Learning Transfer Using Deep Convolutional Features for Remote Sensing Image Retrieval

Ahmad Alzu'bi, Abbes Amira, and Naeem Ramzan

***Abstract*— Convolutional neural networks (CNNs) have recently witnessed a notable interest due to their superior performance demonstrated in computer vision applications; including image retrieval. This paper introduces an optimized bilinear-CNN architecture applied in the context of remote sensing image retrieval, which investigates the capability of deep neural networks in learning transfer from general data to domain-specific application, i.e. remote sensing image retrieval. The proposed deep learning model involves two parallel feature extractors to formulate image representations from local patches at deep convolutional layers. The extracted features are approximated into low-dimensional features by a polynomial kernel projection. Each single geographic image is represented by a discriminating compact descriptor using a modified compact pooling scheme followed by feature normalization. An end-to-end deep learning is performed to generate the final fine-tuned network model. The model performance is evaluated on the standard UCMerced land-use/land-cover (LULC) dataset with high-resolution aerial imagery. The conducted experiments on the proposed model show high performance in extracting and learning complex image features, which affirms the superiority of deep bilinear features in the context of remote sensing image retrieval.**

***Index Terms*—image retrieval, remote sensing, deep learning**

I. Introduction

With the rapid advances of observation satellite technologies, a large amount of high-resolution remote sensing images is generated and become available. As a result, a demand is highly concentrated on providing an automatic and accurate representation of images to gain a timely access to their informative contents, e.g. spatial and spectral responses. Over the last decade, a dramatic change is witnessed in the use of remote sensing tools for satellite image analysis, indexing, retrieval, or even for broadcasting and forecasting purposes [1][2]. Initially, most of the existing approaches utilized a tag-based retrieval based on some keywords attached to image contents [3]. This technique is impractical due to the expensive manual annotation required as well as the potential inaccurate tags generated. Consequently, research efforts have recently been shifted to the content-based image retrieval (CBIR), i.e. using visual contents for image similarity matching.

The typical CBIR approaches consist of two essential processes: feature extraction and similarity/dissimilarity matching. Several global and local image features have been successfully applied to a range of computer vision applications; including geographic image retrieval. Among the best performing approaches is local invariant image features, e.g. scale-invariant feature transform (SIFT) [4], speeded-up robust features (SURF) [5], and histogram of oriented gradients (HOG) [6]. These features are invariant to many image deformations and robust against different viewpoints and illumination changes, which is important for the remote sensing images observed under such these conditions. However, low-level handcrafted features, i.e. local invariant features and global color and texture features, have been outperformed by deep features extracted from neural network architectures in computer vision applications. This motivated us to investigate the capability of deep neural networks for learning transfer from general purpose images to domain specific images such as remote sensing images. Additionally, the existing work in the context of remote sensing, based on deep learning, is very limited and mainly focused on remote sensing image classification [7].

Based on the promising performance obtained by convolutional neural networks (CNNs) in various computer vision tasks, we propose and apply a deep bilinear CNN architecture to investigate the capability of such deep learning models in extracting high discriminative image representations for remote sensing retrieval. This work is inspired by the effective bilinear CNN model applied to fine-grained image categorization [8]; however, their model forms high-dimensional image descriptors (~262K), which is unfavorable and impractical in terms of retrieval speed and memory requirements. Therefore, we employ a modified version of compact pooling, firstly introduced by Gao et al. [9], to generate low-dimensional image representations. The bilinear pooling is optimized to improve the retrieval accuracy using compact but high discriminative features. Specifically, any image features are directly extracted from the convolutional layers by two parallel CNN extractors then combined into a single image vector by inner product calculations. The resulting vectors of all geographic images are finally indexed into one dataset to be efficiently retrieved according to certain query images initiated during the retrieval process. The architecture performance is tested and evaluated on the standard UCMerced image dataset. This dataset is one of the largest publicly available datasets and it consists of high-resolution aerial imagery.

In Summary, this work contributes in three main aspects:

Manuscript received January 01, 2019; Revised August 30, 2019. This work was financially supported by the Middle East University, Amman, Jordan.

A. Alzu'bi is an assistant professor of computer science, Faculty of Information Technology, Middle East University, Amman, Jordan, e-mail: aalzoubi@meu.edu.jo.

A. Amira is a professor of computer engineering, Faculty of Computing, Engineering and Media, De Montfort University, Leicester, United Kingdom, e-mail: abbes.amira@dmu.ac.uk.

N. Ramzan is a professor of computer engineering, School of Computing, Engineering, and Physical Sciences, University of the West of Scotland, Paisley, United Kingdom, e-mail: naeem.ramzan@uws.ac.uk.

(1) We propose an improved bilinear CNN-based architecture for remote sensing image retrieval;
(2) We largely reduce the dimension of image descriptors generated by an optimized root-based bilinear pooling, which improves the performance in terms of retrieval speed and storage consumption;
(3) We draw an instructive conclusion on how neural networks could be able to transfer learning from generalized data to remote sensing domain.

The remaining part of this paper is organized as follows: Section II presents the related works in the domain of remote sensing image retrieval; Section III introduces the proposed architecture, image dataset, and evaluation protocol; Section IV discusses the experimental results; and Section V concludes this work.

