Onat Vuran, Oguzhan Akcin, Mahdyar Ravanbakhsh, Bülent Sankur, Begüm Demir

Deep learning driven content-based image time-series retrieval in remote sensing archives

DEEP LEARNING DRIVEN CONTENT-BASED IMAGE TIME-SERIES RETRIEVAL IN REMOTE SENSING ARCHIVES

Onat Vuran[1], Oguzhan Akcin[2], Mahdyar Ravanbakhsh [3], Bülent Sankur[1] and Begüm Demir[3]

[1]Department of Electrical and Electronics Engineering, Bogazici University, Istanbul, Turkey
[2]Cockrell School of Engineering, The University of Texas at Austin, Austin, TX
[3]Remote Sensing Image Analysis Group, Technische Universitaet Berlin, Berlin, Germany

ABSTRACT

The rapid evolution of satellite imaging systems has resulted in sharp increases of image archive volumes. Multitemporal images constitute a sizeable portion of these time-series databases. Accordingly, development of accurate content based time-series retrieval (CBTSR) methods in massive archives of RS images attracts much research interest. Given a user-defined query time series, CBTSR aims at identifying within a massive archive image time series that show characteristics similar to those of the query time series. In this paper, we focus our attention to CBTSR in pairs of RS images, aiming to search and retrieve bi-temporal image pairs containing changes similar to those modeled in the query. To this end, we introduce two deep learning-based methods in the framework of CBTSR. The first method, called deep change vector retrieval (DVCR), is based on selected deep features extracted from the change vector analysis. The second method, called autoencoder with early fusion (AEEF) uses an autoencoder architecture to recreate the time difference images and the latent codes produced by this network. Experimental results show the effectiveness of the proposed methods for CBTSR problems. The code of the proposed methods is available at : https ://github.com/OnatV/ChangeRetrieval.

Index Terms— Content based time-series retrieval, deep learning, remote sensing.

1. INTRODUCTION

Recent advances in satellite technology have led to a regular, frequent, and high-resolution monitoring of Earth at the global scale, delivering an unprecedented amount of data on the state of our planet and changes occurring on it. As an example, through the Copernicus programme (which is the European flagship satellite initiative in EO), Sentinel satellites reach the scale of more than 10TB data per day and the total size of the Copernicus data archives makes up almost a volume of 20 PB. As a result, information retrieval from massive archives has become an important topic of research in RS. In this direction, the development of content-based single-date image retrieval is widely investigated in RS. However content-based time-series retrieval (CBTSR) has been seldom considered. Given a user-defined image time series (i.e., the query time series), CBTSR is devoted to identify from the archive those portions of time series that are associated to the spatial, spectral and temporal content similar to the query.

CBTSR approaches can be divided into two categories : 1) retrieval of long-term changes (e.g., seasonal changes or time varying phenomena that can be observed at different time resolution) ; 2) retrieval of short term (i.e., abrupt) changes (e.g., forest fires, floods) [1]. The approaches in different categories exploit different strategies to model the change information. In this paper, we concentrate on the retrieval of short term changes. For short term change retrieval in the framework of CBTSR, there are mainly three approaches in the literature. The first approach is devoted to extract and exploit common patterns of pixel neighborhood evolutions, where data mining techniques are used to extract a dictionary of prototypical change sequences [2, 3]. For example, Julea et al. [2] identify groups of pixels with similar evolutionary history and which cover more than a prescribed minimum surface area. After that, such groups of pixels are expressed with symbolic sequences, data-mining techniques are applied to find interesting patterns and extract data models. Radoi and Burileanu [3] extend the multi-temporal query-by example problem to a Bayesian time sequence matching and use dynamic time warping. The second approach is based on feature-based retrieval of image sequences, where feature engineering is applied to multitemporal image tuples [4]. RS images are partitioned into patches and multi-feature vectors consisting of color and texture components are extracted. Retrieval is achieved by similarity matching between the change feature vectors of the target pair of images and the change feature vector of the archive pairs. The third approach is defined based on spectral change vector analysis [1]. In detail, in [1] the spectral change vectors (SCV), which is the difference of multispectral pixel values for a bitemporal image pair, is used. The SCV values are grouped into a large number of clusters, and prototypical clusters are selected (each attributed to one change type). The query image can be associated with more than one cluster, thus accommodating more than one change type.

