WEAKLY SUPERVISED SEMANTIC SEGMENTATION FOR REMOTE SENSING HYPERSPECTRAL IMAGING

Eloi Moliner, Luis Salgueiro Romero and Verónica Vilaplana

Department of Signal Theory and Communications, Universitat Politècnica de Catalunya

ABSTRACT

This paper studies the problem of training a semantic segmentation neural network with weak annotations, in order to be applied in aerial vegetation images from Teide National Park. It proposes a Deep Seeded Region Growing system which consists on training a semantic segmentation network from a set of seeds generated by a Support Vector Machine. A region growing algorithm module is applied to the seeds to progressively increase the pixel-level supervision. The proposed method performs better than an SVM, which is one of the most popular segmentation tools in remote sensing image applications.

***Index Terms*—** Weakly-supervised segmentation, remote sensing, hyperspectral image

1. INTRODUCTION

Semantic segmentation, or pixel-level classification of very high resolution aerial images is of great importance for the task of automated monitoring in various remote sensing applications, like environmental monitoring, urban planning, autonomous agriculture, and others. Two of the biggest difficulties faced by automated methods are the lack of ground-truth annotations defining accurately all the classes, and how hard it is to generate the ground truth manually. A common practice is to annotate a small set of representative Regions of Interest (RoIs) for every class, and use models like Support Vector Machines (SVM) to perform pixel-wise classification [1, 2] .

Deep learning has become the state of the art methodology for semantic segmentation. Most models (like U-Net[3], SegNet[4] or PSPNet[5]) require full annotations of training images, which is unfeasible for remote sensing imagery. Recently, methods based on weaker annotations have started to appear. The concept of weakly supervision means that the model is not trained with fully labeled images but with partial or much simpler annotations, which require much less time to generate.

The objective of this work is to train a semantic segmentation network from weak annotations. Our model is based on the work of Huang et al. [6], where a system named Deep Seeded Region Growing (DSRG) is proposed. This algorithm trains a semantic segmentation neural network starting from a set of small labeled areas known as seeds, and progressively increases the pixel-level supervision using region growing. We adopt the DSRG idea but, instead of generating the seeds with the activation maps of a deep classification network, we generate them with an SVM. The whole system is explained in detail in the following sections.

We apply the technique to hyperspectral aerial images (120 bands) from the Teide National Park in the Canary Islands in order to distinguish autochthonous vegetation, including some endangered species.

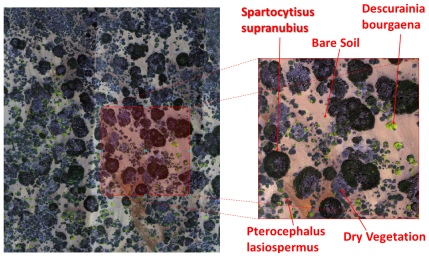


Fig. 1. Teide National Park with the representative vegetation species captured by Pika-L Drone

2. METHOD

2.1. Seed generation with SVM

Figure 2 shows the structure of the proposed system. The first module is the seed generator. In [6], the initial seeds are extracted from the class activation maps of a deep classification neural network. Due to the different nature of the problem, we use a SVM to generate seeds, defining them as regions classified by the SVM with high confidence levels.

The SVM was trained with a small set of fully annotated RoIs, using all the 120 bands. We use a radial basis function kernel with penalty parameter 1.0 and width 0.1. Probability calibration based on isotonic regression has been applied afterwards to obtain estimates of class probabilities.

In the seed generation step, the procedure is the following. Firstly, all the pixels of the hyperspectral input image are fed to the SVM and the calibrated probabilities are extracted. A different probability threshold has been defined for

This work has been partially supported by the ARTEMISAT-2 (CTM2016-77733-R) and MALEGRA TEC2016-75976-R projects financed by the Spanish Ministerio de Economía y Competitividad. L.S.R. would like to acknowledge the BECAL scholarship for the financial support.

