XL CILAMCE
IBERO-LATIN AMERICAN
CONGRESS ON
COMPUTATIONAL
METHODS IN
ENGINEERING
NOVEMBER
11-14, 2019
Praiamar Natal Hotel & Convention
Natal, RN-BRAZIL

MASK R-CNN APPROACH TO DETECT HEALTHY VEGETATION AREAS IN NIR IMAGES

Rhynner Hugo Richelly Silva Santos
Flávio Garcia Pereira
Daniel Cruz Cavalieri
rhynner.santos@gmail.com
flavio.garcia@ifes.edu.br
daniel.cavalieri@ifes.edu.br
Instituto Federal do Espírito Santo - IFES
Rodovia ES-010, km 6,5. CEP: 29173 - 087. Manguinhos, Serra - ES, Brazil

Abstract. With the development of new technologies in the field of deep learning, sectors such as agriculture have been benefited from the application of intelligent systems allied to the use of UAVs (Unmanned Aerial Vehicles) in crop monitoring to quickly and accurately detect specific areas of vegetation and optimize decision making to ensure the quality of planting. Some researchers inspired from the deep learning advance and its success in many areas have been studied solutions in detection of healthy vegetation areas, but, how better are the performances of advanced techniques compared to traditional techniques? In this context, this paper presents a comparison between a traditional technique (K-Means Clustering) and advanced technique (Mask R-CNN) applied to detect different vegetation areas in NIR images. The database of this work consists of NIR images provided by a modified RGB camera installed in a UAV. Basically, as an input were used NDVI (Normalized Difference Vegetation Index), an important index of vegetation healthy, obtained from the NIR images. Finally, a comparison between the proposed algorithms for detection of healthy vegetation areas is presented, showing the improvements of the proposed Mask R-CNN algorithm.

Keywords: Machine Learning, Deep Learning, Healthy Vegetation Index, Image Processing, Precision Agriculture

1 Introduction

With the development of new technologies, several sectors, such as agriculture, have been benefiting from the application of new methods to optimize their processes. Among these technological innovations, imaging crops monitoring has made it possible to identify crop variations more quickly and efficiently. In this concept, the term Precision Agriculture (PA), allied to remote sensing techniques, has been widely used with Unmanned Aerial Vehicles (UAVs) equipped with cameras (RGB, thermal, multispectral and hyperspectral) allowing large planting areas monitoring in a short time [1].

As observed in [1], the use of UAVs in agriculture has been intensified mainly with the use of embedded sensings, focused in reflectance spectroscopy analysis, that is, electromagnetic radiation (EMR) interaction measurement with surfaces (soil, vegetation, among others). Having knowledge of the incident energy magnitude and reflected energy on the area of interest, and that different EMR wavelengths have different surface interactions, it is possible to measure the planting health. Among the many types of spectra used in the planting properties analysis is NIR (Near Infrared).

According [1] NIR has been used to calculate vegetation indices, which may indicate planting physiological (structural and biochemical) characteristics such as vigor and water stress. Since a leaf cellular structure is made up of photosynthetic pigments, the use of vegetation indices, such as the NDVI (Normalized difference vegetation index), can lead to quantification of plant health.

In order to detect healthy areas in planting, along with NDVI, deep learning techniques such as convolutional neural networks (CNN) have been used to learn to discern between a healthy and deficient region, besides demarcating planting areas of interest [2]. Other papers make comparisons between techniques such as Mask R-CNN and classical methods proving efficiency in object detection and image segmentation [3]. As for applicability, in [4] the Mask R-CNN technique was used to classify and detect road damage through images collected by a smartphone, obtaining good results, which makes this technique one of the last generation algorithms capable of performing tasks quickly and effectively. Thus, this paper will use the NDVI to compare the healthy vegetation areas detection performance in NIR images, using K-Means and Mask R-CNN as a classical and modern clustering techniques respectively, showing some advantages and disadvantages of each method for the chosen application, analizing and discussing the results.

1.1 Problem definition

The problem addressed in this paper is based on the comparison of classical and advanced clustering techniques performance in the healthy vegetation areas detection through NIR images analisys collected by UAVs.

