Communications
2021; 9(1): 1-5
http://www.sciencepublishinggroup.com/j/com
doi: 10.11648/j.com.20210901.11
ISSN: 2328-5966 (Print); ISSN: 2328-5923 (Online)

Multi-class Object Detection Model in Satellite Images Using Convolutional Neural Network

Ibrahim Goni, Asabe Sandra Ahmadu, Yusuf Musa Malgwi

Department of Computer Science, Faculty of Physical Science, Modibbo Adama University, Yola, Nigeria

Email address:
algonis1414@gmail.com (I. Goni), aasandy@gmail.com (A. S. Ahmadu), yumalgwi@mautech.edu.ng (Y. M. Malgwi)

To cite this article:
Ibrahim Goni, Asabe Sandra Ahmadu, Yusuf Musa Malgwi. Multi-class Object Detection Model in Satellite Images Using Convolutional Neural Network. *Communications.* Vol. 9, No. 1, 2021, pp. 1-5. doi: 10.11648/j.com.20210901.11

Received: December 10, 2021; **Accepted**: January 4, 2022; **Published**: January 12, 2022

Abstract: The accurate multi-detection of objects in satellite images has become very essential due to the high criminal activities that posed security threat to humanity all over the world. However, there are significant limitations of traditional methods of multi-object detection such as matching based techniques and object based image analysis. Although Convolutional neural network and image processing techniques has been proved to be essential fields in so many applications of computer vision specifically multi-object detection, multi-object classification, object retrieval, object recognition and object segmentation in a digital image or video, however, multi-object detection especially in satellite images suffer from problems such as shadow, camouflage, and occlusion. The aim of this research work was to design a robust multi-class object detection model in satellite images using image processing techniques and convolutional neural network with a particular concern on image preprocessing, image denoising and image enhancement to enable address the issue of noise in satellite images. The Satellite image that are propose for this model is LandSat-8, because it is free access for research and have a tract record in terms of consistency. The proposed model applied supervised learning algorithm for training different samples of labeled data for the model to enable the system detect vegetation, water bodies, road networks and building. This research will enable the government to know the positions as well as the coordinates of every thick forest, drainage, road networks and buildings in the forest for security reasons. It is at the heart of this research to pave away for the full implementation of this model using either MATLAB or Python Programming.

Keywords: Convolutional Neural Network, Computer Vision, Object Detection, Satellite, Image Processing, Digital Image, Camouflage and Occlusion

1. Introduction

Object detection has attracted so much attention from both academia and industries in recent time. It has become a major concerned in the field of computer vision because of it wide range of applications in Security, Robotic Vision, Drone Scene, Autonomous Driving, Complex Transportation System, Agriculture, Remote Sensing, Surveillance System, Geography information system and health sector. Object detection is a subset of computer vision that deals with the detection or locating the position of semantic object that belong to a certain class or classes (Such As Building, Road Network, Water Body and Human) in a digital image or video. Moreover, this detection can be multi-class detection, edge detection, camouflage, facial detection and pedestrian detection [1]. The aim of this research work is to design a multi-class object detection model in satellite images using convolutional neural network and image processing techniques.

2. Literature

Feng et al [2] proposed a robust hybrid water body extraction model from very high resolution satellite image using deep U-net and super pixel-based conditional random field. Li et al. [3] applied fully connected convolutional neural networks for water body extraction from a very high spatial resolution remote sensing database. Guoji et al. [4] proposed a densely connected CNN Model for water body detection from high resolution remote sensing image in which their results shows a significant improvement compared with the Normalized difference water index.

Mengya et al. [5] proposed a dense local feature compression network for extraction of water body in high resolution remote sensing images and they also constructed a datasets with Gaofen-2 (GF-2) satellite images for their research. The model were tested on different satellite images thus; Sentinel-2 and ZY-3 satellite images which has perform very well as compared with the traditional water body extraction methods.

Li et al. [3] applied deep learning technique for road segmentation from a very high resolution remote sensing image their network avoid background interference and used semantic features to segment multistate roads. Gao, et al. [6] proposed a refined deep Residual convolutional Neural network model for road extraction in a very high resolution satellite image that solve the problems of shadow and noise to enable the model detect road network more accurately. Alexander, et al. [7] developed a semantic segmentation model using deep learning technique to extract road and building in satellite images finally they compared their results with the state-of-the-art.

