Convolutional neural network features based change detection in satellite images

Conference Paper · July 2016

DOI: 10.1117/12.2243798

3 authors, including:

Mohammed El Amin Larabi
Space Techniques Center/ Algerian Space Agency
27 PUBLICATIONS **183** CITATIONS
SEE PROFILE

Qingjie Liu
Beihang University (BUAA)
67 PUBLICATIONS **2,275** CITATIONS
SEE PROFILE

Convolutional Neural Network Features Based Change Detection in Satellite Images

Arabi Mohammed El Amin*, Qingjie Liu*, Yunhong Wang
State Key Laboratory of Virtual Reality Technology and System
School of Computer Science and Engineering, Beihang University, Beijing 100191

ABSTRACT

With the popular use of high resolution remote sensing (HRRS) satellite images, a huge research efforts have been placed on change detection (CD) problem. An effective feature selection method can significantly boost the final result. While hand-designed features have proven difficulties to design features that effectively capture high and mid-level representations, the recent developments in machine learning (Deep Learning) omit this problem by learning hierarchical representation in an unsupervised manner directly from data without human intervention. In this letter, we propose approaching the change detection problem from a feature learning perspective. A novel deep Convolutional Neural Networks (CNN) features based HR satellite images change detection method is proposed. The main guideline is to produce a change detection map directly from two images using a pretrained CNN. This method can omit the limited performance of hand-crafted features. Firstly, CNN features are extracted through different convolutional layers. Then, a concatenation step is evaluated after an normalization step, resulting in a unique higher dimensional feature map. Finally, a change map was computed using pixel-wise Euclidean distance. Our method has been validated on real bi-temporal HRRS satellite images according to qualitative and quantitative analyses. The results obtained confirm the interest of the proposed method.

Keywords: Convolutional Neural Network (CNN), Change Detection (CD), High Resolution Remote Sensing (HRRS)

1. INTRODUCTION

Change detection (CD) is the heart process of many applications utilizing remote sensing images. It leads to the identification of changes that has occurred on the Earth's surface by processing two (or more) images acquired at different times that cover the same geographical area. CD has a wide range of uses, including land use and land cover change monitoring, risk assessment, urban growth studies and environmental investigation.

Based on hand-engineered features, a variety of algorithms have been proposed to solve the CD problem, such as image differencing (ID) [1], image rationing (IR) [1], principal component analysis (PCA) [2], change vector analysis (CVA) [3], expectation maximization (EM) [4], graph cut [5], the Parcel-based method [6] and Markov random field [7]. To calculate these hand-engineered features, parameters such as sizes, scales and directions should be prudently and elaborately selected. Also, features selection and combination is another obstacle for HR imagery CD.

Our approach is inspired by the recent success of CNN model [8, 9]. This model transform sequentially a given input to the expected output through a sequence of processing steps [8, 9], producing a hierarchy of feature maps via learned filters. CNNs trained on specific tasks are capable to automatically learn complex features from images and achieve superior performance compared to hand-crafted features [8, 9].

This work follows a similar line of thought, with question in mind: "Can we re-use a pre-trained deep CNN to detect changes in Bi-Temporal HRRS satellite images?"

Several recent researches prove that the upper layers of CNNs encode highly-abstract information about the input image [10]. However, they used the CNN features for generic tasks such as classification [11], where the goal is categorizing a holistic representation of the image without considering the object location within it. Lower levels features are good for correspondence, but higher levels are needed for semantic information. To get the best of both worlds, a feature fusion strategy is employed by stacking all feature maps in a high dimension hyper feature (figure 1).