1.2 Motivation

The agricultural market has great competitiveness, having as one of the main factors the guarantee of the planting quality. Being able to know the crop state quickly and localized, coupled with intelligent systems for healthy areas detection or not, is directly linked to the estimation of crop productivity.

1.3 Goals

Performance comparison between classical and advanced techniques (eg, K-Means and Mask R-CNN, respectively) for healthy areas detection in planting by segmentation, feature extraction and areas of interest delimitation in images. Specific goals include:

- NIR image database creation;
- NDVI extraction through an algorithm based on NIR and RED layers;
- Analyse the results and discuss some advantages and disadvantages of each method for a chosen application.

2 Methodology

This chapter describes the equipments and materials used to do this work, along with the classical and advanced clustering techniques proposed for healthy areas detection.

2.1 Materials and methods

Was used an UAV (Phantom 3 Professional from DJI) with a camera whose internal optical filters were modified in order to transform the original RGB image standard to NGR. Thus, an open area (as shown in Fig. 1) from IFES (Instituto Federal do Espírito Santo) were recorded, aquiring short videos at 24 frames per second (fps), givin more than a thousand NIR vegetation images. Listed below are the other materials and equipments used:

- UAV (DJI Phantom 3 Professional model), with a modified camera.
- Laptop Intel i7 8GB RAM .
- Matlab 2019a.
- Google Colaboratory.

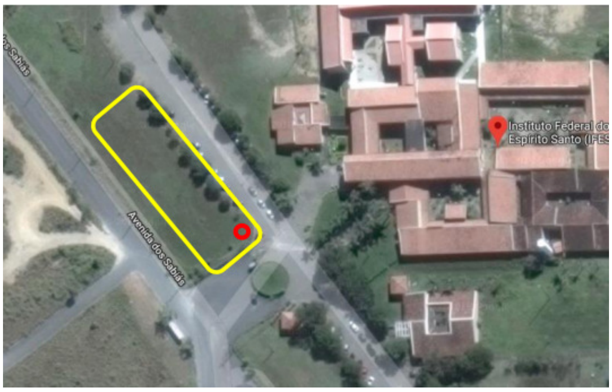


Figure 1. Mapped Area [5].

2.2 Healthy vegetation area detection using K-Means Clustering

In academic papers it is common to find several works using classic clustering methods in a large branch of applications. The K-Means technique is considered a classic method and this section will discuss about the work presented in [5], that used a Color-Based NIR images segmentation using k-means clustering. Figure 2 shows an summary flowchart discribing the main processes used in an automatic vegetation monitoring proposed by the author.

Firstly, the images were extracted from the video recorded by the modified camera of the UAV using Matlab 2019a software (as shown in Fig. 3). The implemented algorithm gives the user freedom to choose, for example, 3 images from the area to be analysed.

After choosing the images, a preprocessing is done that consists in image three channel extraction, as NGR. Then, the NDVI is calculated by the Eq. (1):

$$\mathrm{NDVI} = \frac{\mathrm{NIR} - \mathrm{Red}}{\mathrm{NIR} + \mathrm{Red}} \tag{1}$$

After applying a pixel-by-pixel NDVI calculation and, subsequently, the values normalization (between 0 and 1), a matrix treatment was performed in the case of values lower then 0, greather than 1 and, finally, undetermined values. In order to facilitate the healthy area detection NDVI matrix analisys, a false color image representation technique was implemented using the Modern LDP NDVI Scale, as shown in Fig. 4.

With the false color image, this NGR image was converted to a new color space known as L*a*b* [6]. Since color information exists in this new space, where each pixel has values ’a *’ and ’b *’ and,

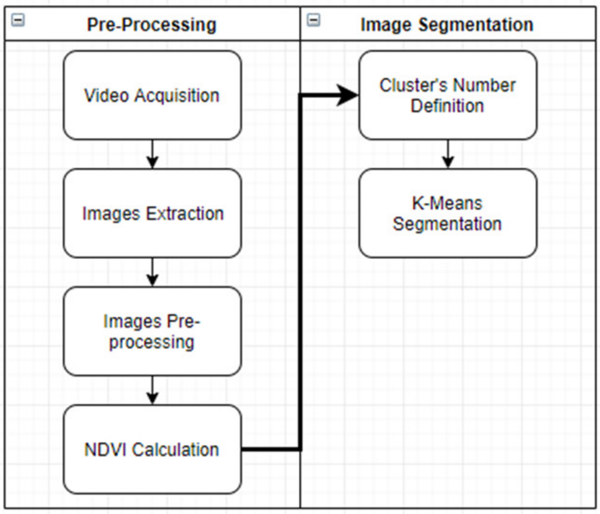


Figure 2. Interest area analysis process flowchart using K-Means.

