Knowledge-Based Systems 99 (2016) 19–27

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Supervised remote sensing image segmentation using boosted convolutional neural networks

Essa Basaeed*, Harish Bhaskar, Mohammed Al-Mualla

Department of Electrical and Computer Engineering, Khalifa University, P.O.Box 127788, Abu Dhabi, United Arab Emirates

ARTICLE INFO

Article history:
Received 14 May 2015
Revised 18 January 2016
Accepted 20 January 2016
Available online 27 February 2016

Keywords:
Image segmentation
Artificial neural networks
Multispectral imaging
Remote sensing

ABSTRACT

In this paper, a region segmentation technique for remote sensing images using a boosted committee of Convolutional Neural Networks (CNNs) coupled with inter-band and intra-band fusion, is proposed. The vast heterogeneity in remote sensing images restricts the application of existing segmentation methods that often rely on a set of predefined feature detectors along with tunable parameters. Therefore, it is highly challenging to design a segmentation technique which could achieve high accuracy while simultaneously maintaining strong generalization particularly for visual data with improved spatial, spectral, and temporal resolutions. The proposed method is a fusion framework consisting of a set of thirty boosted networks that derive individual probability maps on the location of region boundaries from the different multi-spectral bands and combines them into one using an averaging inter-band fusion scheme. The boundaries are then thinned, connected, and region segmented using a morphological intra-band fusion scheme. Qualitative and quantitative results, on publicly-available datasets, confirm the superiority of the proposed segmentation method over existing state-of-art techniques. In addition, the paper also demonstrates the effect of some variations in design-choices of the proposed method.

1. Introduction

Image segmentation has gained significant recognition as an autonomous, efficient yet complementary approach to the human evaluation of high-resolution and multi-spectral remotely sensed data. The segmentation of high-resolution remotely sensed images is fundamental to object-oriented classification which in turn has an impact on applications such as change detection [14], monitoring [32], urbanization [6], and target detection [20], among many others. Image segmentation of remotely sensed images is complicated by the decreased heterogeneity across different land covers, increased spectral variability within the same land-cover due to high-resolution [15], and complex characterization caused by multidimensionality of the data [35]. Despite complexities, there is a growing demand for newer segmentation techniques in order to cope with the rapid developments in sensor technology exploiting enhanced spectral and spatial resolutions [18].

In the recent years, several methods have been proposed to segment remotely sensed images. Such techniques can be conveniently summarized into four broad categories: pixel-based, edge-based, region-based, and knowledge-based segmentation [31,34]. Pixel-level segmentation methods exploit the gray-level of the individual pixels in order to derive the segmentation (a.k.a., threshold methods). Edge-based methods aim to find discontinuities (i.e., edges) in the image and then assume that these discontinuities represent region boundaries. Region-based methods start with a seed pixels (or multiple seeds) and then split or grow the seed until the image is fully segmented [31]. Finally, knowledge-based methods incorporate semantic knowledge about the image content and/or the domain within the segmentation process [34]. In spite of the development of a large number of techniques, ideal segmentation is far from being achieved. Here, ideal segmentation is considered a form of segmentation from the perspective of object-based image analysis, wherein each segment represents a distinct object. However, such a definition makes image segmentation less generic and more application-dependent. Two fundamental limiting assumptions of image segmentation techniques that defy its generic applicability are: 1) the use of hand-crafted measures of homogeneity, in the feature space, that set the criteria for the separation of parts of one object from another, and 2) the selection of optimal parameters; both of which are often difficult and subjective [15]. In order to mitigate the first limitation and to fit different applications, one approach is to deduce several spectral, texture, shape, temporal, and context homogeneity measures [10] along with tunable parameters. However, a countless number of homogeneity measures can become necessary to cope with the

* Corresponding author. Tel: +971 24018155, fax: +971 24472442.
E-mail address: ebasaeed@kustar.ac.ae, essa.basaeed@gmail.com (E. Basaeed).

http://dx.doi.org/10.1016/j.knosys.2016.01.028
0950-7051/

development of new applications and/or new imagery. Further, the addition of new metrics leads to extra parameterization that subsequently builds complexity of selecting optimal parameters. Another approach is to use statistical analysis or machine learning techniques in order to detect features [2,19,36]. Therefore, there is a need for segmentation methods that are easy to use, parameter-free, and generic but highly adaptive in order to fit different applications. Such methods shall bridge the gap between high-level semantic information and raw pixels to make new applications possible.

