REMOTE SENSING IMAGE RETRIEVAL USING CONVOLUTIONAL NEURAL NETWORK FEATURES AND WEIGHTED DISTANCE

Rose Mariya Abraham*1, Treesa Thomas*2, Chippy Babu*3, Dinu Mol Philip*4

*1,2,3 MG University, Department of Computer Science, De Paul Institute of Science & Technology, Angamaly, Kerala, India
*4 Asst. Professor, Department of Computer Science, De Paul Institute of Science & Technology, Angamaly, Kerala, India

ABSTRACT

Remote sensing image retrieval (RSIR) could also be a basic task in remote sensing. Most content-based RSIR approaches take an easy distance as similarity criteria. A retrieval technique supported weighted distance and basic options of convolutional neural network (CNN) are planned throughout this letter. This contains 2 stages. First, in offline stage, the pre-trained CNN is fine-tuned by some tagged pictures from the target information set, then accustomed extract CNN options, and tagged the pictures inside the retrieval information set. Second, in on-line stage, we tend to use the fine-tuned CNN model to extract the CNN feature of the question image and calculate the burden of each image category and apply them to calculate the gap between the question image and thus the retrieved pictures. Experiments square measure conducted on 2 RSIR information sets. Compared with the state-of-the-art ways, the planned technique is simplified however economical, considerably rising retrieval performance.

Keywords: Convolutional Neural Network (CNN), Remote Sensing Image Retrieval (RSIR), weighted distance.

I. INTRODUCTION

Remote sensing images become very important in the fields like in geological analysis, urban planning, natural disaster monitoring and assessment, weather prediction, resource investigation, and so on. How to automatically and efficiently retrieve the remote sensing images that users need from large image databases becomes one of the challenging and emerging research topics in the field of remote sensing. Content based image retrieval (CBIR) is a useful method to solve more problems based on the features of image, such as intensity, shape, texture, and structure. Content based RSIR (CBRSIR) is an active and challenging research topic that has attracted the attention of researchers around the world.

During the years, an explained literature has been developed on high resolution remote sensing (HRRS) image retrieval problem. Many of the researchers have applied different forms of descriptors on remote sensing image retrieval purpose, and these descriptors can be classified into two main categories. These include local features extracted at interest point locations, such as intensity features, spectral features, shape features, structural features, texture features, etc. and also global features, obtained by encoding local features into a single vector representation using bag-of-words (BoW), vector locally aggregated descriptors (VLAD), or their variants.

Remote sensing community has applied deep learning techniques to extract high level semantic features for CBRSIR. To overcome the existing problems of HRRS image retrieval methods, we implement deep features from Convolutional Neural Networks (CNNs) to displace conventional representations. CNNs proved to have a very powerful feature extractor. Deep features are dominant in solving various visual problems, such as image classification, object detection, fine-grained recognition, and image instance retrieval.

This paper uses a weighted distance and CNN features to retrieve image. Here we fine-tune the pre-trained CNN model to calculate the weight of each class in the retrieved data set for the query image. We also give more importance to the retrieved images in more similar classes with the query image. The framework of our proposed approach in remote sensing image retrieval is shown in Fig. 1.
II. LITERATURE REVIEW

This paper [1] the author describes the performance of various image illustration schemes used for image search problems for the aim of remote sensing image retrieval. Here it compare the foremost wide adopted technique of the bag-of-words (BoW) approach with the additional recently introduced vector of domestically mass descriptors (VLAD). It show with the experiments on a in public accessible 21-class land-use/land-cover information set that the VLAD-based illustration outperforms BoW at the worth of redoubled question time, but the additional compact VLAD-PQ illustration achieves terribly similar performance as VLAD while not the redoubled time demand.