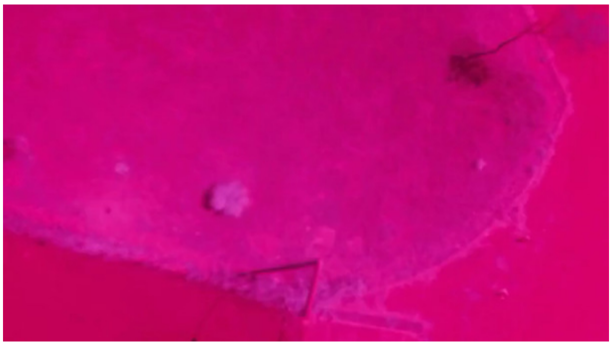


Figure 3. NIR image example.

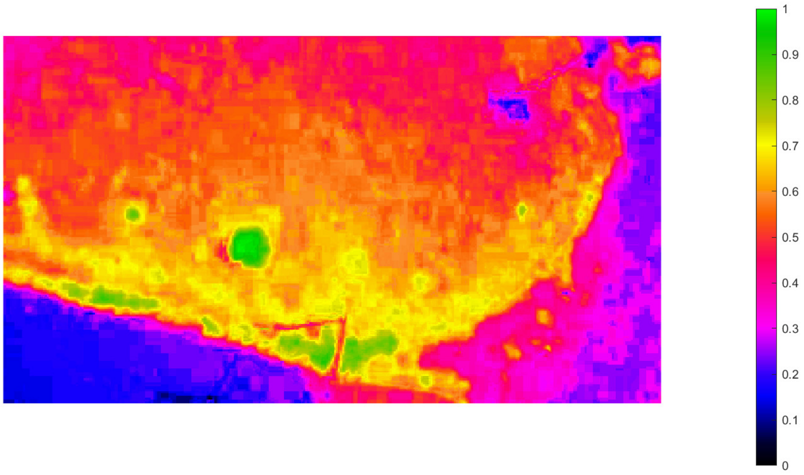


Figure 4. NIR false color representation using Modern LDP NDVI Scale [5].

as K-Means method basically treats each object as a point in space, choosing the number of groups (clusters), the euclidean distance is calculated from each object to its neighbors and vice versa.

2.3 Healthy vegetation area detection using Mask R-CNN

For the application of Mask R-CNN aproach, softwares such as Matlab R2019a and Google Collaboratory in python programing language were used, following as an example one of the practical works described by [7] that used the Mask R-CNN technique in python 3, Keras and TensorFlow to produce a model capable of generating bounding boxes and segmentation masks for each object in an image. One of the main reasons of Google Collaboratory choice was due to the excellent and free GPU available, making it possible to turn days into hours of training. In the sections below a brief explanation of Mask R-CNN technique will be described showing the step by step used to apply the technique to the problem addressed.

Mask R-CNN

Successor to the Faster R-CNN [8] technique (object detection model, whose term R-CNN comes from regional convolutional neural network), this method has the ability to segment object instances, ie differentiate at the pixel level, to which object they belong or not. In other words, given an input image, Mask R-CNN can output bounding boxes, classes, and masks for each object [3].

Basically, Mask R-CNN has, in parallel with the existing Faster R-CNN bounding box detection branch, a branch for predicting an object mask. Conceptually, Mask R-CNN can have its application divided into 2 stages, the first reading the image and generating candidate regions (areas that probably contain an object of interest), and a second stage that classifies the candidate and generates bounding boxes and object masks, both stages connected by a structure known as a backbone (a standard convolutional neural network, such as ResNet-101 and ResNet-50). Fig. 5 shows the Mask R-CNN basic structure.