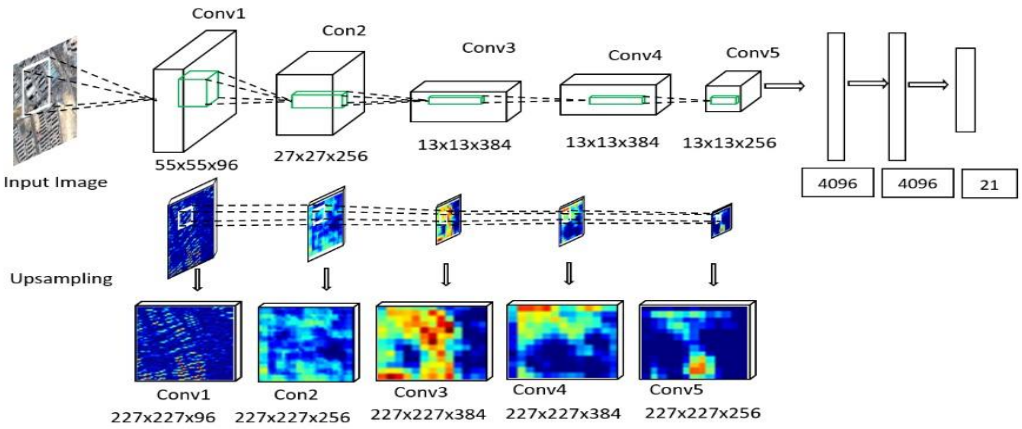

Figure 1. The hyper feature representation on top of AlexNet [8] architecture. Five convolutional feature maps are extracted and concatenated where a pixel feature is the vector of activations of all units above that pixel. The pseudo-colors represent intensity.

In this work, we propose HRRS images CD approach based on CNN features extracted from CaffeNet [12] model, pretrained on a large auxiliary dataset [13]. By using a simple feature fusion strategy, we stack all convolutional feature map levels to get a very high dimension feature [14].

The main contributions of this paper is reflected in the presentation of a novel CD method, based on the fusion of features extracted from deep CNN trained for a classification task. This is the first time using CNN features in the field of HRRS image CD. The remainder of this letter is organized as follows. Section II reviews the CaffeNet CNN model. Section III describes the proposed CD method in details. Section IV reports the experiments and results, and Section V concludes our work.

2. VISUALIZATION OF CONVOLUTIONAL NEURAL NETWORKS FEATURES

CNN is organized into alternating convolutional and max-pooling layers followed by a number of fully-connected layers, which transforms an input image from original pixel values to the final class scores in a feedforward manner. A typical CNN is shown in Figure 1 (AlexNet architecture).

Convolutional layers output a set of feature maps where each element of which is obtained by computing the dot product between a set of learned weights (filters) and the local regions (receptive field). The pooling layers perform a downsampling operation of feature maps by computing the maximum/minimum/average on a local region. Finally, the fully-connected layers follow several stacked convolutional and pooling layers, and the last fully-connected layer is a Softmax layer that computes the scores for each defined class.

Meanwhile, there are many famous and popular models for CNNs, such as AlexNet and CaffeNet. These models have proved to be effective for object detection and scene recognition, and obtained the state-of-the-art performance on ImageNet datasets [13]. However, their performance on CD has not been fully explored before. In the next section, we will exploit the CaffeNet model for CD.

CaffeNet. Based on AlexNet architecture [8], trained on the ILSVRC-2012 dataset, using Convolutional Architecture for Fast Feature Embedding (Caffe framework) [12], which is an open-source deep learning framework that is clean, modifiable and fast. The network architecture is shown in Figure 1. There are five convolutional layers, three max-pooling layers, and three fully-connected layers. CaffeNet differs from AlexNet in two small modifications: (1) trained without data augmentation, (2) exchanging the order of normalization and pooling layers. Through the visualization of the features from various levels of network layers, we can have a better understanding of the features learned by this model.

The information from the first layer are spatially more precise than the deeper layers, but it lacks the semantic information while the information from the last convolutional layer is too coarse spatially (due to sequences of

maxpooling, etc.). To get the best of both worlds, we concatenate features from different layers, to get a hyper vector for each pixel, as the vector of activations of all CNN units "above" that pixel. We will introduce it in the next section.