In this paper [2] the author represents native invariant options for geographic image retrieval. Native invariant options are with success applied to a broad vary of pc vision issues and it helps for detection and classification. Here during this paper it performs a comprehensive analysis of native invariant options for image retrieval of land-use/land-cover (LULC) categories in high-resolution aerial imaging. The author reports on the results of a spread of style parameters on a bag-of-visual-words (BOVW) illustration together with feature extraction, the dimensions of the visual codebook, and therefore the difference live wont to compare the BOVW representations. Here it conjointly performs the comparisons with normal options like color and texture.

This paper [3] the author investigates bag-of-visual-words (BOVW) approaches to land-use classification in high-resolution. They consider about a customary non-spatial illustration during which the frequencies of measure image options area unit accustomed discriminate between categories analogous to however words area unit used for text document classification while not respect to their order of prevalence. The author conjointly think about 2 spatial extensions, the established spatial pyramid match kernel that considers absolutely the placement of the image options, as well as a novel method which we have a tendency to term the spatial co-occurrence kernel that considers the relative arrangement. These extensions area unit actuated by the importance of spatial structure in geographic knowledge. Here it conjointly performs intensive analysis of various configurations like the dimensions of the visual dictionaries.

This concept [4], the author explains a well-known Bag-of-Features (BoF) model that is generalized and developed as a neural network that's composed of 3 layers: a Radial Basis perform (RBF) layer, associate accumulation layer, and a completely connected layer. This formulation permits for decoupling the illustration size from the quantity of used code words, still as for higher modeling the feature distribution employing a separate trainable scaling parameter for every RBF vegetative cell. The ensuing network, known as Retrieval-oriented Neural BoF (RN-BoF), is trained victimization regular back propagation and permits for quick extraction of compact image representations.
In this paper [6] authors explain regarding feature representations for image retrieval that has continuously been a difficult task within the field of remote sensing. Ancient strategies specialize in extracting low-level overhead options that aren't solely long however conjoinly tend to attain disappointing performance thanks to the quality of remote sensing pictures. In this paper, we have a tendency to investigate a way to extract deep feature representations supported convolutional neural networks (CNNs) for high-resolution remote sensing image retrieval (HRRSIR). As a result many effective schemes are planned to come up with powerful feature representations for HRRSIR. The novel CNN has fewer parameters than the pre-trained and fine-tuned CNNs and might learn low dimensional options from restricted labeled pictures. The schemes are evaluated on many difficult, in public obtainable datasets.

This paper [7] deals with deep feature extraction from Convolutional Neural Networks (CNNs). The author describes that this approach have well-tried to possess robust ability to transfer for varied visual recognition tasks, like image classification, object detection, fine-grained recognition, and image instance-level retrieval. During this paper, they target the problem of high-resolution remote sensing (HRRS) image. They introduce multi-scale concatenation and multi-patch pooling strategies for performance improvement. Authors' experimental results indicate that fine-tuning is effective to create progress on CNN and supply outstanding accuracy that outperforms previous progressive strategies.

In this paper [8] authors planned one in all the difficult issues in understanding high-resolution remote sensing pictures that's aerial scene classification. A well-designed feature extractor and classifier will improve classification accuracy. During this paper, author constructs three completely different convolutional neural networks with different sizes of receptive field. They additional propose a structure fusion methodology, which may create judgment by incorporating completely different levels' info. The effectiveness of the planned methodology is tested on a tougher information set-AID that has ten thousand high-resolution remote sensing pictures with thirty classes. Experimental results show that our structure fusion model gets a big classification accuracy improvement.

Here [9] an improved ViBe algorithmic rule is planned by the authors for strong and correct detection of multiple vehicles. It uses the gray-scale spacial data to make lexicon of element life length to create ghost shadows and object's residual shadows quickly integrated into the samples of the background. During this paper, authors additionally style a technique mistreatment 2 classifiers to additional attack the matter of failure to trace vehicles with occlusions and interference. The 2 classifiers technique has each time potency advantage of SVM and high accuracy advantage of CNN. Examination with many existing strategies, the qualitative and quantitative chemical analysis of their experiment results showed that the planned technique not solely effectively removed the ghost shadows, and improved the detection accuracy and period performance, however additionally was strong to agitate the occlusion of multiple vehicles in varied traffic scenes.

