Electronic Letters on Computer Vision and Image Analysis 0(0):1-7, 2020

Increasing the Segmentation Accuracy of Aerial Images with Dilated Spatial Pyramid Pooling

Manuel Eugenio Morocho-Cayamcela[*,†]

[*] *Dept. of Electronic Engineering, Kumoh National Institute of Technology, Gumi-si, Republic of Korea.*
[†] *Currently at CIDIS, Escuela Superior Politécnica del Litoral, Guayaquil, Ecuador.*
Date and location of PhD defense: May 30, 2020 at Kumoh National Institute of Technology.
Doctoral Advisor: Wansu Lim[+].

Received 1 January 2000; revised 1 January 2000; accepted 1 January 2000.

Abstract

This thesis addresses the environmental uncertainty in satellite images as a computer vision task using semantic image segmentation. We focus in the reduction of the error caused by the use of a single-environment models in wireless communications. We propose to use computer vision and image analysis to segment a geographical terrain in order to employ a specific propagation model in each segment of the link. Our computer vision architecture achieved a segmentation accuracy of 89.41%, 86.47%, and 87.37% in the urban, suburban, and rural classes, respectively. Results indicate that estimating propagation loss with our multi-environment model reduced the root mean square deviation (RMSD) with respect to two publicly available tracing datasets.v

Key Words: Computer Vision, Image Analysis, Pattern Recognition Image Segmentation, Supervised Learning, Wireless Communications.

1 Introduction

A radio propagation model is a mathematical formulation that characterizes the radio-wave condition as a function of the environment between the antennas in a wireless system. Because each wireless link exists under different conditions, it is difficult to express a mathematical equation that includes all the link environments. In this thesis, we used computer vision and image segmentation to estimate the environment using satellite images. As with image classification, convolutional neural networks (CNNs) have shown potential on image segmentation problems [1, 2]. Despite the up-convolutional layers, fully connected neural networks (FCNs) produce coarse segmentation maps due to loss of information during pooling [3].

In this thesis, we generated a categorical matrix by segmenting aerial images into urban, suburban, and rural classes, enabling the estimation of an environment-specific propagation loss for each segment of the link. We

Correspondence to: <eugeniomorocho@ieee.org>
This thesis is fully available online at: http://dcollection.kumoh.ac.kr/srch/srchDetail/000000015745
This extended abstract is inspired on a thesis chapter published in IEEE Transactions on Wireless Communications: https://ieeexplore.ieee.org/document/9066907 © 2020 IEEE.
Recommended for acceptance by <name>
ELCVIA ISSN:1577-5097
Published by Computer Vision Center / Universitat Autònoma de Barcelona, Barcelona, Spain

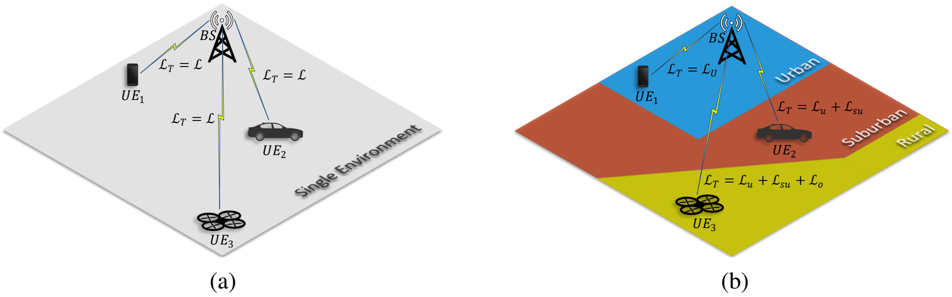
BS
$\mathcal{L}_T = \mathcal{L}$
$\mathcal{L}_T = \mathcal{L}$
$\mathcal{L}_T = \mathcal{L}$
UE_1
UE_2
UE_3
Single Environment
(a)
BS
$\mathcal{L}_T = \mathcal{L}_U$
Urban
UE_1
$\mathcal{L}_T = \mathcal{L}_u + \mathcal{L}_{su}$
Suburban
Rural
UE_2
$\mathcal{L}_T = \mathcal{L}_u + \mathcal{L}_{su} + \mathcal{L}_o$
UE_3
(b)

Figure 1: Total propagation loss assuming (a) single environment, and (b) a multi-environment scenario. © 2020 IEEE.

increased the semantic segmentation accuracy of an encoder-decoder architecture by expanding the receptive field of view of a pre-trained CNN in the encoder, and introducing low-level features and fast bi-linear up-sampling at the decoder.

2 Methodology

We built an image segmentation network to recognize the three environments: urban, suburban, and rural. In Fig. 1(a), the total propagation loss in the user equipment UE_1, UE_2, and UE_3 is estimated using a single environment model $\mathcal{L}$. Fig. 1(b) shows our proposed method, where the propagation loss is estimating the segments of the three classes (urban $\mathcal{L}_u$, suburban $\mathcal{L}_{su}$, and rural $\mathcal{L}_o$). Our system is trained with multiple examples (satellite images) $\boldsymbol{x}$ of a class, along with their pixel-by-pixel label $\boldsymbol{y}$. To guarantee that our segmentation model can be generalized to any city, we have used the INRIA aerial image dataset [4]. INRIA is a collection of 810 km^2 high-resolution aerial ortho-rectified color images with a spatial resolution of 0.3 m/pixel from different European and American cities. The classes were labeled pixel-by-pixel according to the following criteria: *Urban* includes dense cities and its streets, residential areas, buildings, and any field where the signal can be blocked by constructions or man-made structures. *Suburban* includes trees and large parks inside cities, forest regions, and any field where the signal can be blocked by foliage, trees, or vegetation. *Rural* includes open lands without signal blockage, rivers, oceans, or wherever a line of sight between the transmitter and receiver antenna can be guaranteed. The areas that did not meet any of the criteria were not labeled.

3 System Model

To exploit multi-scale features in the dataset, we employ an *encoder-decoder* structure [3,5] to perform feature-dense extraction (Fig. 2). The encoder downscale the satellite images to a feature vector containing dense location information, and the decoder expand the summarized feature vector back into a categorical matrix with the original input dimension. The backbone of the encoder architecture is based on ResNet-101, a 101-layer CNN pre-trained on ImageNet, composed of five convolution (*Conv*) modules. We modify the last convolutional block into parallel versions to impose dilated spatial pyramid pooling at different scales to guarantee the robustness to changes in the environment size, as they encode multi-scale contextual information. As the sampling rate increases, the number of valid filter weights decreases. The resulting features from all branches are then concatenated and passed through another convolution and batch normalization before the final 1 × 1 convolution. The decoder computes feature responses by incorporating low-level features of the encoder, regular convolutions, and fast bilinear interpolation by a factor of eight before generating the final categorical matrix. Cross-entropy was selected as the segmentation model performance metric because as the predicted probability diverges from the ground truth label, the loss increases. To damp oscillations in directions of high curvature,

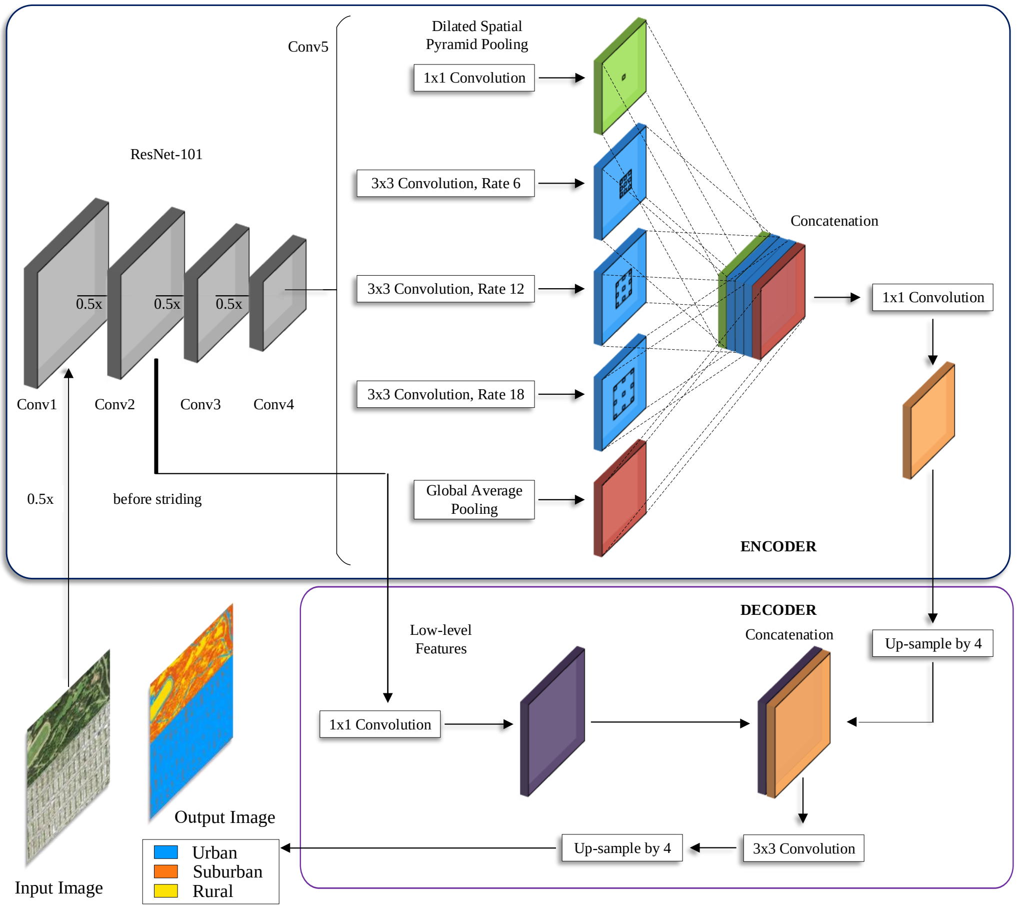

Figure 2: The model architecture employing an encoder-decoder structure. © 2020 IEEE.

we combine the gradients with opposite signs, building up speed in directions with a consistent gradient. We add a fraction of the update of the previous step to amplify the speed and the correct direction.

