Building Extraction from Remote Sensing Images Using Deep Learning Techniques

Harsha Chaudhary[1*], Harshit Tyagi[2], Kshitij Bajpai[3]

[1,2,3]*Student, Department of computer Science and Engineering, Galgotias College of Engineering and Technology, Greater Noida, India*

*Corresponding author: theheroisbackagain@gmail.com

***Abstract*: An approach has been proposed for building extraction from remote sensing imagery using Deep Learning techniques i.e. transfer learning based semantic segmentation and customized decoder network. This approach uses spatial properties of an image scene for building detection. We have implemented our own decoder on a pre-trained model of Microsoft called RESNET, as an encoder. The architecture takes input satellite images of 512x512 resolution and predicts a building mask of the same resolution. The efficiency of this architecture is calculated upon the accuracy, precision, recall, F1-score and mean intersection over union. Also, this architecture will be compared with other state-of-art deep learning techniques for building extraction from satellite images.**

***Keywords*: PSP NET, Convolutional Neural Networks, Deep Learning, Image classification, ImageNet, Machine Learning, Semantic segmentation.**

1. Introduction

Automatic extraction of objects that are fabricated from various remote sensing images can be something very useful for present world. Extracting objects from various images can be little bit typical task but it has many applications. While working for this project we would be dealing with high-level algorithms and with various high-level interactive approaches to extract objects like roads, rivers, forests, buildings and much more.

An approach has been proposed for building extraction from remote sensing imagery using Deep Learning techniques i.e. transfer learning based semantic segmentation and customized decoder network. This approach uses spatial properties of an image scene for building detection. We have implemented our own decoder on a pre-trained model of Microsoft called RESNET, as encoder. The architecture takes input satellite images of 512x512 resolution and predicts a building mask of same resolution. The efficiency of this architecture is calculated upon the accuracy, precision, recall, F1-score and mean intersection over union. Also, this architecture will be compared with other state-of-art deep learning techniques for building extraction from satellite images.

A Convolutional Neural Network (CNN) is a Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other. The pre-processing required in a CNN is much lower as compared to other classification algorithms. While in primitive methods filters are hand-engineered, with enough training, CNN have the ability to learn these filters/characteristics.

A CNN is able to successfully capture the Spatial and Temporal dependency in an image through the application of relevant filters. The architecture performs a better fitting to the image dataset due to the reduction in the number of parameters involved and reusability of weights. For our purpose, we used a Fully Convolutional Network instead of a CNN.

Semantic Segmentation means to classify each pixel within an image i.e. we will have a label for each pixel in an image.

Similar to what us humans do all the time by default, when are looking then whatever we are seeing if we think of that as an image then we know what class each pixel of the image belongs to. Essentially, Semantic Segmentation is the technique through which we can achieve this in Computers.

We implement our own decoder as FCN for semantic segmentation of the remote images using a pre-trained RESNET model as encoder.

2. Literature Survey

Automatic building extraction from various images is both scientifically challenging and of major practical importance for data acquisition and update of Geographic Information System (GIS) databases or site models. Remote sensing data include both air and space borne data that may vary in many different aspects some of which are spatial, radiometric, spectral, and temporal resolutions.to apply a better method for extraction, complete knowledge of data is essential, which is measured on different factors. the user’s need, scale and characteristics of a study area, availability of various image data and their characteristics, cost and time constraints and the analyst’s experience in using the selected image are some factors to begin with [1].

Literature review reveals a great deal of application and approaches used for feature extraction. Many approaches including automatic and semi-automatic methods were proposed for building extraction. This literature review summarizes major events in the field of object extraction. This literature review has two sections. The first section focuses on primitive building extraction methods that have been used for a long time. The next section takes one step ahead and focuses on the methods that use deep learning for image segmentation.

Early methods for image segmentation include thresholding the image which is proposed by Salem and Kalyankar (2010). thresholding is a common approach which divides the image gray scale information processing based on the gray value of different targets. this approach can be divided in local threshold and global threshold [2]. Salem and kalyankar used different algorithm like mean, Histogram Dependent Technique (HDT), Edge Maximization Technique (EMT) and many more for the purpose [3]. The most commonly used for threshold segmentation is largest interclass variance method [4].

Lin and Nevatia (1998) used edge detection-based technique. In this method edge of the object is detected by its discontinuous behavior. there is always a gray edge between two adjacent regions with different gray values in the image and this discontinuity can often be detected derivatives [5]. Lin and nevatia technique construct 3D model for the purpose but can only detect rectilinear structures [6].

Wei, Zhao, and Song (2004) has proposed a supervised clustering and edge detection based technique.in clustering technique pixels are clustered based on seed point chosen randomly. Wei, Zhao, and Song used HRS QUICKBIRD panchromatic imagery that has shadow evidence. However, the technique failed in extracting any data of buildings having small or no shadow [7]. This problem was solved by jinn and Davis (2005). they used spectral information, structural and contextual details to come up with a technique where they can distinguish buildings from parks and roads [8].

Lefevre, Weber, and Sheeran (2007) used Advanced morphological operators, like Hit or Miss transformation with varying size and shape. They worked on HRS QUICKBIRD panchromatic imagery and their methodology acquired 88% precision rate with 63% of kappa value [9].to improve further they introduced Morphological Shadow Index (MSI) and Morphological Building Index (MBI) in 2012 [10].

Huang, Lu, and Zhang (2014) has proposed multi- index learning (MIL) method for HRS images. To improve classification results over urban areas they used set of indices such as MBI, MSI, and Normalized Difference Vegetation Index (NDVI) [11]. Huang et al. (2016) introduced Generalized Differential Morphological Profile (GDMP), which proved useful and advantageous over traditional Differential Morphological Profile (DMP) [12].

Dey et al. (2011) have proposed a context-based multi-level segmentation method. their methodology uses shadow object geometry with an accuracy of 71%.use of pan-sharpened and Multi Spectral (MS) GeoEye-1 imagery have shown that it cannot extract buildings having little or no shadow, although, these buildings have the considerable area on the ground [13].

The use of deep learning in image segmentation and classification gets challenging with the introduction of high-resolution imagery. So to segment we add as many hidden layers and can modify each layer according to our requirement, this practice has boosted the need for deep learning in image segmentation.

The first model was LeNet-5, proposed in 1998, a pioneering 7-level convolutional network by LeCun et al. it can classify digits from a handwritten number and being used in several banks [14]. Baeda and felea used the oldest model where each convolution layer were followed by pooling layer.in the implementation each layer ended with Relu layer.

Chen and Yan (2013) studied other architectures like NiN. Like the rest of CNN, this architecture consists of multiple blocks containing convolutional and pooling layers. The difference occurs in the usage of multilayer perceptron between the two layers of block. Its role is to act as a nonlinear function approximator that can augment the network's abstraction capability [15].

Manoj and neelima (2012) used Alex Net to classify four data sets. This significantly outperformed all the prior competitors in image segmentation. This network is very similar in architecture with LeNet but the difference was deeper, with more filters per layer and with stacked convolutional layers. It consisted of convolutions, max pooling, dropout, data augmentation, ReLU activations, SGD with momentum. ReLU activation layer was attached after every convolutional and fully-connected layer [16]. Alex Net was designed by the SuperVision group, consisting of Alex Krizhevsky, Geoffrey Hinton, and Ilya Sutskever [19].

In 2014 GoogLeNet (Inception V1) was proposed by Google. in ILSVRC 2014 It achieved a top-5 error rate of 6.67%! this accuracy was very close to human level performance.so as to beat its accuracy even humans were needed in numbers. This network uses CNN novel elements which are dubbed an inception module, use of batch normalization, image distortions and RMSprop. This module is based on several very small convolutions in order to drastically reduce the number of parameters. Their architecture consisted of a 22-layer deep CNN but reduced the number of parameters from 60 million (AlexNet) to 4 million [17].

Xiaolong Liu, Zhidong Deng, Yuhan Yang (2019) reviewed many methods like FCN, DNN, and others like Residual Neural Network (ResNet) which was proposed by Kaiming He et al[21]. it introduced a novel architecture with “skip connections” and features heavy batch normalization. in his paper they considered and defined a building block as

$$y = F(x, \{Wi\}) + x$$

the x and y are the input and output vectors of the layers considered. The function $F(x,\{Wi\})$ represents residual mapping [18].

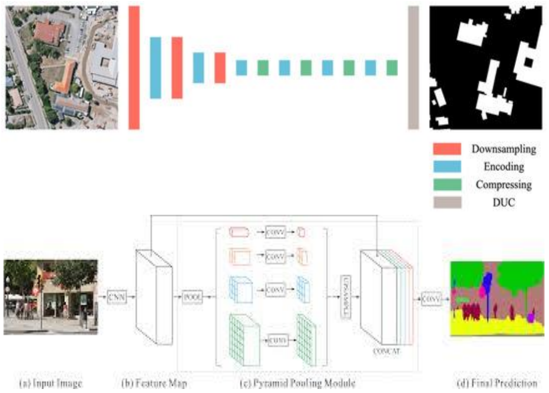
3. Methodology

In this paper we have used pspnet for semantic segmentation.

The pspnet architecture takes into account the global level predictions while fcn was not able to classify/capture the context of whole image in data sets like PASCAL VOC 2012.Most semantic segmentation modules contain one encoder which is responsible for extracting features and decoder which predicts the class of the pixel at the end. While the pspnet encoder contains CNN backbone with dilated convolutions and pyramid pooling module [20].

The replacement of traditional convolutional layers from dilate convolution layer helped in increasing the receptive field. The features received at the end were richer in content. The dilation value specifies the sparsity while doing convolution.in pspnet dilation values are 2 and 4 respectively. Pyramid pooling is the main part of this model. It helps to capture the global context in the image. The feature map which is the backbone of the model is pooled at different sizes before passing through a convolution layer and after which up sampling takes place.to pass to the decoder all the up sampled maps are concatenated to aggregate all features. After all the features are extracted out from the encoder it is the turn of decoder to take those features and convert them into predictions by passing them into its layers.to get the high resolution output from the model use of feature pyramid network (FPN) decoder is done which is similar to U-Net [22].

A. *Method Design*



B. *Dataset*

WHU building dataset. The aerial dataset consists of more than 220, 000 independent buildings extracted from aerial images with 0.075 m spatial resolution and 450 km2 covering Christchurch, New Zealand. The satellite imagery dataset consists of two subsets. One of them is collected from cities all over the world.

The other satellite building sub-dataset consists of 6 neighboring satellite images covering 550 km2 on East Asia with 2.7 m ground resolution.

One used here is collected from cities over the world and from various remote sensing resources including Quick Bird, Worldview series, IKONOS, ZY-3, etc. It contains 204 images (512 × 512 tiles with resolutions varying from 0.3 m to 2.5 m). Besides the differences in satellite sensors, the variations in atmospheric conditions, panchromatic and multispectral fusion algorithms, atmospheric and radiometric corrections and season made the samples suitable yet challenging for testing robustness of building extraction algorithms.

C. *Evaluation Matrices*

Intersection-over-union, precision and recall and F1 score (Dice-score) were used at pixel level to evaluate the performance of the model as they have been shown to perform well in segmentation jobs.

IoU (Intersection over Union): it is also known as the 'Jaccard index' and is one of the most straightforward and effective evaluation matrices. IoU is defined as the area of overlap between the predicted segmentation and the ground truth divided by the area of union between the predicted segmentation and the ground truth.

$$IoU \ = Area\ of\ intersection\ or\ overlap/Area\ of\ Union$$

There are four classifying conditions: true prediction on a positive sample (TP), false prediction on a positive sample (FP), true prediction on a negative sample (TN) and false prediction on a negative sample (FN).

$$precision \ = TP/(TP + FP)$$
$$recall \ = TP/(TP + FN)$$

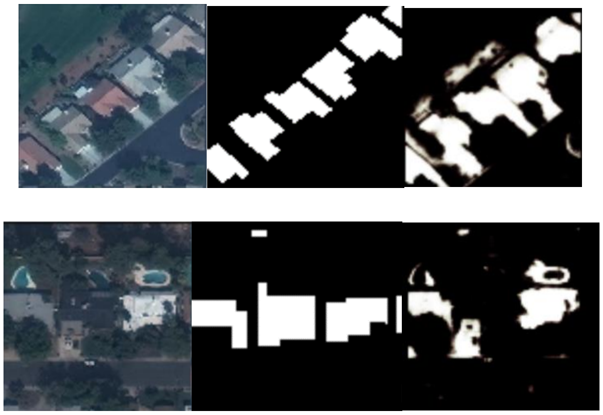
Precision represents the percentage of TP in total positive prediction and recall indicates the percentage of TP over the total positive samples.

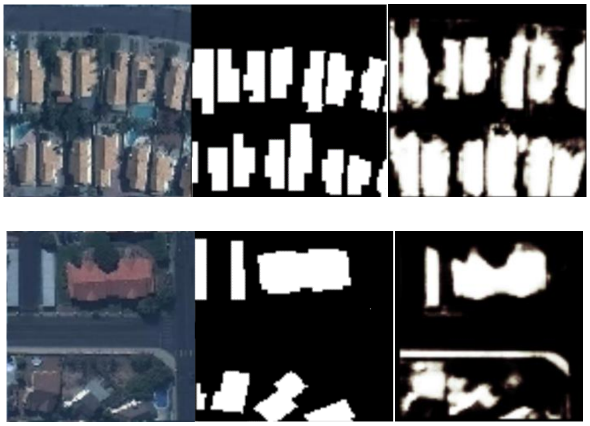
$$F1\ score \ = 2 * precision * recall/(precision + recall)$$

The F1- score is the weighted average of precision and recall, which considers both FP and FN.

4. Processed Images

Images shown below are the test images from WHU building dataset (on the left), their corresponding ground truths (in the middle) and predicted mask (on the right).