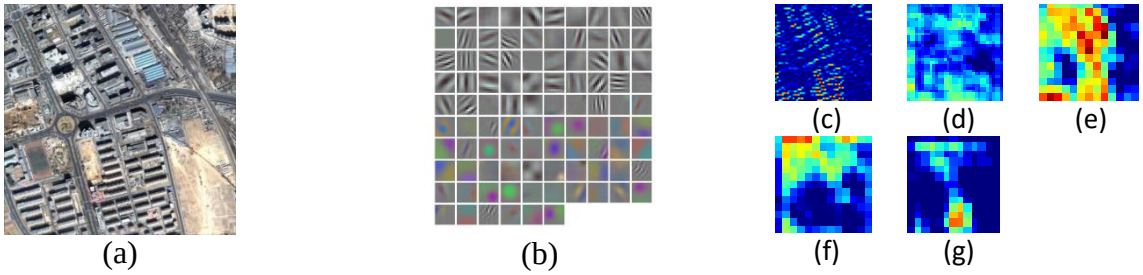

Figure 2. Visualization of features extracted from different layers, (a) Reference Image. (b) kernels learnt from ImageNet datasets by the first convolutional layer on the 227×227×3 input image. As it can be seen, most learned filters are edges detectors at different positions and orientations. (c), (d), (e), (f) and (g) are features extracted from: conv1, conv2, conv3, conv4, conv5 consecutive. It is clear that features extracted from the first convolutional layer are more precise spatially and it lose spatial details and get more sematic information increasingly along with deeper layers.

3. PROPOSED METHODOLOGY AND MODULE DESIGN

In the proposed framework, a rich set of feature maps is extracted from different layers with different abstraction levels in order to evaluate the CD task. In this section, details of the proposed methodology as well as brief experimental analysis of the proposed framework will be presented. The pipeline of our approach is demonstrated in Figure 3.

Images used as input to CNN in most current works are resized to a fixed size depending on the network input size (227×227 for CaffeNet). This process may suffer from information loss during image downsampling (pooling). To tackle this problem, we divide the input images I_1 and I_2 to N regular square grids of sizes $g \times g$ equal or less than the network input size $s \times s$. Regardless of the aspect ratio of the grid, we wrap all pixels in a tight bounding box around it to the required input size.

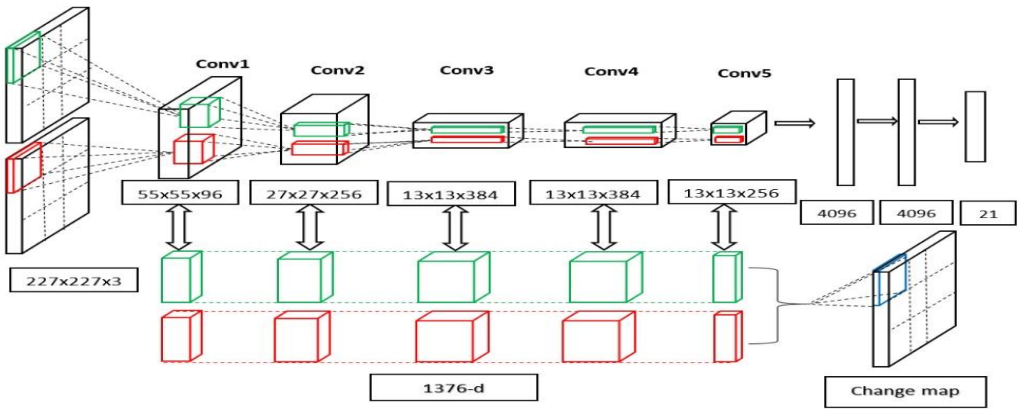

Figure 3. The workflow of the proposed CD framework.

3.1 Feature Extraction, Upsampling and Fusion

Given an input grid (as a result of the last step), we extract features from all convolutional layers, i.e., $Conv1, Conv2, \ldots, Convn$. These features are not of the same size due to downsampling operations. To combine multi-level maps, an upsampling step is necessary by a simple bilinear interpolation. We compute each output yij from the

nearest four inputs. Resulting in a set of upsampled features of the same size. Finally, after an L_2 normalization, a simple feature concatenation is employed where the dimension of the final representation F is calculated as follow:

$F = 96 + 256 + 384 + 384 + 256 = 1376$

3.2 Change Map generation

Given a hyper feature, each pixel is represented as a vector of activations of all CNN units "above" that pixel. To get the pixel wise distance between each two pixels in the same position in the Bi-temporal images I_1 and I_2. Euclidian distance is performed in the feature space of k-dimension the hyper vector as follows:

$$d_{ij} = \sum_{k=1}^{k} \left(\mu_i^k - \mu_j^k \right)^2 \qquad (1)$$

Where, k is the feature dimension. μ_i^k and μ_j^k are the features values at dimension k_{th} of the positions i and j respectively. A thresholding step is performed to get the final change map. Otsu segmentation method [15] becomes widely used in image processing for its better segmentation results, easy computation and wide scope of application. The algorithm assumes that any image contains two classes of pixels (e.g. foreground and background). Then, it calculates the optimum threshold separating those two classes so that their combined spread (intra-class variance) is minimal.