This paper addresses the abrupt change retrieval problems in the framework of CBTSR in RS using deep learning (DL) based methods. In particular, we investigate two DL based methods and compare their performance on the SECOND dataset presented in [7]. The first method is deep change vector retrieval (DVCR) method that exploits the difference of features extracted from layers of convolutional neural networks (CNNs) for bitemporal images. The second method is the autoencoder with early fusion (AEEF) that extracts autoencoder codes from concatenated image pairs, and uses these latent space codes to synthesize the difference image of bitemporal images.

2. METHODOLOGY

Let $\mathcal{D}$ be an archive that consists of M bi-temporal image pairs (i.e., samples), where $\mathcal{D} = \{\mathbf{X}^i\}_{i=1}^{M}$. In $\mathcal{D}$ each pair $\mathbf{X}^i$ is described as $\mathbf{X}^i = \{X_1^i, X_2^i\}$, where X_1^i and X_2^i are two co-registered RS images acquired using the same sensor over the same geographical area at times t_1 and t_2. Both DCVR and AEEF aim at retrieving bi-

temporal image pairs that represent similar changes with respect to a given query pair. To this end, the DCVR and AEEF methods compute a change descriptor vector F^i to represent the changes between X_1^i and X_2^i. The change descriptor vector F^i then is used to retrieve the most similarly changed pairs with regard to query images. The selection is based on the comparison of their change descriptors using the k-nn algorithm.

2.1. Deep Change Vector Retrieval (DCVR)

DCVR is an unsupervised method that exploits the middle layer features of a pre-trained CNN network to compute F^i. DCVR is inspired by [5], however, our approach differs in terms of feature selection strategy. More specifically, we extend the feature selection strategy proposed in [5] to address the change retrieval problem. We consider the output of each convolutional layer $l = 1..., L$ as a pixel-wise deep feature vector, where l is the chosen layer including ch number of channels (filters). Let $f_{\theta:l}(.)$ indicate the output of l^{th} layer activations (features) corresponding to a given input of CNN. The difference of multi-channel layer activations constitutes the deep change feature vector δ_l^i for the l^{th} is computed as :

$$\delta_l^i = f_{\theta:l}(X_2^i) - f_{\theta:l}(X_1^i) \tag{1}$$

where $f_{\theta:l}(X_1^i)$ and $f_{\theta:l}(X_2^i)$ indicate the l^{th} layer features for images X_1^i and X_2^i, respectively. We stack the channel-wise differences δ_l^i over the selected layers $l = 1, ..L$ to obtain the initial change descriptor vector $\hat{F}_i = [\delta_l^i]_{l=1}^L$. Since δ_l^i can become quite large for convolutional deep networks $\hat{F}_i$ has very large dimension. To obtain the final change descriptor vector F_i, we reduce the dimensionality of $\hat{F}_i$ by applying two feature selection strategies :
(i) Max pooling : selecting a fixed number m of largest elements in each channel $n = 1..., ch$ to obtain F_i with dimensions $m \times ch$,
(ii) Histogram : computing a histogram with b bins for each channel and use channel-wise histogram descriptors to obtain F_i with dimensions $b \times ch$.
It worth noting that the CNN network can be used as off-the-shelf. We can use any pre-trained CNN for semantic labeling in which a semantic label is predicted for each pixel, separately [6]. Without further fine-tuning we can compute the change descriptor vector $\hat{F}_i$ by extracting features from a number of selected layers.