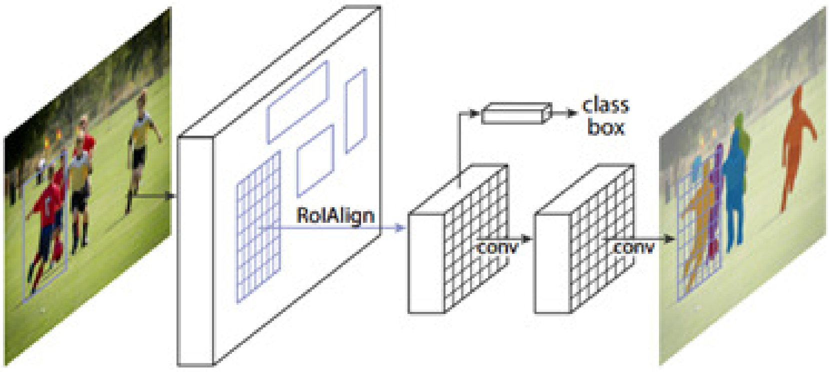


Figure 5. Mask R-CNN basic structure [3].

As mentioned earlier, in a very summarized way, Mask R-CNN model can be divided into two stages (as shown in Fig. 6), encompassing steps such as feature extraction, classification and bounding box regression, and segmentation masks generation.

In the first stage, while the ResNet-101 backbone (used in this paper) performs well generating a standard feature extraction pyramid, a feature pyramid network (FPN) layer is added to optimize this process, by adding a second pyramid with access to the high level features of the first pyramid and sending them to lower layers, enabling interaction between different feature levels.

Still in the first stage, a region proposal network (RPN) analyzes each level of the FPN, proposing regions that may contain objects. The regions analyzed by the RPN are called anchors, which are as a set of boxes with predefined locations and sizes relative to the images. Basically the RPN generates for each anchor the following outputs: Class (can be background or foreground, the latter being the probable class that contains the object) and bounding box to adjust its location and size, since more than one anchor can exist for same candidate object. If many anchors have high overlap, the one with the highest foreground score is maintained and the others are discarded, producing regions of interest (ROI).

In the second stage, a new neural network is applied to the regions of interest produced by RPN generating two other outputs for each region, being class (unlike the first stage, it has the ability to classify the regions into specific classes such as person, car, chair, etc.) and bounding box to refine the interest object encapsulation. Importantly, after adjusting the bounding box performed in the previous steps,

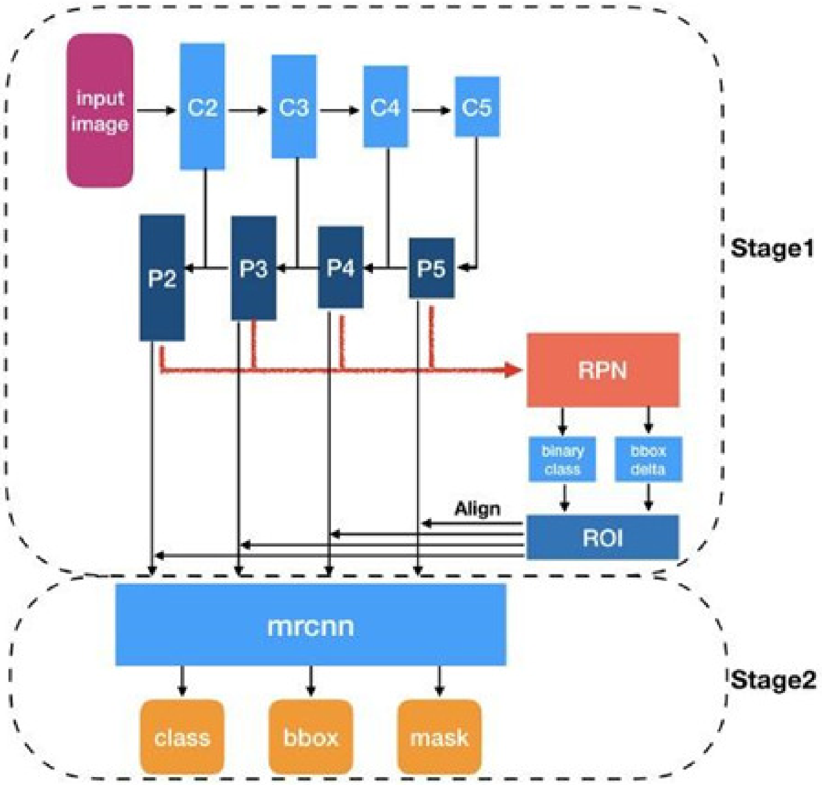


Figure 6. Mask R-CNN basic structure in two stage illustration [9].

