Abstract—In this paper, we propose a novel unsupervised optical remote sensing change detection (CD) based on pre-trained convolutional neural network (CNN) on ImageNet dataset and superpixel (SLIC) segmentation technique. The proposed approach can be divided into three steps. First, bitemporal images are stacked, and Principal Component Analysis (PCA) is applied to extract three higher uncorrelated channels, which will be later segmented into superpixels. Second, we zoom out each region into three levels and fit them separately into a pre-trained CNN. Third, we extract features of different zooming levels that represent the same region (superpixel) and concatenate them. We compare the concatenated features to get the final change map. The experimental results demonstrate the efficacy of the proposed approach.

Keywords-component; remote sensing; change detection; deep learning; convolutional neural network

I. INTRODUCTION

“Change detection (CD) is the process of identifying the changes in remote sensing images that cover the same area of the earth surface in two different times” [1]. Remote sensing CD methods can be divided into three classes depending on the analysis unit: pixel, kernel, and object-based approaches [2]. Pixel based approaches are the well-known methods, which are the most used for low and medium resolution [3].

Recently, with the appearance of a new generation of high dimensions remotely sensed imagery, pixel-based approaches become inappropriate [3]. Kernel and object-based approaches were appeared, which can better highlight the spatial neighborhood and the contextual information, and they are less sensitive to noises caused by the geometric and radiometric corrections [4]. Object-based CD has become increasingly popular due to its significant advantages over pixel-based approach. Object-based approaches offer more possibilities for the extraction of highly distinctiveness features that can better highlight the changed regions. Object-based approaches can be classified into three broad categories. 1) Object overlay. 2) Image-object direct comparison. 3) Multitemporal images objects. Image object overlay can be approached by segmenting one of the multi-temporal images and superpositioning the extracted boundaries on the second image; later the objects are compared. The main disadvantage of this method is that the geometry of the image-objects imitates only one of the bitemporal images [2]. Image objects direct comparison starts by segmenting each image separately, and then compare objects from the same location. The main downside of this approach is that it introduces the problem of ‘silver’ objects under inconsistent segmentations [5]. The last method is object overlay, where images are stacked and co-segmented in one-step. Our work is based on the object-overlap approach, which demonstrates its effectiveness besides the other approaches. The method is robust to deformation caused by the comparison of two different objects [2].

Currently, most of the object-based CD methods are subjected by handcraft features [3]. Recently, deep learning (DL) has become a standard way of automatically learning features from data. Convolutional Neural Networks (CNNs) can learn hierarchical feature from images. CNNs have demonstrated its effectiveness in many computer vision applications, such as face detection [6], scene parsing [7], facial point detection [8], and image classification [9].

Applying and training deep CNNs from scratch is not applicable in remote sensing field due to the unavailability of data. CNNs require a huge amount of data to be trained from scratch. An excellent property of CNNs that open the door for a wide range of applications in remote sensing field is its transferability from task to another. CNNs trained for a classification task can be transferred to object detection [10], object tracking [11], semantic segmentation [12], scene parsing [7], etc.

Based on the propriety above, CNNs were successfully transferred and implemented in remote sensing field, and new state-of-the-art were achieved in optical remote sensing classification [13]–[15], and pixel-wise labeling [16]. Beside this success, small amounts of works were applied DL for remote sensing CD, the most of them are based on unsupervised autoencoders or restricted Boltzmann machines [17]. Authors in [18] applied unsupervised patch based approach for SAR images CD, which is a lightweight CNN model where convolutional filters presented as PCA filters, they achieved remarkable results. Arabi et al. in [19], employed pixel-based approach for optical remote sensing CD by transferring a CNN model pre-trained on ImageNet dataset [20]. They achieved remarkable results, while the approach suffers from the overlapping boundaries and the higher computational requirement.

In this paper, a novel optical remote sensing CD approach based on pre-trained CNNs’s features and superpixel segmentation technique [21]. A superpixel is a set of pixels, which have similar spectrum and are adjacent in space [21]. It is highly uniform, homogeneous, and compact. An object is typically composed of numerous super-pixels.
The approach consists of four steps. 1) Superpixel segmentation, to avoid discords in segmentation boundaries, images are first fused using PCA, only the first three principle components are chosen for segmentation. 2) Zooming out each region (superpixel) into multiple levels (multiple scales) and feeding them into a pre-trained CNN separately. 3) Features extraction and fusion. 4) Dissimilarity measure and change classification. The suggested approach is estimated to take advantages of both CNN features discrimination ability and the superpixel ability to ignore noises.