In this paper [10] authors discusses on road detection from the attitude of moving vehicles. Recently, several deep learning strategies grow for this task as a result of they will extract high-level native options to search out road regions from raw RGB information, like Convolutional Neural Networks (CNN) and totally Convolutional Networks (FCN). However, the way to observe the boundary of road accurately continues to be associate recalcitrant downside. During this paper, authors planned a siamese totally convolutional network supported VGG-net design, that is in a position to think about RGB-channel, linguistics contour and site previous at the same time to phase road region in an elaborate way. Experiments demonstrate that the planned s-FCN-loc will learn a lot of discriminative options of road boundaries. Finally, the planned approach is evaluated on KITTI road detection benchmark, and achieves a competitive result.
This paper [11] deals with street scene understanding. One necessary step toward this direction is scene labeling that annotates every pixel within the pictures with an accurate category label. This paper proposes a joint technique of priori convolutional neural networks at super pixel level and soft restricted context transfer. Authors contributions area unit threefold: a priori s-CNNs model that learns priori location info at super pixel level is projected to explain varied objects discriminately; a stratified information augmentation technique is given to alleviate information set bias within the priori s-CNNs coaching stage, that improves foreground objects labeling significantly; and a soft restricted MRF energy perform is outlined to enhance the priori s-CNNs model's labeling performance and cut back the over smoothness at constant time.

In this paper [12] the author proposed Band choice, by selecting a collection of representative bands in a very hyper spectral image is planned, that is a good methodology to scale back the redundant data while not compromising the first contents. This paper focuses on clustering based band choice and proposes a brand new framework to resolve the higher than perplexity, claiming the subsequent contributions: a best clustering framework, which might acquire the best clustering result for a selected sort of objective operate underneath an affordable constraint; a rank on clusters strategy, that provides a good criterion to pick out bands on existing clustering structure; and an automatic methodology to see the amount of the specified bands, which might higher judge the distinctive data created by sure range of bands. The planned formula is strong and considerably outperforms the opposite ways on varied information sets.

In this work [13] authors examines the result of the convolutional network depth for its accuracy within the large-scale image recognition setting. Authors main contribution could be a thorough analysis of networks of skyrocketing depth victimization design with terribly little convolution filters, that shows that a major improvement on the prior-art configurations are often achieved by pushing the depth to 16-19 weight layers. These findings were the idea of their ImageNet Challenge 2014 submission, wherever their team secured the primary and therefore the second places within the localization and classification tracks severally. They need created 2 best-performing ConvNet models publically offered to facilitate any analysis on the employment of deep visual representations in computer vision.

This paper [14] authors has given on deep neural networks that has been driven by a general observation wherever increasing depth will increase the performance of a network. Recently, however, proof has been tried that merely increasing depth might not be the most effective ways to increase performance. Investigations into deep residual networks have conjointly urged that they'll not actually be in operation as one deep network. Authors examine these problems, and make a brand new interpretation of the unraveled read of deep residual networks that explains a number of the behaviors that are ascertained by experimentation. As a result, they were able to derive a brand new, Shallower, design of residual networks that considerably outperforms a lot of deeper models on the ImageNet classification dataset. The design that they propose therefore outperforms its comparators, as well as terribly deep ResNets, and however is additional economical in memory use and generally conjointly in coaching time.

Here [15] authors describes benchmark datasets that are unit vital for developing, evaluating, and examination remote sensing image retrieval (RSIR) approaches. However, current benchmark datasets are unit deficient in this they were originally collected for land use/land cowl classification rather than RSIR. Authors so gift a replacement large-scale remote sensing dataset termed "PatternNet" that was collected specifically for RSIR. PatternNet was collected from high-resolution imagination and contains thirty eight categories with 800 pictures per category. Considerably, PatternNet's giant scale makes it appropriate for developing novel, deep learning primarily based approaches for RSIR. They use PatternNet to gauge the performance of over thirty five RSIR ways starting from ancient handcrafted feature primarily based ways to recent, deep learning primarily based ones.

