Types of small width feature extractor that can be used by CNNs to improve image segmentation

Honours Literature Review

Charl Ritter

RTTCHA002

Department of Computer Science

University of Cape Town

RTTCHA002@myuct.ac.za

ABSTRACT

With the surge in popularity of Artificial Intelligence and Neural Networks, there exists an interest to investigate whether the use of Artificial Intelligence and more specifically Convolutional Neural Networks could improve the performance of image segmentation techniques and how the different types of Convolutional Neural Networks could be optimised to improve the performance of image segmentation even further. From previous research in the field, Convolutional Neural Networks were found to perform slightly better than Random Tree Models. Due to the findings of this research, more investigation needs to be put into Convolutional Neural Networks. There are three types of Convolutional Neural Networks that are explored in the paper, namely U-Nets, Fully Convolutional Networks and Atrous Convolutional Networks. Thereafter, two different types of small width feature extractors are examined, namely Edge detectors and Mean shift. Small width feature extractors are one type of extractor inputs that can provided to a Convolutional Neural Network. It focuses on the smaller details in an image, but not so small that it's on a pixel to pixel basis. The other two types of feature extractors that will be looked at in the larger paper is per-pixel transforms and large-scale feature extractors. Finally, it was concluded that each type of Convolutional Neural Network can outperform the rest in certain cases, that U-Nets and Atrous Convolutional Networks have a higher likelihood to outperform the rest in general and that there may be no optimal Small width feature extractor and that different models can be combined and expanded to improve their performance.

CCS CONCEPTS

- Machine Learning
- Image Processing
- Computer Vision

KEYWORDS

Image segmentation, Convolutional Neural Networks, U-Net, Fully Convolutional Network, Atrous Convolutional Network, Edge detectors, Mean shift.

1. INTRODUCTION

The development of image segmentation techniques has had a big impact on the world we live in. It has been used from identifying diseases in images of organic matter, to facial recognition. One field that it shows a lot of promise for is the farming sector. By using image segmentation, farmers can have drones survey their farms and relay information back to the them, on a plant to plant basis. Usually models such as Random Forest Models have been used to analyse the data, but with the surge in popularity of Artificial Intelligence and Neural Networks, there is interest in researching what Artificial Intelligence models could be used to perform image segmentation on imagery of the farmlands (orchards) and how these models could be optimised for imagery of the farmlands (orchards).

The topic of this paper is based off, and a partial continuation of, the findings made by Finnis J. in their paper "*Random Forest Classification of Tree Crops on Farming Land*" and Motsumi N. in their paper "*III-CNN: Image-to-Image Inception CNN for Pixel-Wise Segmentation to extract tree and tree boundaries*". Their experiments showed that Convolutional Neural Networks performed slightly better overall with regards to accuracy, precision and recall than a Random Forest model. Thus, it was concluded that using a Convolutional Neural Network would be better for image-to-image segmentations than a Random Forest model [4] [9].

This paper also leads into a larger paper, namely "*Tree segmentation by combining CNNs with engineered features*".

Convolutional Neural Networks are classes of deep neural networks that are most commonly used to analyse imagery. This is done by assigning importance to different objects in an image and the ability to differentiate one object from another. Convolutional Neural Networks make strong and mostly correct assumptions about the nature of images, which has been proven by Krizhevsky et al., in their paper "*ImageNet Classification with Deep Convolutional Neural Networks*". Their results showed that a large, deep convolutional neural network can achieve record breaking results on highly challenging datasets, using purely supervised learning [8].

Convolutional Neural Networks can take different engineered features as inputs, to improve their segmentations performance. In the "*Tree segmentation by combining CNNs with engineered features*" project, we will be analysing how these inputs can be used to improve the segmentations performance for tree data.

These feature inputs can be split into three different levels, namely per-pixel transforms, small width feature extractors and large-scale feature extractors. Per-pixel transformations are modifications that only look at one pixel at a time, like colour space transforms and decompositions. Small width feature extractors transformations are modifications that look at smaller objects in an image or parts of an image, like Edge detectors and Mean shifts. Large scale feature extractors transformations are modifications that looks at larger objects in an image, parts of an image or the whole image, like Hough transformations and Template matching.