different ROI boxes sizes are generated, which becomes a problem, as classifiers work with fixed input sizes. Thus, the method proposed by Mask R-CNN authors, called ROIAlign becomes important since it performs the feature map relevant areas sampling at different points by applying a bilinear interpolation. In the end, a branch contained by a convolutional network is responsible for generating pixel-level masks, one for each object.

Preparing the training dataset

Firstly, in Matlab R2019a, were extracted the video images from the NIR camera and then, with the NDVI calculation, generated the regions of interest masks (binary images, as shown in Fig. 7). It is important to remember that NDVI, in this case, indicates at levels from 0 to 1, the analyzed vegetation health, being considered “healthy” levels closer to 1. Thus, the use of 0.8 as a NDVI threshold generated masks in order to facilitate network learning.

Thus, some of the images were separated into training (160 images) and test (40 images) group. Despite the relatively low number of images used in the training phase, the idea of using a sample gap of 1 image every 10 observed aimed to ensure greater data variability. With the help of the free online software VGG Image Annotator, ROIs were manually demarcated, as shown in Fig. 8. Another important point was to have carried out this demarcation in an attempt to involve most objects of interest precisely. The result of the annotations is saved in a file capable of storing and transmitting information in text format, of extension .json (JavaScript Object Notation), to be loaded in the training model.

Due to the high graphic processing demand and in order to obtain a quick Mask R-CNN technique performance analisys in a solution little addressed so far, as in the [7] paper, was applied the transfer learning method, which consists of using a predefined dataset in order to optimize the different present objects segmentation in an image. For this paper, were used the Microsoft COCO dataset [10] which, although not containing a ”healthy vegetation” class, contains several other images (approximately 120K) with weights that have been trained and whose network has already learned several characteristics about

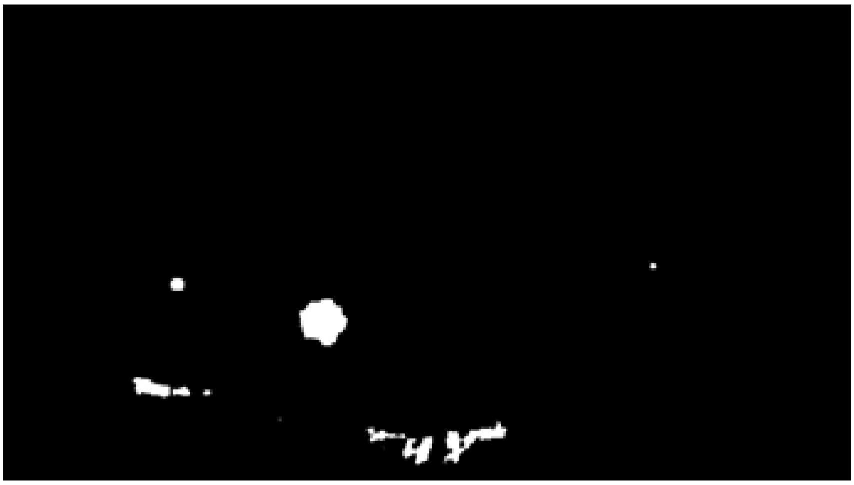


Figure 7. Mask example (binarized image generated by a NDVI factor).