In this paper [16] author explains the results of applying morphological texture descriptors. Mathematical morphology offers a range of multi-scale texture descriptors, capable of computing translation, rotation and illumination invariant options. They specialize in the circular variance bar chart and therefore the rotation invariant points approaches, and take a look at them with the UC Merced Land Use dataset. They are compared against different far-famed descriptors like LBP and Dennis Gabor filters, and area unit shown to produce either comparable or superior performance despite their shorter feature vector length.
This paper [17] discusses high resolution remote sensing image captured by the satellites or the craft is of nice facilitating for military and civilian applications. In recent years, with associate increasing quantity of high resolution remote sensing pictures, it becomes additional and additional pressing to seek out the way to retrieve them. During this case, some ways supported the applied mathematics info of the native options are projected, that have achieved sensible performances. During this paper, authors propose a brand new technique to represent these pictures, by taking the structural info into thought. the most contributions of this paper include: mapping the options into a manifold area by a Lipchitz swish operate to reinforce the illustration ability of the features; coaching associate anchor set with many regularization constrains to urge the intrinsic manifold structure. Within the experiments, the strategy is applied to 2 difficult remote sensing image datasets: UC Merced land use dataset and Sydney dataset. Compared to the progressive approaches, the projected technique can do an additional strong and commendable performance.

**Table 1. Comparison Table**

| TECHNIQUE                      | ADVANTAGES                                                                                             | DISADVANTAGES                                                                                   |
|--------------------------------|--------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------|
| 1. Bag-of -words with VLAD representations. | Very simple to understand and implement.                                                              | Leads to a high dimensional feature vector due to large size of vocabulary and doesn’t leverage co-occurrence statistics between words. |
| 2. Local invariant features     | Locality: features are local, so robust to occlusion. and clutter (no prior segmentation)              | It will not evaluate methods based on a single image.                                             |
|                                | Distinctiveness: individual features can be matched. To a large database of objects.                  |                                                                                                  |
|                                | Quantity: many features can be generated for even small objects.                                       |                                                                                                  |
|                                | Extensibility: can easily be extended to wide range.                                                   |                                                                                                  |
| 3. Bag-of-visual-words (BOVW)  | Invariance to scale and orientation. Offline computation. Promising to adopt existing algorithms in text domain. | Size of vocabulary. Efficiency of generating visual words. Feature selection and reduction.      |
| 4. Bag-Of-Feature Model        | Scalability Generalization                                                                             | The lack of spatial information in traditional BoF representations seems to make them poor choices for systems that localize objects in images or describe relationships among objects. |
| 5. CNN Features                | It automatically detects the important features without any human supervision.                       | A Convolutional Neural Network is significantly slower due to an operation.                       |
| 6. Feature representation based on CNN | Retain more local features and spatial information of the image.                                       | It will not be highly effective in scenarios where ordering of features does not describe the   |
| Model Description                                                                 | Details                                                                                           |
|----------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------|
| 7. Multi-scale concatenation and Multi-Patch pooling Method                      | Helps to learn more robust convolutional filters, and thus the fusion accuracy can be advanced from the current state-of-the-art level. The pooling operations will lose some information although the pyramid is used to obtain multi-scale information. |
| 8. Multi-level fusion method in aerial scene classification                      | Capable of preserving vital Information by extracting all important Information from the images without producing any inconsistencies in the output image. This model provides lack of flexibility. |
| 9. ViBe Algorithm for accurate detection of multiple vehicles.                  | Simple and easy to implement. High operation efficiency. The noise problem under dynamic background. The target fixing problem. |
| 10. Siamese Fully Convolutional Network                                           | More robust to class imbalance. Detect boundary of road accurately. It is slower than normal classification type of learning. |
| 11. Priori Superpixel-CNN Model                                                  | Adaptsively changes the number of super pixels according to the given images. Used to segment road and non-road parts. Limitations of computational efficiency |
| 12. Optimal Clustering Framework                                                 | Obtain the optimal clustering result for a particular form of objective function under a reasonable constraint. Correlation among bands is ignored |
| 13. ConvNet Model for large scale image recognition                             | It automatically detects the important features without any human supervision. Requires a large Dataset to process and train the neural network. |
| 14. ResNet Model for visual recognition                                         | Strengthen feature propagation. Encourage feature reuse. Reduce the number of parameters. Increased complexity of architecture. Implementation of Batch normalization layers since ResNet heavily depends on it. |
| 15. PatternNet Model for performance evaluation                                 | Provides an order of magnitude speedup. Severeley restricted the development of novel feature representations for RSIR. |
| 16. Global Morphological Texture Descriptors                                    | Capable of computing translation, rotation and illumination invariant. Higher retrieval accuracy. Enhance entire structures in a medical image without discrimination. |
| 17. Local Structure Learning                                                     | It possesses high indexing capability and low dimensionality. Inaccuracy in notation due to the subjectivity of human perception. |
III. EXPERIMENTS AND ANALYSIS