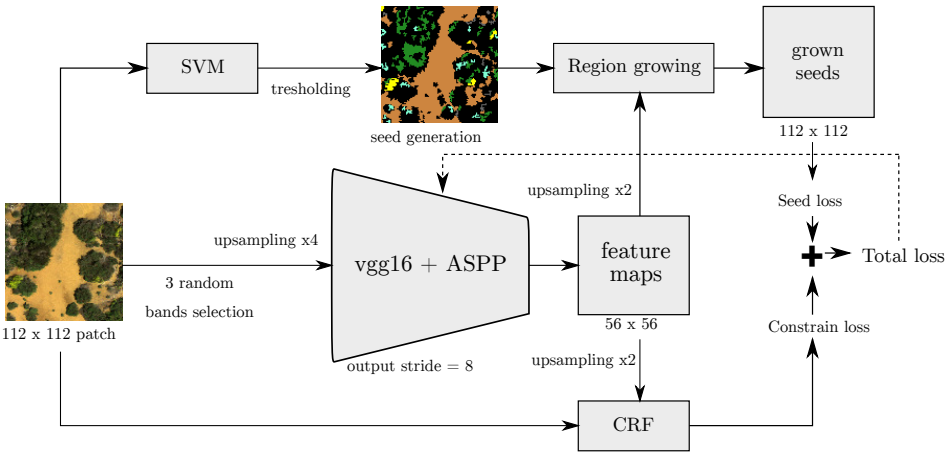


Fig. 2. Main scheme of the proposed system

each one of the classes; threshold values have been chosen in order to have a comparable number of seeds per class. Subsequently, the class probability maps extracted from the SVM are thresholded by these factors. After that, morphological filters are applied with the purpose of filling holes of less than 3 pixels and removing connected components smaller than 5 pixels.

2.2. Semantic segmentation deep neural network

The neural network for semantic segmentation is DeepLabv2-ASPP [7]. It is a modified version of VGG-16 [8] incorporating an Atrous Spatial Pyramid Pooling (ASPP) at the end. The aim of this pooling is to explore multi-scale features by using multiple parallel atrous convolutional layers with different rates. The features at each rate are processed in parallel branches and merged at the end to generate the final result.

Tiles of 112x112 pixels were used for training input images with three input channels, each channel of the tile was selected from the group of bands corresponding to the red, green and blue portions of the spectrum, the selection was done randomly according to a uniform distribution. So, as for each iteration, a triplet of RGB bands were fed to network as an input image. The weights and biases of the network were pretrained on Imagenet[9] and fine-tuned with the Teide dataset.

At the end of the ASPP, a softmax layer is applied in order to normalize the resulting feature maps, which will be used for the region growing step.

2.3. Deep seeded region growing

The idea behind the region growing block is to expand the initial seeds using the feature maps (output from the segmentation network) as references. This way, we will have more pixel-level annotations that will become more accurate as the network training progresses.

To formulate the seed cues growing problem, [6] use a classical algorithm named Seeded Region Growing (SRG) [10]. The SRG algorithm defines as seeds small homogeneous regions based on simple color, intensity or texture criteria. Then, regions are grown from the seeds based on a similarity criterion that decides whether a candidate pixel should be added or not to a specific region. Our model grows the SVM seeds using a similarity criterion P, the probability threshold value of a pixel in the segmentation feature map H generated by the segmentation network.

$$\begin{cases} P(H_{u,c}, \theta_c) = \text{True} & H_{u,c} \geq \theta_c \;\&\; c = \text{argmax}_c[H_{u,c}] \\ P(H_{u,c}, \theta_c) = \text{False} & \text{otherwise} \end{cases}$$

where $H_{u,c}$ is the probability that a pixel at position u belongs to class c. This criterion basically expands the seeds to only those adjacent pixels in which the segmentation network has a very high confidence in the decision, higher than a given threshold θ_c. Note that at the early stages of the training process, the feature maps may induce the region growing block to expand the seeds to erroneous pixels. But, as the training progresses, the feature maps are more accurate and the region expanding process improves. The algorithm is iterative, at each iteration it visits every single pixel in the image for every class, computes the similarity criteria $P(H_{u,c}, \theta_c)$ with every 8-connectivity pixel neighbour, and generates a set of newly labeled pixels S_c. After that, the algorithm revisits the new set S_c and updates the seeds. Once S_c is changed, it will visit S_c again, otherwise it stops. The resulting seeds S_c at the end will be used to compute the seed loss and train the segmentation network.

2.4. Seeding and constrain losses

The total loss is computed from the sum of two separate losses: seeding loss and constrain-to-boundary Loss.

The seeding loss is proposed to encourage predictions of the segmentation network to match only seed cues given by the classification network while ignoring the rest of the pixels in the image. The seed loss is defined as follows:

$$l_{seed} = -\frac{1}{\sum_{c \in C} |S_c|} \sum_{c \in C} \sum_{u \in S_c} \log H_{u,c}$$

Where C is the set of classes, Sc is a set of locations that are classified to class c and $H_{u,c}$ is the probability that a pixel in position u belongs to class c. This seeding loss could be defined as a cross-entropy between the seeds and the feature maps generated by the segmentation network.