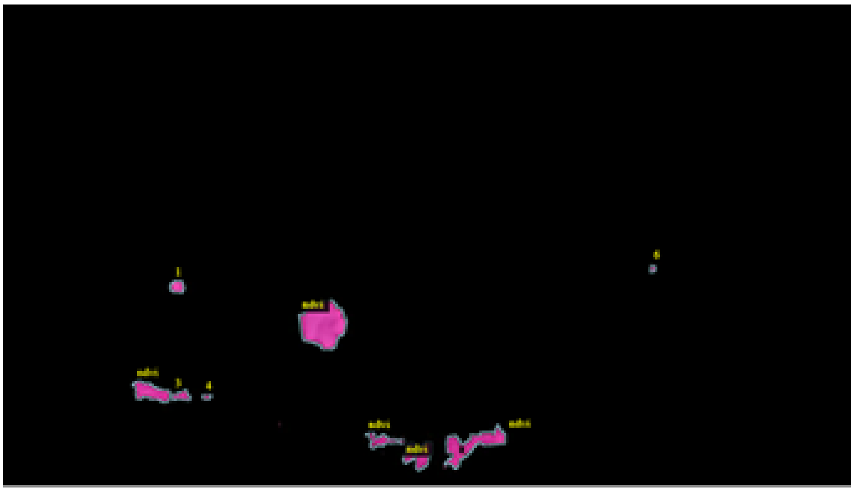


Figure 8. Highlighted ROI image example.

each object.

Loading the training dataset

In the dataset loading step for the training model, the generated masks (composed by images with manually annotated polygons) were imported through a function algorithm, responsible for extracting each point coordinates from the .json file. In this same algorithm, a function is then responsible for generating bitmap masks on each image object, referencing the input polygons.

Configuring and training the model

To perform network training, the number of epochs was set to 150 with a number of 32 training steps per epoch. This ensured that the 160 images were divided into 32 steps with 5 input groups each. Since Mask R-CNN is a highly complex model and, naturally, due the use ResNet-101 and FPN (Feature Pyramid Network) in this approach, a modern GPU become necessary.

3 Experimental Results

In this section were described and discussed the results found in the application of classical and modern clustering techniques, chosen in this paper. It is possible to verify how good or bad the performance of each technique was, showing the strategies and points adopted. As shown in Fig. 9, a NIR image from dataset, adapted to show regions with a NDVI greater than 0.8, was used to compare both techniques.

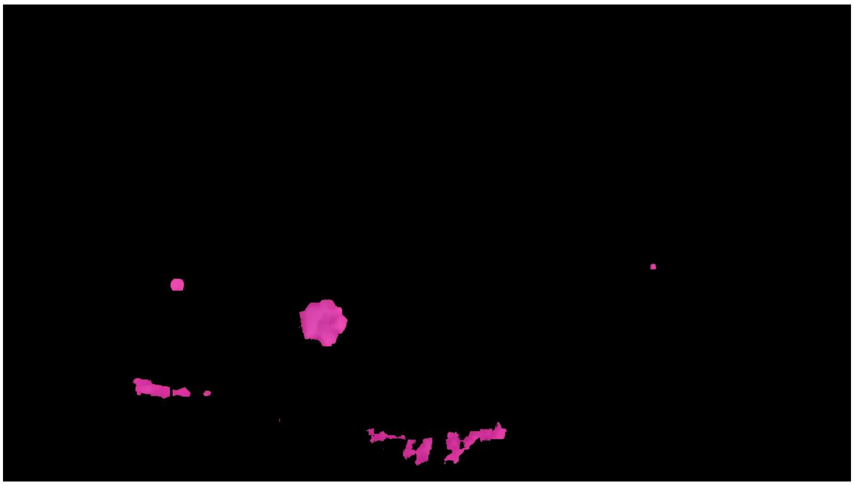
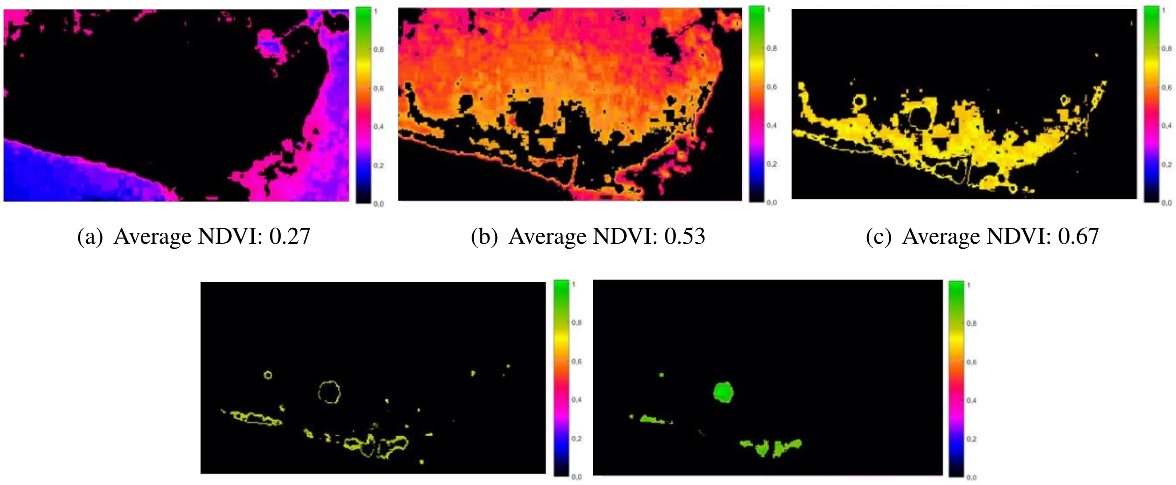


