International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-9, Issue-6, November 2021
https://doi.org/10.55524/ijircst.2021.9.6.63
Article ID IRPS120374, Pages 285-289
www.ijircst.org

Automatic Object Detection on Aerial Images Using Convolutional Neural Networks

Jasdeep Singh

RIMT University, Mandi Gobindgarh, Punjab, India
Correspondence should be addressed to Jasdeep Singh; jasdeepsingh@rimt.ac.in

ABSTRACT- Large quantities of aerial and satellite images are being acquired on a daily basis. Many practical applications may benefit from the analysis of such huge amounts of data. We propose an automated content-based analysis of aerial photography in this letter, which may be used to identify and label arbitrary objects or areas in high-resolution pictures. We developed a convolutional neural network-based approach for automated object identification for this purpose. In the tasks of aerial picture classification and object identification, a new two-stage method for network training is developed and validated. First, we used the UC Merced data set of aerial pictures to evaluate the suggested training method, and we were able to obtain an accuracy of about 98.6%. Second, a technique for automatically detecting objects was developed and tested. For GPGPU implementation, a processing time of approximately 30 seconds was needed for one aerial picture of size 5000 x 5000 pixels.

KEYWORDS- Aerial images, Automatic, Convolutional Neural Networks (CNNs), Convolutional neural network, Object detection.

I. INTRODUCTION

A vast number of aerial photos are now available thanks to a recent development in remote sensing technology [1-3]. A vast number of aerial photos necessitated quick analysis and classification in order to efficiently allow their use in civic and military applications. The analysis of remote sensing imagery helps a wide range of land use tasks, including urban planning, surveillance, crop monitoring, flood and fire control, and so on. The need for automatic extraction of useful data from aerial photographs prompted the invention and refinement of a variety of processing algorithms tailored to that aim [4-7]. In practice, automatic content-based aerial imagery description is frequently used in land use applications [8-9]. Local and global features were used in seminal research to define visual content. Because texture is a good discriminator, it is frequently utilized in aerial image classification. Feature extraction methods were successfully applied to colour, multispectral, and hyper spectral pictures, and post-processing feature data was used to improve classification accuracy and represent semantic image information more efficiently. Even though feature-based techniques performed well, they did not take advantage of higher-order local features or their intricate spatial relationships. Convolutional neural networks (CNNs) allowed researchers to overcome past limitations, such as high time and computational consumption and the need for a large amount of training data.

Under some circumstances, both feature-based and network-based techniques to aerial picture classification can achieve state-of-the-art performance. Extracting low-level local features, feature post-processing, and statistical classifier are some of the most recent feature-based techniques [10-13]. When descriptors with low dimensionality are utilized, feature-based techniques can learn from tiny sets of training data and obtain good results. Network-based techniques, on the other hand, take advantage of local features' spatial dependence and recombine low-order features into higher-order features in each succeeding layer. One of the most intriguing features of CNNs is their ability to learn features from a data set of a single modality and then apply them efficiently on a data set of a second modality after some further training. However, when compared to networks that were trained from scratch, employing a pre-trained network produced greater results.

Automatic object detection is another crucial task in aerial imaging. Automatic object recognition uses traditional image classification to process aerial photos piece by piece in order to discover regions with specific visual content automatically. Because modern technology can provide vast amounts of high-resolution aerial and satellite imagery, automatic processing that is quick and accurate could be very useful in practical applications [14-17].

The purpose of this study is to develop a system for automatically detecting objects in aerial photos using CNNs, as well as to analyze network training methods. We look into how fine-tuning CNN can help capture local textural properties of aerial photos more efficiently. In addition, we present a novel two-stage strategy for teaching a CNN to solve two traditional remote sensing tasks, namely land use classification and object recognition. First, we put our fine-tuned CNN to the test on the UCMerced data set1 in the land use classification task, and compared the results against state-of-the-art feature- and network-based techniques. Following the validation of the training approach, we used the suggested fine-tuned network to construct an algorithm for automatic object detection on high-resolution aerial photos. We looked at the broader case of automatically detecting various sorts of objects on a USGS picture data collection, which was inspired by the approach to detect major facilities presented in.