II. Related work

Visual content-based image retrieval has been a challenging and active computer vision task since decades. Motivated by the promising results obtained using CBIR approaches in the context of remote sensing, a remarkable attention paid on utilizing them to provide an efficient and reliable access to the rich informative contents of high-resolution geographic images. Several works have applied different global low-level image features for remote sensing image retrieval; including spectral features [10], global texture features [11-13], shape features [3], and combined features [14-16]. However, local invariant features, e.g. SIFT and its variants, extracted from certain regions/patches have shown a better performance in the domain of remote image sensing, as their effectiveness has been initially investigated by Chen et al. [17] for scene classification using satellite images. In the domain of remote sensing image retrieval, the recent thorough investigation using local invariant features is introduced in [18], and they made the UCMerced LULC image dataset public.

Recently, Aptoula [19] has also applied a couple of global morphological texture descriptors, e.g. Fourier power spectrum. More recent works have also examined compound structures [20] for satellite image classification, and active learning in relevance feedback [21] for remote sensing image retrieval. Despite the successful use of handcrafted low-level image features, the performance of these adopted approaches is readily affected by several factors, e.g. geographical scene, sensor types, and acquisition environment. Other extensive experiments [22-27] are carried out to improve the retrieval accuracy of remote sensing systems including the combination of image features, relevance feedback, similarity metrics, and learning optimization. Though the existing approaches have tackled a plenty of application-level problems, they have a limited performance due to the complexity and content diversity of high-resolution remote sensing images [28][29].

One the other hand, deep learning models based on neural networks show a promising performance in many computer vision tasks. Several works [30-32] proved that CNN models sufficiently trained for computer vision tasks, e.g. image classification and recognition, can be extended to some domain-specific retrieval tasks such as remote sensing image retrieval. However, the learning transfer and adaptation to a target application are affected by several factors that limit the performance of deep learning retrieval methods [33]. This motivated us to investigate the effectiveness of deep CNN-based architectures in the domain of content-based remote sensing image retrieval. Penatti et al. [7] evaluated the generalization capability of deep features from general object images to aerial and remote sensing image classification. They have applied some pre-trained deep models along with several global and quantized image descriptors, e.g. BOW. Medjahed et al. [34] also applied binary search algorithms for remote sensing imagery and hyperspectral image classification. Although deep features generalize well in both aerial and remote sensing classification, they do not outperform some low-level color descriptors.

Recently, Xiong et al [35] proposed a multi-task learning network structure to extract learning-based features for remote sensing image search. Their CNN-based structure accumulates feature representations from convolutional layers to make them as dispersed inter-class and compact intra-class as possible. Abe et al [36] investigated a set of ensemble (Random forest (RF) and bagging) and non-ensemble (neural networks) classifiers for land cover classification (LCC) using the generalized reduced gradient approach on hyperspectral dataset.

In this paper, we investigate the effectiveness of deep convolutional features extracted by the bilinear CNN architecture we propose in the domain of content-based remote sensing retrieval. This work is largely distinguished from the aforementioned works in many directions.

Firstly, it performs and investigates deep bilinear CNNs for remote sensing image retrieval, which is different from the models introduced for aerial and remote sensing classification and retrieval. We perform a fine-tuning of deep CNNs on a remote sensing image dataset. Additionally, our bilinear CNN model generates very compact image representations using a low-dimensional space projection through an optimized bilinear pooling.

The resulting compact image descriptors are high discriminative at characterizing and recognizing complex geographic images observed under different viewpoints and illumination conditions. More critically, a low storage space is required to index images achieving higher retrieval speed. Moreover, the whole end- to-end training of the deep CNN model is conducted without any image tags/annotations.

III. The proposed framework

In this section, we present the proposed deep bilinear CNN model, the process of feature extraction, the UCMerced LULC image dataset, and the evaluation protocol of model performance.

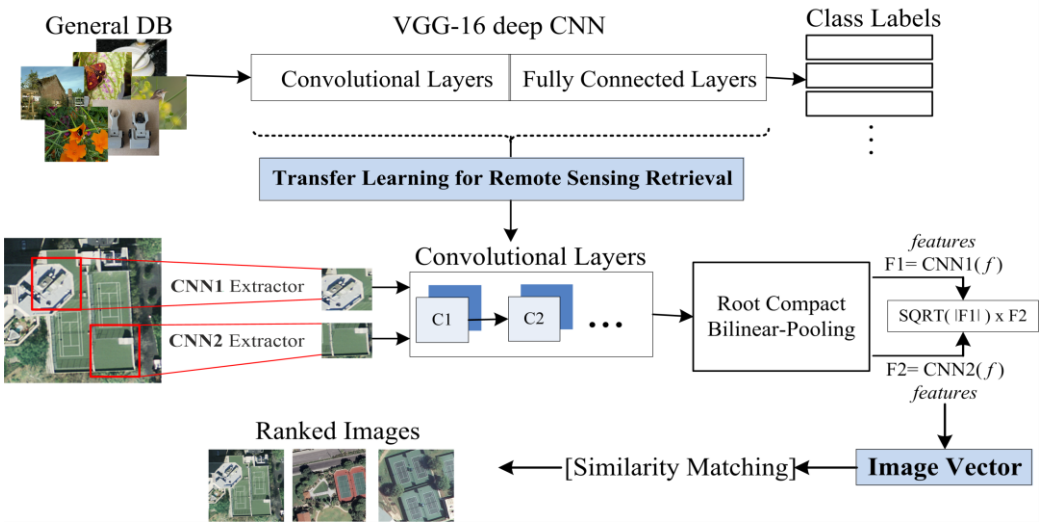

Fig. 1. A schematic representation of the proposed CNN architecture.

A. *Deep Bilinear CNN Architecture*

The proposed approach consists of three essential steps: (1) fine-tuning a pre-trained deep CNN model; (2) training the bilinear CNN architecture on the remote sensing image dataset, i.e. learning transfer from generic to specific domain; and (3) extracting image features using the trained model for image retrieval. As shown in Figure 1, the CNN architecture is based on one of the successful deep neural networks known as imagenet-vgg-verydeep-16 [37] and denoted here as VGG-16 for simplicity, which is sufficiently pre-trained on millions of general-purpose images. The VGG-16 network consists of 35 convolutional, pooling, and fully connected layers. The input layer takes images of size 224×224 pixels. To simplify the experiments, the fully connected layers are discarded from the fine-tuned bilinear CNN architecture so that all image features are extracted from the activations of convolutional feature maps.

Firstly, the original CNNs are truncated at the last convolutional layer in the network, i.e. layer 30 where the output feature size is 512. Secondly, three additional layers are added to the end of the resulting CNN architecture as follows: (1) root compact bilinear pooling to project the data into a compact size N; (2) signed square-root layer; and (3) L2 normalization. The low-dimensional bilinear features extracted from this network are formed into a single generic image descriptor (i.e. vector) using the inner product between every two descriptors obtained by the two parallel CNN extractors. Finally, the trained model is used to extract the features of all queries and dataset images.

B. *Feature Extraction*

Given the VGG-16 CNN network with 35 layers, an input image I is wrapped into 224×224 square to fit the size of training images and then passed through the network in a forward pass of E epochs after applying the filters to the input image. In the last i-th convolutional layer, i.e. C_{30}, we obtain 512 output size followed by ReLU layer. However, the bilinear architecture first introduced by Lin [6] has a high dimensional output size at bilinear pooling layer where each single CNN generates features of size d×512 so that the output size of pooling layer formulated by the outer product of two bilinear CNNs generates 512×512 image descriptor, i.e. 262,144 vector size. This unfavorable size of image descriptor is unwieldy in the context of remote sensing image retrieval where indexing complexity, retrieval speed, and memory size are critically considered.

As a result, a low-dimensional projection on the extracted features is applied using a modified compact pooling based on [9][38]. Additionally, our proposed model computes the square root of the extracted features for only one of two bilinear extractors to break the potential symmetry property between them. Specifically, given two sets of local features $X = \{x_1, \ldots, x_{|SP|}, x_{sp} \in R^c\}$ from image I_1 and $Y = \{y_1, \ldots, y_{|SP|}, y_{sp} \in R^c\}$ from image I_2. All features are extracted using the last convolutional layer of the CNN network, where SP is a set of spatial locations. An image vector is formed into $(c \times c)$ using our modified root bilinear pooling as follows:

$$RB(X) = \sum_{sp \in SP} r(x_{sp}) x_{sp}^T \tag{1}$$

where $r(x_{sp}) = sqrt(|x_{sp}|)$.

The kernelized version of bilinear pooling is then applied to compare X and Y of two images using the second order polynomial kernel as follows:

$$\langle RB(X), RB(Y) \rangle = \Big\langle \sum_{sp \in SP} r(x_{sp}) x_{sp}^T, \sum_{sp \in SP} r(y_{sp}) y_{sp}^T \Big\rangle$$

$$= \sum_{sp \in SP} \sum_{sp \in SP} \langle r(x_{sp}), y_{sp} \rangle^2 \tag{2}$$

Then, any low-dimensional projection function applied to approximate image features into $\phi(x) \in R^{dim}$ and $\phi(y) \in R^{dim}$, where $dim << c^2$, by calculating:

$$\langle RB(X), RB(Y) \rangle \equiv \langle RB(X)_{compact}, RB(Y)_{compact} \rangle$$

$$\approx \sum_{sp \in SP} \sum_{sp \in SP} \langle \phi(x_{sp}), \phi(y_{sp}) \rangle \tag{3}$$

Our modified compact pooling layer employs the random Maclaurin (RM) [9] as a low-dimensional projection function. The resulting image descriptor, of size 16 to 512 dimensions, is then passed to the next layers, i.e. signed

Fig. 2. Representative samples of UCMerced dataset.

square root and *L2* normalization, as shown in Figure 1. The final architecture is used to extract the image representations of all queries and dataset images to compute the distance scores, ranking, and retrieval accuracy.

C. Image Dataset

The UCMerced ground truth dataset is used to evaluate the retrieval performance, which is the largest data set of its kind made publicly available. It consists of 2100 high-resolution images of 21 LULC classes selected from aerial orthoimagery with a pixel resolution of 30 cm. The dataset images are downloaded from the USGS National Map of different 20 US regions [16]. Each geographic category contains 100 images of size 256×256 pixels. A sample of each category is shown in Figure 2.

D. Performance Evaluation

The retrieval accuracy of the proposed model is evaluated using several standard measures. Firstly, average normalized modified retrieval rank (ANMRR) is computed and compared with related works in the field of remote sensing image retrieval. A range of standard dissimilarity measures is evaluated under different vector lengths. Secondly, the average precision (AP) at position k is computed and compared with the reported results recent state-of-the-art approaches. Finally, the mean average precision (mAP) is computed for the overall image dataset as a standard measure used in the domain of image retrieval.