For the work being done in this literature review, three different types of Convolutional Neural Networks will be explored. These three types are Fully Convolutional Networks, U-Nets and Atrous Convolutional Networks. Thereafter, two different types of small width feature extractors will be defined and explored, namely: Edge detectors and Mean shifts.

2. TYPES OF CONVULUTIONAL NEURAL NETWORKS

2.1. Fully Convolutional Networks

2.1.1. Overview

A Fully Convolutional Network is defined by Guo et al., as an extension of the Convolutional Neural Network model, where the basic idea is to make the Convolutional Neural Network take an input of arbitrary-sized images [6]. The main difference between a Fully Convolutional Network and a Convolutional Neural Network, is that the last fully connected layer is substituted by another convolution layer with a large receptive field.

2.1.2. Characteristics

One of the drawbacks of Fully Convolutional Networks is that because of the fixed nature of the size of the receptive field, if an object is substantially larger or smaller than the size of the receptive field, it could be mislabelled or fragmented [10]. Another drawback is that because of all the convolutions and pooling layers that the data goes through, the resolution of the feature map is down-sampled. This leads to low resolution predictions, which in turn leads to fuzzy object edges [6].

2.2. U-Nets

2.2.1. Overview

The U-Net model is built on a Fully Convolutional Network model that has been modified to yield better image segmentation. It was originally designed to segment medical imagery. The name "U-net", comes from the shape of the U-net architecture, as can be seen in figure 1.

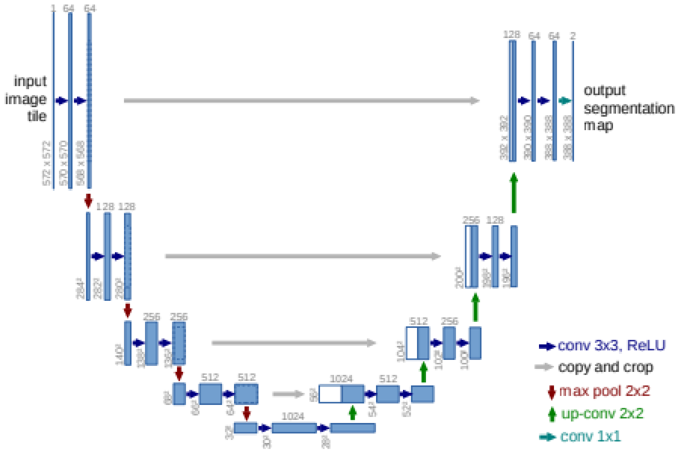

Figure 1: U-net architecture.

This architecture can be split into two parts. These two parts are the left half of figure 1, where down sampling occurs, and the right half of figure 1, where up sampling occurs. In the left half, the architecture follows a structure very similar to a Fully Convolutional Network. In the right half, up sampling of the feature map occurs, which entails the layers increasing the resolution of the output, which are then combined [13].

2.2.2. Characteristics

One of the advantages of using a U-net model, is that it utilises identity mapping, which makes it a lot easier to train [16]. Another advantage is that it combines low level detail information and high-level semantic information, which increases the performance of the model [16].

2.3. Atrous Convolutional Networks

2.3.1. Overview

Lastly, Atrous Convolutional Networks will be examined. It is defined by Chen et al., as allowing us "to extract denser feature maps by removing the down sampling operations from the last few layers and up sampling the corresponding filter kernels, equivalent to inserting holes between filter weights" [3].

2.3.2. Characteristics

One of the good characteristics of this type of Convolutional Neural Network, is that it allows one to configure the resolution at which feature responses are computed, without requiring any more parameters [3].

There are however two problems when it comes to the implementation of an Atrous convolutional network, namely signal down sampling, and spatial invariance. Signal down sampling occurs when the signal resolution gets reduced, due to the down sampling and max-pooling that occurs at every layer of the repetition. Spatial invariance occurs when object-centric decisions are fetched from a classifier, which requires invariance to spatial transformations. This in turn then greatly limits the spatial accuracy of the model [2].

3. TYPES OF SMALL WIDTH FEATURE EXTRACTORS

3.1. <u>Edge detectors</u>

3.1.1. Motivation

The first type of small width feature extractor that will be discussed is Edge detectors. In their paper Abdou and Pratt, defines Edge detectors as “primitive features of an image that are widely used in image classification and analysis systems to outline the boundaries of objects” [1]. They also define an image edge, as “a local change or discontinuity in image luminance” [1].