II. LITRATURE REVIEW

The first and second-order local features are typically used in feature-based approaches, and they are designed to capture local textural aspects with varied invariant requirements [18]. Each spectral band is normally regarded as a greyscale image when colour characteristics are collected. In scale-invariant feature transform histogram of directed gradients, etc., highly effective local feature extraction algorithms are recognized. Local features are frequently combined with the bag-of-visual-words paradigm to create final descriptors, which comprise a histogram of the occurrence of various local characteristics detected in a single image. Low-level local feature descriptive characteristics are more effectively utilised when combined with their spatial properties, therefore both local features and their position in the image are considered when the final descriptor is calculated. Texture descriptor extraction can also be applied to color and multispectral pictures with ease. Additional post-processing of local features improves feature-based techniques even further. The network-based approach uses a deep learning technique, which is extensively utilized in machine learning and hierarchically extracts high-order local information [19]. CNNs typically have a number of convolutional layers for detecting features and a classification layer, all of which are capable of learning deep features based on the training data. When opposed to the feature-based approach, this method has two key advantages. It can learn higher order local features by integrating low-level information and leveraging spatial dependence between them inherently. CNNs have recently been used to classify aerial photos, after promising results in object recognition [20,21]. These publications provided new approaches for using CNNs to classify aerial images and achieved impressive results. In order to uncover corresponding hierarchical structures in photos, a network-based method capable of learning certain spatial properties from remote sensing images was described. The scientists hypothesized that using a multi-scale image input would improve the representation of size-varying objects, and they used pre-trained networks to extract descriptors and classify high-resolution remote images. In, a comprehensive description of CNNs and their application to aerial picture categorization is provided. The authors explored numerous methods for teaching a CNN and found that extra training of networks that have already been trained on a large data set can increase classification accuracy.

A. *Methodology*

We devised an unique two-stage technique to train a CNN and applied it to the land use categorization problem and the object identification problem, inspired by the work provided in and the network training methodologies presented in. We utilized the GoogleNet architecture as the foundation for our CNN and trained it with the Caffe framework and the NVidia DIGITS training system. Instead of using randomly initialized weights to train the network, we utilized weights from the same architecture that had been trained to convergence on the ImageNet database. Because the number of classes changed between the sets, all weights were copied from the pre-trained network, with the exception of the final completely connected layer. This method was used to prevent the network from over fitting on the UCMerced data set. During training, no layers were frozen, and the network was allowed to converge. The learning rate was 0.01% at the start of the training and dropped to 35% after 70 repetitions (10 epochs). 45 epochs were used to train the network (3150 iterations). As a learning approach, stochastic gradient descent was utilized [22-24]
The network was fine-tuned on the same data set using the same set of training pictures in the second round of training. We utilized an adaptive gradient method with an exponentially declining learning rate with a gamma value of 0.998, starting at 0.001 and decaying to 0.0002 at the 20th epoch, where the highest accuracy was attained, instead of stochastic gradient descent. Because the adaptive gradient technique differs from standard stochastic gradient descent, we utilized fine-tuning with it. Because the network had already converged and over fitted in the first stage, we decided to lower the learning rate and apply a more flexible method to allow for modest changes in the fitted function. Our objective was to obtain a high prediction rate across all classes so that the network may be utilized for item recognition on high-resolution aerial pictures in the future.

A high-resolution aerial image was divided into overlapping patches of size 256 x 256 pixels, the same size as the trained CNN's input image, to build the object detection technique. The CNN categorized each of these changes using the DIGITS server's REST API. The server delivers the top-5 class predictions for each patch. The supplied picture is effectively divided into a finer grid since overlapping patches are utilized. Because we divided the image into patches in 128 stages, the grid has 128 x 128 cells. The total of probability for patches overlapping the cell is determined for each cell.