5. Experimentation

This section shows how changing some parameters affects the performance of our model. Experimentation was carried out by changing learning rate and epochs.

Learning rate: 0.0003
Epochs: 20

Training loss	Validation loss	IoU	Precision	Recall	F1 score
0.36512	0.37855	82.8%	82.56%	89.4%	88.10%

Learning rate: 0.00003
Epochs: 50

Training loss	Validation loss	IoU	Precision	Recall	F1 score
0.406886	0.338357	78.22%	85.16%	84.14%	82.12%

6. Result

Our method is able to solve the problem of building extraction from remote sensing low resolution images, in this research a new way was proposed that used PSPNET and resnet50 to obtain required results. The values evaluation matrices are shown below.

Learning rate	IoU	precision	recall	F1 score
0.0003	82.8%	82.56%	89.44%	88.10%

7. Conclusion

Performing this experiment showed that using PSPNet as segmentation model and RESnet50 as base model can perform well on WHU building dataset and the values of various evaluation matrices have shown how accurate and efficient the model is.

Currently, our experiment is implemented for building extraction only and in future, it would be able to perform multi-class extraction tasks like land area, water bodies, roads etc.

References

[1] Giftlin, J.G, Jennica, D.S., & Tucker, S. (2017). A Survey Paper on Buildings Extraction from Remotely Sensed Images.

[2] Davis L S, Rosenfeld A, Weszka J S. (1975) Region extraction by averaging and thresholding [J]. IEEE Transactions on Systems, Man, and Cybernetics.

[3] Al-amri, Salem & Kalyankar, N. & Khamitkar, SD. (2010). Image Segmentation by Using Threshold Techniques. J. Compute.

[4] S. Yuheng and Y. Hao, “Image Segmentation Algorithms Overview,” 2017.

[5] C. Lin and R. Nevatia, “Building Detection and Description from a Single Intensity Image", CVIU 72:101-121, 1998.

[6] Senthil Kumaran N, Rajesh R. Edge detection techniques for image segmentation–a survey of soft computing approaches [J]. International journal of re- cent trends in engineering, 2009, 1(2): 250-254.

[7] Wei, Y., Zhao, Z., & Song, J., 2004, September. Urban building extraction from high-resolution satellite pan- chromatic image using clustering and edge detection. In Geoscience and Remote Sensing Symposium, 2004. IGARSS’04. Proceedings. 2004 IEEE International.

[8] Shan, J., & Lee, S.D. (2005). Quality of building extraction from IKONOS imagery. Journal of Surveying Engineering, 131(1), 27–32.

[9] Lefevre, S., Weber, J., & Sheeren, D. (2007, April). Automatic building extraction in VHR images using advanced morphological operators. In Urban remote sensing joint event, 2007.

[10] Huang, X., & Zhang, L. (2012). Morphological building/ shadow index for building extraction from high-resolution imagery over urban areas. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 5(1), 161–172.

[11] Huang, X., Lu, Q., & Zhang, L. (2014). A multi-index learning approach for classification of high-resolution remotely sensed images over urban areas. ISPRS Journal of Photogrammetry and Remote Sensing, 90.

[12] Huang, X., Han, X., Zhang, L., Gong, J., Liao, W., & Benedicts, J.A. (2016). Generalized differential morphological profiles for remote sensing image classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 9(4), 1736–1751.

[13] Dey, V., Zhang, Y., & Zhong, M., 2011, August. Building detection from pan-sharpened -+ GeoEye-1 satellite imagery using context based multilevel image segmentation. In Image and Data Fusion (ISIDF), 2011 International Symposium on (pp. 1–4). IEEE.

[14] M. Badea, I. Felea, C. Vertan, L. Florea, The Use of Deep Learning in Image Segmentation Classification and Detection, Computer Vision and Pattern Recognition 62(5), 2016, pp. 187–198.

[15] M. Lin, Q. Chen, and S. Yan. Network in network. arXiv:1312.4400, 2013.

[16] Krishna, M & Neelima, M & Mane, Harshali & Matcha, Venu. (2018). Image classification using Deep learning. International Journal of Engineering & Technology.

[17] Szegedy, Christian & Liu, Wei & Jia, Yangqing & Sermanet, Pierre & Reed, Scott & Anguelov, Dragomir & Erhan, Dumitru & Vanhoucke, Vincent & Rabinovich, Andrew. (2015). Going deeper with convolutions. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 1-9.

[18] X. Liu, Z. Deng, and Y. Yang, Recent progress in semantic image segmentation, Artif. Intell. Rev., vol. 52, no. 2, pp. 1089–1106, 2019.

[19] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In NIPS, 2012.

[20] Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR, 2017.

[21] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In CVPR, 2016.

[22] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In MICCAI. Springer, 2015.