The ANMRR measure is commonly used to evaluate the MPEG-7 retrieval performance, but it has become widely accepted and used in the CBIR domain. It considers both the relevancy and ranking of retrieved images to the query image, and it also addresses the problem of having different number of ground-truth images. In this work we follow the standard ANMRR definitions as introduced in [18][19] for fair comparisons. ANMRR is calculated as follows: given a query image q, $NG(q)$ is the size of the ground truth images and let the k^{th} ground-truth image is retrieved at the position $Rank(k)$. Then, the image ranks that considered feasible in terms of retrieval are denoted as $K(q)$, which is a number given a value of twice $NG(q)$ so that images with a higher rank are assigned a constant penalty as follows:

$$Rank(k) = \begin{cases} Rank(k), & if \quad Rank(k) \le K(q) \\ 1.25K(q), & if \quad Rank(k) > K(q) \end{cases} \tag{4}$$

Then the average rank of query AVR(q) is defined as:

$$AVR(q) = \frac{1}{NG(q)} \sum_{k=1}^{NG(k)} Rank(k) \tag{5}$$

To overcome the impact of having different number of ground truth to the query image, the normalized score NMRR is computed and averaged over all query images NQ to calculate ANMRR as follows:

$$ANMRR = \frac{1}{NQ} \sum_{q=1}^{NQ} \frac{AVR(q) - 0.5(1 + NG(q))}{1.25K(q) - 0.5(1 + NG(q))} \tag{6}$$

The score value of ANMRR ranges between zero (i.e. all ground truth images are found and retrieved) and one (i.e. no relevant images are retrieved). Accordingly, lower values of ANMRR indicate better retrieval accuracy.

IV. Experiments and Results

In all experiments, a fine-tuning procedure is performed on the training dataset using a number of epochs. Once the final deep CNN model is generated, it is used to extract and index image representations of both queries and images in UCMerced dataset for retrieval task.

A. End-to-End Training Setup

End-to-end deep learning is applied using the stochastic gradient descent (SGD) optimization algorithm that estimates the error gradient for the current state of the proposed model using the training dataset, then it updates the model weights using the error backpropagation.

The challenging task of training deep CNN models involves hyperparameter tuning. Selecting the learning rate hyperparameter is very important since it controls how quickly the model is adapted to the specific problem; i.e. remote sensing image retrieval. In our model, the learning rate is carefully configured and set to 0.001. Specifically, it controls the change rate of model weights in response to the estimated error. Figure 3 depicts the relationship between

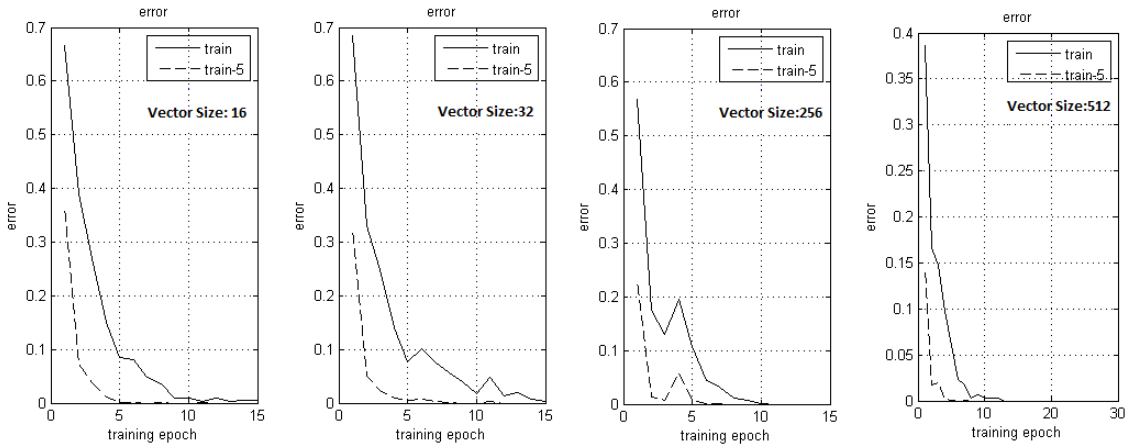

Fig. 3. Model training error against number of epochs.

error change and the number of epochs needed for our model to converge. As shown, small vector sizes of 16 and 32 need around 15 epochs to converge even with a small learning rate, where this takes less number of epochs (around 10) while forming larger vector sizes; i.e. 256 and 512. This confirms that the model is quickly adapting in learning an optimal set of weights and following a stable training process.

B. Retrieval Process

After generating the final trained model, the dataset images are indexed into one archive, i.e. database of image vectors. To record distance scores, every single vector of selected query images is compared against all database vectors using Euclidean (*L2*) distance measure. We evaluate the retrieval performance using 2100 queries, i.e. every image in the dataset is used as a query. Finally, all images are ranked according to their obtained scores. Figure 4 shows a sample of top-20 retrieval results ranked from left-to-right and top-down.

Table 1 also shows some performance results obtained by the model in terms of time elapsed to index any image and to search the whole database for query matching as well as the average memory size of images. The dimensionality of final image descriptors is reduced using the root compact bilinear pooling to a range of compact sizes: 16, 32, 64, 128, 256, and 512. As shown, a low retrieval time is achieved by reporting about 250*ms* in average to index and find the most relevant images of any submitted query. Moreover, a large save on memory space is acquired by generating compact vectors, which is beneficial for large-scale image repositories. For example, only 125KB of the actual disk storage is needed to index the whole dataset (i.e. 2100 images) when 16-vector size is used for image representation.

TABLE 1
THE AVERAGE PERFORMANCE RESULTS ON UCMERCED DATASET

Vector Size	**512**	**256**	**128**	**64**	**32**	**16**
Image vectorization (Time in millisecond)	224	223	223	223	220	220
Query search (Time in millisecond)	45	33	33	25	24	17
Indexed Image Size (Memory in KB)	1.80	0.90	0.45	0.20	0.12	0.06

C. Retrieval Results and Discussion

In addition to the experimental setup illustrated in the previous section, selecting a proper similarity measure

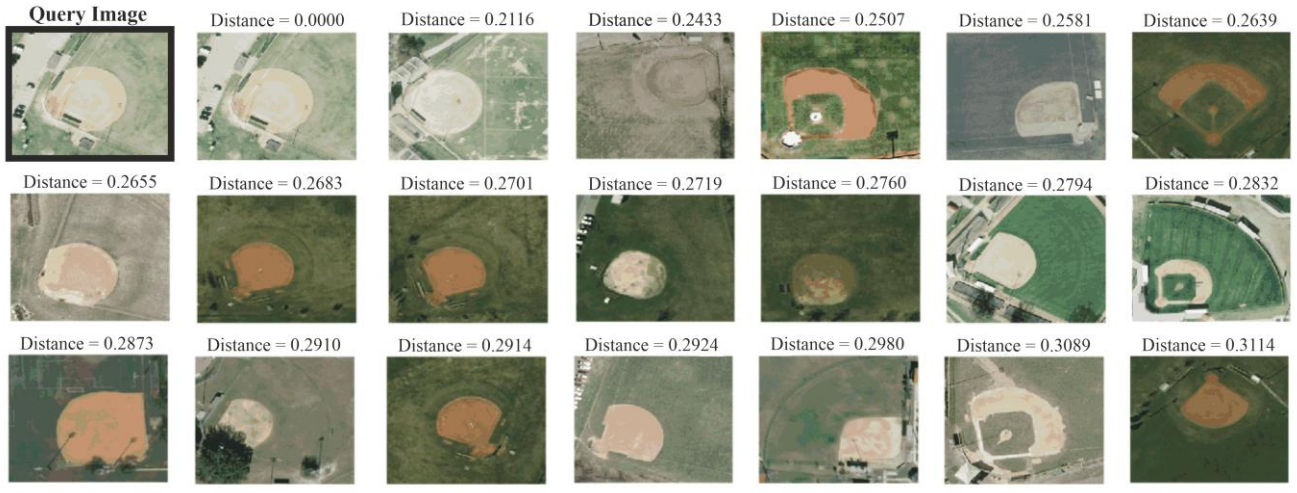

Fig. 4. Top-20 image ranking using 16-vector size. Images ordered left-right and top-down.

TABLE 2
ANMRR SCORES OBTAINED BY DIFFERENT DISSIMILARITY METRICS

Dissimilarity	16	32	64	128
L1	0.4103	0.1557	0.1177	0.1090
L2	0.0430	0.0632	0.1111	0.1043
Chebychev	0.0717	0.2405	0.1501	0.1268
Cosine	0.0430	0.0632	0.1111	0.1043
Correlation	0.0462	0.0653	0.1118	0.1046
CityBlock	0.0510	0.0659	0.1118	0.1060
Hamming	0.9687	0.9687	0.9687	0.9687
Spearman	0.0511	0.0728	0.1164	0.1046
Mahalanobis	0.0941	0.2405	0.4454	0.6721

between queries and indexed dataset is crucial for retrieval performance. A remote sensing retrieval system should employ effective matching metrics that accurately quantify and characterize the underlying perceptual similarities. A range of standard dissimilarity metrics is evaluated here on several sizes of image vectors, i.e. 16, 32, 64, and 128. Table 2 summarizes the ANMRR retrieval accuracy.

As shown in Table 2, it's easy to notice the performance disparities of selected metrics in term of accuracy. The Euclidean L2 is the best performing distance metric between others over all vector dimensions. In contrast to L2, Hamming distance metric is not able to characterize features properly as seen from the poor ANMRR scores obtained by this metric. It is worthy of attention that the other metrics provide better and comparable ANMRR scores with L2 on large image vector dimensions, e.g. 128, compared to the smaller ones, e.g. 16.

Figure 5 shows how the retrieval system performs under different matching distances using vector size 16. The top 5 relevant images are ranked according to the computed dissimilarity score that's shown above every retrieved image. For instance, Dist=0.000 means no differences between the retrieved image and the query image while this score is increasing in the following ranked images. This thorough evaluation of dissimilarity metrics made L2 the best choice for our remote sensing retrieval system.

Different ranking algorithms including rank learning approaches can be adopted in order to get more accurate image ranking. Additionally, relevance feedback methods can be applied as a postprocessing step either by getting the user involved in the retrieval process or by making the feedback an automatic procedure. However, relevant image ranking and excluding false retrieved images is beyond of this work focus.

TABLE 3
ANMRR SCORES OBTAINED FOR EACH IMAGE CATEGORY

Category	16	32	64	128	256	512
Agricultural	0.023	0.017	0.017	0.044	0.017	0.003
Airplane	0.000	0.015	0.014	0.023	0.006	0.021
Baseball	0.011	0.003	0.057	0.023	0.041	0.028
Beach	0.002	0.001	0.047	0.000	0.001	0.005
Buildings	0.059	0.146	0.172	0.229	0.259	0.209
Chaparral	0.000	0.000	0.006	0.000	0.000	0.000
Dense resid.	0.120	0.106	0.278	0.329	0.380	0.487
Forest	0.100	0.002	0.025	0.001	0.000	0.001
Freeway	0.007	0.027	0.096	0.029	0.249	0.244
Golf course	0.149	0.103	0.222	0.239	0.199	0.340
Harbor	0.000	0.004	0.027	0.006	0.022	0.010
Intersection	0.099	0.157	0.426	0.277	0.211	0.323
Med. Resid.	0.052	0.171	0.290	0.203	0.224	0.201
Mobile home	0.003	0.022	0.085	0.092	0.017	0.075
Overpass	0.010	0.070	0.064	0.032	0.210	0.212
Parking lot	0.000	0.010	0.007	0.008	0.001	0.002
River	0.070	0.117	0.088	0.148	0.068	0.089
Runway	0.001	0.001	0.072	0.046	0.020	0.017
Sparse resid.	0.168	0.291	0.250	0.277	0.242	0.315
Storage tanks	0.022	0.039	0.072	0.146	0.082	0.123
Tennis courts	0.000	0.002	0.014	0.031	0.002	0.000