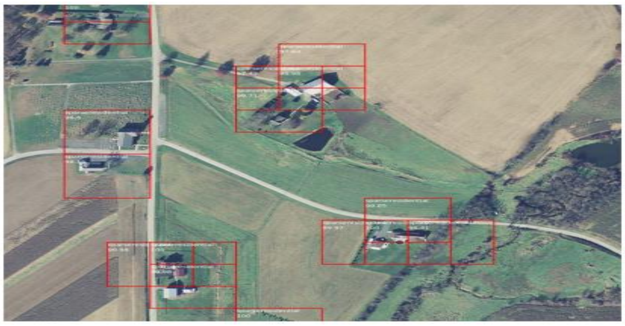


Figure 1: Part of the image from Fig. 3(a) in which the example of object detection is given. In this example, the blocks which have a high probability to contain a sparse residential area are highlighted with red (best view in color)

A cell is labelled as the associated class if its class probability is greater than a preset threshold. Fig. 1 shows an example of querying areas.

This method lets you to search a high-resolution picture for any class that the CNN has been trained on. The method does not impose any restrictions on the size or resolution of the input image (as long as the image format permits it). It is always possible to split a picture into cells. The patching step can be reduced to offer a finer grid for improved accuracy. On the server, several GPUs may be utilized to classify numerous patches at the same time.

1) Experimental Setup

a) Used Data

In this study, two sets of data were used. The first is the well-known UCMerced aerial ortho-image data set, which contains pictures carefully removed and labelled from USGS national maps [18-22]. It is divided into 21 categories, each with 100 pictures measuring 256 by 256 pixels. Because it comprises visually similar but conceptually diverse groups, this data set is frequently used to assess the effectiveness of categorization techniques. Figure 2 shows samples of pictures from the UCMerced categories.

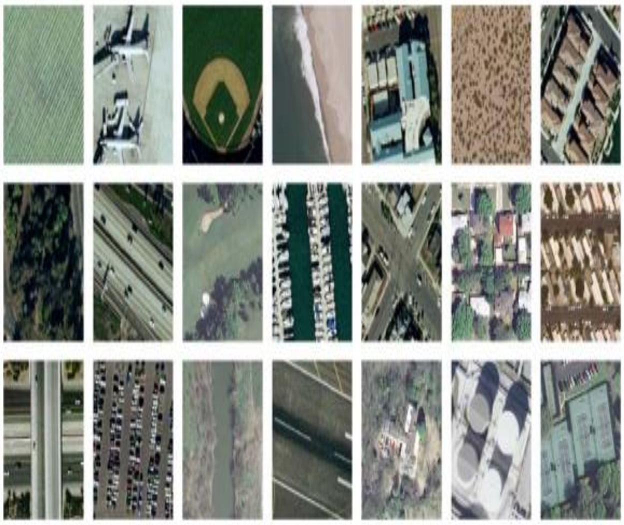


Figure 2: From top to bottom, left to right, examples of pictures from the UCMerced data collection is illustrated: agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbour, intersection, medium residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis court (best view in colour)

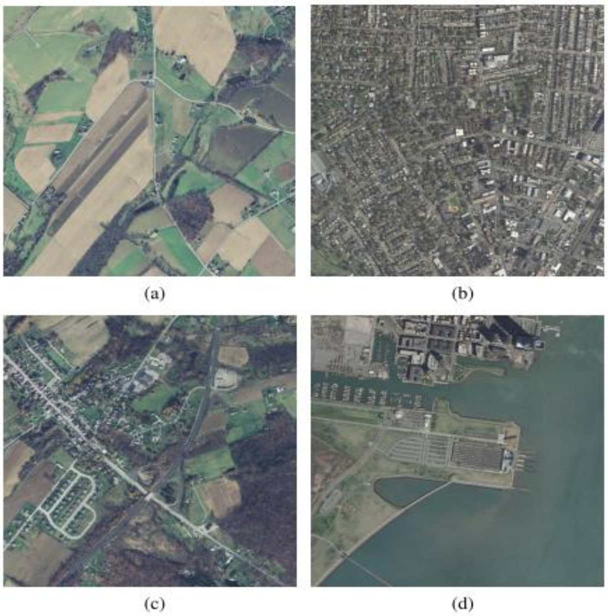


Figure 3: Examples of high-resolution USGS images. All images are resized (best view in colour). (a) Image from the USGS data set showing countryside with isolated farms. (b) Image from the USGS data set showing densely populated urban area. (c) Image from the USGS data set showing medium populated countryside area. (d) Image from the USGS data set showing harbour and urban area along with parking lots and fields