Ji et al. [8] proposed improved convolutional neural network architecture to extract buildings from high-resolution aerial and satellite images. They applied two dilated convolutions at the first two layers for increasing the sight-of-view and adding the semantic information of large buildings for improving segmentation accuracy of the model. Geesara et al [9] developed a deep learning model for building detection. They have preprocessed their datasets using 2-sigma percentile normalization technique. They also applied binary distance transformation for better data labeling process to obtained nearly perfect results.

Nahhas *et al.* [10] they combined LIDAR with orthophoto data and deep learning for building extraction feature were extracted from LIDAR data which includes; the boxy fit, shape index and density which are used for training the model and the results obtained are 93% and 90.2% respectively. Arshitha et al [11] developed a building detection model from satellite images using convolutional neural network, the method in this work detect building with an accuracy of 83%. Vakalopoulou et al. [12] developed an automatic building detection framework from a very high resolution remote sensing data using deep convolutional neural network. The technique applied supervised classification method during training. Large number of data was used in their work to enable obtained perfect results. Li et al [13] proposed a two stage CNN model to detect rural buildings from a very high resolution images. This method applied a robust CNN architecture that automatically locate villages and recognized buildings using the concept of human vision. Moreover there are some recent surveys and reviews that contributed so much in the area such as Ma et al [14] meta-analysis, Li et al [15] survey research that applied deep learning in image classification.

3. Proposed Model

The following model shows pictorial representation of how an image is going to be collected from LandSat 8 satellite, followed by image preprocessing to image enhancement, dataset preparation and training and testing of the model and finally objects detection. But this are going to be discuss in detail in the next section.

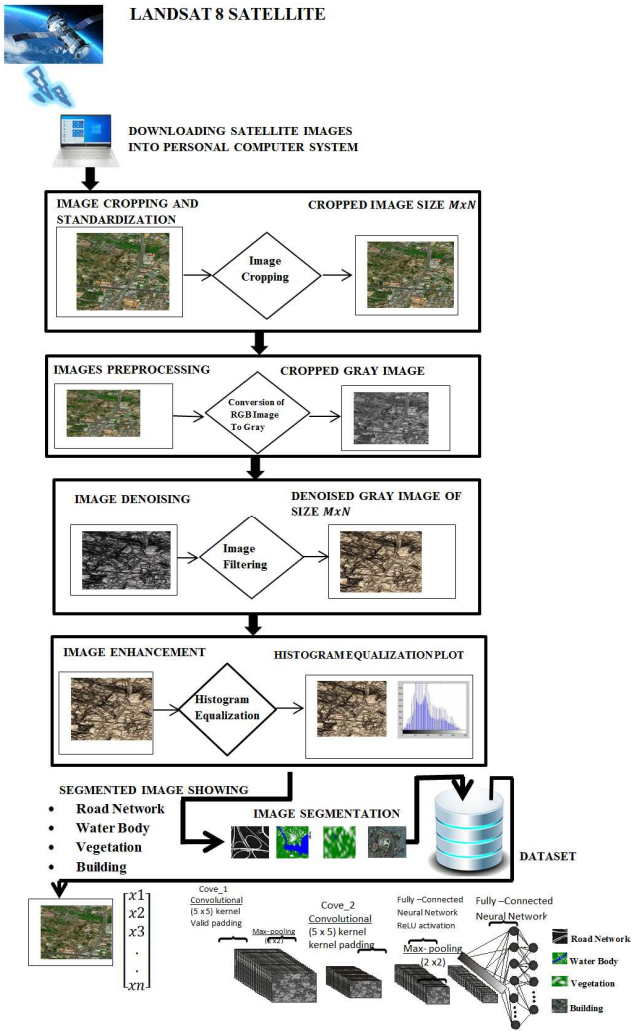

Figure 1. *Multi-class detection model.*

3.1. Satellite Image

Mathematically image is represented by a two dimensions function: $f(x,y)$ where the value of f at the spatial coordinates (x,y) are usually positive and are determined by the source of the image, hence $f(x,y) \neq 0$ and finite;

$$0 < f(x,y) < \infty \tag{1}$$

Where the image function f has two major components.

$$fi < (x,y)\ and\ r(x,y) hence$$

$$f(x,y) = 1(x,y) r\ (x,y) \tag{2}$$

where $0 < i\ (x,y) < \infty\ and\ 0 < (x,y) < 1$

Thereby $i(x,y)$ relies on the illumination source and $r(x,y)$ depend on the texture of the image.