Table 3 shows the ANMRR retrieval results for 21 image categories using different descriptor's dimensions. As shown, the bilinear CNN model performs efficiently over all image categories and achieves a high retrieval accuracy with respect to the top relevant images. Several image categories have been identified and retrieved with high accuracy such as beach, chaparral, harbor, parking lot, and tennis courts. It is also obvious that some image types are retrieved more accurately on large vectors, e.g. Agricultural images, while others performs better on small vectors, e.g. airplane images.

The ANMRR scores for each vector size are also computed and compared to several related works. As shown in Table 4, our deep bilinear CNN model largely

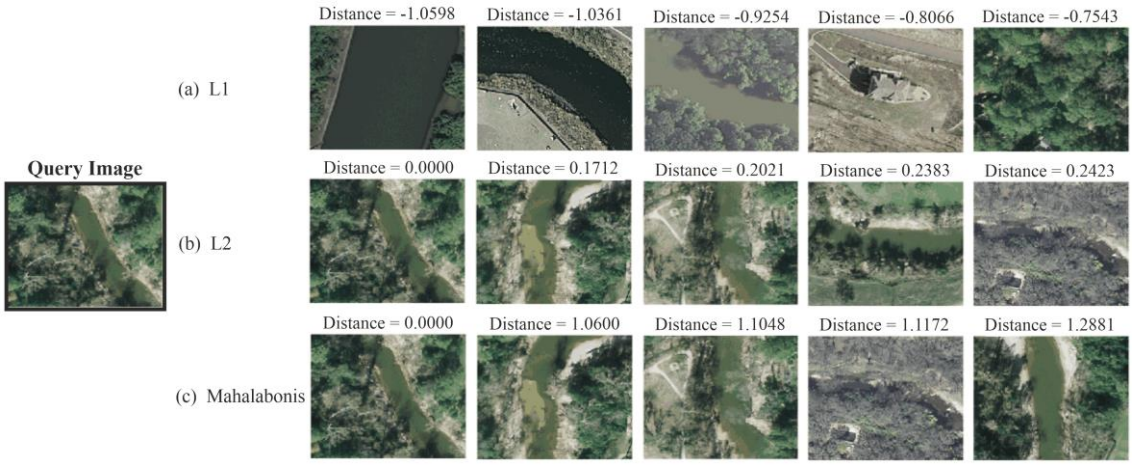

Fig. 5. Top-5 image retrieval results using different dissimilarity metrics: (a) L1 distance, (b) L2 distance, (c) Mahalanobis distance.

TABLE 4
COMPARISON OF ANMRR SCORES WITH RELATED WORKS

Zhou [39]	ReCNN+					
ANMRR	0.264					
Yang [18]	v-150	v-15000				
ANMRR	0.601	0.591				
Apto. [19]	62-d_M	62-d_B	62-d_{χ^2}			
ANMRR	0.589	0.586	0.575			
Xiong [35]	Vgg16_CL		Vgg19_CL		Res50_ CL	
ANMRR	0.081		0.114		0.089	
Proposed	16	32	64	128	256	512
ANMRR	**0.043**	**0.063**	0.111	0.104	0.107	0.129

outperforms the best scores reported using convolutional features [35][39], local invariant features [18], and morphological texture features [19]. The noticeable results obtained by the compact image descriptors, i.e. 16, 32, 64, emphasizing the efficiency of this model in preserving high discrimination levels and keeping the most characterized data of images even with low-dimensional image representations. To our knowledge, these are among the smallest vector used for retrieval by CNN-based models in the domain of remote sensing image retrieval.

In addition to ANMRR, two standard evaluation measures are also computed: average precision (AP) and mean AP (*m*AP). The reason is that sometimes ANMRR can be misleading. AP retrieval accuracy is calculated for 630 queries using 30 images randomly selected from every image category. AP is reported for the top-20 ranked images. The AP scores are also averaged over all categories to find the *m*AP on the whole image dataset.

Table 5 shows the AP and *m*AP accuracy scores of our proposed model compared to some state-of-the-art approaches for all image categories. The accuracy scores in [21] and [40] are reported from a column chart except some categories for which AP scores are specified. Therefore, there is a margin of ±0.5 AP approximation. However, *m*AP exact scores are specified in [40] and this validates the approximation of their AP scores.

Obviously, our deep CNN model outperforms other approaches for all image categories. It also achieves higher retrieval accuracy than others works in some challenging categories such as Dense Residential (ours 94.7% and others 41.0%, 84.5%, 80.0%). We also achieve 99.03% *m*AP score. This retrieval accuracy outperforms the best *m*AP results recently achieved by many state-of-the-art approaches as shown in Figure 6.

In terms of memory storage, only 125KB, 247KB, 488KB, 973KB, 1.5MB, and 3.89MB are needed to store all indexed dataset images using a vector size of 16, 32, 64, 128, 256, and 512, respectively. The low-dimensional image representations, generated by the proposed model, are necessary for large-scale image retrieval and processing systems. The nature of rich data embedded in remote sensed or satellite images usually need to be processed and handled in real-time systems and environments. For instance, it takes only 58.6MB to index and store one million of satellite images on the actual disk by our compact CNN model.