The second data set comes from the USGS national map, but this time high-resolution pictures with a size of 5000 x 5000 pixels are utilized to solve the item recognition problem. The first image (Fig. 3(a)) depicts a rural landscape with farms, agricultural, and woodland regions. It may be used for agricultural monitoring, forest protection, and other purposes. It might be regarded a simple item detection problem because there are single objects of the same type. The second picture, shown in Fig. 3(b), depicts a densely populated metropolitan region. Even with more time, finding items with a defined purpose, such as tennis courts or baseball diamonds, by visual examination is difficult. The presence of dense residential and building types may be seen in pictures from the UCMerced data collection. The challenge of automatically detecting items in this image might be deemed difficult. The picture in Fig. 3(c) depicts a medium residential area with countryside, where distinguishing between medium and sparse residential areas is difficult. Finally, Fig. 3(d) depicts a portion of the urban area along the seashore, complete with a harbour, parking lot, and green fields.

b) Land Use Classification

The first experiment uses the UCMerced data set to classify land uses, allowing us to compare the classification accuracy attained with the proposed fine-tuned CNN to findings published in the literature. Images from all categories are randomly split into training and test sets with an 80:20 ratio, according to standard technique. All pictures from the training set and images from the test set are utilized to teach a convolutional network.

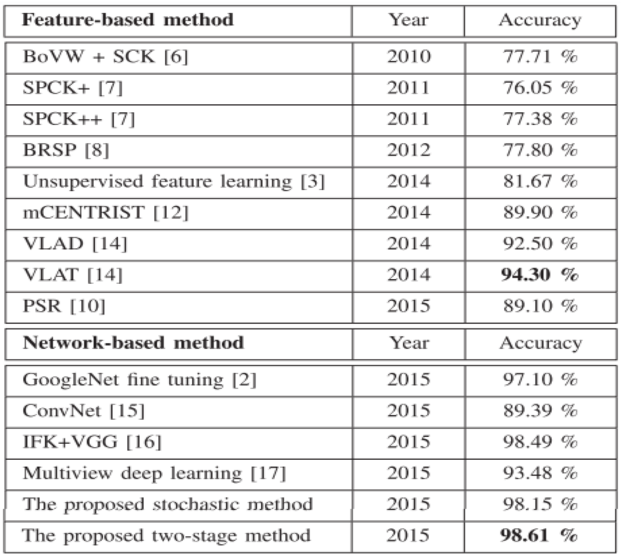
c) Automatic Object Detection

The second experiment uses a pre-trained neural network to automatically recognize a specific object in high-resolution pictures. In general, we wanted to see if the CNN-based object identification algorithm could effectively find regions semantically comparable to pictures from the categories used to train the CNN, in this case, the UCMerced data set. Our object identification method was evaluated on high-resolution aerial pictures of lonely residences, buildings, baseball diamonds, medium residential neighbourhoods, boats in harbour, and automobiles in parking lots, as shown in Fig. 3. In other words, the identified item might be anything that is semantically comparable to the UCMerced data set's categories.

III. DISCUSSION

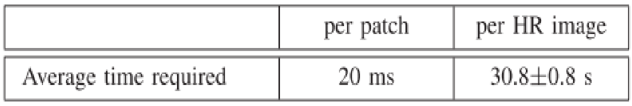
On a machine with an Intel i7 CPU and a single NVidia GeForce 970 GTX GPU, we tested the suggested method. To begin, we classified pictures from the UCMerced data set using fine-tuned CNN. The acquired results are compared to those found in the literature, which are presented in Table 1. When second-order local features are integrated into compact texture representations termed vectors of locally aggregated tensors, the best performance is attained for the feature-based method. Another aspect to consider is that "deeper" local features provide better discriminative texture representation. Given that the network method integrates low-level characteristics and attempts to learn higher-order features, it is reasonable to anticipate CNN to outperform the feature-based approach. The networks shown in and were adjusted to extract aerial picture characteristics in the most discriminative way possible, and they outperformed feature-based methods, as predicted. The suggested technique was created to collect local characteristics more effectively in order to enhance classification accuracy. After the first stage of network training, we assessed the performance and found that it was 98.15 percent accurate. The two-stage fine-tuned network then obtained a 98.61 percent accuracy. We also tried single-stage training using ImageNet initialized weights, which yielded a maximum accuracy of 98.38 percent after 100 epochs and a learning rate of 0.01 declining to 50% every 16 epochs.

