ISPRS Journal of Photogrammetry and Remote Sensing 144 (2018) 48–60

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Deep multi-task learning for a geographically-regularized semantic segmentation of aerial images

Michele Volpi[a,*], Devis Tuia[b]

[a] Swiss Data Science Center, ETH Zurich, Switzerland
[b] Laboratory of GeoInformation Science and Remote Sensing, Wageningen University and Research, The Netherlands

ARTICLE INFO

Keywords:
Semantic segmentation
Semantic boundary detection
Convolutional neural networks
Conditional random fields
Multi-task learning
Decimeter resolution
Aerial imagery

ABSTRACT

When approaching the semantic segmentation of overhead imagery in the decimeter spatial resolution range, successful strategies usually combine powerful methods to learn the visual appearance of the semantic classes (e.g. convolutional neural networks) with strategies for spatial regularization (e.g. graphical models such as conditional random fields).

In this paper, we propose a method to learn evidence in the form of semantic class likelihoods, semantic boundaries across classes and shallow-to-deep visual features, each one modeled by a multi-task convolutional neural network architecture. We combine this bottom-up information with top-down spatial regularization encoded by a conditional random field model optimizing the label space across a hierarchy of segments with constraints related to structural, spatial and data-dependent pairwise relationships between regions.

Our results show that such strategy provide better regularization than a series of strong baselines reflecting state-of-the-art technologies. The proposed strategy offers a flexible and principled framework to include several sources of visual and structural information, while allowing for different degrees of spatial regularization accounting for priors about the expected output structures.

1. Introduction

This paper deals with parsing decimeter resolution abovehead images into semantic classes, relating to land cover and/or land use types. We will refer to this process as *semantic segmentation*. For a successful segmentation, one requires visual models able to disambiguate local appearance by understanding the spatial organization of semantic classes (Gould et al., 2008). To this end, machine learning models need to exploit different levels of spatial continuity in the image space (Campbell et al., 1997; Shotton et al., 2006). Accurate land cover and land use mapping is an active research field, growing in parallel to developments in sensors and acquisition systems and to data processing algorithms. Applications ranging from environmental monitoring (Asner et al., 2005; Giménez et al., 2017) to urban studies (Zhong and Wang, 2007; Jat et al., 2008) benefit from advances in processing and interpretation of abovehead data.

Semantic segmentation of sub-decimeter aerial imagery is often tackled by Markov and conditional random fields (MRF, CRF) (Besag, 1974; Lafferty et al., 2001) combining local visual cues (the *unary* potentials) and interaction between nearby spatial units (the *pairwise* potentials) (Kluckner et al., 2009; Hoberg et al., 2015; Zhong and Wang, 2007; Shotton et al., 2006; Volpi and Ferrari, 2015). By maximizing the posterior joint probability of a CRF over the labeling (i.e. minimizing a Gibbs *energy*), one retrieves the most probable *labeling* of a given scene, i.e. the most probable configuration of local label assignments over the whole image space. These frameworks allow to model jointly bottom-up evidence, encoded in the unary potentials, together with some domain specific prior information encoded in the spatial interaction pairwise terms.

The idea behind the proposed model is that, when dealing with urban imagery (and in general decimeter resolution imagery), both the content of the image and the classes are highly structured in the spatial domain, calling for data- and domain-specific regularization. To follow such intuition, we model two key aspects of spatial dependencies: *input* and *output* space interactions. The former are usually encoded by operators accounting for the spatial autocorrelation of pixels in their spatial domain. The latter are encoded by different kinds of pairwise potential, favoring specific configurations issued from a predefined prior distribution.

* Corresponding author.
E-mail addresses: michele.volpi@sdsc.ethz.ch (M. Volpi), devis.tuia@wur.nl (D. Tuia).
URL: https://sites.google.com/site/michelevolpiresearch/ (M. Volpi).

https://doi.org/10.1016/j.isprsjprs.2018.06.007
Received 11 December 2017; Received in revised form 4 June 2018; Accepted 7 June 2018

– To extract information about local *input* relations, we combine state-of-the-art convolutional neural networks (CNN, LeCun et al., 1998; Simonyan and Zisserman, 2015; Krizhevsky et al., 2012) providing data-driven cues for multiple tasks: We employ a CNN to not only provide approximate class-likelihoods, but also to predict semantic boundaries between the different classes. The latter coincide usually with natural edges in the image, but also corresponding to changes in labeling. Then, we build a segmentation tree using the semantic boundaries predicted by the CNN. Such tree represents hierarchy of regions spanning from the lowest level defined by groups of pixels (or superpixels) to the highest level, the whole scene. The region partitioning depends jointly on shallow-to-deep visual features *and* the semantic boundaries learned by the multi-task CNN.
– To account for the *output* relations between regions, we combine the information within each region in a hierarchy using a top-down graphical model including different key aspects of the spatial organization of labels, given the observed inputs. This second modeling step is based on a CRF that aims at reducing the complexity (i.e. regularizing) of the pixel-wise maps, by semantically and spatially parsing consistent regions of the image, likely to belong to given classes, at different scales. Specifically, the CRF model takes into account evidence from the CNN (class-likelihoods, learned visual features and presence of class-specific boundaries) and spatial interactions (label smoothness, label co-occurrence, region distances, elevation gradient) within the hierarchy. In other words, it learns the extent *and* the labeling of each segment simultaneously, by minimizing a specifically designed energy.