The idea behind Edge detectors stems from the assumption that boundaries of objects are manifested in distinguishable, extreme changes between different objects. This approach is a small feature extractor, because it does not look at images on a pixel to pixel bases, but rather looks at a part of the image to identify where the edges are of the objects on the image. This works well, because it can be hard to identify edges on a pixel to pixel bases, as not all edges have extreme enough changes on such a small scale.

According to Heath et al., there are four well-known edge detectors, namely Canny, Nalwa–Binford, Sarkar–Boyer, and Sobel [7]. Of these four edge detectors, both Canny and Sobel are also discussed in the articles by Gonzalez et al. and, Petrou and Kittler [5] [12].

3.1.2. Presentation

In Heath et al.’s, work, the researchers were comparing different types of Edge detectors with each other. They first performed an experiment, to determine the edge detector parameter settings. The second experiment conducted by these researchers was to compare the ratings of these edge detectors.

These two experiments lead the researchers to make three major observations. Firstly, they observed that there are statistically significant differences between the ratings of the edge detectors. The average rating values for the detectors lied in a relatively small range on a 7-point scale. Thus, while they could measure progress in the quality of the edge detector output, there is substantial room for further improvement. Secondly, the optimal parameter settings of an edge detector are strongly dependent on the image. There is a need however for strategies of adaptively choosing the parameters of the edge detectors based on domain and image characteristics. Thirdly, the relative performance of the edge detectors varied statistically significantly across the images. This seems to indicate that there is something about each of the edge detectors that makes it the best option for some type of image. This is contrary to the assumption that edge detection is a context independent, purely bottom-up process. This suggests that it may be worthwhile to incorporate context information into the edge detection process [7].

Gonzalez et al., wrote a paper about optimising fuzzy edge detectors based on the traditional Sobel technique, combined with interval type-2 fuzzy logic. By using the interval type-2 fuzzy logic in edge detection methods, the researchers wish to give the edge detectors the ability to handle uncertainty in processing real world images.

The researchers ran multiple simulations, where the researchers applied Cuckoo Search Algorithms and Genetic Algorithms to optimize the antecedent design parameters of optimised type-2 fuzzy logic systems. The researchers found that when these algorithms were applied, the optimised type-2 fuzzy logic systems, achieved similar results. However, the results achieved by non-optimized type-2 fuzzy logic systems, optimized type-1 fuzzy logic systems, non-optimized type-1 fuzzy logic systems and traditional Sobel were improved. Cuckoo Search also has the advantage over Genetic algorithms, of having less parameters that need customisation [5].

In their work Petrou and Kittler, attempted to find the optimal mathematical function for edge detectors. In their paper they found that the function depicted as function 1, models real image edges well, that the Gaussian filter’s performance varies significantly and is worse than the performance of the optimal filters for ramp edges, that the performance of the Gaussian filter improved as edges became more step like (this does not mirror real world images), and that some optimal filters for noisy and blurred images were better than the optimal filter for ideal step edges and much better than the commonly used Gaussian filter [12].

$$c(x) = \begin{cases} 1 - e^{-sx}/2 & \text{for } x \geq 0 \\ e^{sx}/2 & \text{for } x \leq 0 \end{cases}$$

Function 1: A mathematical function, found that models real image edges well

3.1.3. Discussions

Both Heath et al. and Gonzalez et al., attempted to find optimal parameters for the respective types of edge detectors that they were researching. Heath et al., found that, in general, there isn’t a single set of optimal parameters, but rather that the optimal parameters will differ from image to image and that an effort should rather be made to improve the process of adaptively choosing parameters for the specific image, rather than finding some optimal case for all. Gonzalez et al., focused on Sobel and fuzzy logic systems specifically, for images and did find parameters that improved the results of all the antecedents of an optimised type-2 fuzzy logic systems. Gonzalez et al., also attempted to optimise the parameters with Cuckoo Search Algorithms and Genetic Algorithms, concluding that their performance is similar, although Cuckoo Search has less parameters that need to be set, making it preferable.

These findings point to the conclusion that parameters will need to be optimised for every image individually. This being the case, it would be better to use algorithms that require less parameters.

Petrou and Kittler demonstrated the importance of keeping work grounded in the real world. In their work, they found that older filters such as the Gaussian filter only works well on well-defined step edges and that filters need to be able to account for edges that aren't well defined. Heath et al., did note in their related work section, that results based on synthetic images have a limited usability.

Thus, researchers agree on the statement, that the images used in the research needs to be grounded in reality, otherwise the results may not be as useful.

3.1.4. Comparison of existing work

One of Petrou and Kittler's conclusions was that edge detectors perform a lot better on edges that change stepwise, such as images with clear borders or synthetic images. Yet, these do not mirror real life images. Gonzalez et al. however, didn't run any simulations on real images, instead they only used synthetic images. They do acknowledge this in their paper and that they would like to include it in their future work, but it is unfortunate that they did not include simulations, run on real images, to ensure that their findings translate to the real world.

There is also a concern expressed by Heath et al., on how they used a different number of parameters on the different types of edge detectors. They are concerned that this may have biased their results, but that they chose to keep it as it would better reflect the real world. The other two papers, however, would suffer from their same potential bias, but their writers made no acknowledgement on these potential biases.

3.2. Mean shifts

3.2.1. Motivation

The second type of small width feature extractor that will be discussed is Mean shifts. In their paper, Wang et al., defines mean shifts, as a nonparametric algorithm that starts at every data point, estimating the local density of similar pixels [15]. They also state that "traditional mean shift-based segmentation uses a radially symmetric kernel to estimate local density which is not optimal in view of the often-structured nature of image" [15].

In more detail, how this image segmentation method works, is that it estimates the local density gradient of similar pixels. In an iterative process, the algorithm then searches for peaks in the local density, using the gradient estimates it produced. Pixels are then grouped in segments together, depending on if they were "drawn upwards" to the same peak [15].

3.2.2. Presentation

In their work, Zheng et al., looked at mean-shift-based colour segmentation of images containing green vegetation. Their goal was to improve the segmentation rate of images containing green vegetation by introducing a mean-shift procedure.

For their work they created an algorithm that combined a back-propagation neural network and mean shift. Each pixel of the image would be represented by six features, and then a mean-shift procedure and a BPNN were used to complete the image segmentation. They tested their proposed algorithm with 100 different images and reported a median of mis-segmentation of about 4.2%. Their method did however have to run for a long time, meaning it would not be suitable for real time segmentation [17].

Another paper, by Tao et al., explored colour image segmentation based on mean shift and normalised cuts. Their goal was to create a new algorithm that incorporated both mean shift and normalised cuts, that doesn't require much computation power, thus making it more feasible for real-time image segmentation.

Their proposed algorithm took the advantages of both the mean shift and normalised cuts segmentation methods, eliminating many of the drawbacks of the two methods on their own. Using the mean shift method allows the formation of segments that preserves the discontinuity characteristic of an image. Then when the region adjacent graph and normalised cuts methods are applied to the resulting segments, instead of directly to the pixels of the image, greatly improves the segmentation performance. Their proposed method also requires significantly less computational complexity and, thus is feasible to real-time image processing [14].

In the last paper being analysed, Paris and Durand, looked at creating a topological approach to hierarchical segmentation using mean shift. Their goal was very similar to the goal set by Tao et al. They also created their own algorithm to decrease the computational power required.

Firstly, by recasting the process in Morse theory, they showed that a hierarchical structure can be computed at a negligible cost. They also observed that their algorithm was as precise as previous work in the field, yet significantly faster [11].

3.2.3. Discussions

In Zheng et al.'s work, they achieved a very low percentage of mis-segmentation, with only 4.2% getting mis-segmented. However, they also reported that using their algorithm to achieve this low mis-segmentation, takes a very long time, making it unsuitable for actual use in a real-world situation. Tao et al. and, Paris and Durand on the other hand had very different results with the algorithms they devised. Neither of these papers experienced an accuracy quite like the algorithm in Zheng et al.'s work. This can be seen for example in Paris and Durand's work, where they had various test images form the Berkeley database, that performed well on colourful images, but struggled to segment camouflaged objects. They found that their algorithm achieved an F value of 61, which is a bit higher than the standard. It did however achieve this much faster than the normal algorithms, without sacrificing accuracy [11]. Paris and Durand did also note that their algorithm was not trained and that the same parameters were used for all the images. Tao et al., reported very similar findings for their algorithm. They reported that due to the combination of mean shift and normalised cuts, the algorithm achieved improved accuracy, as it's based on region nodes versus pixel nodes, and offered a considerable reduction of computational complexity, since the number of basic image entities is far smaller than that of the pixels.