The manuscript is organized as follow. In Section II, we present the CNN model used and the experiment and define the transfer learning strategy. Section III, present our method. Sections IV and VI presents the results and the conclusion.

II. CONVOLUTIONAL NEURAL NETWORK

The concept of CNN was first proposed in 1980 by Fukushima [22] with the name of NeoCognitron and later refined by LeCun [23]. Unlike traditional feed-forward networks, CNN has neurons with limited receptive fields that only process a local image region. The advent of affordable Graphics Processing Units (GPUs), together with the availability of large databases of annotated images make it possible to train and employ CNNs.

A typical CNN is composed of several cascaded layers with various types. First, convolutional layers: they convolve the learned weight with the input to produce a feature map that will be propagated to the next layer, which is usually the pooling layer. Second, pooling layers, which is a downsampling layer, its role is reducing the size of the input. Since the images’ sizes are reduced continuously, CNNs have the capacity of abstracting the low-level information generated in previous layers for final results. Third, the fully connected and Softmax layers, which are the classifier of the network. The parameters of CNN are usually trained with stochastic gradient descent (SGD).

A. Transfer Learning

In remote sensing field, there are no sufficient images to train a useful deep CNN [14]. Instead, we tend to transfer successfully pre-trained CNNs to remote sensing CD tasks. Features learned from natural images [20], and remote sensing images are similar, such as edges and blobs. These features are useful for remote sensing tasks. High-level features in deep CNNs computed from daily nature images may contain powerful representations for remote sensing images. To employ a CNN model pre-trained on ImageNet dataset [20] for remotely sensed imagery, two paradigms can be used, which are: direct application and fine-tuning. Direct application implies using the CNN models without any further retraining on remote sensing datasets, while the fine-tuning strategy involves retraining the pre-trained CNNs models on new remote sensing dataset to adjust its learned weights to the new datasets. Unfortunately, due to the lack of training data, we opt for using the first method to demonstrate the effectiveness of pre-trained CNNs for the unsupervised optical remote sensing change detection.

B. Pretrained CNNs Models

The most successful CNN model pre-trained on ImageNet is evaluated in our work, which is the leading VGG model [24]. In the literature, there exist many CNNs models. However, these successful deep CNNs do not achieve good performance for remote sensing classification task because they are tasks oriented and cannot be generalized to another dataset [15].

VGG model is a very deep CNN including VGG-16 (13 convolutional layers and three fully-connected layers) and VGG-19 (16 convolutional layers and three fully-connected layers). These models were developed to demonstrate the effect of network depth on ImageNet classification task. Recently, many works prove that the representations learned by VGG-16 and VGG-19 can generalize well to other datasets and tasks [14], [15], [25].

III. METHODOLOGY

In the proposed approach, the reference and changed images are pre-registered and radiometric corrected. The proposed method can be simply divided into four steps: 1) Multiresolution superpixel segmentation. 2) Zooming out superpixel into multiple levels (multiple scales). 3) Feature extraction and fusion. 4) Dissimilarity measure and change classification.

A. Superpixels Segmentation

The simple iterative cluster (SLIC) algorithm is employed to obtain the superpixel map. Superpixels preprocessing have become crucial components for computer vision tasks. The superpixels can replace the pixel-based structure of the original image by a set of regions that capture the local redundancy of information. They offer a useful way to represent the image and thus significantly reduce the complexity of the following image analysis process. One of the significant advantages of this method is the simplicity of its settings. Only two parameters are required, which are: 1) the index of compactness for the superpixels and, 2) the number of required superpixels. Functioning at the superpixel level rather than the pixel level can accelerate algorithms in comparison with the pixel-based approach. Furthermore, it improves their results [21].

Figure 1. Presents the SLIC segmentation results on bi-temporal images.
To escape any discordances of the superpixel boundaries obtained from the image pairs, a simple concatenation by stacking is employed. Then, PCA is applied to remove the redundant information, and only the first three principle components are chosen. This results in a new three spectral channels PCA image. Later, two-steps filtering stage is applied aiming at improving the superpixel segmentation. First, a median filter is applied to smooth the image and to eliminate the noise. Then, we use a bilateral filter [26] to preserve the edges. The result of this preprocessing and segmentation steps is shown in Fig. 1.

