Revista de Informática Teórica e Aplicada - RITA - ISSN 2175-2745
Vol. 27, Num. 04 (2020) 20-33

RESEARCH ARTICLE

Comparison of Classical Computer Vision vs. Convolutional Neural Networks for Weed Mapping in Aerial Images

Comparação de Visão Computacional Clássica vs. Redes Neurais Convolucionais para Mapeamento de Daninhas em Imagens Aéreas

Paulo César Pereira Júnior[1,4], Alexandre Monteiro[2,4], Rafael da Luz Ribeiro[4], Antonio Carlos Sobieranski[3,4], Aldo von Wangenheim[1,4*]

Abstract: In this paper, we present a comparison between convolutional neural networks and classical computer vision approaches, for the specific precision agriculture problem of weed mapping on sugarcane fields aerial images. A systematic literature review was conducted to find which computer vision methods are being used on this specific problem. The most cited methods were implemented, as well as four models of convolutional neural networks. All implemented approaches were tested using the same dataset, and their results were quantitatively and qualitatively analyzed. The obtained results were compared to a human expert made ground truth, for validation. The results indicate that the convolutional neural networks present better precision and generalize better than the classical models.
Keywords: Convolutional Neural Networks — Deep Learning — Digital Image Processing — Precision Agriculture — Semantic Segmentation — Unmanned Aerial Vehicles

Resumo: Neste artigo apresentamos uma comparação entre redes neurais convolucionais e abordagens clássicas de visão computacional, para o problema específico da agricultura de precisão de mapeamento de plantas daninhas em campos de cana-de-açúcar a pertir de imagens aéreas. Uma revisão sistemática da literatura foi realizada para descobrir quais métodos de visão computacional estão sendo usados para este problema. Os métodos mais citados foram implementados, bem como quatro modelos de redes neurais convolucionais. Todas as abordagens implementadas foram testadas, usando o mesmo conjunto de dados, e seus resultados foram analisados quantitativa e qualitativamente. Os resultados obtidos foram comparados com um padrão ouro gerado por um especialista humano, para validação. Os resultados indicam que as redes neurais convolucionais apresentam melhor precisão e generalizam melhor que os modelos clássicos.
Palavras-Chave: Redes Neurais Convolucionais — Aprendizagem Profunda — Processamento de Imagens Digitais — Agricultura de Precisão — Segmentação Semântica — Veículos Aéreos não Tripulados

[1] *Computer Science Post-Graduate Program - PPGCC, Federal University of Santa Catarina, Brazil*
[2] *Automation and Systems Engineering Post-Graduate Program - DAS, Federal University of Santa Catarina, Brazil*
[3] *Department of Computing - DEC, Federal University of Santa Catarina, Brazil*
[4] *Brazilian Institute for Digital Convergence - INCoD, Federal University of Santa Catarina, Brazil*
*__Corresponding author__: aldo.vw@ufsc.br
DOI: http://dx.doi.org/10.22456/2175-2745.97835 • **Received:** 01/11/2019 • **Accepted:** 24/07/2020

1. Introduction

Precision agriculture is a relatively new application field characterized by the use of technology to increase productivity and quality of cultures while making use of policies to preserve the environment [1]. There are several examples of precision agriculture, varying according to their application, such as decision support systems for farm management, data management, pesticide/nutrient use optimization, crop marketing, telematics services, unmanned aerial vehicles (UAV), and others [2]. Specifically for UAVs, it adds an advantage, the ability to inspect a wide area of monitoring, providing images in high-resolution with varied multispectral channels (visible light and near, medium and far infrared channels).

The automation of agricultural production is turning into a large field of research. The first works related to this theme appeared in the decade of 1980. Automation of the processes

of planting, fertilization, protection, and harvesting, are becoming more and more important in the global context, as it helps human labor to become less intensive and improves precision, generating an increase in productivity [3].

There are several applications for the use of aerial images in precision agriculture. For example, locating the center of plantation rows facilitates to optimize the spread of chemicals. It is also an important step to further tasks such as counting plants, skips detection and weed management. All that information can be used to estimate the productivity of specific plantation zones, estimating the vigor, coverage, and density. The autonomous navigation of ground vehicles (tractor) can also be assisted by this kind of information.

This work focuses mainly on weed detection. Invasive plants are responsible for a great part of the productive loss. To prevent this problem, herbicides are spread all over the plantation field, even knowing that weeds are distributed in patches. Herbicides represent a considerable part of production cost on a farm, and its excessive application represents a risk for both human and environmental health. To reduce this over-application problem, site-specific weed management can be used with the aid of aerial image information [4]

To be able to support precision agriculture through aerial imaging, precise and reliable automatic quantitative analysis of agricultural aerial images is necessary, to provide relevant and significant decision making information. To perform the weed detection, some categories of algorithms were proposed, such as object-based image analyses (OBIA), classical machine learning algorithms (Random Forest, SVM, KNN, etc), and even modern convolutional neural network (CNN). All these methods are used to perform the task of classifying each pixel of an input image into 3 possible classes (soil, crop, and weed). The main difficulty in such a task comes from the small variance inter-class (crop and weed).

It can be noticed that there is a huge range of Computer Vision and Machine Learning techniques and diverse applications in the agricultural field. However, there are still a few works that perform semantic segmentation applied in cultures from RGB images. This is due to the fact that the use of multi-spectral cameras greatly facilitates the separation of soil and planting, especially using infrared information (NIR) [5]. Nevertheless, the cost of multi-spectral cameras is expensive when compared to traditional acquisition devices, which often becomes a limitation for producers and researchers.

The objective of our work is to compare and analyze the results of classical computer vision methods and modern convolutional neural networks. First, we present some classical methods. For this purpose, we used three classical machine learning supervised classifiers, with several combinations of features in their feature space. And for the deep learning approach, we use four CNN's semantic segmentation models to compute the weed map. Two of them are simple networks in the field of semantic segmentation, and the others are state-of-the-art networks that have brought great results over challenging semantic segmentation data sets like PASCAL VOC, ADE20K [6, 7].

The remainder of this paper is structured as follows: section 2 shows the state of the art related to the weed mapping on aerial images. Section 3 exposes the dataset and machine learning models used in the experiments, as well as the conducted comparison strategy. Section 6 shows all the founded results. And at last, section 7 presents the final conclusions.

2. State of the Art

To obtain the state of the art related to weed mapping, we performed a systematic literature review (SLR) [8]. The review was performed accordingly to the method for systematic literature reviews proposed in [9]. The review was conducted in 2018 and searched on the following digital databases: Science Direct, IEEE Xplore, Springer Link and Arxiv. The criteria restricted to articles written in English, published between 2008 and 2018, that proposed some algorithm capable of mapping the weeds in aerial images. In this SLR we could identify 19 relevant papers in the field of automated weed mapping.

We identified several methods specifically designed to address the weed mapping problem using aerial images. The majority of these methods employ a supervised classification algorithm to classify each pixel on the image. Half of the revised works used a strategy know as object-based image analysis (OBIA). Those methods perform a classical image segmentation procedure, dividing the input image into blocks of pixels (objects), and then classify each image pixel, using those objects as minimum classifying elements.

Revised works showed a good variety of classifiers. Support Vector Machines (SVM), as a popular supervised classification algorithm, was used on [10, 11, 12, 13, 14, 15]. Another recurrent classification method found in our review was decision trees, like the random forest or C4.5 [16, 16, 10, 14].