Moreover, if an image is obtained via a transmission the same equation will be given as

$$L_{min} \leq 1 = f(x,y) \leq L_{max} \quad (3)$$

Hence $I = f(x,y)$ is the gray level at coordinates (x,y).

3.2. Image Digitalization

Generally image are converts into digital in two ways digitalization of the coordinates value or digitalization of the amplitude values, hence digital image are represented by M X N matrix.

Such that

$$f = \begin{cases} f_{0,0} \; f_{0,1} \, f_{1,2} \; f_{0,N-1} \\ f_{1,0} \; f_{1,1} \, f_{1,2} \; f_{1,N-1} \\ f_{2,0} \; f_{2,1} \, f_{2,2} \; f_{2,N-1} \\ f_{3,0} \; f_{3,1} \, f_{3,2} \; f_{3,N-1} \\ f_{m-1,0} \; f_{m-1,1} \, f_{m-1,,2} \; f_{m-1,n-1} \end{cases} \quad (4)$$

3.3. Image Cropping and Standardization

Consider an input image of size $m \, X \, n$ and the output size $(H_{0\,X} W_O)$ therefore cropped image can be written as;

$$(H_{c\,X} W_c) = p(H_0 \,|a, a\, W_O) \min \{a\, H|H(H_0\, W|(aW_O)\} \quad (5)$$

3.4. Converting of Red Green Blue (RGB) Image to Gray

$If \; x(u,v) \in R^3$ represent an RGB image located in (U,V) image then the total average color Subtracted from the brightness of the image can be expressed as;

$$x(U,V) \leftarrow x(u,v) + b + \mathrm{u} \quad (6)$$

After the shift is applied and the contrast of the image will change to

$$n(u,v) \leftarrow \gamma x(u,v) + (1-\gamma) avg[x(u,v)] \quad (7)$$

Where $\gamma \sim u \; ([1-c, 1+c])$ and γ is uniformly in the interval $[1-c, 1+c]$ and C is the contrast deviation coefficient but $if \; \gamma > 1$ then the contrast can also increase.

Finally the overall gray scale image can be written as;

$$x(u,v) \leftarrow \left(\delta \, I + \tfrac{1-\delta}{3}\right) 11^T \left(\gamma x(u,v) + (1-\gamma) avg\big(x(u,v)\big) + Bw - \mu\right\} \quad (8)$$

3.5. Image Denoising

From the degradation model $g(x,y) = f(x,y) = n(x,y)$ where $f(x,y)$ is the real image and $g(x,y)$ is the degraded image and $n(x,y)$ is the noise. However there are different form of noise and are represented mathematically as:

Gaussian noise in relation to probability distribution function is define as

$$p(r) = \tfrac{1}{\sqrt{2\Omega\sigma}} \, e^{-(r-\mu)/2\sigma} \quad (9)$$

Where μ is the mean and σ is the standard deviation.

Uniform noise in relation to probability distribution function is given by

$$p(r) = f(x) = \begin{cases} \frac{1}{B-A} If \; A \leq r \leq B \\ O \; otherwise \end{cases} \quad (10)$$

Impulse noise in relation to probability distribution function is given by

$$\begin{cases} PA \; if \; r = A \\ PB \; if \; r = B \\ O \; otherwise \end{cases} \quad (11)$$

3.5.1. Mean Filters of Random Noise Removal

Let g be the input noisy image and f be the output that is denoised image if also $S(x,y)$ is the neightborhood of the pixel $S(x,y)$ defined as;

$$S(x,y) = \{(x+s, y+t), -a \leq s \leq a, -b \leq t \leq b\} \quad (12)$$

of the image size $mn \; where \; m = 2a+1 \; and \; n = 2b+t \; and \; are \; postive \; integer$.