**B. Zoom-out Feature**

Zoom-out features were introduced in [27]. They achieved new state-of-the-art results in semantic segmentation task. The “secret” behind this technique is that “the information used in the feedforward classification is not computed from a small local region in isolation, but collected from a sequence of levels, obtained by “zooming out” from the close-up view of the superpixel” [27]. Starting from the superpixel itself, going to a large region surrounding it, extending to a larger region around them, we extract a rich feature representation for each level [27]. The technique permits us to exploit statistical structure and relations between image regions at different resolutions in a simple way. The main idea of employing the zoom-out architecture is to allow features extracted from various stages of spatial context nearby the superpixel segment to contribute to the categorization decision (change vs. no change) at that superpixel, (see Fig. 2).

**C. Features Extraction**

Convolutional layer with K filters produce K feature maps, which is later downsampled by the pooling layers. Usually, these feature maps have lower resolution than the original image, due to subsampling induced by the pooling operation and possibly in filtering operations. Features extracted from different layers are of the different level of abstractions, which can play a dominant role for the transfer learning approaches. The most of the works that applied the transfer learning proved that the convolutional layers are the most appropriate. For this end, we will employ and concatenate features from different convolutional layers. To improve performance, some preprocessing steps are presented in this section.

**D. Features Fusion**

In this work, we have two levels of fusions, which are intra-fusion and inter-fusion. First, intra-fusion is the aggregation step of CNNs features extracted from different convolutional layers that belong to the same zoom level, which represents the same center of mass (center pixel). Second, the inter-fusion between various levels of zoom, which is achieved by integrating the Spatial Pyramid Pooling (SPP) method [28], which can generate a fixed size representation, for all levels of zoom.

1) **Intra-Fusion**

By feeding a single level of zoom into the pre-trained CNN, features extracted from different convolutional layers are of a different depth and spatial dimensions. To unify theses dimensions, two steps are considered. First, fusing features of various depth dimensions may be risky for decision-making, where the high dimensional features may dominate the lowers. For this end, we unify and reduce the features depth dimensionality. Second, we unify the spatial dimension by upsampling the lower dimension features to meet the higher ones using bilinear interpolation.

a) **Bilinear interpolation**

This is the mostly used method for upsampling lower dimensional features [29], (see Fig. 3).

![Bilinear interpolation procedure.](image)

b) **Depth dimensionality reduction**

Can be approached by various methods. In our case, as we opt for highlighting the super-pixel boundaries, we try to
convolve the extracted features with the corresponding superpixel mask and extract X-best feature maps that represent better the superpixel mask. $X$ is the number of feature maps, (see Fig. 4).

2) Inter-Fusion

Features from different levels of zoom are of different spatial resolution, which may highly affect the results where the higher dimensional features that represent a higher dimensional zoom level may dominate the lower level one. SPP method can leverage this problem by fixing all features to the same depth dimension $D$.

a) Spatial pyramid pooling

The SPP algorithm pools the features and generates fixed depth outputs for each level of zoom separately, which are then concatenated. It is about performing features “aggregation” extracted from convolutional layers, (please refer to the reference [28] for more details).

E. Change Feature Extraction

After features extraction and fusion, Euclidian distance will be computed to compare them followed by unsupervised clustering.

1) Dissimilarity measure

The differences of each superpixel are characterized by the Euclidean distance. The Euclidean distance between two features ($X^t$ and $X^s$), expressed as the square root of the band-wise sum of the squares of the differences

\[
d_{\text{Euclidean}} = \sqrt{\sum_{i=1}^{N} (X^t_i - X^s_i)^2}
\]

(1)

2) Fuzzy C-means clustering

Fuzzy C-means (FCM) algorithm, one of the most popular fuzzy clustering techniques [30]. FCM can determine, and iteratively update the membership values of a data point with the pre-defined number of clusters.

IV. EXPERIMENTAL SETTING

To validate the effectiveness of the proposed methods, we challenge the high-resolution optical CD problem on two real data sets. The two bi-temporal images with their ground-truth are displayed in Fig. 5.

Three state-of-the-art CD methods, which are the EM-based method and MRF-based method [31], the PCA-based method [32]. These methods are implemented using the default configurations provided in the original implementations. The experiments will demonstrate the superiority of the proposed technique based on zooming-out superpixel segmentation and CNNs features.

A. Data Sets

The study area is located in Beijing Province, China. Two datasets, including a pair of QuickBird images, and another pair cropped from Google Earth as shown in Fig. 5, the ground truth images are produced manually.

The first image was taken in September 2002 and the second one was captured in November 2003. The second image pair were captured on September 2012 and March 2013. They have a spatial resolution of 1 m.