A. Per Class mAP
It measures the per class mean average precision (mAP) values for the CNN features (Fc6, Fc7, and Pool5) on each data set. An average value of the ways using the pre-trained features (Fc6_Pre, Fc7_Pre, and Pool5_Pre) without the weighted distance are poor performance for buildings, storage tanks, and tennis court, however our ways with the weighted distance (Fc6_W, Fc7_W, and Pool5_W) achieve noticeably better results so do basketball court, nursing home, swimming pool. Average values of the methods using the pre-trained feature are approximately 60%, whereas our methods are approximately 90%. It can conclude that our methods can get better results than the pre-trained features. Moreover, it can be seen that Pool5_W is the best in the three features with the weighted distance.

B. Performance of a Different Training Image Number
The weight is affected by the image class prediction exactness of the fine-tuned CNN models. The number of the training images using in the fine-tuned method is an important factor of the class prediction precision. Therefore, it creates the experiment to investigate the result of training data volume on retrieval performance. It takes a series of images per image class to fine-tune the pre-trained CNN model, which these images are randomly split into training and testing data sets with an 80%/20% split. It is discovered that the performance of the proposed methods improves as the number of training image increases in terms of mAP value. And the highest mAP value is up to almost 95% on the two data sets. Their ways perform better than the pre-trained CNN feature when the number of training images per image class is bigger than 10. Pool5 feature have a better performance on both data sets. Fc6 and Fc7 feature leads to similar performance. The results of pre-trained CNN features with weight (Fc6-pre_W, Fc7-pre_W, and Pool5-pre_W) are similar to those of fine-tuned features with weight (Fc6_W, Fc7_W, and Pool5_W). It means that these methods will get good performance by the two kinds of features.

C. Comparisons with the State-of-the-Art Methods
To verify the effectiveness of the proposed methodology, here we compare our methodology with the state-of-the-art methods. Since most connected works are supported by UCMD. The fine-tuned CNN model is to extract image features and label the category of image in the retrieval data set. In the online stage, here we calculate the weight of each class with the class probability of the query image and used it to adjust the distance between the query image and the retrieved images. The experimental results on the UCMD and PatternNet data sets demonstrate that the proposed methodology achieves higher performance compared with that of the state-of-the-art methods.

IV. CONCLUSION
Automatically and efficiently to retrieve the remote sensing pictures that users would like from massive image databases becomes one in every of the difficult and rising analysis topics within the field of remote sensing. Within the offline stage, we have a tendency to use the fine-tuned CNN models to extract image options and label the category of image within the retrieval knowledge set. Within the online stage, we have a tendency to calculate the load of every category in step with the category likelihood of the question image and used it to regulate the gap between the question image and therefore the retrieved pictures. The experimental results on the UCMD and PatternNet knowledge sets demonstrate that the planned methodology achieves higher performance compared therewith of the progressive ways.