Table 1: A comparison of the proposed method's aerial picture classification accuracy on the UCMerced data set with published findings



Feature-based method	Year	Accuracy
BoVW + SCK [6]	2010	77.71 %
SPCK+ [7]	2011	76.05 %
SPCK++ [7]	2011	77.38 %
BRSP [8]	2012	77.80 %
Unsupervised feature learning [3]	2014	81.67 %
mCENTRIST [12]	2014	89.90 %
VLAD [14]	2014	92.50 %
VLAT [14]	2014	**94.30 %**
PSR [10]	2015	89.10 %
Network-based method	Year	Accuracy
GoogleNet fine tuning [2]	2015	97.10 %
ConvNet [15]	2015	89.39 %
IFK+VGG [16]	2015	98.49 %
Multiview deep learning [17]	2015	93.48 %
The proposed stochastic method	2015	98.15 %
The proposed two-stage method	2015	**98.61 %**

Table 2: Time consumption for automatic object detection



	per patch	per HR image
Average time required	20 ms	30.8±0.8 s

The efficacy of our approach in automatically detecting diverse items on high-resolution aerial pictures was assessed in the second experiment. Processing, client–server communication, and loading image patches from disk to memory and transferring to GPU memory all took time per high-resolution picture. Table 2 shows an overview of time consumption. We used the picture in Fig. 3(a) to assess the accuracy of automated solitary item detection. Regions semantically comparable to the sparse residential category are sought based on the UCMerced classifications. Figure 4 shows the upper right corner of the output image (a). We may infer that all isolated items were recognized based on visual verification, i.e., patches visually most comparable to the sparse residential category from UCMerced were discovered with perfect matching. In the second case, Fig. 4 shows the bottom right portion of the original picture from Fig. 3(b) (b). We coloured all of the buildings yellow and the baseball diamond red in this example. The picture from Fig. 3(c) is used in the third example, and the resulting image is shown in Fig. 4. (c). Blue denotes medium residential, whereas red denotes sparse residential. The picture from Fig. 3(d) is used in the fourth example, and the upper left portion of the resultant image is displayed in Fig. 4. (d). Blue denotes the boats in the harbour, whereas red denotes the whole parking lot. These targets may be identified using the UCMerced data set's harbour and parking lot categories, which only give pictures of the complete harbour and parking lot, respectively. All semantic categories used to train the network can be easily recognized and tagged in high-resolution aerial pictures, it may be inferred. Furthermore, CNNs trained with enough data and a wide variety of categories might be utilized to discover items that are difficult to find by eye examination quickly and efficiently.

IV. CONCLUSION

We looked into the potential of using CNNs for aerial picture analysis in this letter. In an image classification challenge, a unique two-stage training technique was proposed and tested. The suggested technique outperforms feature-based approaches as well as other network-based alternatives, according to the findings. The network-based approach for automated content-based item recognition on high-resolution aerial pictures was then built. The accuracy of the suggested technique is demonstrated for various sorts of targeted objects in many instances. Finally, we show how a CNN can be effectively integrated into an object identification algorithm to reliably find regions in pictures that match to the categories on which the CNN was trained. We want to use the suggested approach on a multi-GPU system in the future to cut calculation time even further. In addition, for a large-scale aerial data analysis, a mix of various CNN architectures will be employed.

REFERENCES

[1] Andres L, Boateng K, Borja-Vega C, Thomas E. A review of in-situ and remote sensing technologies to monitor water and sanitation interventions. Water (Switzerland). 2018;

[2] Wei L, Zhang Y, Zhao Z, Zhong X, Liu S, Mao Y, et al. Analysis of mining waste dump site stability based on multiple remote sensing technologies. Remote Sens. 2018;

[3] Singh D. Robust controlling of thermal mixing procedure by means of sliding type controlling. Int J Eng Adv Technol. 2019;

[4] Ghai W, Kumar S, Athavale VA. Using gaussian mixtures on triphone acoustic modelling-based punjabi continuous speech recognition. In: Advances in Intelligent Systems and Computing. 2021.