A visual summary of the proposed pipeline is presented in Fig. 1.

We evaluate all the components of the system and show that spatial regularization is indeed useful in simplifying class structures spatially, while achieving accurate results. Since spatial structures are learned and encoded directly in the output map, we believe our pipeline is a step towards systems yet based on machine learning, but not requiring extensive manual post-processing (e.g. local class filtering, spatial corrections, map generalization, fusion and vectorization Crommelinck et al., 2016; Höhle, 2017), at the same time employing domain knowledge and data specific regularization, tailoring it to specific application domain and softening black-box effects. Specifically, the contributions of this paper are:

– A detailed explanation on our multi-task CNN, building on top of a pretrained network (VGG).
– A strategy to transform semantic boundaries probabilities to superpixels and hierarchical regions.
– A CRF encoding the desired space-scale relationships between segments.
– The combination of different energy terms accounting for multiple input-output relationships, combining bottom-up (outputs and features of the CNN) and top-down (multi-modal clues about spatial arrangement) into local and pairwise relationships.

In the next section, we summarize some relevant related works. In Section 3, we present the proposed system: the multi-task CNN architecture (Section 3.1), the hierarchical representation of image regions (Section 3.2) and the CRF model (Section 3.3). We present data and experimental setup in Section 4 and the results obtained in Section 5. We finally provide a discussion about our system in Section 6, leading to conclusions presented in Section 7.

2. Related works

2.1. Mid-level representations

To generate powerful visual models, traditional methods compute local appearance models mapping locals descriptors to labels, over a dense grid covering the image space. Then, the relationships between output variables are usually modeled by MRF and CRF. Standard approaches to local image descriptors involve the use of local color statistics, texture, bag-of-visual-words, local binary patterns, histogram of gradients and so on (Kluckner et al., 2009; Hoberg et al., 2015; Zhong and Wang, 2007; Volpi and Ferrari, 2015; Shotton et al., 2006).

However, the use of fixed shape operators causes an inevitable loss of geometrical accuracy, in particular on objects borders. The most common solution to this problem is to employ a locally adaptive spatial support, to either summarize precomputed dense descriptors or to retrieve new ones. This strategy is usually implemented by the use of superpixels, which are defined to be small spatial units, uniform in appearance, while matching the natural edges in the image (Felzenszwalb and Huttenlocher, 2004). Moreover, superpixels significantly reduce the number of atomic units to be processed and consequently the computational time.

2.2. Deep representations

Convolutional neural networks (CNN), learn a parametric mapping from inputs to outputs, sidestepping the definition of (i) an explicit processing resolution, (ii) the type of appearance descriptors best representing the data, (iii) a classifier to map descriptors to output labels (LeCun et al., 1998; Simonyan and Zisserman, 2015; Krizhevsky et al., 2012). The fact that a CNN learns an end-to-end mapping from data to outputs directly *within* the network makes deep neural networks more than valid alternatives to classical models, since complex feature engineering is avoided. However, most of these methods still rely on a fixed spatial support, either in the form of a patch to be classified (for patch classification) or in the form of the convolution kernel (for fully convolutional architectures) (Long et al., 2015). This can cause boundary blur and loss of definition on small image details. But despite

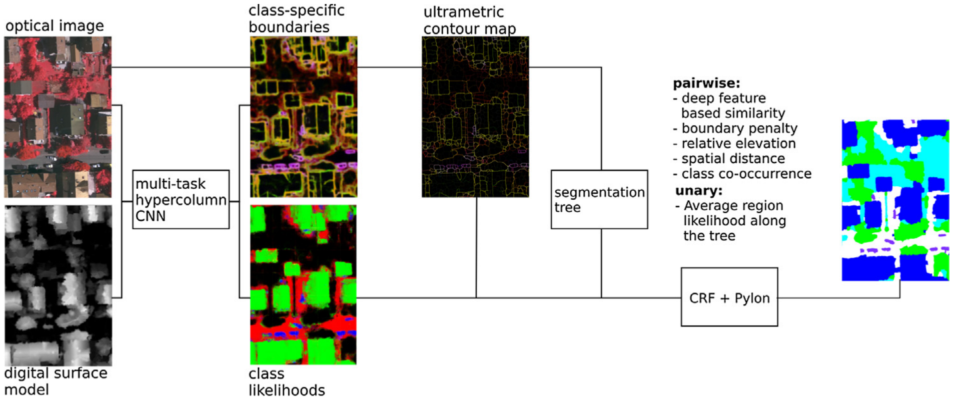

Fig. 1. Flowchart of the proposed pipeline: from multi-modal inputs and multi-task learning to multi-modal and geographically regularized semantic segmentation.