4. EXPERIMENTS AND DISCUSSION

To assess the effectiveness of the proposed approach, experiments were conducted on a bi-temporal HRRS satellite images. The proposed method is compared with image differencing ID and image rationing IR methods [1], block PCA method [2], EM-based method, MRF-based method [4] and Parcel-based method [6]. An image registration and correction preprocessing steps is necessary for the experiments. Due to space limitation, only two pairs of images are shown in this paper.

4.1 Study area and dataset

Bi-temporal images were used for performance comparison, which were taken by QuickBird-2 satellite over Beijing, China. The sizes of them are 1024×1024 pixels, as shown in Figure 4. Reference, target and ground truth images are (a), (b) and (c) respectively.

(a) (b) (c)

Figure 4. (a) reference image, (b) target image, (c) ground truth.

For better understanding of CNN feature maps in CD task, we evaluate a series of experiments by computing the change map from each layer (Figure 5).

(a) (b) (c) (d) (e) (f)

Figure 5. Visualization of CD results, (a) (b) (c) (d) (e): represent CD results based on conv1, conv2, conv3, conv4, conv5. And (c) is the CD results based on hyper feature.

For visual comparison, the CD from the first convolutional layer is more structured spatially, but can't detect precise changes due to its low level features. The upper layers demonstrate more precise changes semantically, but wrong boundaries. However, the hyper features have the best results.

4.2 Visual results

The change detection results generated from the three methods have been shown in Figure 6. Where (GT) is the ground truth image, (ID) and (IR) are image subtraction and image rationing methods, (PCA), (EM), (MRF), (Parcel) are the images results from: block PCA, EM-based method, MRF-based method and Parcel-based method respectively. (CaffeNet) image is the result of the proposed method.

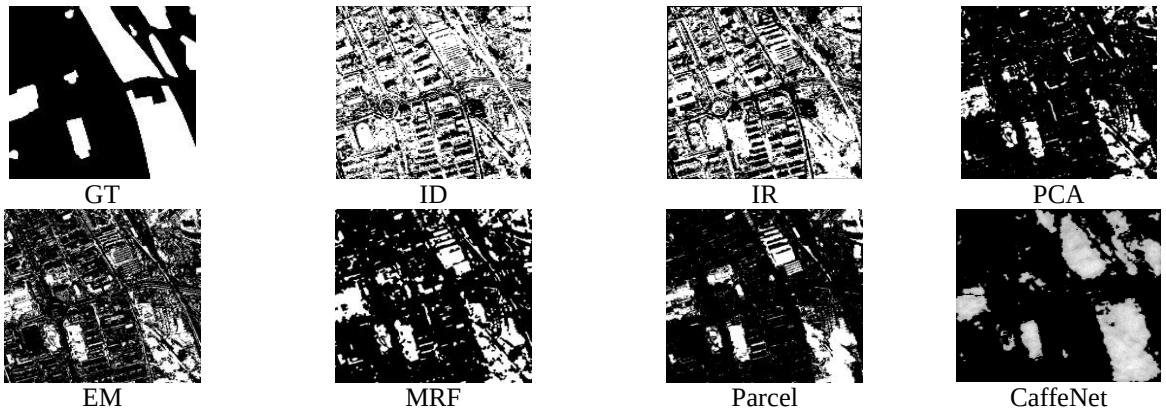

Figure 6. Change results: (GT) ground truth, (ID) image subtraction, (IR) image rationing, (CaffeNet) hyper features from CaffeNet. (PCA) block PCA method [2], (EM) EM-based method, (MRF) MRF-based method [4] and (Parcel) Parcel-based method [6]. black pixels are classified as 'No Change' and white pixels as 'Change'.

From the precision of view, this method has the best results, followed by the Parcel and MRF methods, the last is the images subtraction method. On an operational perspective, images subtraction and image ratio methods have relatively simple operations and cost less time, followed by PCA, Parcel, MRF and EM methods, while the proposed method is more complicated and time-consuming.

4.3 Quantitative results

In order to assess the effectiveness of our approach, we make a quantitative comparison against the best three methods of the aforementioned methods, (PCA) block PCA method, (EM) EM-based method and (MRF) MRF-based method [4] by computing false alarms, missed alarms, total error rate and kappa coefficient.

