Counting the uncountable: deep semantic density estimation from space

Andres C. Rodriguez, Jan D. Wegner
Photogrammetry and Remote Sensing, ETH Zürich

We propose a new method to count objects of specific categories that are significantly smaller than the ground sampling distance (GSD) of a satellite image. This task is hard due to the cluttered nature of scenes where different object categories occur. Target objects can be partially occluded, vary in appearance within the same class and look alike to different categories. Sentinel-2 satellite configuration provides since 2015 multi-spectral images of up to 10 meters GSD.

However, for applications that need more high-resolution evidence like detecting and counting objects (e.g, supply chain management, financial industry), spatial resolution is too poor to apply traditional object detectors like Faster R-CNN (Ren et al., 2017). In this work, we thus propose to circumnavigate explicit object detection by turning the counting problem into a density estimation task. Furthermore, we add semantic segmentation to be able to count objects of very specific object categories embedded in cluttered background.

Semantic Segmentation is a standard task in computer vision and has seen significant performance gains since the comeback of deep learning. The major goal is predicting a class label per pixel over an entire image. Typical objects classes of interest are buildings, persons and vehicles that are (i) clearly visible in the image, (ii) large in size (usually several hundreds of pixels) (iii) and can be distinguished from background and further classes primarily relying on shape and RGB texture.

Objects like cars, trees, and buildings constitute a single pixel or less in remote sensing imagery of Sentinel-2 (10m GSD) and Landsat (15m GSD). While this resolution is sufficient for semantic segmentation of large, homogeneous regions like crops, counting individual object instances in cluttered background becomes hard. Here, high spectral resolution comes to the rescue. Deep learning techniques can greatly benefit from high spectral resolution that conveys object information invisible to human sight. It can learn complex relations between spectral bands to identify object-specific spectral signatures that support pixel-accurate semantic segmentation.

To distinguish objects of different classes, our approach combines density estimation with semantic segmentation in an end-to-end learnable convolutional neural network (CNN) to count objects of down to 1/10 the size of the GSD. We compare our proposed architecture with state-of-the-art semantic segmentation methods for terrestrial images that use among other ideas atrous convolutions to prevent lowering the resolution of the learned features keeping a large receptive field (Chen et al., 2017).

It is impossible to manually segment individual trees or cars in the Sentinel-2 images of 10 meters GSD. We thus resort to Google Maps overhead images of much higher resolution for groundtruth labeling. We apply the Faster R-CNN Object detector to very high-resolution (1m GSD) Google Maps images to identify and count reference objects. The detector is tuned to achieve high recall to then manually remove false positive predictions. This allows us to obtain a highly detailed count per area that is used as ground truth to test our model on lower resolution satellite imagery.

Our approach consists of using ResNet blocks, which contain mainly convolutional layers, to obtain features from the input image without any downsampling. Then, for each task, we add an independent convolutional layer at the end of each architecture. For all training experiments, our Loss is defined as the sum of (1) cross entropy of the estimated and reference class (object of interest or backround) and (2) the euclidian distance of the estimated and reference count.

In order to test how robust deep semantic density estimation is to changes in texture, object density and size, we create four different datasets. Three datasets contain trees (olive, coconut, oil palm) with different planting patterns and one dataset contains cars. The tree and car sizes with respect to the pixel area size were approximately of 1/3 and 1/10.

Experiments on four different objects show that deep semantic density estimation can robustly count objects of various classes at sub-pixel scales in cluttered scenes. In our tests, we were able to predict the total object count with an error of less than 5%. For the semantic segmentation task of Olive Trees, our architecture obtained Intersection over Union of 0.86 and precision of 0.90 in our test set; in contrast to other architectures that showed poor performance. See Figure 1 for a visualization or our results. Experiments with trees (Olives, Coconuts and Palms) show the importance of infrared bands in the prediction. In contrast, Cars benefited mostly from the high spatial resolution of the RGB bands.

We have proposed end-to-end learnable deep semantic density estimation for counting object instances of fine-grained classes in cluttered background. Results show that counting objects of sub-pixel size is possible for 10m GSD satellite images. Experimental evaluation with four datasets shows that our method is robust to change in object types, background, and object density. It turns out that a shallow network specifically designed for satellite imagery of 10m GSD and sub-pixel objects outperforms more sophisticated, state-of-the-art architectures from computer vision. This signifies that direct application of networks tailored for vision to remote sensing images should be done with care. In our experiments, we find that any down-sampling operation inside the network risks loosing precious details. We should thus always keep in mind the particularities of remote sensing imagery in terms of object scale, GSD, (nadir) perspective and (high) spectral resolution. If carefully considered during the network design process, these specific properties offer new possibilities in network design.

—

Chen, Liang-Chieh, et al. (2018) Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence 40(4):834–848

Ren, Shaoqing, et al. (2015) Faster R-CNN: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems.