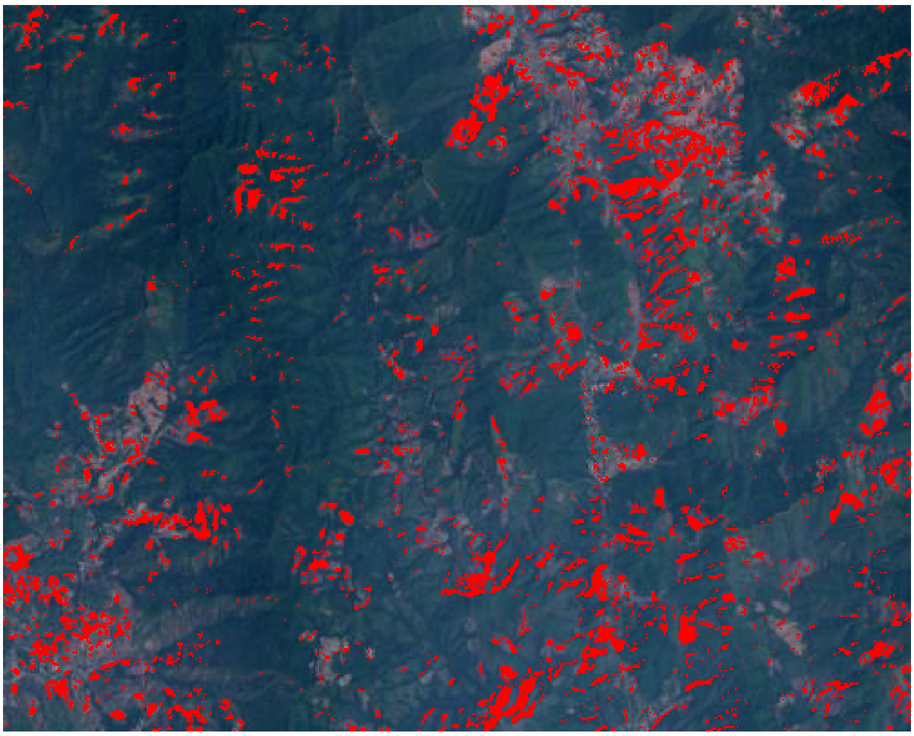
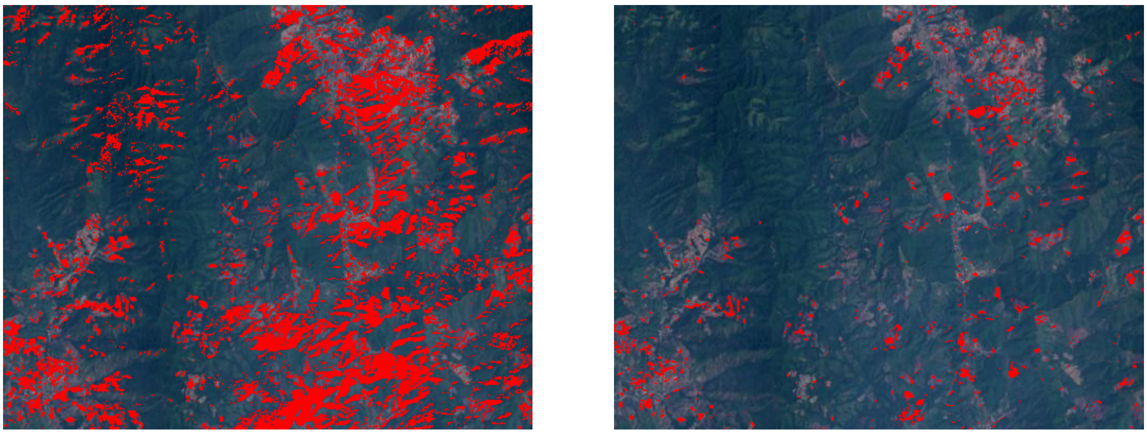


Figure 5: Change detection result of the image in Figure 2. The change pixels are labelled with red color.

(a) Subarea of the image in Figure 2.a (b) Subarea of the image in Figure 2.a (c) Change detection result

Figure 6: The change detection result on the subarea of the image shown in Figure 2. The change pixels are labelled with red color.

In other words, unchanged pixels are mis-classified as "change" because its threshold value is too low. The cause of low threshold value is from the random sampling process drawing unchange pixels from the image. Conversely, the Figure 7.b shows under-detection. That is, some change pixels were not detected. Using the majority vote can hence reduce these effects.



(a) Over-detection (b) Under-detection

Figure 7: Change detection results from two different single detectors.

4. Conclusion

In this paper, a novel change detection method is presented. Since it can handle the non-linear relation in the data, the kernel PCA is used as a change detector. However, using all satellite image pixel at once to train the change detector cannot be possible because the image size can be very large. To reduce the size of computational problem, we adopted the ensemble approach i.e., bagging. Therefore, a set of change detectors is used in stead of a single detector. The process begins with the histogram matching in order to reduce the effect of sensor response which can change over time. A set of change detectors is then generated by randomly sampling image pixels with replacement i.e., bootstrap sampling. The bootstrap samples are then used to train different change detectors. The final change detection result of an instance is then the majority voting of the result from all detectors. The performance of the proposed method is evaluated with satellite image.

In the future work, a new sampling scheme will be developed in order to sample more importance data. Moreover, a method for determining weight of weight voting will be invented.

References

[1] Breiman, L. Bagging predictors. *Machine Learning 24*, 2 (1996), 123–140.

[2] Bruzzone, L., and Prieto, D. Automatic analysis of the difference image for unsupervised change detection. *Geoscience and Remote Sensing, IEEE Transactions on 38*, 3 (May 2000), 1171–1182.

[3] Celik, T. Unsupervised change detection in satellite images using principal component analysis and k -means clustering. *Geoscience and Remote Sensing Letters, IEEE 6*, 4 (Oct 2009), 772–776.

[4] Deng, J. S., Wang, K., Deng, Y. H., and Qi, G. J. Pca-based land-use change detection and analysis using multitemporal and multisensor satellite data. *Int. J. Remote Sens. 29*, 16 (Jan. 2008), 4823–4838.

[5] Dietterich, T. G. Ensemble methods in machine learning. In *Multiple Classifier Systems*, vol. 1857 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2000, pp. 1–15.

[6] Erbek, F. S., zkan, C., and Taberner, M. Comparison of maximum likelihood classification method with supervised artificial neural network algorithms for land use activities. *International Journal of Remote Sensing 25*, 9 (2004), 1733–1748.

[7] Huang, C., Song, K., Kim, S., Townshend, J. R., Davis, P., Masek, J. G., and Goward, S. N. Use of a dark object concept and support vector machines to automate forest cover change analysis. *Remote Sensing of Environment 112*, 3 (2008), 970 – 985.

[8] Hussain, M., Chen, D., Cheng, A., Wei, H., and Stanley, D. Change detection from remotely sensed images: From pixel-based to object-based approaches. {*ISPRS*} *Journal of Photogrammetry and Remote Sensing 80* (2013), 91 – 106.

[9] Nackaerts, K., Vaesen, K., Muys, B., and Coppin, P. Comparative performance of a modified change vector analysis in forest change detection. *International Journal of Remote Sensing 26*, 5 (2005), 839–852.

[10] Nielsen, A. A., and Canty, M. J. Kernel principal component analysis for change detection. In *SPIE Europe Remote Sensing Conference* (sep 2008), F. P. Lorenzo Bruzzone, Claudia Notarnicola, Ed., vol. 7109, SPIE.

[11] Schölkopf, B., Smola, A., and Müller, K.-R. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput. 10*, 5 (July 1998), 1299–1319.

[12] Tomowski, D., Ehlers, M., and Klonus, S. Colour and texture based change detection for urban disaster analysis. In *Urban Remote Sensing Event (JURSE), 2011 Joint* (April 2011), pp. 329–332.