Table 1. Quantitative results of the proposed method

Accuracy	PCA [2]	EM [4]	MRF [4]	Proposed Method
False alarms	0.0565	0.2235	0.2093	0.495
Missed alarms	0.4952	0.3488	0.2050	0.316
Overall Alarms	0.1149	0.2402	0.2087	0.066
Kappa coefficient	0.4737	0.2913	0.3931	0.876

The proposed method is compared against three other CD methods, as shown in the table 1, the proposed method achieved the best Kappa coefficient 0.876 which mean that is an excellent classification. Our approach presents a general method that can be applied to detect change in land use and lad cover areas due to CNN features characteristics.

5. CONCLUSION AND FUTURE WORK

In this paper, a novel CD is presented, based on CNN hyper features of two registered images, covering the same area took at different times, t1 and t2. Experiments has demonstrated the effectiveness of the CNN features on CD task. By fusing the information from several layers, the CD performances are greatly enhanced. The technique has a drawback of requirement of more computational time and the necessity of the registration step. In future studies, we plan to investigate more sophisticated strategies to high-level spatial information and shape based features encoding process to improve the invariance of representations. would also like to use state-of-the-art deeper CNNs like VGG-19 [9] and ResNet [16], and finetune these networks on more specific dataset for HRRS images like UC Merced Land Use Dataset [17] and WHU-RS Dataset [18].

REFERENCES

[1] Coppin, P.R., Bauer, M.E., 1996. Digital change detection in forest ecosystems with remote sensing imagery. Remote Sensing Reviews 13, 207–234.

[2] Celik, T. Unsupervised change detection in satellite images using principal component analysis and k-means clustering. IEEE Geosci. Remote Sens. Lett. 2009, 6, 772-776.

[3] Johnson, R.D., Kasischke, E.S., 1998. Change vector analysis: a technique for the multispectral monitoring of land cover and condition. International Journal of Remote Sensing 19, 411–426.

[4] Bruzzone, Z.; Prieto, D.F. Automatic analysis of the difference image for unsupervised change detection. IEEE Trans. Geosci. Remote Sens. 2000, 38, 1171-1182.

[5] Chen, C. Huo, Z. Zhou and H. Lu, "Unsupervised Change Detection in SAR Image using Graph Cuts," Geoscience and Remote Sensing Symposium, 2008. IGARSS 2008. IEEE International, Boston, MA, 2008, pp. III - 1162-III - 1165.

[6] Bovolo, F. A multilevel parcel-based approach to change detection in very high resolution multitemporal images. IEEE Geosci. Remote Sens. Lett. 2009, 6, 33-37.

[7] Kasetkasem and P. K. Varshney, "An image change detection algorithm based on Markov random field models," in IEEE Transactions on Geoscience and Remote Sensing, vol. 40, no. 8, pp. 1815-1823, Aug 2002.

[8] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks. In NIPS, 2012.

[9] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556, 2014.

[10] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In ECCV, pages 818–833. Springer, 2014.

[11] Penatti, O.A.; Nogueira, K.; dos Santos, J.A. Do Deep Features Generalize from Everyday Objects to Remote Sensing and Aerial Scenes Domains? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Boston, MA, USA, 12 June 2015; pp. 44–51.

[12] Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; Darrell, T. Caffe: Convolutional Architecture for Fast Feature Embedding. In Proceedings of the ACM International Conference on Multimedia, Orlando, FL, USA, 3–7 November 2014. 29.

[13] Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 2015, doi: 10.1007/s11263-015-0816-y.

[14] B. Hariharan, P. Arbelaez, R. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization. In CVPR, 2015.

[15] N. Otsu, “A Threshold Selection Method from Gray-Level Histograms”, IEEE Transactions on Systems, Man, and Cybernetics, vol. 9, no. 1, pp. 62-66, Jan. 1979.

[16] K. He, X. Zhang, S. Ren and J. Sun, “Deep Residual Learning for Image Recognition,” arXiv:1512.03385v1.

[17] Yang, Y.; Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010; pp. 270–279.

[18] Xia, G.S.; Yang, W.; Delon, J.; Gousseau, Y.; Sun, H.; Maitre, H. Structrual High-Resolution Satellite Image Indexing. In Processings of the ISPRS, TC VII Symposium Part A: 100 Years ISPRS—Advancing Remote Sensing Science, Vienna, Austria, 5–7 July 2010.