This leads to the questions of; 1) if traditional mean shift algorithms are better to use versus a combination of algorithms and 2) if it's better to have a highly accurate algorithm, that has a high computable complexity, versus having a less accurate, but highly usable algorithm? From the papers evaluated it's clear that a combination of algorithms can improve on speed and accuracy, over just using a traditional mean shift algorithm. The papers also lead to the conclusion that it is better to have a less accurate algorithm, that can be used in real time, as, although the inverse is important for research purposes, if an algorithm takes to long to produce meaningful results, it isn't worth much.

3.2.4. Comparison of existing work

With Tao et al. and, Paris and Durand did very similar work, it's interesting to see how they approached their respective papers differently. Both papers came from American institutes and were published in the same year, thus coming from a very similar point of view. Tao et al. however, seemed to have gone into much more detail regarding the math behind their algorithm and went into detail on the outputs of their simulations and how this was impacted by the region nodes versus pixel nodes. They made conclusions regarding the speed and accuracy of their algorithm, but sadly did not go into much details regarding it or making comparisons with other similar algorithms. There were some details on the speed, but the accuracy details were very bare. Paris and Durand on the other hand went into less details on the math of their algorithm but went into more detail when it came to the algorithms speed and accuracy versus other works and algorithms. They also wrote about related work, which drastically helps ground their work in the field.

4. CONCLUSIONS

After exploring the different types of Convolutional Neural Networks, a few conclusions can be made. Although it must be reiterated that these conclusions are not based on any experimentation done in this paper, but rather by reviewing existing literature regarding the different types of Convolutional Neural Networks.

The first conclusion that can be made is that every type of Convolutional Neural Network has its own set of advantages and disadvantages. This means that there exists a possibility that although some of the types of Convolutional Neural Networks may usually underperform the rest, there may be specific cases in which they outperform the rest. This will however be investigated and determined in the future work on this project. Another conclusion that can be made is that the U-Net model and the Atrous Convolutional Network model have a higher possibility of outperforming the Fully Convolutional Network model and the Convolutional Neural Network model. This is because the U-Net model and the Atrous Convolutional Network model are both built on and improvements of the Fully Convolutional Network model, which in turn is built on and an improvement of the Convolutional Neural Network model.

As for the two types of Small width feature extractors that were examined, the following conclusions were made.

In the case of Edge detectors, it was observed that there is no optimal set of parameters that will work for all images. The optimal set of parameters would differ for every image, meaning that instead of investing research into finding optimal parameters for all images, research should rather be made into how to effectively choose optimal parameters for every image and that possible improvements on the Edge detector models that allows for a similar speed and accuracy of the model, while needing less parameters.

In the case of Mean shifts, there was a plethora of researchers changing and expanding on the Mean shift model to get better results. It was apparent however that the usefulness of some high accuracy models is very poor as they are computationally very taxing, making them hard to use in real time. Models that chose to emphasize speed over accuracy seemed much more promising. These models all had an accuracy that was considered standard for a Mean shift model, but drastically increased the speed at which these models run, making them usable for real time.

Thus, when it comes to the types of Small width feature extractors, although the standard models should be tested, it may be worthwhile to combine, build on and expand these models, as this may prove to be more beneficial.

REFERENCES

[1] Abdou, I. and Pratt, W., 1979. Quantitative design and evaluation of enhancement/thresholding edge detectors. *Proceedings of the IEEE*, 67(5), pp.753-763. DOI:https://ieeexplore-ieee-org.ezproxy.uct.ac.za/stamp/stamp.jsp?tp=&arnumber=1455594

[2] Chen, L., Papandreou, G., Kokkinos, I., Murphy, K. and Yuille, A., 2018. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4), pp.834-848. DOI:https://ieeexplore-ieee-org.ezproxy.uct.ac.za/stamp/stamp.jsp?tp=&arnumber=7913730&tag=1