Figure 9. NIR image with regions of interest (ROIs).

Results using K-Means

As described before, the analyzed NGR images were converted to a new color space known as L*a*b*, to facilitate user visualization and interpretation. In this case, a label is given to every pixel in the image, according to the k number of clusters chosen. As can be shown in Fig. 10, choosing k=5, the different regions of interest were well separated. As [5] describes, other number of clusters such as 2, 3, 4 and 6 were tested, although none of these had better results.



(a) Average NDVI: 0.27

(b) Average NDVI: 0.53

(c) Average NDVI: 0.67

(d) Average NDVI: 0.76

(e) Average NDVI: 0.87

Figure 10. K-Means segmentation using k=5.

Results using Mask R-CNN

After the training period was performed and the weights file for the new class was generated, the model was imported and one of the test images was applied to verify the network performance. As can be seen in Fig. 11, for objects with a well-defined shape, without discontinuities (holes, for example), the ROIs detection performed well and accurately.

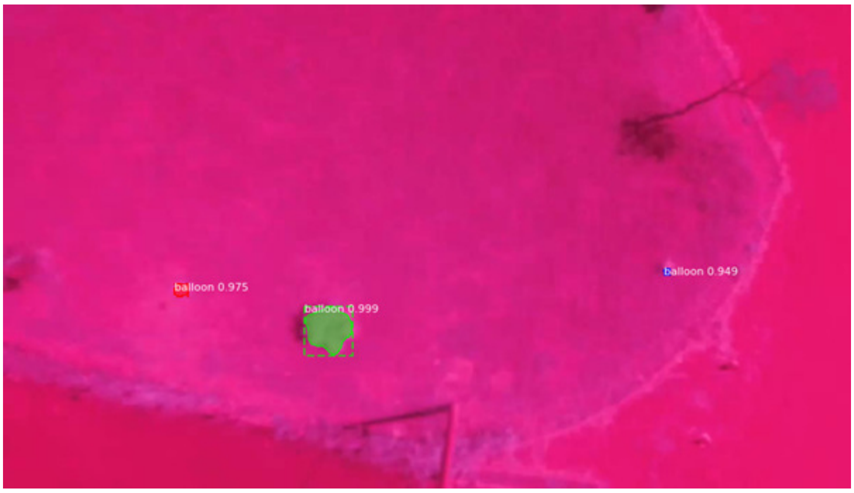
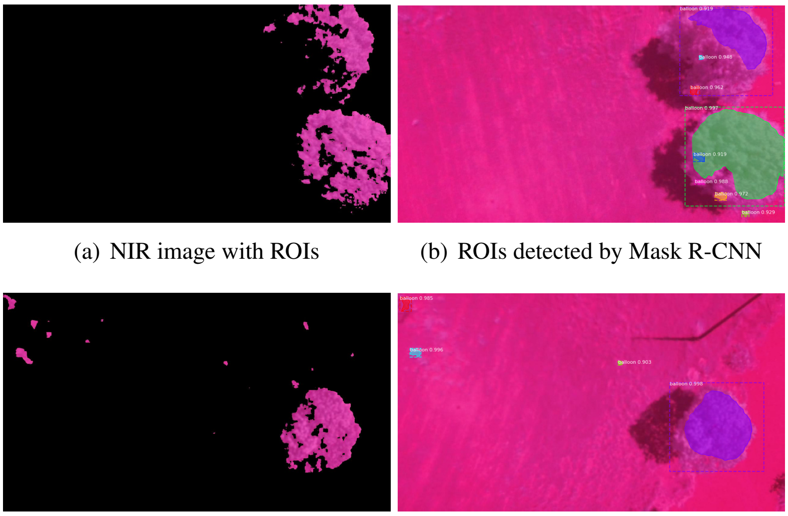


Figure 11. Mask R-CNN healthy vegetation area detection in a NIR image with uniform shapes.