TABLE 5
COMPARISON OF (AP) AND (MAP) OVER 21 IMAGE CATEGORIES

Category	TCAL [21]	ConvNet [40]	CapsNet [41]	Proposed
Agricultural	99.23	99.00	100.0	100.0
Airplane	72.00	100.0	96.00	100.0
Baseball	30.00	97.00	98.00	100.0
Beach	82.00	100.0	100.0	100.0
Buildings	62.00	94.00	98.00	95.50
Chaparral	99.50	100.0	100.0	100.0
Dense resid.	41.00	84.50	80.00	94.67
Forest	97.33	99.00	100.0	100.0
Freeway	87.00	97.00	98.00	100.0
Golf course	40.00	98.00	100.0	99.50
Harbor	99.50	100.0	100.0	100.0
Intersection	74.00	96.50	98.00	98.33
Med. Resid.	58.00	92.00	100.0	95.50
Mobile home	96.00	96.00	100.0	100.0
Overpass	58.00	97.50	98.00	99.67
Parking lot	99.50	100.0	100.0	100.0
River	55.00	98.00	98.00	99.67
Runway	95.00	98.50	100.0	100.0
Sparse resid.	33.00	99.00	94.00	96.67
Storage tanks	25.00	95.00	96.00	100.0
Tennis courts	38.00	97.50	98.00	100.0
***m*AP**	**68.62**	**97.10**	**97.71**	**99.02**

V. CONCLUSION

This paper introduced a deep bilinear CNN architecture applied for content-based remote sensing image retrieval. The proposed model generates compact but high discriminative image representations using deep convolutional layers. It has shown its high ability to learn and discriminate remote sensing images based on the visual contents only. The small size of resulting features is an advantage in terms of retrieval time and memory storage required for rich informative geographic images. It has shown to outperform all best scores achieved by low-level image features with a noticeable improvement using our deep learning architecture applied on a standard dataset. In future, data augmentation and fully connected layers with quantization methods, e.g. BOW, VLADs and Fisher vectors, can be considered.

Fig. 6. *m*AP comparisons with related works on UCMerced dataset.

ACKNOWLEDGMENT

We gratefully acknowledge the Middle East University, Jordan, for their support for this research work. We also acknowledge the NVIDIA for their generous donation of the GPU used in this research.

REFERENCES

[1] Zhong Z. and Pi D., "Forecasting Satellite Attitude Volatility Using Support Vector Regression with Particle Swarm Optimization," *IAENG International Journal of Computer Science*, vol. 41, no. 3, pp. 153-162, 2014.

[2] Hagag A., Fan X., and El-Samie F.E.A., "Satellite Images Broadcast Based on Wireless SoftCast Scheme." *IAENG International Journal of Computer Science*, vol. 44, no. 1, pp. 1-7, 2017.

[3] Scott G. J., Klaric M. N., Davis C. H., and Shyu C. R., "Entropy-balanced bitmap tree for shape-based object retrieval from large-scale satellite imagery databases." *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 5, pp. 1603-1616, 2011.

[4] Lowe D. G., "Distinctive image features from scale-invariant keypoints." *International journal of computer vision*, vol. 60, no. 2, pp. 91-110, 2004.

[5] Bay H., Ess A., Tuytelaars T., and Van Gool L., "Speeded-up robust features (SURF)." *Computer vision and image understanding*, vol. 110, no. 3, pp. 346-359, 2014.

[6] Dala N. and Triggs B., "Histograms of oriented gradients for human detection." *In CVPR*, pp. 886-893, 2008.

[7] Penatti O., Nogueira K., and Santos J., "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?" *In Proceedings of CVPR Workshops*, pp. 44-51, 2016.

[8] Lin T.Y., RoyChowdhury A., and Maji S., "Bilinear CNN models for fine-grained visual recognition." *In Proceedings of the IEEE ICCV*, pp. 1449-1457, 2015.

[9] Gao Y., Beijbom O., Zhang N., and Darrell T., "Compact bilinear pooling." *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 317-326, 2016.

[10] Bretschneider T., Cavet R. and Kao O., "Retrieval of remotely sensed imagery using spectral information content." *In Geoscience and Remote Sensing Symposium*, no. 4, pp. 2253-2255, 2002.

[11] Ferecatu M. and Boujemaa N., "Interactive remote-sensing image retrieval using active relevance feedback." *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 4, pp. 818-826, 2007.

[12] Li Y. and Bretschneider T., "Semantics-based satellite image retrieval using low-level features." *In Geoscience and Remote Sensing Symposium*, no. 7, pp. 4406-4409, 2004.

[13] Hongyu Y., Bicheng L., and Wen C., "Remote sensing imagery retrieval based-on Gabor texture feature classification." *In 7th International Conference on Signal Processing*, pp. 733-736, 2004.

[14] Xu S., Fang T., Li D. and Wang S., "Object classification of aerial images with bag-of-visual words." *IEEE Geoscience and Remote Sensing Letters*, vol. 7, no. 2, pp. 366-370, 2010.

[15] Gleason S., Ferrell R., Cheriyadat A., Vatsavai R. and De S., "Semantic information extraction from multispectral geospatial imagery via a flexible framework." *IEEE Geoscience and Remote Sensing Symposium*, pp. 166-169, 2010.