[3] Chen, L., Zhu, Y., Papandreou, G., Schroff, F. and Adam, H., 2017. Rethinking Atrous Convolution for Semantic Image Segmentation. *Rethinking Atrous Convolution for Semantic Image Segmentation*(2017), 1–14. DOI:http://dx.doi.org/arXiv:1706.05587

[4] Finnis, J., 2018. Random Forest Classification of Tree Crops on Farming Land. , pp.1-13. DOI:http://projects.cs.uct.ac.za/honsproj/cgi-bin/view/2018/finnis_motsumi.zip/deliverables/Final%20Paper%20(Final).pdf

[5] Gonzalez, C., Melin, P., Castro, J., Castillo, O. and Mendoza, O., 2016. Optimization of interval type-2 fuzzy systems for image edge detection. *Applied Soft Computing*, 47, pp.631-643. DOI:https://www.sciencedirect.com/science/article/pii/S1568494614006425

[6] Guo, Y., Liu, Y., Georgiou, T. and Lew, M., 2017. A review of semantic segmentation using deep neural networks. *International Journal of Multimedia Information Retrieval*, 7(2), pp.87-93. DOI:https://link.springer.com/content/pdf/10.1007%2Fs13735-017-0141-z.pdf

[7] Heath, M., Sarkar, S., Sanocki, T. and Bowyer, K., 1998. Comparison of Edge Detectors. *Computer Vision and Image Understanding*, 69(1), pp.38-54. DOI:https://www.sciencedirect.com/science/article/pii/S1077314297905877

[8] Krizhevsky, A., Sutskever, I. and Hinton, G., 2017. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), pp.84-90. DOI:https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[9] Motsumi, N., 2016. : III-CNN: Image-to-Image Inception CNN for Pixel-Wise Segmentation to extract tree and tree boundaries. , pp.1-12. DOI:http://projects.cs.uct.ac.za/honsproj/cgi-bin/view/2018/finnis_motsumi.zip/deliverables/mstnto015Final.pdf

[10] Noh, H., Hong, S. and Han, B., 2015. Learning Deconvolution Network for Semantic Segmentation. *2015 IEEE International Conference on Computer Vision (ICCV)*(2015). DOI:http://dx.doi.org/10.1109/iccv.2015.178

[11] Paris, S. and Durand, F., 2007. A Topological Approach to Hierarchical Segmentation using Mean Shift. *2007 IEEE Conference on Computer Vision and Pattern Recognition*(2007). DOI:http://dx.doi.org/10.1109/cvpr.2007.383228

[12] Petrou, M. and Kittler, J., 1991. Optimal edge detectors for ramp edges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(5), pp.483-491. DOI:https://ieeexplore-ieee-org.ezproxy.uct.ac.za/stamp/stamp.jsp?tp=&arnumber=134047

[13] Ronneberger, O., Fischer, P. and Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Lecture Notes in Computer Science Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*(2015), 234–241. DOI:http://dx.doi.org/10.1007/978-3-319-24574-4_28

[14] Tao, W., Jin, H. and Zhang, Y., 2007. Color Image Segmentation Based on Mean Shift and Normalized Cuts. *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, 37(5), pp.1382-1389. DOI:https://ieeexplore-ieee-org.ezproxy.uct.ac.za/stamp/stamp.jsp?tp=&arnumber=4305291

[15] Wang, J., Thiesson, B., Xu, Y. and Cohen, M., 2004. Image and Video Segmentation by Anisotropic Kernel Mean Shift. *Lecture Notes in Computer Science Computer Vision - ECCV 2004*(2004), 238–249. DOI:http://dx.doi.org/10.1007/978-3-540-24671-8_19

[16] Zhang, Z., Liu, Q. and Wang, Y., 2018. Road Extraction by Deep Residual U-Net. *IEEE Geoscience and Remote Sensing Letters*, 15(5), pp.749-753. DOI:https://ieeexplore-ieee-org.ezproxy.uct.ac.za/stamp/stamp.jsp?tp=&arnumber=8309343

[17] Zheng, L., Zhang, J. and Wang, Q., 2009. Mean-shift-based color segmentation of images containing green vegetation. *Computers and Electronics in Agriculture*, 65(1), pp.93-98. DOI:https://www.sciencedirect.com/science/article/pii/S0168169908001828