Besides that, a constrain-to-boundary loss is included in the system, which was firstly proposed in [11]. The idea is to construct a Conditional Random Field (CRF) $Q(X, H)$ as in [12]. The CRF unary potentials are given by the feature maps from the output of the segmentation network and the pairwise potentials depend only on the original image pixels. The loss is defined as the mean KL divergence between the outputs of the network and the outputs of the CRF.

$$l_{constrain} = \frac{1}{n} \sum_{u=1}^{n} \sum_{c \in C} Q_{u,c}(X, H) \log \frac{Q_{u,c}(X, H)}{Hu, c}$$

This loss encourages the network to match up with object boundaries. The total loss is computed as the sum of both:

$$l_{total} = l_{seed} + l_{constrain}$$

3. EXPERIMENTS

3.1. Teide dataset

The dataset consists of an aerial image captured by a Pika-L sensor mounted on a drone. The sensor is a VNIR (Visible and Near-Infrared) hyperspectral sensor of high spectral coverage (400 a 1000 nm) [13].

The image data has a total of 120 bands from 420 nm to 900 nm, and due to the fact that the drone flies at low altitude, the resulting images have a very high spatial resolution of 0.1 m/pixel. This characteristic makes our problem differ significantly from working with satellite imagery or even images captured by a plane, because the useful information for training the model is not only the spectral information at pixel level, but also the spatial shape of the species we want to segment. Our aim is to design a model that explores both spatial and spectral information.

The drone swept the study area forming tiles that were georeferenced and concatenated generating a raster image of 1327x1497 pixels covering the center of Teide National Park.

In general, the park has an arid appearance but several species can be identified; the most representative are *Descurainia bourgaena*, *Pterocephalus lasiospermus* and *Spartocytisus supranubius*. Two more classes were included, Bare soil and Dry vegetation (see Figure 1).

Fig. 3. Representation of the selected cuts, for SVM training (blue), network training (red) and testing (green)

3.2. Experiment setup

Three differentiated cuts are extracted from the original raster (see figure 3). The first cut (blue) contains a small set of annotated RoIs that are used for the SVM training. The second cut (red) is used for training the DSRG segmentation network. It was divided in patches of 112x112 pixels. This cut does not contain any annotated RoI, but its pixels are meant to be classified with the SVM trained with the first cut. The original seeds are extracted after thresholding the SVM output probabilities. The last and smallest cut (green) is used for validation, it contains a smaller set of ROI annotations for every class.

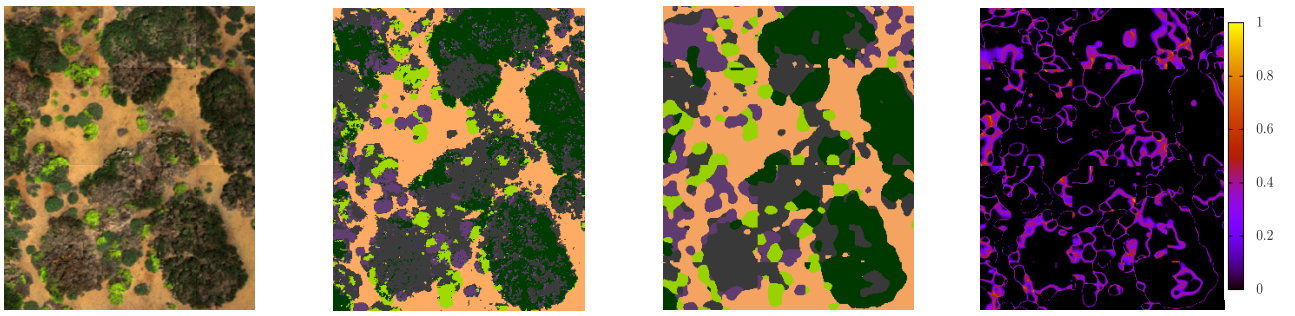
The input tiles that were fed to the proposed system are 112x112 patches of only 3 bands (channels). The bands change in each iteration. They are selected randomly but keeping each channel in the spectral range corresponding to the R, G and B portions of the spectrum. This change adds some diversity to the data and helps to regularize the network.

The model has been trained for 40 epochs, using momentum optimizer with a parameter of 0.95. The batch size is 4 and the gradient accumulation number is 4, meaning that the gradients are updated after 4 batches of 4 images each. The learning rate is 10^{-5}, quite low but enough to guarantee that the algorithm converges even if it requires a large number of iterations.

3.3. Testing and majority voting

The test set is also tiled with 112x112 pixels, and 3 random bands are chosen in the same way as in the training. The patches are fed as inputs into the network and an argmax layer is added at the output in order to obtain the segmented image. Then, all the patches are concatenated reconstructing the entire test region.