2.2. Auto Encoder with Early Fusion (AEEF)

The proposed AEEF is an unsupervised method based on the autoencoder with early fusion. We assume that a training set $\mathcal{D}_T \subset \mathcal{D}$ is available, where $\mathcal{D}_T = \{X_1^i, X_2^i\}_{i=1}^{M_{\text{train}}}$. We first compute the difference image X_{diff}^i as :

$$X_{\text{diff}}^i = X_2^i - X_1^i. \tag{2}$$

The AEEF aims at learning to generate the difference image from the input bi-temporal images. To this end, the AEEF consists of two consecutive networks : 1) encoder network f_{enc}, and 2) decoder network f_{dec}. A block scheme of the proposed architecture is shown in Fig. 1. The encoder f_{enc} yields the one-dimensional latent space representation z^i, and the decoder f_{dec} gives an approximation to the difference image. To train the AEEF we used the training set $\mathcal{D}_T$. The input of the encoder network is the concatenation of the bi-temporal images X_1^i and X_1^i along their b channels (i.e., spectral bands), which results in $2b$-channel images $X_{\text{con}}^i = [X_1^i, X_2^i] \in \mathbb{R}^{W \times H \times 2b}$, where W and H are the width and height of the image

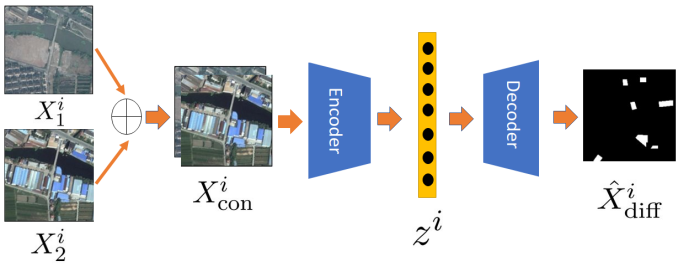

Fig. 1. A block scheme of the proposed AEEF method. $\oplus$ denotes the channel concatenation operation.

and b is the number of bands of the input images. The encoder generate latent space representation $z^i \in \mathbb{R}^d$ of image pairs X_1^i, X_2^i, where d is the dimension of the latent space representation z^i. Then, the decoder network inputs z^i and learn to reconstruct the difference image $\hat{X}_{\text{diff}}^i \in \mathbb{R}^{W \times H \times b}$ as follows :

$$\hat{X}_{\text{diff}}^i = f_{\text{dec}}(z^i) \ , \ z^i = f_{\text{enc}}(X_{\text{con}}^i) \tag{3}$$

We train networks f_{enc} and f_{dec} wrt the reconstructed difference image $\hat{X}_{\text{diff}}^i$ and the target difference image X_{diff}^i with the loss function $\mathcal{L}$. A weighted combination of ℓ_1 and ℓ_2 distance is computed between reconstructed difference image and actual difference image. The loss function $\mathcal{L}$ is defined as :

$$\mathcal{L} = \lambda_{\ell_1} \|\hat{X}_{\text{diff}}^i - X_{\text{diff}}^i\|_1 + \lambda_{\ell_2} \|\hat{X}_{\text{diff}}^i - X_{\text{diff}}^i\|_2. \tag{4}$$

Table 1. Land cover class (LCC) occurrence percentages in the SECOND dataset.

LCC	LV	NV	TR	WT	BL	PG
LV	-	63.1%	26.0%	9.34%	51.8%	1.44%
NV	57.4%	-	39.0%	8.40%	83.0%	3.34%
TR	16.9%	32.9%	-	2.84%	22.8%	0.70%
WT	8.30%	10.4%	3.52%	-	4.65%	0.10%
BL	32.7%	68.6%	21.0%	2.01%	-	1.27%
PG	1.03%	2.75%	0.37%	0.07%	1.04%	-

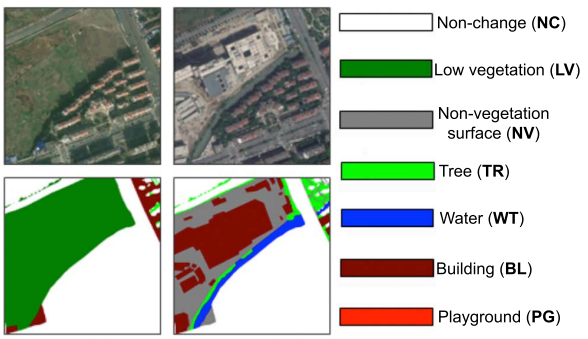

Fig. 2. Examples of a pair of images and their reference maps for the SECOND dataset.