[16] Liu T., Zhang L., Li P., and Lin H., "Remotely sensed image retrieval based on region-level semantic mining." *EURASIP Journal on Image and Video Processing*, no. 1, pp. 1-11, 2012.

[17] Chen L., Yang W., Xu K., and Xu T., "Evaluation of local features for scene classification using vhr satellite images." *In Urban Remote Sensing Event (JURSE)*, pp. 385-388, 2011.

[18] Yang Y. and Newsam S., "Geographic image retrieval using local invariant features." *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 2, pp. 818-832, 2013.

[19] Aptoula E., "Remote sensing image retrieval with global morphological texture descriptors." *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 5, pp.3023-3034, 2014.

[20] Gueguen, L., "Classifying compound structures in satellite images: A compressed representation for fast queries." *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 4, pp.1803-1818, 2015.

[21] Demir B. and Bruzzone L., "A novel active learning method in relevance feedback for content-based remote sensing image retrieval." *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 5, pp. 2323-2334, 2015.

[22] Chaudhuri B., Demir B., Chaudhuri S., and Bruzzone L., "Multilabel Remote Sensing Image Retrieval Using a Semisupervised Graph-Theoretic Method." *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 2, pp. 1144-1158, 2018.

[23] Shao Z. F., Zhou W. X., and Cheng Q. M., "Remote sensing image retrieval with combined features of salient region." *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 40, no. 6, pp. 83, 2014.

[24] Sebai H., Kourgli A., and Serir A., "Dual-tree complex wavelet transform applied on color descriptors for remote-sensed images retrieval." *Journal of Applied Remote Sensing*, vol. 9, no. 1, pp. 095994, 2015.

[25] Bao Q., and Guo P., "Comparative studies on similarity measures for remote sensing image retrieval." *IEEE International Conference on Systems, Man and Cybernetics*, no.1, pp. 1112-1116, 2004.

[26] Gueguen L., and Datcu M., "A similarity metric for retrieval of compressed objects: Application for mining satellite image time series." *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 4, pp. 562-575, 2008.

[27] Samal A., Bhatia S., Vadlamani P., and Marx D., "Searching satellite imagery with integrated measures." *Pattern Recognition*, vol. 42, no. 11, pp. 2502-2513, 2009.

[28] Zhou W., Newsam S., Li C., and Shao Z., "Patternnet: a benchmark dataset for performance evaluation of remote sensing image retrieval." *ISPRS Journal of Photogrammetry and Remote Sensing*, no. 145, pp. 197-209, 2018.

[29] Xia G. S., Tong X. Y., Hu F., Zhong Y., Datcu M., and Zhang L., "Exploiting Deep Features for Remote Sensing Image Retrieval: A Systematic Investigation." *arXiv preprint arXiv*: 1707.07321, 2017.

[30] Li Y., Zhang Y., Huang X., Zhu H., and Ma J., "Large-scale remote sensing image retrieval by deep hashing neural networks." *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 2, pp. 950-965, 2018.

[31] Zhou W., Newsam S., Li C., and Shao Z., "Learning low dimensional convolutional neural networks for high-resolution remote sensing image retrieval." *Remote Sensing*, vol. 9, no. 5, pp. 489-508, 2017.

[32] Jiang T. B., Xia G. S., Lu Q. K., and Shen W. M., "Retrieving aerial scene images with learned deep image-sketch features." *Journal of Computer Science and Technology*, vol. 32, no. 4, pp.726-737, 2017.

[33] Bunrit S., Chanklan R., Boonamnuay S., Kerdprasop N. and Kerdprasop K., "Neural network-based analysis of precipitation and remotely sensed data." Lecture Notes in Engineering and Computer Science: Proceedings of The International MultiConference of Engineers and Computer Scientists 2016, IMECS 2016, 16-18 March, 2016, Hong Kong, pp. 40-45.

[34] Medjahed S.A., Saadi T.A., Benyettou A. and Ouali M., "Binary cuckoo search algorithm for band selection in hyperspectral image classification." *IAENG International Journal of Computer Science*, vol. 42, no. 3, pp. 183-191, 2015.

[35] Xiong W., Lv Y., Cui Y., Zhang X. and Gu X., "A Discriminative Feature Learning Approach for Remote Sensing Image Retrieval." *Remote Sensing*, vol. 11, no. 3, pp. 281, 2019.

[36] Abe B.T., Olugbara O.O., and Marwala T., "Hyperspectral Image Classification using Random Forests and Neural Networks." Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2012, WCECS 2012, 24-26 October, 2012, San Francisco, USA, pp. 522-527.

[37] Chatfield K., Simonyan K., Vedaldi A., and Zisserman A., "Return of the devil in the details: Delving deep into convolutional nets." *arXiv preprint arXiv*: 1405.3531, 2014.

[38] Alzu'bi A., Amira A. and Ramzan N., "Compact root bilinear cnns for content-based image retrieval." *In Proceedings of IEEE International Conference on Image, Vision and Computing*, pp. 41-45, 2016.

[39] Zhou W., Deng X., and Shao Z., "Region Convolutional Features for Multi-Label Remote Sensing Image Retrieval." arXiv *preprint arXiv*: 1807.08634, 2018.

[40] Castelluccio, M., Poggi, G., Sansone, C. and Verdoliva, L., "Land use classification in remote sensing images by convolutional neural networks." *arXiv preprint arXiv*:1508.00092, 2015.

[41] Zhang, W., Tang, P. and Zhao, L., "Remote Sensing Image Scene Classification Using CNN-CapsNet." *Remote Sensing*, vol. 11, no. 5, p.494, 2019.