Approaches that employ classical machine learning classifiers need to build a feature space to describe each pixel (or object, in OBIA cases). The feature spaces identified in our review are composed of some categories of information: color/spectral, texture, geometric/shape, and position. For color information, it is possible to notice that vegetation indices are used on almost every revised work, being the excess green index (ExG) and the normalized difference vegetation index (NDVI) the most common ones. Some approaches rely only on color/spectral information, for example, [15] used 33 different spectral bands as feature space. Gray level co-occurrence matrix (GLCM) was the most used method to extract texture information [13, 14, 4]. This method consists of a matrix that is defined over an image to be the distribution of co-occurring pixel values at a given offset. Features that represent geometrical/shape information are common on object-oriented approaches. They can consist of features like area, border length, the ratio of length and width, radius of largest/smallest enclosed/enclosing ellipse, asymmetry, and others [4]. And for position information, works like [11, 17, 16, 18, 4] employed the Hough lines transform to detect the plantation rows, and used the pixel/object distance to

the closest line as a feature, to improve differentiation between intra-row and inter-row weed clusters.

Approaches like [14, 19, 20] make use of state of the art computer vision and deep learning methods, that is convolutional neural networks (CNNs). Convolutional neural networks have as a great advantage, the capability of learning (in training time) how to extract good features from the input image, in order to minimize the training error (loss function), removing from the researcher the responsibility of discover what features (color, texture, shape, etc) best represent each class.

It is important to note that deep learning technology has gained a notorious space on the computer vision research community in this decade, paving the way to great advances in the machine learning area. With this in mind, it is safe to assume that in the coming years there will be a growth in the number of works using CNN technology.

A comparison between classical computer vision methods and CNN was already performed in [14]. Both methods were implemented using the OBIA strategy. The superpixels simple linear iterative clustering (SLIC) algorithm was used to generate the objects. The classical classifiers used in their experiments were: SVM, C4.5 decision tree and random forest. The feature spaces used on these methods were: GLCM, local binary patterns, histogram of oriented gradients (HOG) and RGB, HSV and CIELab color spaces. The CNN architecture used was the AlexNet classification network. Each object generated from the SLIC procedure was cropped from the input image and individually fed in the network for classification. Different from [14], in our work, only semantic segmentation CNN architectures were used for the deep learning methods. Semantic segmentations models are specifically designed for applications of this kind, where each pixel on the image needs to be classified. Further using semantic segmentation models eliminates the need generic segmentation algorithms (like SLIC) and features extractor (like GLCM).

3. Material and Methods

To investigate which of the methods (traditional computer vision or CNN) gives the best performance to the problem of weed segmentation in aerial images, we implemented the most used supervised classifications algorithms, image features extractors, and four state-of-the-art models of CNNs. To compare the results, we perform the following experiment. A UAV image of a sugarcane field was applied as input for both methods. Each traditional classifier was tested with different combinations of features in their feature vector. For the CNN approach, each model was trained with the same parameters and conditions. t

3.1 Study Design

We performed a multi-approach study where several traditional computer vision approaches were compared to a set of state-of-the-art CNN approaches. The motivation for this study was to gather evidence to answer the following question: Which approach leads to better results, classical computer vision, or deep-learning?

For the classical computer vision methods, we chose three classifiers that are commonly used in preview works and tested them with various combinations of information in their feature space. The selected classifiers were: Support Vector Machine (SVM), Mahalanobis classifier (MC) and Random Forest (RF). In the deep learning methods, we apply four models of Convolutional Neural Networks (CNN) to segment the pixels into weeds, plants, and soil: SegNet, UNet, Full-Resolution Residual Network (FRRN) and Pyramid Scene Parsing Network (PSPNet). The figure 1 indicates a diagram showing our study design.

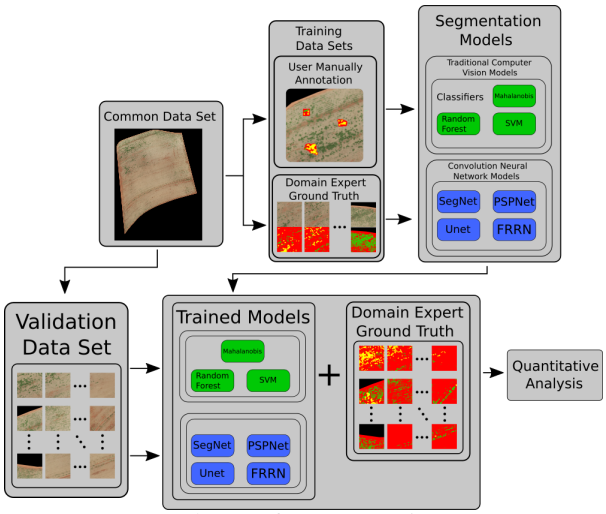

Figure 1. Scheme of our comparison strategy.

3.2 Orthomosaic Sugar Cane Image

The data set used in this work is composed of two sugar cane field orthorectified images. Only one of these images presents a plantation with weed infestation, the other image presents a complete weed-free plantation. The images were captured employing a Horus Aeronaves[1] fixed-wing UAV with a camera model Canon G9X with a resolution of 20.4 megapixels. The UAV captured the data following a flight altitude of 125 to 200 meters, resulting in a resolution of approx. $5cm/pixel$. The flights occurred in the Amazon region, Brazil. The images were rectified with the drone mapping photogrammetry Pix4dMapper software[2].

3.3 Ground Truth

From the data set, an expert biologist produced a human-made ground truth (GT). The expert classified each pixel of the original orthomosaic manually, employing the GNU image manipulation program (GIMP)[3], into three classes: *crop*, *soil* and *weed*. The manual classification was made with