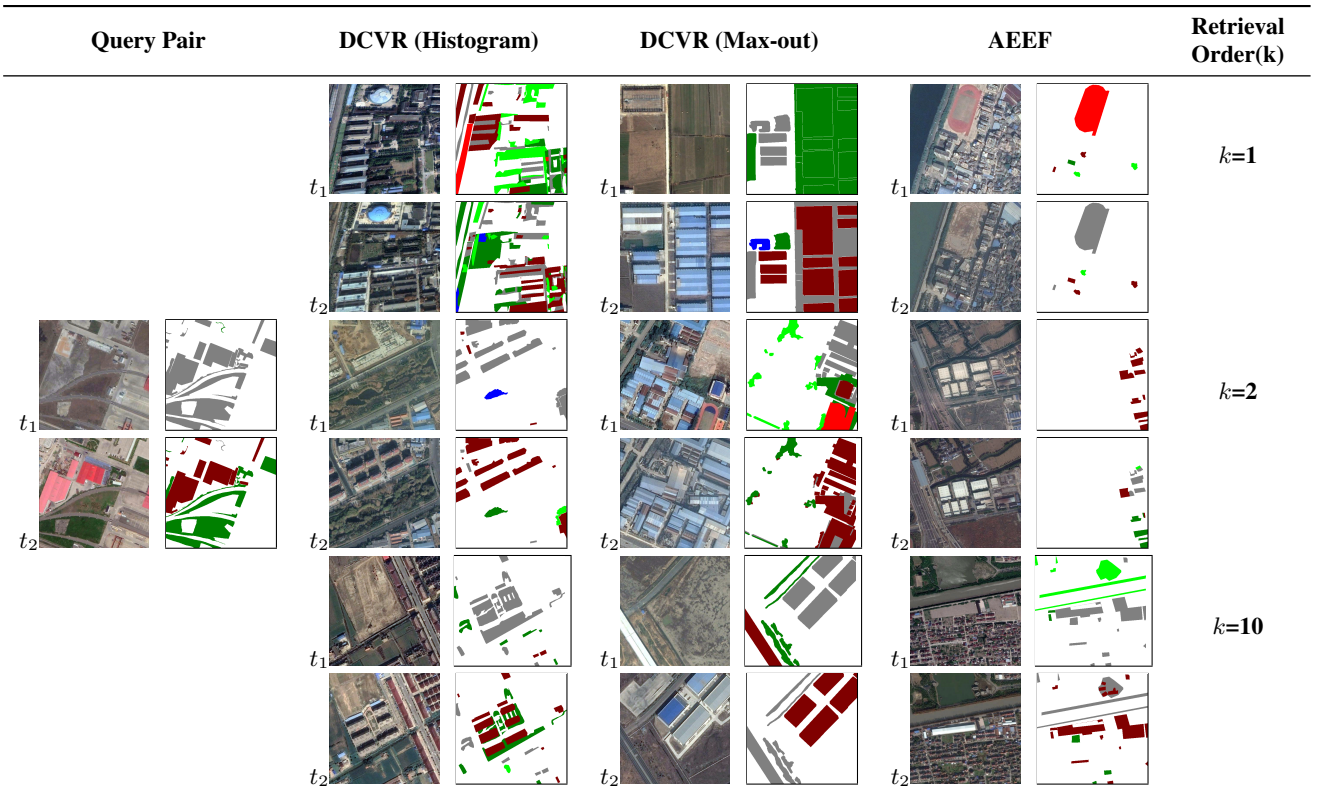

Fig. 3. Visual retrieval results in the order of retrieval obtained from DCVR method with Histogram and max-out feature selection strategies and AEEF method.

For change retrieval, only the latent space representations z^i's are generated and stored into the set Z such that $Z = \bigcup_{i=1}^{M_{\text{test}}} z^i$. To retrieve the most similar k image pairs R^i, k image pairs with the closest latent space representations to query image pair are selected.