We wanted to measure the consistency in the results when selecting different bands, checking that the segmentation results do not differ too much. To do so, we have conducted the following experiment. We have executed the test procedure several times changing the RGB bands randomly and we have



(a) Original image (b) SVM (c) CNN after major voting (d) Entropy consistency result

Fig. 4. Segmentation results on a sample patch. The color code used is: gray for dry vegetation, dark green for *Spartocytisus*, light green for *Descurainia*, purple for Pterocephalus and beige for Bare soil.

computed the class histogram for each pixel. Consequently, we have computed the information entropy of each of the histograms as a metric of uncertainty. Lower values of entropy would mean higher consistency. Repeating the test procedure several times, choosing randomly different bands each time, leads us to a set of similar results with small differences between them. To take profit of this variance we obtain the final decision by performing a majority voting with the set of all predictions. Each pixel in the image is classified according to the most voted class from all of the predictions. Specifically, we have performed a total of 125 parallel tests.

4. RESULTS

Figure 4 shows a small portion of the test cut along with the result from an SVM and the proposed method. The figure also includes the entropy results from the consistency test explained in 3.3. Looking at the images, it can be seen that the SVM, despite being able to distinguish quite well the different classes, makes a lot of small mistakes due to the pixel level classification. On the other side, the DSRG method is able to generate softer and cleaner regions.

Looking at the entropy image, we can see that most of the values are close to 0, this means that the network always classifies the pixels in the same way independently of the chosen bands. Even though, there are some smaller regions with higher entropy, specially in region borders and regions which are easier to confuse and there is a discordance between two or more classes.

Tables 1 and 2 show the results obtained by the classification using the SVM and DSRG algorithm, after being compared with the ROIs set in the test cut (1568x672 pixels). The support column in both tables shows the amount of ground-truth pixels for each class with three traditional metrics for multi-class classification: Precision, Recall and F1-score.

As can be noticed, the results of DSRG work well for the test set, improving the metrics in some classes like Bare Soil or *Spartocytisus*. In the other classes, the results were similar to SVM results, improving it a little bit in average.

	Precision	Recall	F1-score	Support
Pterocephalus l.	0.98	0.93	0.95	9771
Dry vegetation	1.00	0.98	0.99	9749
Descurainia b.	1.00	0.98	0.99	7119
Spartocytisus s.	0.89	0.98	0.93	10195
Bare Soil	0.92	0.90	0.91	12103
Accuracy			0.95	48937

Table 1. SVM classification results

	Precision	Recall	F1-score	Support
Pterocephalus l.	0.92	0.93	0.92	9771
Dry vegetation	1.00	0.94	0.97	9749
Descurainia b.	0.99	0.97	0.98	7119
Spartocytisus s.	0.94	0.96	0.95	10195
Bare Soil	0.98	1.00	0.99	12103
Accuracy			0.96	48937

Table 2. DSRG classification results

5. CONCLUSIONS

Initial experiments with a weakly-supervised approach present promising results. The objective metrics are better than SVM metrics and, moreover, the segmented image looks much cleaner. Even though, there is much more work to do. The proposed system is able to distinguish quite well the five classes, but in some cases has a lack of precision, e.g. it is not able to distinguish the smaller regions. It is a future task to train the model with all the available bands in order to make use of all the spectral information. It is also pending to apply the proposed model to other datasets with different classes, which would be useful to evaluate how adaptive the system is to different kinds of data.

6. REFERENCES

[1] Anabella Medina Machín, Javier Marcello, Antonio I. Hernández-Cordero, Javier Martín Abasolo, and Francisco Eugenio, "Vegetation species mapping in a coastal-dune ecosystem using high resolution satellite imagery," *GIScience & Remote Sensing*, vol. 56, no. 2, pp. 210–232, 2019.

[2] Alberto Signoroni, Mattia Savardi, Annalisa Baronio, and Sergio Benini, "Deep learning meets hyperspectral image analysis: A multidisciplinary review," *Journal of Imaging*, vol. 5, no. 5, pp. 52, 2019.

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[4] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[5] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[6] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang, "Weakly-supervised semantic segmentation network with deep seeded region growing," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7014–7023, 2018.

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, Apr 2018.

[8] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[10] Rolf Adams and Leanne Bischof, "Seeded region growing," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 16, no. 6, pp. 641–647, 1994.

[11] Alexander Kolesnikov and Christoph H. Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," *CoRR*, vol. abs/1603.06098, 2016.

[12] Philipp Krähenbühl and Vladlen Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," *CoRR*, vol. abs/1210.5644, 2012.

[13] Resonon Inc., "Hyperspectral imaging cameras - datasheed," "https://resonon.com/content/products-files/ResononHyperspectralCameras.Datasheet-3.pdf", Accessed: 21-10-2019.