Such that arithematic mean filter can be defined as;

$$f(x,y) = \tfrac{1}{mn} \sum_{(s,t) \in S_{(x,y)}} g(s,t) (16) \quad (13)$$

therefore is very good In denoising gaussian noise and uniform noise.

Geometric mean filter is also defined as;

$$f(x,y) = \left(\Downarrow_{(s,t) \in S_{(x,y)}} g(s,t) \right)^{1/min} \quad (14)$$

Contraharmonic mean filter is also defined as;

$$f(x,y) = \frac{\sum_{(s,t)} \in S_{(x,y)} g(s,t)^{Q=1}}{\sum_{(s,t)} \in S_{(x,y)} g(s,t)^{Q}} \quad (15)$$

Modern filter is also called order statistics filter where $f(x,y)$ depend on the ordering of the pixel value of the image g in the window $s_{(x,y)}$ which is given as;

$$f = (x,y) = modern \, \{g(s,t), (s,t) \in S(x,y)\} \quad (16)$$

Midpoint filter is seen to be a hybrid filter that combine statistical filter and averaging filter and is good for denoising Gaussian noise and uniform noise, it is given by;

$$f(x,y) = \tfrac{1}{2} [(s,t) \underset{es(x,y)}{max} \{g(s,t)\} + \underset{es(x,y)}{min} \{g(s,t)\}] \quad (17)$$

3.5.2. Alpha-Trimoned Mean Filter

From the first order mn pixel values of an input image g in the window $S(x,y)$ and then we remove image g in the lowest $d/2$ and then largest $d/2$ and also we denote the remaining mn-d value by grprovided that $d \leq 0$ be an even integer such that $o \leq d \leq mn-1$ it is given by;

$$f(x,y) = \tfrac{1}{mn-d} = \sum_{(s,t)} \in S(x,y) \, gr \quad (18)$$

3.6. Histogram Equalization

Histogram equalization is an advanced statistical technique in image processing for improving the construct of either dark image or light image, the final output of the processing image is given by $g(x, y)$.

In a direct form let rk= f(x, y) the the histogram equalization of the image g(x, y) is given by;

$$g(x,y) = Sk = (l-1)^n = \sum_{j=0}^{k} p(rj) \quad (19)$$

Theoritcally let r represent the gray levels of the image and s is the random variables in relation to probabilty function Pr(r) and Ps (s) respectively then from (19) if we use the cummulative distribution function it becomes;

$$S = T(r) = (l-1)\int_0^T pr(w)dw \quad (20)$$

Such that $Pr(r) > 0$ on $[0, L-1]$ then $\boldsymbol{T}$ is increasing from $[0, L-1]\ to\ [o, l-1]$ but $\boldsymbol{T}$ is invertible.

However, if we differentiate T and used the formula from probability if $T\ (r)$ then

$$Ps\ (s) = Pr(r)\left[\frac{\partial s}{\partial r}\right] \quad (21)$$

Since we refer $s = s(r) = T(r)$ on a formula of r where $r = r(S) = T^{-1}(s)$ as a function S with reference to the diffraction of s then we obtained

$$\frac{\partial r}{\partial s} = (l-1)\,\mathrm{pr}(r) = T^1(r) \quad (22)$$

Therefore the equation (22) becomes;

$$\frac{\partial r}{\partial s} = \frac{\partial}{\partial s}\left(T^{-1}(s)\right) = \frac{1}{T^{1(s)}} = \frac{1}{(L-1)\mathrm{pr}(r(s))} \quad (23)$$

But

$$Ps(s) = pr(r)\left|\frac{\partial s}{\partial r}\right| = pr(r)\left|\frac{1}{(l-1)pr(r)}\right| = \frac{1}{l-1} \quad (24)$$

The equation (24) now becomes the uniform distribution function on the interval $[o, l-1]$ is also correspond to the flat histogram equation of the satellite image in a discrete form.

3.7. Image Segmentation

Segmented image is an image without noise and sometime sharp which can be an input f and the output could be an image g or not even an image but would be an attribute set of point representing the edges of f boundaries of objects but, segmentation based on our proposal are water body, vegetation, road network and building.