3. EXPERIMENTAL RESULTS

In the experiments, we used the SECOND dataset [7] which includes 512x512-sized 2968 bi-temporal RGB aerial images (H, W, and d parameters are 512, 512, and 3, respectively) acquired over cities such as Hangzhou, Chengdu, and Shanghai in China. We randomly separated the dataset into training and test sets with an 80% - 20% ratio. Image pairs were already co-registered and experts labeled each pixel of X_1^i and X_2^i according to the land cover types. Six land cover classes have been annotated in the dataset including : 1) low vegetation (denoted as LV), 2) non-vegetated ground surface (denoted as NV), 3) tree (denoted as TR), 4) water (denoted as WT), 5) building (denoted as BL), and 5) playground (denoted as PG). Examples of images and their reference map are shown in Fig. 2, where white pixels correspond to unchanged areas and otherwise semantic labels are represented by the corresponding color. Since there were six types of labels, this resulted in 30 possible types of changes. Occurrence percentages of changes in the dataset are shown in Table 1, where land cover types indicated in the rows are changed to land cover types in the columns. In our experimental setup, we exclude the change types observed less than 40% because such low occurrence cases do not enable proper training of the AEEF network.

For the DCVR method, the number of layers L was chosen experimentally 3 as suggested in [5] and we select layers $l = 2, 5, 8$, to extract deep features. We used a pre-trained 33-layer CNN network proposed in [6]. The network is trained on an RS areal images. For the max-out and the histogram aggregation strategies, the choice of m and the number of bins (h) in the histogram construction were kept constant for each channel. We performed experiments for $m = 1, .., 6$ and $h = 1, .., 6$ and empirically picked lowest values where performance was satisfactory, i.e., $m = 5$ and $h = 6$.

Table 2. Mean average precision (mAP) scores associated to different change in land cover classes (LCC).

LCC Change	**DCVR** Histogram	**DCVR** Max-out	**AEEF**	**Occurrence Ratio (%)**
LV→ NV	0.51	**0.62**	0.56	61.8%
LV→ BL	0.39	**0.50**	0.49	51.0%
NV→ LV	0.57	0.59	**0.61**	58.4%
NV→ TR	**0.46**	0.43	0.34	41.2%
NV→ BL	**0.90**	0.84	0.78	85.2%
BL→ NV	**0.82**	0.67	0.60	66.7%
Average	**0.65**	0.64	0.59	-

For the AEEF method, the encoder network f_{enc} is composed of 16 convolutional layers and the decoder network f_{dec} consist of 16 deconvolutional layers. The latent space representation size d was chosen experimentally and set to 1024. The training was performed for 100 epochs with the Adam optimizer, using a learning rate of 0.001. Both hyperparameters λ_{ℓ_1} and λ_{ℓ_2} (weights for ℓ_1 and ℓ_2 distances) were treated equally and set to 1.

In order to evaluate both DL-based methods DCVR (both Histogram and Max-out feature selection strategies) and AEEF, we have tested the methods with the nearest-neighbor retrieval based upon cosine distance [8]. Since bi-temporal images in our dataset possess multiple change labels and since the differential between label abundances can be quite large, we measured the retrieval performance (precision) for each change type separately, as well as their weighted average. In detail, for a label λ, we denote all the queries in the test set with label λ as $\{Q_\lambda\}$. For R images retrieved per query, the mean average precision (mAP) is computed as :

$$mAP_\lambda = \frac{1}{Q_\lambda} \sum_{q_\lambda=1}^{Q_\lambda} \frac{1}{R} \sum_{r=1}^{R} s_{r_\lambda} \tag{5}$$

where $s_{r_\lambda} = 1$ if r has label λ, and 0 otherwise. We have adopted in all methods $R = 5$.

Fig. 3 compares cases of bi-temporal images retrieved by the DCVR method with histogram and max-out feature selection strategies and by the AEEF method when query images contain changes from "non-vegetated ground surface" to "low vegetation" and "building" to land cover class (NV→LV and NV→BL). The retrieval order of images is given in the rightmost column. One can observe that the change type NV→LV is successfully retrieved by DCVR in most cases, while AEEF can only retrieve them at lower orders, such as k=10. In some cases, AEEF has retrieved the opposite change type (LV→NV). When DCVR with histogram strategy is compared with max-out strategy, the former strategy retrieves semantically more similar images, which include both change types. One of the reasons is that the histogram can retain more of the semantic content in the deep representation, which can be more statistically representative than a simple max-out feature selection.