V. REFERENCES
[1] S. Özkan, T. Ates, E. Tola, M. Soysal, and E. Esen, “Performance analysis of state-of-the-art representation methods for geographical image retrieval and categorization,” IEEE Geosci. Remote Sens. Lett., vol. 11, no. 11, pp. 1996–2000, Nov. 2014.
[2] Y. Yang and S. Newsam, “Geographic image retrieval using local invariant features,” IEEE Trans. Geosci. Remote Sens., vol. 51, no. 2, pp. 818–832, Feb. 2013.
[3] Y. Yang and S. Newsam, “Bag-of-visual-words and spatial extensions for land-use classification,” in Proc. ACM Int. Conf. Adv. Geogr. Inf. Syst., 2010, pp. 270–279.
[4] N. Passalis and A. Tefas, “Learning neural bag-of-features for large-scale image retrieval,” IEEE Trans. Syst., Man, Cybern. Syst., vol. 47, no. 10, pp. 2641–2652, Oct. 2017.
Y. Ge, S. Jiang, F. Ye, C. Jiang, Y. Chen, and Y. Tang, “Aggregating CNN features for remote sensing image retrieval,” Remote Sens. Land Resour., to be published.

W. Zhou, S. Newsam, C. Li, and Z. Shao, “Learning low dimensional convolutional neural networks for high-resolution remote sensing image retrieval,” Remote Sens., vol. 19, no. 5, p. 489, May 2017.

F. Hu, X. Tong, G.-S. Xia, and L. Zhang, “Delving into deep representations for remote sensing image retrieval,” in Proc. Int. Conf. Image Process., Nov. 2017, pp. 198–203.

Y. Yu and F. Liu, “Aerial scene classification via multilevel fusion based on deep convolutional neural networks,” IEEE Geosci. Remote Sens. Lett., vol. 15, no. 2, pp. 287–291, Feb. 2018.

W. Min, M. Fan, X. Guo, and Q. Han, “A new approach to track multiple vehicles with the combination of robust detection and two classifiers,” IEEE Trans. Intell. Transp. Syst., vol. 19, no. 1, pp. 174–186, Jan. 2018.

Q. Wang, J. Gao, and Y. Yuan, “Embedding structured contour and location prior in siamesed fully convolutional networks for road detection,” IEEE Trans. Intell. Transp. Syst., vol. 19, no. 1, pp. 230–241, Jan. 2018.

Q. Wang, J. Gao, and Y. Yuan, “A joint convolutional neural networks and context transfer for street scenes labeling,” IEEE Trans. Intell. Transp. Syst., vol. 19, no. 5, pp. 1457–1470, May 2018.

Q. Wang, F. Zhang, and X. Li, “Optimal clustering framework for hyperspectral band selection,” IEEE Trans. Geosci. Remote Sens., to be published, doi: 10.1109/TGRS.2018.2828161.

K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in Proc. Int. Conf. Learn. Represent., Sep. 2015, pp. 1–14.

Z. Wu, C. Shen, and A. van den Hengel. (2016). “Wider or deeper: Revisiting the ResNet model for visual recognition.” [Online]. Available: https://arxiv.org/abs/1611.10080

W. Zhou, S. Newsam, C. Li, and Z. Shao, “PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval,” ISPRS J. Photogramm. Remote Sens., to be published, doi: 10.1016/j.isprsjprs.2018.01.004.

E. Aptoula, “Remote sensing image retrieval with global morphological texture descriptors,” IEEE Trans. Geosci. Remote Sens., vol. 52, no. 5, pp. 3023–3034, May 2014.

Z. Du, X. Li, and X. Lu, “Local structure learning in high resolution remote sensing image retrieval,” Neurocomputing, vol. 207, pp. 813–822, Sep. 2016.