[5] Khatri M, Kumar A. Stability Inspection of Isolated Hydro Power Plant with Cuttlefish Algorithm. In: 2020 International Conference on Decision Aid Sciences and Application, DASA 2020. 2020.

[6] Solanki MS, Goswami L, Sharma KP, Sikka R. Automatic Detection of Temples in consumer Images using histogram of Gradient. In: Proceedings of 2019 International Conference on Computational Intelligence and Knowledge Economy, ICCIKE 2019. 2019.

[7] Anand V. Photovoltaic actuated induction motor for driving electric vehicle. Int J Eng Adv Technol. 2019;8(6 Special Issue 3):1612–4.

[8] Paravolidakis V, Ragia L, Moirogiorgou K, Zervakis ME. Automatic coastline extraction using edge detection and optimization procedures. Geosci. 2018;

[9] Guo W, Yang W, Zhang H, Hua G. Geospatial object detection in high resolution satellite images based on multi-scale convolutional neural network. Remote Sens. 2018;

[10] Deng Z, Sun H, Zhou S, Zhao J, Lei L, Zou H. Multi-scale object detection in remote sensing imagery with convolutional neural networks. ISPRS J Photogramm Remote Sens. 2018;

[11] Kumar A, Jain A. Image smog restoration using oblique gradient profile prior and energy minimization. Front Comput Sci. 2021;

[12] Gupta N, Vaisla KS, Jain A, Kumar A, Kumar R. Performance Analysis of AODV Routing for Wireless Sensor Network in FPGA Hardware. Comput Syst Sci Eng. 2021;

[13] Kumar Gola K, Chaurasia N, Gupta B, Singh Niranjan D. Sea lion optimization algorithm based node deployment strategy in underwater acoustic sensor network. Int J Commun Syst. 2021;

[14] Kumari N, Kr. Bhatt A, Kr. Dwivedi R, Belwal R. Hybridized approach of image segmentation in classification of fruit mango using BPNN and discriminant analyzer. Multimed Tools Appl. 2021;

[15] Jain A, Kumar A. Desmogging of still smoggy images using a novel channel prior. J Ambient Intell Humaniz Comput. 2021;

[16] Goel S, Dwivedi RK, Sharma A. Analysis of social network using data mining techniques. In: Proceedings of the 2020 9th International Conference on System Modeling and Advancement in Research Trends, SMART 2020. 2020.

[17] Younis S, Ahsan A. Know Your Stars before They Fall Apart: A Social Network Analysis of Telecom Industry to Foster Employee Retention Using Data Mining Technique. IEEE Access. 2021;

[18] Sevo I, Avramovic A. Convolutional neural network based automatic object detection on aerial images. IEEE Geosci Remote Sens Lett. 2016;

[19] Lu J, Ma C, Li L, Xing X, Zhang Y, Wang Z, et al. A Vehicle Detection Method for Aerial Image Based on YOLO. J Comput Commun. 2018;

[20] Lin TY, Cui Y, Belongie S, Hays J. Learning deep representations for ground-to-aerial geolocalization. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2015.

[21] Qayyum A, Saad NM, Kamel N, Malik AS. Deep convolutional neural network processing of aerial stereo imagery to monitor vulnerable zones near power lines. J Appl Remote Sens. 2018;

[22] Kumar S, Kumar K, Pandey AK. Dynamic Channel Allocation in Mobile Multimedia Networks Using Error Back Propagation and Hopfield Neural Network (EBP-HOP). In: Procedia Computer Science. 2016.

[23] Banerjee K, Prasad RA. Reference based inter chromosomal similarity based DNA sequence compression algorithm. In: Proceeding - IEEE International Conference on Computing, Communication and Automation, ICCCA 2017. 2017.

[24] Verma S, Biswas R, Singh JB. Extension of superblock technique to hyperblock using predicate hierarchy graph. In: Communications in Computer and Information Science. 2010.

[25] Yang C, Liu H, Wang S, Liao S. Remote sensing image classification using extreme learning machine-guided collaborative coding. Multidimens Syst Signal Process. 2017;