4 Brief Results

The performance measure used to compare our model with other systems is *segmentation accuracy*. For comparison purposes, our multiple-environment propagation loss estimator under study is referred to as *SegNet-Prop* (Segmentation Network Propagation model). In our simulations, the initial learning rate hyperparameter α controls how much the weights in the convolutional encoder-decoder change in response to the pixel wise cross-entropy estimation error. Fig. 3 shows the three classes sectioned with our improved encoder-decoder architecture. The sections in blue, orange, and yellow, correspond to the urban, suburban, and rural classes, respectively. The accuracy values achieved with *SegNetProp* in the segmentation of the INRIA dataset are presented in Table 1. We have increased the segmentation accuracy of an FCN segmenter by 3.17%, 4.03%, and 4.28%, for urban, suburban, and rural classes, respectively.

5 Conclusions

This thesis summary shows that fundamental ideas borrowed from computer vision and image analysis can be useful in wireless communications problems. First we have improved a convolutional encoder-decoder architecture to produce semantic maps of aerial images, and then we have employed the generated map to

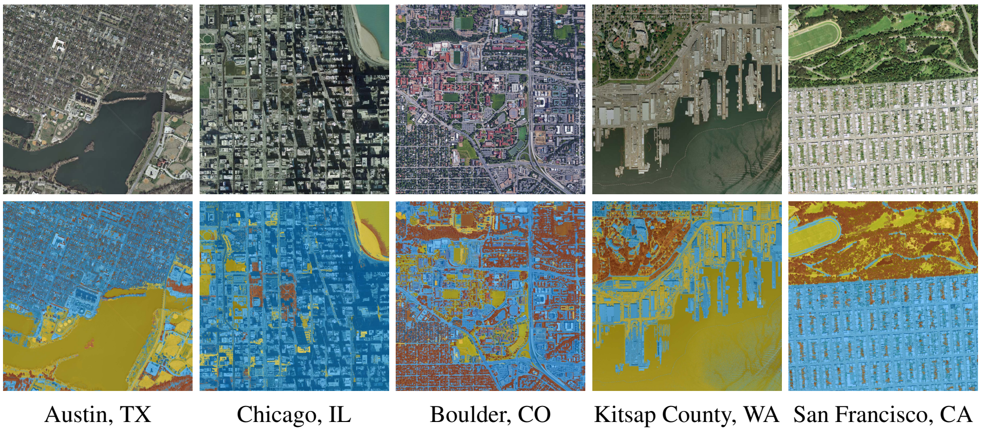

Figure 3: A subset of images from the dataset used for validating and testing the segmentation neural network (top), and their corresponding automatic *pixel-segmented* data (bottom). © 2020 IEEE.

Table 1: Segmentation Accuracy obtained with different Models*. © 2020 IEEE.

Environment Classes	Texton Forests [6]	Patch Based [7]	FCNs [2]	U-Net [8]	**Proposed Model**† ‡
Urban	45.82%	84.03%	86.24%	87.28%	**89.41%**
Suburban	41.07%	82.12%	82.44%	84.03%	**86.47%**
Rural	43.63%	83.65%	83.09%	85.15%	**87.37%**

*Trained using 4 NVIDIA GTX 1080Ti with local parallel pool.

†Convol. encoder-decoder architecture, optimized with stochastic gradient descent with mom.

create a model that estimates the propagation loss by segments. In addition, the proposed architecture may be retrained with data from new cities or extended to include more specific environments, such as urban micro, urban macro, rural macro, streets, open air festival, stadium, etc.

References

[1] M. E. Morocho-Cayamcela, M. Maier, and W. Lim, "Breaking wireless propagation environmental uncertainty with deep learning," *IEEE Transactions on Wireless Communications*, vol. 19, no. 8, pp. 5075–5087, 2020.

[2] E. Shelhamer, J. Long, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 640–651, 4 2017.

[3] V. Badrinarayanan, A. Kendall, and R. 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, 2015.

[4] The National Institute for Research in Computer Science and Automation, "Inria Aerial Image Labeling Dataset," 2016.

[5] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "High-Resolution Aerial Image Labeling with Convolutional Neural Networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, pp. 7092–7103, 12 2017.

[6] J. Shotton, M. Johnson, and R. Cipolla, "Semantic texton forests for image categorization and segmentation," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, IEEE, 6 2008.

[7] D. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Deep Neural Networks Segment Neuronal Membranes in Electron Microscopy Images," in *NIPS Proceedings: Advances in Neural Information Processing Systems*, pp. 2843–2851, 2012.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9351, pp. 234–241, 2015.