[1] https://horusaeronaves.com/

[2] https://www.pix4d.com/product/pix4dmapper-photogrammetry-software

[3] https://www.gimp.org/

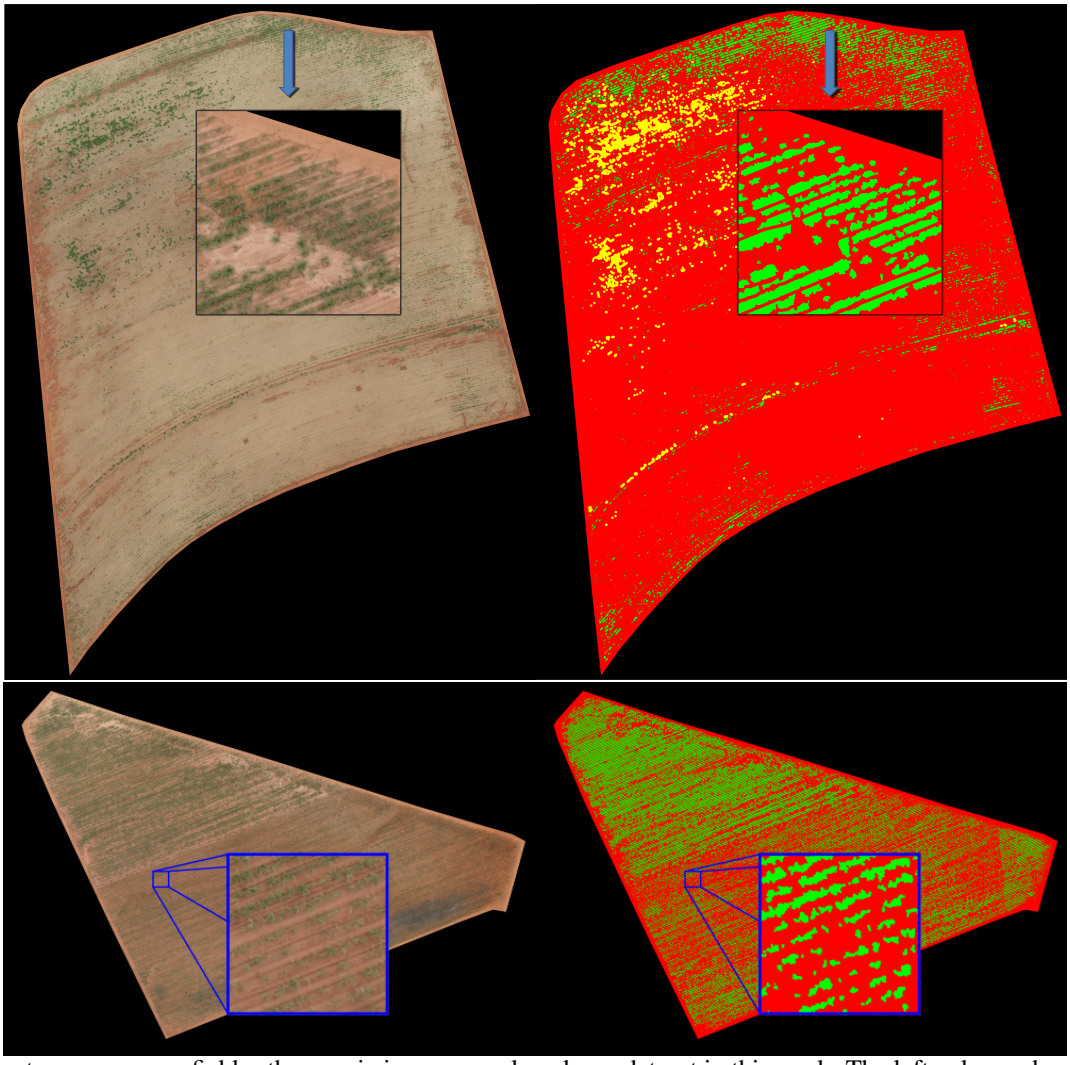

Figure 2. The two sugar cane field orthomosaic images employed as a dataset in this work. The left column shows the sugar cane field orthomosaic. The right column shows the manually generated GT, where *green* = crop, *yellow* = weed and *red* = soil.

the GIMP's pencil tool, where each plant in the orthomosaic was manually segmented. Figure 2 presents the orthomosaic and the corresponding GT[4]. It also shows an example of an image field and its GT. The field in the first row contained weeds. The second one only sugar cane.

3.4 Train Dataset

Since there are fundamental differences between CNNs and classical methods, we needed to build two different training datasets. The classical classifiers, used in our study, segment the image in a pixel-wise classification manner. On the other hand, the CNNs needs to receive an entire image as input, to classify all their pixels in once. It means that the classical methods can receive, as the training dataset, a set of unordered pixels and their respective GT's, and the CNNs require a set of images and the respective GT of all pixels in all those images.

To build the CNN training dataset, the orthomosaic images were subdivided into non-intersecting image slices of size 512x512 pixels. Slices containing only background (black) pixels were discarded. The orthomosaic image containing weeds generated 99 image slices, which is not enough to train a semantic segmentation CNN model. For that reason we used the free weed orthomosaic image as well, to provide more data, so the CNN models can converge. The dataset provided a total of 228 images slices with actual content. We randomly divided these image fields into training set (n=161, 70%) and validation set (n=67, 30%). For the classical methods GT, we manually annotated the pixels of nine small areas in the orthomosaic image, using the GIMP software. These small areas contain, approximately, a total of 140 thousand pixels. Some of these areas are highlighted in figure 3.

3.5 Classical Classifiers

Our literature review showed that Support Vector Machine (SVM), Random Forest (RF) and Mahalanobis Classifier (MC)

[4]Dataset available for download at http://www.lapix.ufsc.br/weed-mapping-sugar-cane and http://www.lapix.ufsc.br/crop-rows-sugar-cane

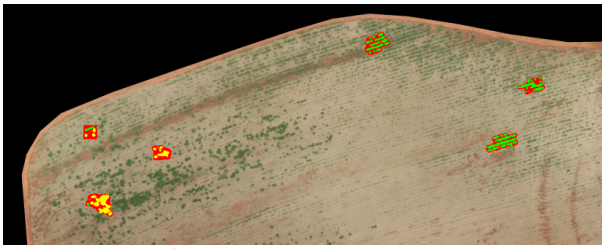

Figure 3. Some of the highlighted annotated pixels employed into the classical classifiers training dataset.

are commonly used supervised algorithms, for this kind of pixel classification problem.

3.5.1 Mahalanobis Distance

Mahalanobis distance is a distance metric based on the correlation between the vector components of a data sample. Given two arbitrary vector coordinates u and g, and a data sample of vector coordinates C with the same dimensionality as u and g, the Mahalanobis distance (MD) is computed by:

$$MD(u,g) = \left((u-g)^T A^{-1} (u-g)\right)^{1/2} = \|u-g\|_A \quad (1)$$

where A^{-1} is the inverse of the covariance matrix obtained from C. The MD is a dual metric: if A is an identity matrix, MD is reduced to the L_2-norm. This statistical distance presents an elliptic topology which surrounds the center of C.

An interesting variation of the MD is the *Polynomial Mahalanobis Distance* (PMD). The PMD was proposed by [21] as a distance metric that can capture the non-linear characteristics of a multivariate distribution as a global metric. The degree of the polynomial (q-order) determines how rigorous the distance will be, based on the samples of the input distribution. A first-order PMD has the same effect as a simple MD.

The Mahalanobis distance, in both its linear and higher-order polynomial variations, has been shown to produce better results than linear color-metric approaches such as RGB or CIELab, when employed as a customized color-metric in various segmentation algorithms [22, 23, 24, 25, 26, 27, 28].

The methodology using the Mahalanobis Classifier consists, in the training step, to generate a different distance metrics (PMD) for each class present in the training set. The classification step consists of finding the closest class mean, using their respective distance metrics. We employed the first three polynomial orders in our experiments.

3.5.2 Support Vector Machines

Probably one of the most popular classifier algorithms, Support Vector Machines incrementally approximate a data classifier trying to create hyperplanes that best separate the data set into their classes. The best hyperplane must maximize the margin between the extreme points in each class. These extreme points that define the hyperplane are called support vectors [29]. Since this method tries to separate the data employing hyperplanes, the classification can only work on linearly separable data. To overcome this limitation, a non-linear kernel function is applied in the data set, transforming the feature space in a nonlinear high-dimensional projection, where it is linearly separable. The most popular kernels are the polynomial and the radial basis function (RBF). In our experiments, we tested two kernels, a simple linear and the RBF kernel. The RBF kernel is described in equation 2.

$$K(\mathbf{X_i}, \mathbf{Y_j}) = EXP(-\gamma||\mathbf{X_i} - \mathbf{Y_j}||^2) \quad (2)$$

3.5.3 Random Forests

Random forest uses ensemble learning (bootstrap aggregating or bagging) on multiples decision trees [30]. Each one of these trees contains internal nodes (condition nodes) and leaf nodes (decision nodes). Each condition node contains a simple rule using one feature from the feature vector, and each decision node contains a class label. The classification of an unlabeled data is done walking down the tree following each condition node until a leaf node is reached, outputting a class label.

The bagging method is used in a slightly different way, decorrelating the trees splitting the feature vector into random subsets of features. Each tree in the forest considers only a small subset of features rather than all of the features of the model (each subset has $\sqrt{n}$ features, where n is the total number of features). This way, highly correlated trees are avoided.

3.6 Feature Space

For the features, we found that vegetation indices are almost unanimously used as color information, and gray level co-occurrence matrices (GLCM) are frequently used for texture features extractor. Since textures are important information to discriminate different objects with the same color (connected parallel crop rows for example), we decided to include another well known texture feature extractor in our experiments, Gabor filters.