Consider a differentiable function $(x, y) \rightarrow f(x, y)$ in two dimensions, to let defined it gradient operator as being the vector of first order partial directives as;

$$\Delta f(x,y) = \left(\frac{\partial f}{\partial x}\,(x,x)\frac{\partial f}{\partial y}\,(x,y)\right) \quad (25)$$

And the gradient magnitude as Euclidean norm of the vector Δf

$$|\Delta f|(x,y) = \sqrt{\left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2} \quad (26)$$

The central finite differences approximation of the gradient are assuming $\Delta fx = \Delta y = 1$

Such that;

$$\frac{\partial f}{\partial x}\,(x,y) \approx \frac{f(x+1,y)-f(x-1,y)}{2}, \frac{\partial f}{\partial y}\,(x,y) \approx \frac{f(x,y+1)-f(x,y-1)}{2} \quad (27)$$

The gradient can be used to detect edges where the image f does not vary the gradient magnitude $|\Delta f|$ is close to zero while in the area where there are strong variation the gradient magnitude $|\Delta f|$ is larger. Since we have define the output image as $g(x, y) = |\Delta f|(x, y)$ which would show while edges or black background or a threshold version of $|\Delta f|$

If we are going to represent the discrete version of the output it would become

$$g(x,y) = |\Delta f|(x,y) \quad (28)$$

Or the threshold gradient of the final output can be represented as;

$$g(x,y) = \begin{cases} 255|\,\nabla f|\ (\mathrm{x,y}) \geq \mathrm{tolerance T} \\ 0|\nabla f|(\mathrm{x,y}) < \mathrm{tolerance\ T} \end{cases} \quad (29)$$

Where the operation $f \rightarrow g|\nabla f|\ is\ non\ linear.$

3.8. Convolutional Neural Network

Convolutional Neural Network is a deep learning technique that inspired the information processing capability of human brain. It consists of three basic layers; Convolutional layer, polling layer and fully connected layer. At the convolutional layers kernel are used to compute features Maps mathematically the values (ij) in the $K-th$ feature Map of $l-th\ layer$ is expressed as;

$$Z^{\iota}_{i,j,k} = W^{\iota\,T}_{k,Xi,j}\ +\ b^{\iota}_{k} \quad (30)$$

Where $w^{\iota}_{k}\ and\ b^{\iota}_{k}$ are the weight vector and bias term of the $k-th$ filter of the $l-th\ layer$ $X^{\iota}_{i,j}$ is the input at the location (I, i) of the$l-th\ layer$, $Z^{\iota}_{i,j,k}$ is the feature map.

The activation value $a^{\iota}_{i,j,k}$ of the convolutional feature $Z^{\iota}_{i,j,k}$ can be obtained using

$$a^{\iota}_{i,j,k} = a(Z^{\iota}_{i,j,k}) \quad (31)$$

At the polling layer resolution of the feature map are reduced to obtained the shift-invariance by denoting the polling function as $pool(.)$ for each of the feature map $a^{\iota}_{i,j,k}$of the input thus;

$$y^{\iota}_{i,j,k} = pool\left(a^{\iota}_{m,n,k}\right), \forall(m,n) \in R_{i,j} \quad (32)$$

Where $R_{i,j}$ is a local neighborhood around a location (i, j).

CNN has several convolutional and pooling layers. Therefore, after several convolutional and pooling layers, the last layer is fully -connected layer which are used for classification, segmentation and detection. At this layer, there

is an important function known as Loss function which can be computed as;

$$L = \frac{1}{N}\sum_{n=1}^{N} e\left(\theta; y^{(n)}, O^{(n)}\right) \quad (33)$$

Where θ denote all the parameters of CNN
N is the desired inputs and outputs relations
$X^{(n)}$ is the n-th input data
$Y^{(n)}$ is the corresponding target label
$O^{(n)}$ I the output of CNN

But softmax loss is commonly used as a loss function in object detection which is a combination of multinomial logistic loss and is can be computed using the expression;

$$L_{softmax} = -\frac{1}{N}\left[\sum_{i=1}^{N}\sum_{j=1}^{K} 1\{y^{(i)} = j\} \log P_j^{(i)}\right] \quad (34)$$

Where $y^{(i)}$ I the target class
j is the prediction class
N is the input/output relations
K is the filter