Since several different regions of interest formats exist, and that can not always present a well-defined shape, in some situations such as in this approach whose NDVI generated masks do not conform to a uniform format, the Mask R-CNN model presents detection capability in some regions of interest with a bounding box, however the generated mask does not always approximate the one implemented in the training dataset, as seen in Fig. 12.



(a) NIR image with ROIs (b) ROIs detected by Mask R-CNN

(c) NIR image with ROIs (d) ROIs detected by Mask R-CNN

Figure 12. Mask R-CNN healthy vegetation area detection in a NIR image with nonuniform shapes.

4 Conclusion

In this paper, classical and modern clustering techniques (such as K-Means and Mask R-CNN respectively) were approached, using the NDVI calculation in order to verify performances, its advantages and disadvantages for the proposed problem.

As could be observed, K-Means proved to be robust in detecting the analyzed vegetation areas, presenting ease of implementation and adaptability. Nevertheless, it was concluded that the need to choose a correctly K value to determine the number of clusters is one of the major disadvantages of the method.

The Mask R-CNN method applied on the proposed problem performed well in the detection of areas whose vegetation, given the chosen NDVI, presented well-defined shape, providing data reliability of over 90% in most scenarios. For areas whose shape did not have an uniform distribution, the results obtained were insufficient, which probably had as its main factor this caracteristic. Another disadvantage observed was the difficult adaptability to build a model capable of detecting other vegetation areas with varying NDVI.

For future work, it is suggested to apply the classic Mean Shift clustering technique for performance verification, as it works interactively, having as its basic principle the density estimate of the chosen kernel (that may be a gaussian kernel, for example). Another technique, which may possibly bring better results to the problem proposed in this paper, is the Faster R-CNN, since this technique works at object detection level using bounding boxes, not looking at the pixel level problem, since the sample area does not have a well-defined shape.

Acknowledgements

The authors acknowledge FAPES (Fundação de Amparo à Pesquisa e Inovação do Espírito Santo) for funding this research (process 65656431/2014).

References

[1] Bernardi, A., Naime, J., Resende, , Bassoi, L., & Inamasu, R., 2014. *Agricultura de Precisão: Resultados de um Novo Olhar*, volume 1. Embrapa.

[2] Scarpa, G., Gargiulo, M., Mazza, A., & Gaetano, R., 2018. A cnn-based fusion method for feature extraction from sentinel data. *Remote Sensing*, vol. 10, n. 2.

[3] He, K., Gkioxari, G., Dollár, P., & Girshick, R. B., 2017. Mask R-CNN. *CoRR*, vol. abs/1703.06870.

[4] Maeda, H., S. Y. S. T. K. T. O. H., 2018. Road damage detection and classification in smartphone captured images using mask r-cnn.

[5] SEGATTO, W. G., 2018. Monitoramento automático de vegetação utilizando câmera nir e o algorítmo k-means. Technical report, INSTITUTO FEDERAL DO ESPÍRITO SANTO.

[6] MathWorks, 2019. Color-based segmentation using the l*a*b* color space.

[7] Abdulla, W., 2017. Mask r-cnn for object detection and instance segmentation on keras and tensorflow. `https://github.com/matterport/Mask_RCNN`.

[8] Ren, S., He, K., Girshick, R. B., & Sun, J., 2015. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, vol. abs/1506.01497.

[9] Zhang, X., 2018. Simple understanding of mask rcnn.

[10] Lin, T., Maire, M., Belongie, S. J., Bourdev, L. D., Girshick, R. B., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L., 2014. Microsoft COCO: common objects in context. *CoRR*, vol. abs/1405.0312.