3.6.1 Vegetation Indices

Since most crops present a green coloration, it is intuitive that color information can be a useful feature for the segmentation. A vegetation index (VI) consists of mathematical manipulation of the image spectral channels (RGB), to measure the greenness of a pixel. Various VI's were proposed on the literature [31], the most present ones in our review were the excess green index (ExG), for RGB cameras, and the normalized difference vegetation index (NDVI), for cameras with near-infrared spectrum. In this work, we only used VI for RGB cameras. The ExG index is expressed by the equation 3, where G, R, and B are the intensity values of each channel normalized by the sum of the three.

$$ExG = 2G - R - B \quad (3)$$

3.6.2 Gabor Filters

Gabor filters are traditional texture descriptors proposed by Denis Gabor [32]. The texture is extracted from the image by a set of base functions, which can be employed to build a *Gabor filter bank*. Each base function is modulated by a specific scale and orientation, and a process of convolution of the filters with an image produces responses where the structure adapts with the scale and orientation analyzed. Gabor filters have been shown to help in performing texture-based image segmentation through integrated color-texture descriptors [33]. Gabor filter-based segmentation approaches have also been shown to be easy to parallelize, implement in GPUs and use to perform fast color-texture-based image segmentation [34]. The Gabor filter can be expressed by the equation 4, where u_0 and ϕ respectively correspond to frequency, and the phase offset (in degrees). The standard deviation σ determines the size of the Gaussian filed.

$$h(x,y) = \exp\left\{-\frac{1}{2}\left[\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right]\right\} cos(2\pi u_0 x+\phi). \quad (4)$$

Different orientations can be obtained employing a rigid rotation of the x-y coordinate system with an angle value predefined by θ, as follows:

$$x = x_0 cos(\theta)+y_0 sin(\theta), y = y_0 cos(\theta)-x_0 sin(\theta) \quad (5)$$

The Gabor filter bank used in our experiments was composed of 4 filters. The guidelines proposed in [35] were used to chose the Gabor features. We utilized frequencies values that generate kernels with a size that matches a crop row width. Only two orientations were used (0° and 90°) to reduce the feature vector size generated from the filter back.

3.6.3 Grey Level Co-Occurrence Matrix

Grey Level Co-Occurrence Matrix are second order statistic matrix used to extract texture information from an image [36]. This information is extracted from the image computing several co-occurrence matrices. A Grey Level Co-Occurrence Matrix $\mathbf{P}n \times n$, where n is the number of gray levels of the image, is defined using the neighborhood of pixels, where $\mathbf{P}_{ij}$ is the probability of two neighbors pixels have intensities of i and j. The neighborhood relationship is defined by an offset from the reference pixel. Different offsets can be used, a vertical GLCM can use $(1,0)$ and $(-1,0)$ offsets, and a horizontal one can use $(0,1)$ and $(0,-1)$. From each GLCM $\mathbf{P}$, different texture features can be extracted, some of them are energy, entropy, contrast, dissimilarity, homogeneity, mean, standard deviation, and correlation.

For the experiments, we downsampled the input image from 8 bits per channel to 4 bits, so our GLCM's have 16×16 size. We used the most referenced features on our review, so we used five features: contrast, energy, mean, standard deviation, and correlation. Vertical and horizontal GLCM's are used as offsets, with a window size of 33 pixels.

3.7 CNN Approaches

Convolutional Neural Networks (CNNs) [37] have been used in many different tasks of image processing and computer vision. When it comes to image interpretation, we can differentiate 3 categories of actions: (a) classification of images [38, 39, 40], (b) detection and location of objects in images [41, 42, 43, 44, 45] and (c) segmentation and classification of objects in images [46, 47, 48]. This last modality of CNNs is called semantic segmentation and is the one useful for our problem. Networks for semantic segmentation classify each pixel of an image, associating individual pixels with the class they represent, performing in practice a segmentation of the image according to the semantics of the objects to which each pixel is associated [49].

To tackle the weed mapping problem with semantic segmentation CNN's, we trained the networks to segment the images into four possibles objects, crop, weed, soil, and dark-background. We used this last class to facilitate the convergence of the networks. Since the background areas are irrelevant for this problem, we perform simple post-processing that turns all predicted background pixels into soil pixels.

We used four known models of CNNs that have achieved very good results in different semantic segmentation applications, e.g. indoor and outdoor scene parsing, road and city scenarios, biomedical imagery of competition datasets [6, 7]. The SegNet and UNet [50, 51] models are more traditional and were the first ones to overcome classical techniques of computational vision in the segmentation challenges such as PASCAL VOC [6]. The other two, the PSPNet and FRRN [52, 53], are used in this work given their innovative techniques of feature extraction, their great potential and great results in very difficult segmentation datasets. They were originally build to be applied to the road scene, autonomous driving scenarios, indoor identification and also medical images. In this work, we used these models in a completely different dataset (agricultural imagery). The model's description and differences are shown in the next subsections. In Table 1 we present a summary of the main characteristics of the CNN models used in this work.

3.7.1 SegNet

The SegNet network, according to Badrinarayanan et al. (2015), was developed to correct a problem, which was the adoption of convolutional networks originally applied for classification being used for pixel segmentation. With the operations of max pooling and sub-sampling, the previous approaches reduced the resolutions of the images and thus, the important characteristics of the objects were not well captured, which generated bad results. The solution found by SegNet relies on mapping the low-resolution characteristics as input to produce an effective pixel rank [50].

In SegNet, the encoder is topologically identical to the layers of the network VGG16 (thirteen convolution layers) [39], except those that are fully connected. Also, the way the upsampling is performed is considered the key component for the authors. Each max-pooling operation saves an index,

which is passed later to the appropriate encoder layers. The operations of max pooling and upsampling consist of reducing and increasing the spatial size of the image, respectively. Max pooling is used to reduce the number of parameters and hence the computational effort required. With upsampling done in this way, it is possible to retrieve the previously mapped features in a practical, fast way and taking up less memory, since it requires fewer parameters. This all together with the other characteristics shown in Table 1, make SegNet efficient and advantageous over other networks that generate very good results, but use more memory and are slower.

3.7.2 UNet

The architecture of the neural network called UNet achieves very good results in several different biomedical targeting applications [51]. In their paper, Ronneberger et al. (2015) have extensively used the practice of data augmentation, which was the main technique that made them successful, since their data set was small.

The initial part of the network - the encoder - is the same as a typical architecture of a convolutional image classification network. However, it does not have the last layers fully connected. The architecture modification made by UNet consists of having a large number of channels with extracted features directly connected also to the step of upsampling. This allows the network to be able to propagate more information about the context of the image to the layers of higher resolution. Other important features about UNet are shown in Table 1 at the end of this section.

3.7.3 Full-Resolution Residual Network (FRRN)

The FRRN model is a very clear example of the multi-scale processing technique. It Combines multi-scale context with pixel-level accuracy by using two processing streams within the network. The two streams are coupled at the full image resolution using residuals. FRRN progressively processes and downsamples the feature maps in the pooling stream. At the same time, it processes the feature maps at full resolution in the residual stream. So the pooling stream handles the high-level semantic information (for high classification accuracy) and the residual stream handles the low-level pixel information (for high localization accuracy). After each max pooling operation, FRRN does some joint processing of the feature maps from the 2 streams to combine their information [53].