Figure 5. Bi-temporal images, the first row presents the first dataset and the second row shows the second dataset.

B. Evaluation Criteria

The CD results are often shown in the form of binary maps, while the white pixels in the maps indicate the changed pixels and the black represent the unchanged, and we need to use some criteria to assess the performance of different methods. To calculate the criteria quantitatively, we compare the CD result with the reference image. In our experiments, false alarms (FAs) denote the number of pixels, which are classified into the changed class but unchanged in the reference image, while missed alarms (MAs) represent the number of pixels, which are classified into unchanged class but changed in the reference image. TP and TN are the number of true changed pixels and unchanged pixels compared with the reference image, respectively. Then, the overall error (OE) can be computed by

\[
OE = FA + MA
\]

(2)

Furthermore, kappa coefficient (KC) is usually applied to evaluate the effect of classification. The higher the value of KC is, the better the classification result one method obtains. KC is calculated as

\[
KC = \frac{PCC - PRE}{1 - PRE}
\]

(3)

where PCC represents the percentage of the correct classification, and PRE represents the proportion of expected agreement, and they are defined as follows

\[
PCC = \frac{TP + TN}{TP + FA + TN + MA}
\]

(4)

\[
PRE = \frac{(TP + FA)(TP + MA)}{(TP + FA + TN + MA)} + \frac{(TN + MA)(TN + FA)}{(TP + FA + TN + MA)}
\]

(5)
V. RESULTS

A. Visual Results

The CD results generated from the three methods have been shown in Fig. 6. (GT) is the ground truth images, (PCA), (EM), (MRF) are the images results from block PCA, EM-based method, MRF-based method respectively. The (VGG-16) image is the result of the proposed method.

From the precision of view, the proposed approach demonstrate the best results and proves its effectiveness besides the other methods, followed by PCA method, while the EM and MRF methods present similar results. On an operational perspective, the site-of-the-arts methods consume less time and energy, while the proposed method requires more execution time.

![Image](image-url)

Figure 6. Change results: The first and the second bi-temporal images are shown in the first and the second rows. (EM) EM-based method, (MRF) MRF-based method [31]. (PCA) block PCA method [32]. (VGG-16) is our proposed method’s results. Black pixels are classified as ‘No Change’ and white pixels as ‘Change.’

B. Quantitative Results

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

We compared our approach against three other CD methods, as shown in the Tables I and II, the proposed method achieved the best Kappa coefficient which mean that methods, as shown in the Tables I and II, the proposed method that can be applied to detect changes in land use and land cover areas due to CNN features characteristics.

![Table Image](table-url)

TABLE I. QUANTITATIVE COMPARISON OF EXPERIMENTAL RESULTS OBTAINED BY DIFFERENT METHODS ON THE DATASET 1

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>False alarms</td>
<td>21887</td>
<td>75462</td>
<td>70536</td>
<td>25433</td>
</tr>
<tr>
<td>Missed alarms</td>
<td>29138</td>
<td>20724</td>
<td>13121</td>
<td>17497</td>
</tr>
<tr>
<td>Overall Alarms</td>
<td>51026</td>
<td>96187</td>
<td>83658</td>
<td>42930</td>
</tr>
<tr>
<td>Kappa coefficient</td>
<td>0.4247</td>
<td>0.2786</td>
<td>0.3815</td>
<td>0.6981</td>
</tr>
</tbody>
</table>

VI. CONCLUSIONS

In this paper, a novel optical remote sensing CD method was presented and demonstrated. Based on features extracted from CNN pre-trained on large natural image dataset. We generate three levels of zoom over each superpixel segment and fit them separately into the pre-trained CNN. Then, features were extracted, concatenated, and compared with the corresponding segment from the second image.

Employing superpixels instead of basic pixel units accelerate the framework while escaping the noisy detections. We demonstrate the high level of distinctiveness property of CNNs features. In future works, we will work on modeling and training a new Siamese Network based on CNNs.

REFERENCES


学霸图书馆

www.xuebalib.com

本文献由“学霸图书馆-文献云下载”收集自网络，仅供学习交流使用。

学霸图书馆（www.xuebalib.com）是一个“整合众多图书馆数据库资源，提供一站式文献检索和下载服务”的24小时在线不限IP图书馆。

图书馆致力于便利、促进学习与科研，提供最强文献下载服务。

图书馆导航：
图书馆首页 文献云下载 图书馆入口 外文数据库大全 疑难文献辅助工具