Rectified Linear Unit (ReLU) is one of the most activation function used in object detection which can be express as;

$$a_{i,j,k} = \max\left(Z_{i,j,k}, O\right) \quad (35)$$

4. Conclusion

In this research work multi-class object detection model was proposed ranging from satellite image acquisition with a special consideration about the satellite image that will be collected that is landSat-8 it is free access for research and has a track record of quality and consistency since 1970s, due to the nature of satellite images with too much noise this research applied special section for image preprocessing and image enhancement to enable the images free from noise completely so that multiclass detection of objects will be easy for the convolutional neural network. This research work has paved away for the implementation of multi-class object detection model using either MATLAB or Python programming language.

References

[1] Lincheng, J., Fan, Z., Fan, L., Shuyuan, Y., Zhixi, F. & Rang, Q. (2019). A survey of deep learning-Based object detection. *IEEE Access Multidisciplinary.* 7, 128837-128867.

[2] Feng, H., Sui, W., Huang, C., Xu, L. & Ki, A. (2019). Water body extraction from very high resolution remote sensing imagery using deep U-net and super-pixel based conditional random field model. *IEE Geoscience. Remote Sensing Letter*, 16 (4), 618-622.

[3] Li, Y., Xu, Rao, J., Guo, L. L., Yan, Z., & Jin, S. A. (2019). Y-Net deep learning method for road segmentation using high-resultion visible remote sensing image. *Remote sensing letter* 10, 381-390.

[4] Guoji W., Wu, M., Wei, X. & Song, H. (2020). Water identification from high resolution remote sensing image based on multidimensional densely connected convolutional neural networks, *Remote Sensing,* 12 (5), 795.

[5] Mengya, L., Penghai, W., Biao, W., Honhlyun, P., Hui, Y. & Yanlan, W. (2021). A Deep learning method of water body extraction from high resolution remote sensing images with multisensors. *IEEE Journal of Selected Tropics in applied Earth observation and remote sensing,* 14, 3120-3132.

[6] Gao, L., Song, W., Dai, J., & Chen, Y., (2019). Road extraction from high-resolution remote sensing imagery using refined Deep Residual Convolutional neural network *Remote sensing*. 11, 553.

[7] Alexander, A. S. G., Ilma, A. & Edy, I. (2020). Semantic segmentation of Aerial imagery for road Extraction with deep learning, ICIC Express letter, 14 (1), 43-51.

[8] Ji, S., Wei, S., & Lu, M. (2019). A scale robust convolutional neural network for automatic.

[9] Geesara, P. & Ilya, A. (2018). Deep Learning Approach for Building Detection in Satellite Multispectral Imagery, *IEE International Conference on Intelligent Systems* Sep. 2018.

[10] Nahhas, F. H., Shafri, H. Z., Sameen, M. I. Pradhan, B. & Mansor, S. (2018). Deep learning approach for building detection using LIDAR-orthophoto fusion. *Journal of Sensor* 3 (6), 1-9.

[11] Arshitha, F. & Biju, K. S. (2020). Accurate detection of building from satellite images using CNN. *In Proceeding of the 2nd International Conference on Electrical, Communication and Computer Engineering (ICECCE) 12-13 June 2020*, Istanbul, Turkey.

[12] Vakalopoulou, M., Karantzalos, K., Komodakis, N., & Paragios, N. (2015). Building detection in Very high-resolution multispectral data with deep learning features. *In Geoscience and Remote Sensing Symposium (IGARSS), IEEE International* 1873-1876.

[13] Li, S., Yuqi, T. & Liangpei, Z. (2017). Rural Building Detection in High-Resolution Imagery Based on a Two-Stage CNN Model, *IEEE Geoscience and Remote Sensing Letters*, 14 (11), 1998-2002.

[14] Ma, L., Liu Y., Zhang, X., Ye, Y. and Honson, J. (2019) Deep learning in remote sensing application a meta-analysis and review. I*SPRS Journal of Photogrammetric and remote sensing magazine* 152, 166-177.

[15] Li, Y., Zhang, H., Hue, X., Jiang, Y. and Shen, Q. (2018) Deep learning for remote sensing image classification a survey. *Wiley Interdisciplinary review data mining and knowledge discovery* 8 (6) 1264.