Table 2 shows the quantitative results of change retrieval obtained from DCVR (with histogram and max-out feature selection strategies) and AEEF methods. The first column shows the different change types, where each change type is represented as change in the land cover label (CLN). For example change from "low vegetation" land cover class (denoted as LV) to "non-vegetated ground surface" is indicated as LV→NV. The table shows the mean average precision (mAP) results computed by (5) over different types of land cover change, where the last row shows the weighted average of the label precisions. To compute the weighted average we used the relative occurrence frequencies as weights for each change type. The last column in Table 2 shows the occurrence percentages of the change type in the database images. By analysing the result in the Table 2, one can observe that the DCVR method (Histogram and Max-out feature selection strategies) perform slightly better than the AEEF method. As an example, DCVR with histogram feature selection strategy obtained 6% higher mAP score than AEEF. Furthermore, for the DCVR method, the histogram feature selection strategy performed almost comparably, though which strategy outperforms depends on the change type. As an example, for BL→NV change type the DCVR method with histogram strategy obtained 15% higher mAP than DCVR with max-out strategy, while for LV→BL obtained 11% lower mAP score than DCVR with max-out strategy.

We have observed that the DCVR method is practically insensitive to the choice of m (the number of ranked maxima chosen in the feature map). The performance, however, improves slightly with the increasing number of quantization steps h. Feature selection seems a hard problem because bi-temporal image pairs can end up with very high dimensional deep features. However, this dimension is reduced dramatically by selecting the m highest components of the deep change vector, or representing their statistics with few bin histograms. For the AEEF method we observe that there is some parallelism between its retrieval performance and the occurrence frequency of change type. The higher the frequency, the better the chances of capturing meaningful information for training (see Table 2). Furthermore, we surmised that the weaker performance of the AEEF method might be due to the semantic gap in the latent space code distances.

4. CONCLUSION

This paper has introduced two deep learning based methods (the DCVR and the AEEF) in the framework of CBTSR for abrupt change retrieval from large archives. The proposed methods aim to retrieve pairs of images in the archive that show the same kind of change associated with a query pair. The DCVR exploits the difference of features extracted from layers of CNNs for pairs of images, while AEEF extracts autoencoder codes from concatenated bitemporal images and exploits these latent space codes to synthesize the difference image. As a future work, we plan to adapt the the DCVR and the AEEF to be applicable for the retrieval of long time series of RS images.

5. ACKNOWLEDGMENT

This work is funded by the European Research Council (ERC) through the ERC-2017-STG BigEarth Project under Grant 759764.

6. REFERENCES

[1] F. Bovolo, B. Demir, and L. Bruzzone, "A cluster-based appraoch to content based time series retrieval," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2015, pp. 2793–2796.

[2] A. Julea, N. Méger, P. Bolon, C. Rigotti, M.-P Doin, C. Lasserre, E. Trouvé, and V. Lazarescu, "Unsupervised spatiotemporal mining of satellite image time series using grouped frequent sequential patterns," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, pp. 1417–1430, 04 2011.

[3] A. Radoi and C. Burileanu, "Query-by-example retrieval in satellite image time series," in *International Conference on Telecommunications and Signal Processing*, 2018, pp. 1–5.

[4] C. Ma, W. Xia, F. Chen, J. Liu, Q. Dai, L. Jiang, J. Duan, and W. Liu, "A content-based remote sensing image change information retrieval model," *ISPRS International Journal of Geo-Information*, vol. 6, no. 10, 2017.

[5] S. Saha, F. Bovolo, and L. Bruzzone, "Unsupervised deep change vector analysis for multiple-change detection in vhr images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 6, pp. 3677–3693, 2019.

[6] M. Volpi and D. Tuia, "Dense semantic labeling of subdecimeter resolution images with convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 2, pp. 881–893, 2017.

[7] K. Yang, G-S. Xia, Z. Liu, B. Du, W. Yang, M. Pelillo, and L. Zhang, "Asymmetric siamese networks for semantic change detection in aerial images," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–18, 2021.

[8] J. Friedman, J. Bentley, and R. Finkel, "An algorithm for finding best matches in logarithmic expected time," *ACM Trans. Math. Softw.*, vol. 3, pp. 209–226, 09 1977.