3.7.4 Pyramid Scene Parsing Network (PSPNet)

PSPNet can work with multi-scale feature maps without applying many convolutions to them. First, it uses the ResNet [54] model to extract the feature map. PSPNet applies four different max pooling operations (multiple scales of pooling), with varying window sizes and strides to capture feature information from four scales without processing each one individually. Then, it does convolution on each scale, followed by upsampling and concatenation of all of them. The result is combined multi-scale feature maps with less convolution, which saves computational effort and gives higher speed. In the end, the output segmentation map is upscaled to the desired size using bilinear interpolation [52]. PSPNet ranked 1st place in ImageNet Scene Parsing Challenge 2016 [7].

4. Comparison Strategy

To better understand which factors are to be taken into consideration when choosing between a classical or a CNN-based CV method for weed mapping, it is important to look into the operational parameters that play a role in the workflow of adapting a machine-learning-based weed mapping application to a new plantation.

Several different factors will influence the decision of which approach to use. It is not only the prediction accuracy of the model that plays a role, but there are also several operational cost factors involved. The factors we identified are person-hours of specialist time for the generation of the ground truth images (for the training dataset), the processing time for the training of the model, and the prediction time of new images. We called them *cost and quality factors*.

To take these factors additionally into account, besides the obvious choice of the *prediction accuracy* as a criterion to be considered when choosing a model, is based upon the following rationales:

- **Train dataset generation** (cost factor - person/hours): classical models need only the GT of a few thousands of pixels, and these can be generated by simply dragging the mouse across a few typical samples of the areas occupied by crops, weed, and soil. For a CNN application training dataset generation is a much more labor-intensive process, it corresponds to the full GT of several slices of the input image. It means that the whole dataset will have to be manually processed and every pixel of the dataset will have to be classified as *crop*, *weed* or *ground*, as we show in our dataset.

- **Training time** (cost factor - processing time): the computational effort and, consequently, training time for training classical models is generally less than CNN's. Even more computation-intensive approaches such as RBFSVM train much faster than CNN. These networks will have the additional requirements of special parallel hardware (GPUs) and RAM, needing either special computer configurations or implicate in having to outsource training to a cloud processing account.

- **Prediction time** (cost factor - processing time): the question of whether a special computing infrastructure is necessary to run trained models to classify new images has also to be taken into account. A CNN can take several seconds to provide a prediction for a single image on a standard computer, and may need to be run either on special hardware or have the run-times also outsourced to a cloud computing provider, but some classical models, such as KNN, which will compare a new image to the whole set of GT images, can also

Table 1. Neural Networks Main Characteristics

Model	Up Sampling	Application	Output Layer	Valuable Feature
SegNet	pooling indices computed at max pooling step	road and indoor scene understanding	classification layer (Softmax)	great performance using less memory
UNet	convolution layers that learn a precise output	biomedical image processing	no fully connected layers	fast
FRRN	bilinear interpolation	autonomous driving systems	concatenation of two streams	combined multi-scale context using two processing streams
PSPNet	bilinear interpolation	scene parsing in general	different levels of features concatenated as the final global feature	pyramid pooling module

take considerable time to classify all pixels of this new image.

To better compare the classification methods, we understand that is necessary to go beyond the simple comparison of *prediction accuracies*, as has been the focus of related works. For this purpose, we compare them under the light of these four *cost and quality factors*.

4.1 Prediction Accuracy Measure

The validation of our results was performed through an automated quantitative comparison of our results against the specialist-generated ground truth (GTs). We evaluated each employed segmentation method by pixel-wise comparing our results to the ground-truths generated by the specialist. The precision measures employed in our experiments are the Jaccard index [6, 55], also known as the intersection-over-union index (IoU), and the F1-score [56], which is defined as the harmonic mean between Precision and Recall. A good result of F1 means that there are low false positives and low false negatives. It is considered to correctly identify correct results and is not disturbed by false results. Those metrics were chosen because they are the most widely employed quantitative GT comparison measures in the area of semantic segmentation approaches that were published in the last years.

The precision was calculated using the image slices in the validation dataset, but only the slices extracted from the orthomosaic that contains weeds. The background pixels were not included in the calculation. Besides this quantitative validation, we also performed a qualitative validation through visual inspection of the results.

5. Experiments

For the classical classifiers, the training and classification step were performed as follows. We computed the feature vector (color and texture features) of each pixel selected for the training dataset, some of them are shown in figure 3. All the classifiers were then trained, with the same train data, and with several combinations of feature vectors. Once all classifiers were trained, the classification step was applied over the whole orthomosaic images, for a qualitative validation, and over the 44 images slices present in the validation dataset. The combinations of features tested were done in an incremental form. First, feature vectors with only color information (RGB channels and vegetation indices), then color plus texture information (Gabor Filters and GLCM), resulting in 6 different feature vectors. All classifiers and feature extractors were implemented using C++ and OpenCV library[5], and executed on a Linux (Ubuntu 18.04) machine, with 32GB RAM and a i7-7700 CPU with 3.60GHz.

The parameter values employed for training the CNNs present some variations between the four models described earlier. We employed different learning rates and batch sizes, starting with typical values found in the literature for each of the models. The values of these parameters were then tuned experimentally by performing a series of tests with varying learning rates and batch sizes from these start values to identify which values would present the best results for each of the models. The best values we identified are presented in table 2. All models were trained for 200 epochs employing the TensorFlow framework [6] running on the Google Colaboratory environment [7]. The GPU available at the instance of the Google platform allocated for our experiments was a Tesla model K80 with computing capability 3.7 and 11 GB RAM. The codes are programmed in the Python language and for the weight updating step, the Adam [57] algorithm was used with the parameters in their default values, except for the learning rate. We made a public description of the codes we used [8].

[5]Code available at: https://codigos.ufsc.br/lapix/Tools-for-Supervised-Color-Texture-Segmentation-and-Results-Analysis

[6]https://www.tensorflow.org/

[7]https://colab.research.google.com/

[8]Available at: https://codigos.ufsc.br/lapix/Weed-Mapping

Table 2. Learning Rates and Batch Sizes that presented best Results

Model	Learning Rate	Batch Size
SegNet	0.001	3
UNet	0.0001	2
FRRN	0.00001	2
PSPNet	0.00001	2

6. Results and Discussion

The results obtained in each *cost and quality factors* will be presented in this section.

6.1 Training Dataset Generation Effort

The total time spent to create the GT for the entire dataset was 83.6 hours, divided over three weeks of work, and only one biologist worked on this task. The training data used in the classical model took half an hour to be created by one person.

6.2 Training Time

The training time spent to train each model is presented in Figure 4. All CNN models have taken hours to train, while the classical models have taken no more than a few minutes. This is the expected result since a CNN model needs several times more data to converge.

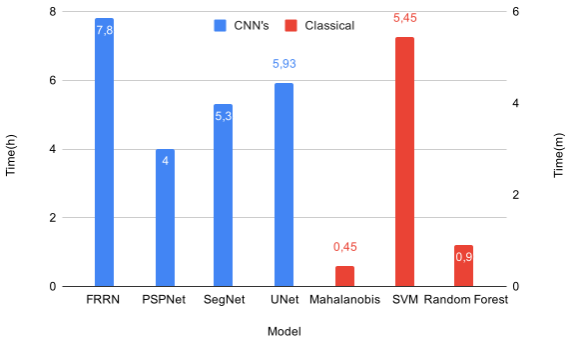

Figure 4. The training times for each tested model. CNN's times are presented in hours and classical ones in minutes.

6.3 Prediction Accuracy

The quantitative precision results obtained from the classical classifiers are demonstrated in table 3. A total of 36 results were obtained, one from each combination of classifiers (with different parameters) and feature vectors. For each classifier, the feature vector that presented the best result has his result marked in bold. In the same way, for each feature vector, the best classifier has his result underlined.

Between the classical methods, the observed best result was archived using Random Forest with $RGB+EXG+GLCM$, reaching an F1 score precision of 0.78. This classifier was the only one that was capable of make good use of GLCM features, showing better results for this kind of feature vector. To compare the classical methods with the CNN approaches, we took the best result of each classifier and put them aside with the CNN results. These quantitative results are presented in table 4, where we show the values of F1-Score in percentage for each class and the mean value. Also, the mean IoU value in percentage. The best result in each column is marked in bold.

The best result for both metrics was achieved with the FRRN model, reaching a mean F1 of 80.66% and a mean IoU of 69.71%. The SegNet model achieved a similar result in both metrics and for all three classes. UNet achieved the poorest results among the four network models. When looking for the weed class F1 scores, the random forest methods presented a slightly better result.

6.4 Prediction Time

The prediction time was calculated measuring the time spent to segment the whole orthomosaic image presented in figure 2. The FRRN model has a faster response than our best classical model, random forest. Also, the SegNet model presented a prediction time almost four times longer than the other CNN models. Figure 5 shows the time spent in each model.

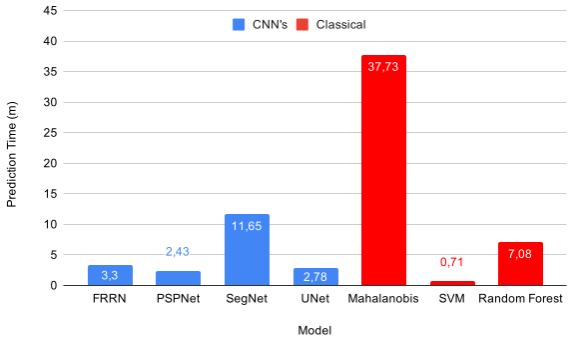

Figure 5. The prediction times, in minutes, of each tested model.

6.5 Qualitative Analysis

To perform a qualitative analysis we visually compared the weed maps generated from the seven models with the GT. To be able to visualize the whole weed map generated by the networks, we re-assembled all the 99 image slices predicted by each network. Each generated weed map can be seen in figure 6.

Due to the slicing process required to use the CNN models, the border pixels of each slice loses a portion of information, that lies on the neighbor image slice. This loss of information generates some undesired classification errors on the border of each image slice, these errors can be seen in detail on the CNNs results in figure 6. Since the classical models can process the entire orthomosaic in once, this kind of prediction error is not present.

To analyze the generalization capabilities of each model, we performed a segmentation with each of our trained models

Table 3. Classical classifier's precision measures.

		RGB	RGB+EXG	RGB+GABOR	RGB+EXG+GABOR	RGB+GLCM	RGB+EXG+GLCM
Order 1	IOU	0.544086	**0.548179**	0.314340	0.409837	0.013137	0.013137
	F1	0.654281	**0.659662**	0.324834	0.476346	0.025278	0.025278
Order 2	IOU	**0.549559**	0.539652	0.519721	0.527287	0.245271	0.245271
	F1	**0.660259**	0.649763	0.626300	0.634082	0.282614	0.282614
Order 3	IOU	0.555844	0.542475	**0.556147**	0.555080	0.245271	0.245271
	F1	**0.667707**	0.653626	0.666652	0.665532	0.282614	0.282614
LSVM	IOU	0.418371	0.383301	0.405959	**0.537555**	0.401189	0.099146
	F1	0.479947	0.437613	0.465656	**0.645496**	0.474317	0.159750
SVMRBF	IOU	0.316514	**0.509402**	0.377596	0.394635	0.368942	0.406156
	F1	0.329108	**0.603964**	0.427998	0.453472	0.426704	0.486883
RF	IOU	0.563429	0.568327	0.620124	0.639770	0.667932	**0.676676**
	F1	0.680331	0.685636	0.738710	0.756993	0.778510	**0.786046**

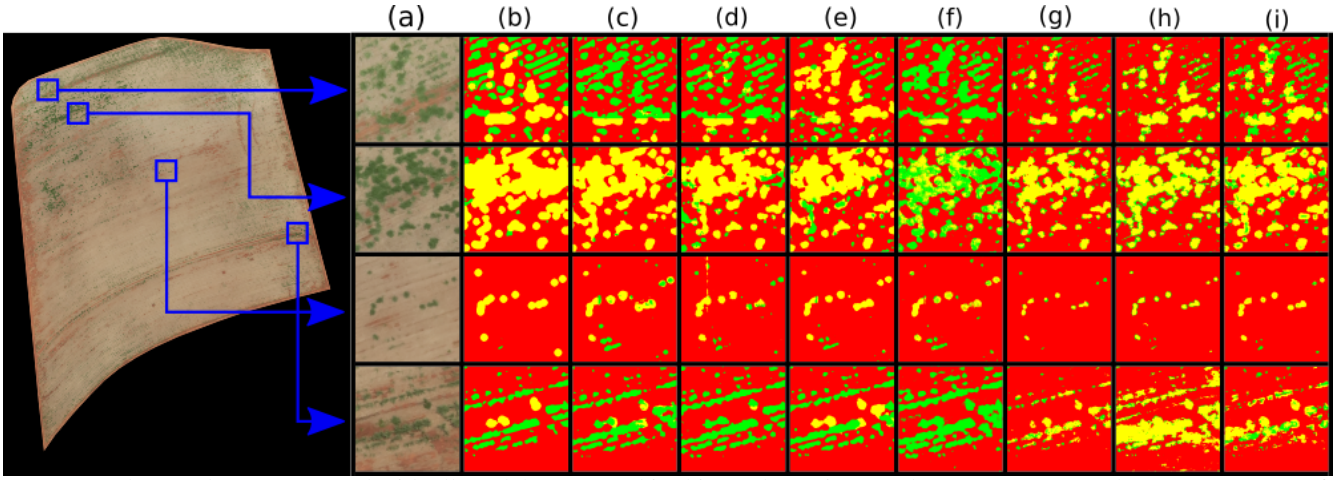

Figure 6. The weed map generated with all models presented in this work. (a) image, (b) GT, (c) FRRN, (d) PSP, (e) Segnet, (f) Unet, (g) Mahalanobis order 3, (h) Linear SVM, (i) Random Forest.

Table 4. Quantitative Validation Results for all models

	F1-Score (%)				IoU (%)
	plant	invasive	soil	Mean	
SegNet	65.67	74.21	98.34	79.41	68.20
UNet	56.60	50.02	97.88	68.17	56.23
FRRN	**67.98**	75.61	98.40	**80.66**	**69.71**
PSPNet	64.17	65.96	98.27	76.13	64.35
Mahala	37.56	64.30	98.13	66.66	55.61
SVM	34.25	61.14	98.25	64.54	53.75
RF	59.83	**77.41**	**98.55**	78.60	67.66

on a different sugarcane orthomosaic image. This orthomosaic has not a human-made ground truth but contains the same species of plant and was captured in the same data with the same equipment, so it is possible to use it for qualitative analysis. The figure 7 shows this orthomosaic image and the weed maps of each model.

All three classical models do not have demonstrated to be able to generalize well to another plantation field. These models presented an over-identification of weeds (false positives), while not detecting in a good manner the actual ones. The CNN model with the best quantitative result (FRRN) did not demonstrate a good generalization. On the other hand, the SegNet model, witch generated a quantitative result close to the FRRN, presented the best generalization from all models.

7. Conclusions

This work performs a comparison between classical computer vision, and CNN approaches, applied to the problem of weed mapping on RGB aerial images. A systematic literature review was performed to find which methods are being used for this task. The review showed several classical computer vision methods that, in some way, consist of supervised classification of the elements in the image. This classification varied with several machine learning classifiers and image features extractors. Some solutions using CNNs models also appeared in our review.

To perform the comparison study, we implemented some

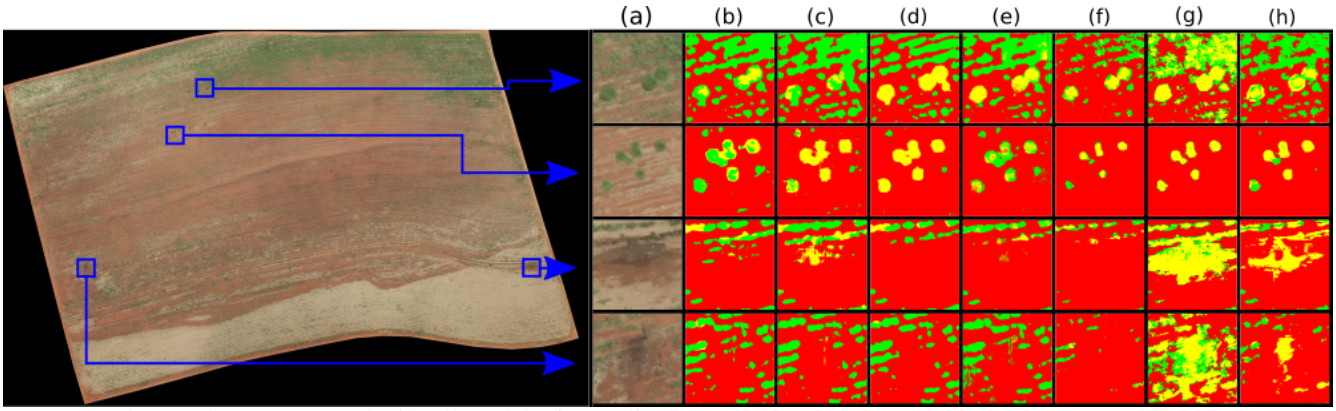

Figure 7. The weed map generated with all models, for qualitative analysis. (a) image, (b) FRRN, (c) PSP, (d) Segnet, (e) Unet, (f) Mahalanobis order 3, (g) Linear SVM, (h) Random Forest.

of the classifiers and features extractors that most appeared on the review and compared their results with the results obtained by four different CNN models. The quantitative validation was made by automatic comparison with ground truth, manually generated by a biologist. We employed the Jaccard index and the F1-score for the quantitative validation of the methods we tested. This will allow researchers using this database, in the future, to also compare these approaches to CNN-based semantic segmentation approaches.

Classical models were trained on different hardware than the CNN models due to the intrinsic differences between these approaches, but all CNN were trained at the same conditions with the same processor unit, and the same was done to the classical methods.

The precision quantitative measures showed that random forest, in combination with GLCM features, presented the best result among the classical methods. The best quantitative result obtained was generated by the FRRN network, which archived a slightly better mean F1-score than the SegNet model. On the other hand, the qualitative result, using a different plantation field image (not used in training), seems to suggest that the classical models do not generalize well to other fields, and the SegNet model generalizes better than the FRRN one.

The results exceeded our expectations since we applied a relatively simple scheme of study, where no preprocessing was applied to the images and only images captured in the visual light (RGB) spectrum channels were employed. In the literature, many authors employ more sophisticated information beyond the one contained in visible light. However, with the results of this work, we could check the good performance of the CNN models using only visible light information for this application, with our dataset specifically.

For the question of which model should we employ on the problem of weed mapping on aerial image, this work presents some evidence. The CNN models presented results with better precision and are more capable to generalize to other plantations. At the same time, these models require more hardware and much more train data, this last one being extremely expensive in terms of person-hour. The classical models, if chosen correctly, can archive results with good precision. Their great advantage lies in the training data required to use them, which can be generated by one person with much less effort compared to the training data used on CNN models.

Further works should include an extensive validation with a richer and vast data set, containing other cultures besides sugarcane. Also, evaluate the same scheme of study using more and newer state-of-the-art CNNs. To enrich the research field of weed mapping, future work should analyze the main advantages of the models which achieved the best result in this kind of study and develop a new CNN designed specifically for this application. Our dataset was made publicly available and can be used by other authors to test different approaches.

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) and by Fundação de Amparo à Pesquisa e Inovação do Estado de Santa Catarina (FAPESC). We also thank Horus Aeronaves for support and data. There are no conflicts of interest.

References

[1] MCBRATNEY, A. et al. Future directions of precision agriculture. *Precision agriculture*, Springer, v. 6, n. 1, p. 7–23, 2005.

[2] REYNS, P. et al. A review of combine sensors for precision farming. *Precision Agriculture*, Springer, v. 3, n. 2, p. 169–182, 2002.

[3] VIDOVIĆ, I.; CUPEC, R.; HOCENSKI, Ž. Crop row detection by global energy minimization. *Pattern Recognition*, Elsevier, v. 55, p. 68–86, 2016.

[4] GAO, J. et al. Fusion of pixel and object-based features for weed mapping using unmanned aerial vehicle imagery. *In-*

ternational Journal of Applied Earth Observation and Geoinformation, Elsevier, v. 67, p. 43–53, 2018.

[5] LOUARGANT, M. et al. Unsupervised classification algorithm for early weed detection in row-crops by combining spatial and spectral information. *Remote Sensing*, Multidisciplinary Digital Publishing Institute, v. 10, n. 5, p. 761, 2018.

[6] EVERINGHAM, M. et al. The pascal visual object classes (voc) challenge. *International journal of computer vision*, Springer, v. 88, n. 2, p. 303–338, 2010.

[7] ZHOU, B. et al. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, Springer, v. 127, n. 3, p. 302–321, 2019.

[8] JÚNIOR, P. C. P.; WANGENHEIM, A. V.; MONTEIRO, A. *Weed Mapping on Aerial Images - A Systematic Literature Review*. [S.l.], 2019. Doi10.13140/RG.2.2.34979.71204. Disponível em: ⟨http://www.incod.ufsc.br/wp-content/uploads/2019/05/INCoD.LAPIX_.01.2019.E.SLR_.Weed_.Mapping.Aerial.Images.pdf⟩.

[9] KITCHENHAM, B. Procedures for performing systematic reviews. *Keele University, Joint Technical Report TR/SE-0401*, v. 33, n. 2004, p. 1–26, 2004.

[10] CASTILLEJO-GONZÁLEZ, I. L. et al. Evaluation of pixel-and object-based approaches for mapping wild oat (avena sterilis) weed patches in wheat fields using quickbird imagery for site-specific management. *European Journal of Agronomy*, Elsevier, v. 59, p. 57–66, 2014.

[11] PÉREZ-ORTIZ, M. et al. A semi-supervised system for weed mapping in sunflower crops using unmanned aerial vehicles and a crop row detection method. *Applied Soft Computing*, Elsevier, v. 37, p. 533–544, 2015.

[12] PÉREZ-ORTIZ, M. et al. Machine learning paradigms for weed mapping via unmanned aerial vehicles. In: IEEE. *Computational Intelligence (SSCI), 2016 IEEE Symposium Series on*. [S.l.], 2016. p. 1–8.

[13] DAVID, L. C. G.; BALLADO, A. H. Vegetation indices and textures in object-based weed detection from uav imagery. In: IEEE. *Control System, Computing and Engineering (ICCSCE), 2016 6th IEEE International Conference on*. [S.l.], 2016. p. 273–278.

[14] FERREIRA, A. dos S. et al. Weed detection in soybean crops using convnets. *Computers and Electronics in Agriculture*, Elsevier, v. 143, p. 314–324, 2017.

[15] ISHIDA, T. et al. A novel approach for vegetation classification using uav-based hyperspectral imaging. *Computers and electronics in agriculture*, Elsevier, v. 144, p. 80–85, 2018.

[16] LOTTES, P. et al. Uav-based crop and weed classification for smart farming. In: IEEE. *Robotics and Automation (ICRA), 2017 IEEE International Conference on*. [S.l.], 2017. p. 3024–3031.

[17] PÉREZ-ORTIZ, M. et al. Selecting patterns and features for between-and within-crop-row weed mapping using uav-imagery. *Expert Systems with Applications*, Elsevier, v. 47, p. 85–94, 2016.

[18] BAH, M. D.; HAFIANE, A.; CANALS, R. Weeds detection in uav imagery using slic and the hough transform. In: IEEE. *Image Processing Theory, Tools and Applications (IPTA), 2017 Seventh International Conference on*. [S.l.], 2017. p. 1–6.

[19] SA, I. et al. weedNet: Dense Semantic Weed Classification Using Multispectral Images and MAV for Smart Farming. *IEEE Robotics and Automation Letters*, IEEE, v. 3, n. 1, p. 588–595, 2018.

[20] BAH, M. D. et al. Deep learning based classification system for identifying weeds using high-resolution uav imagery. In: SPRINGER. *Science and Information Conference*. [S.l.], 2018. p. 176–187.

[21] GRUDIC, G.; MULLIGAN, J. Outdoor path labeling using polynomial mahalanobis distance. In: *Robotics: Science and Systems*. [S.l.: s.n.], 2006.

[22] LOPES, M. D. et al. A web-based tool for semi-automated segmentation of histopathological images using nonlinear color classifiers. In: *2016 IEEE 29th International Symposium on Computer-Based Medical Systems (CBMS)*. [S.l.: s.n.], 2016. p. 247–252.

[23] CARVALHO, L. E. et al. Improving graph-based image segmentation using nonlinear color similarity metrics. *International Journal of Image and Graphics*, v. 15, n. 04, p. 1550018, 2015. Disponível em: ⟨https://doi.org/10.1142/S0219467815500187⟩.

[24] CARVALHO, L. E. et al. Hybrid color segmentation method using a customized nonlinear similarity function. *International Journal of Image and Graphics*, v. 14, n. 01n02, p. 1450005, 2014. Disponível em: ⟨https://doi.org/10.1142/S0219467814500053⟩.

[25] SOBIERANSKI, A. C. et al. Color skin segmentation based on non-linear distance metrics. In: BAYRO-CORROCHANO, E.; HANCOCK, E. (Ed.). *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. Cham: Springer International Publishing, 2014. p. 143–150.

[26] SOBIERANSKI, A. C.; COMUNELLO, E.; WANGENHEIM, A. von. Learning a nonlinear distance metric for supervised region-merging image segmentation. *Computer Vision and Image Understanding*, v. 115, n. 2, p. 127 – 139, 2011. Disponível em: ⟨http://www.sciencedirect.com/science/article/pii/S1077314210002006⟩.

[27] SOBIERANSKI, A. C. et al. Learning a color distance metric for region-based image segmentation. *Pattern Recognition Letters*, v. 30, n. 16, p. 1496 – 1506, 2009. Disponível em: ⟨http://www.sciencedirect.com/science/article/pii/S0167865509002098⟩.

[28] SOBIERANSKI, A. C. et al. Learning a nonlinear color distance metric for the identification of skin immunohisto-

chemical staining. In: *2009 22nd IEEE International Symposium on Computer-Based Medical Systems*. [S.l.: s.n.], 2009. p. 1–7.

[29] JAKKULA, V. Tutorial on support vector machine (svm). *School of EECS, Washington State University*, v. 37, 2006.

[30] BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001.

[31] WANG, A.; ZHANG, W.; WEI, X. A review on weed detection using ground-based machine vision and image processing techniques. *Computers and Electronics in Agriculture*, Elsevier, v. 158, p. 226–240, 2019.

[32] GABOR, D. Theory of communication. part 1: The analysis of information. *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, IET, v. 93, n. 26, p. 429–441, 1946.

[33] ILEA, D. E.; WHELAN, P. F. Image segmentation based on the integration of colour–texture descriptors—a review. *Pattern Recognition*, Elsevier, v. 44, n. 10-11, p. 2479–2501, 2011.

[34] SOBIERANSKI, A. C. et al. A fast gabor filter approach for multi-channel texture feature discrimination. In: BAYRO-CORROCHANO, E.; HANCOCK, E. (Ed.). *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. Cham: Springer International Publishing, 2014. p. 135–142.

[35] JAIN, A.; FARROKHNIA, F. Unsupervised texture segmentation using gabor filters. *PR*, v. 24, n. 12, p. 1167 – 1186, 1991.

[36] HARALICK, R.; SHANMUGAM, K.; DINSTEIN, I. Textural features for image classification. *IEEE Trans Syst Man Cybern*, SMC-3, p. 610–621, 01 1973.

[37] LECUN, Y. et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, IEEE, v. 86, n. 11, p. 2278–2324, 1998.

[38] KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2012. p. 1097–1105.

[39] SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[40] SZEGEDY, C. et al. Going deeper with convolutions. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2015. p. 1–9.

[41] GIRSHICK, R. Fast r-cnn. In: *Proceedings of the IEEE international conference on computer vision*. [S.l.: s.n.], 2015. p. 1440–1448.

[42] SERMANET, P. et al. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

[43] SZEGEDY, C. et al. Scalable, high-quality object detection. *arXiv preprint arXiv:1412.1441*, 2014.

[44] LIN, T.-Y. et al. Feature pyramid networks for object detection. In: *CVPR*. [S.l.: s.n.], 2017. v. 1, n. 2, p. 4.

[45] GIDARIS, S.; KOMODAKIS, N. Object detection via a multi-region and semantic segmentation-aware cnn model. In: *Proceedings of the IEEE International Conference on Computer Vision*. [S.l.: s.n.], 2015. p. 1134–1142.

[46] DAI, J.; HE, K.; SUN, J. Instance-aware semantic segmentation via multi-task network cascades. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2016. p. 3150–3158.

[47] HARIHARAN, B. et al. Hypercolumns for object segmentation and fine-grained localization. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2015. p. 447–456.

[48] SHELHAMER, E.; LONG, J.; DARRELL, T. Fully convolutional networks for semantic segmentation. *arXiv preprint arXiv:1605.06211*, 2016.

[49] THOMA, M. A survey of semantic segmentation. *arXiv preprint arXiv:1602.06541*, 2016.

[50] BADRINARAYANAN, V.; KENDALL, A.; CIPOLLA, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561*, 2015.

[51] RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: Convolutional networks for biomedical image segmentation. In: SPRINGER. *International Conference on Medical image computing and computer-assisted intervention*. [S.l.], 2015. p. 234–241.

[52] ZHAO, H. et al. Pyramid scene parsing network. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2017. p. 2881–2890.

[53] POHLEN, T. et al. Fullresolution residual networks for semantic segmentation in street scenes. *arXiv preprint*, 2017.

[54] HE, K. et al. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2016. p. 770–778.

[55] RAHMAN, M. A.; WANG, Y. Optimizing intersection-over-union in deep neural networks for image segmentation. In: SPRINGER. *International symposium on visual computing*. [S.l.], 2016. p. 234–244.

[56] RIJSBERGEN, C. J. V. *Information Retrieval*. 2nd. ed. Newton, MA, USA: Butterworth-Heinemann, 1979.

[57] KINGMA, D. P.; BA, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.