In this paper, an image fusion framework that exploits a committee of boosted-CNNs is proposed for remote sensing image segmentation. The main innovation is in the image fusion framework along with the boosted-CNNs that seamlessly integrates the complementarity of the individual multi-spectral bands for delineating the boundary between two regions without explicitly setting features and parameters. Some contributions of this paper include: the removal of the pooling layer from the CNN architecture in order to improve efficiency and effectiveness in addition to the introduction of morphological intra-band fusion which is achieved using H-minima and watershed transform. Also, at the application-level, it is anticipated that this work is one of the first supervised segmentation techniques that exploit deep learning within remote sensing.

The rest of the paper is organized as follows. Section 2 introduces CNNs along with related work. Section 3 formulates the problem and Section 4 explains in detail the proposed method. Section 5 presents experimental results carried out on a set of multi-spectral remote sensing images. It presents subjective and objective comparisons of the proposed method with existing state-of-the-art segmentation techniques. In addition, it discusses different variations of the proposed method and their respective effect on the performance. Section 5 shall also clearly highlight the shortcomings of the proposed method. Finally, Section 6 concludes with some remarks on future directions.

2. Background and related work

The purpose of this study is to demonstrate the use of CNNs for a supervised remote sensing image segmentation technique. Therefore, a brief overview of CNNs is provided as follows.

CNN is an end-to-end learning technique which means that feature detectors are learned from raw data rather than hard-coded. CNNs showed superior performance over existing state-of-art techniques in many fields such as: image classification and segmentation [21,30], natural language processing [9], and speech recognition [1]. A CNN is a hierarchy of layers of three different types: convolutional, pooling, and fully-connected layers. A convolutional layer is a set of filters (acting as feature detectors). Convolutional layers derive the classification decision by detecting low-level features from raw pixels, higher-level features from lower-level features, and so on. Pooling layers aim to introduce spatial-invariance into the network by combining low-level spatially-localized features from a previous level into a fewer set for the next level [29]. They also aim to force next convolutional layers to derive higher-level features. Finally, fully-connected layers combine the outputs of a previous layer into a one-dimensional vector. The final layer is usually a fully-connected layer with each element in its output vector representing a class [7].

Of particular interest in this paper are the applications of CNNs in image classification and segmentation across different fields. As an object classifier, Krizhevsky et al. in [17] used CNNs in order to classify images from ImageNet Large Scale Visual Recognition Challenge 2012 [26] into different object classes. Furthermore, the use of CNNs is proposed for the purpose of identifying semantic regions of objects in [30] and [21]. In medical imaging, Ciresan et al. in [8] proposed a technique to detect mitosis in breast cancer histology images using a CNN-based pixel classifier. A CNN is used also in order to classify infant brain tissue images into three classes: white matter, gray matter, and cerebrospinal fluid in the work presented in [38]. In remote sensing, the author of [25] proposed the use of CNNs for the classification of roads and buildings in aerial images. Moreover, CNN is used for the purpose of classifying remote sensing hyper-spectral data in [13] whereas it is used in order to detect vehicles in high-resolution satellite images in [5].

The work, presented in this paper, is motivated by [7]. The authors of [7] used CNN in order to segment an electron microscopy image through classifying its pixels into membrane and non-membrane pixels. Their technique starts by dividing each plane in the electron microscopy image into a set of overlapping patches using a moving window. Then, four CNNs are trained independently in order to deduce a decision of whether the center pixel of each patch is a membrane or a non-membrane pixel. The decision confidences of all networks are averaged and then a threshold is applied on the average in order to deduce the final decision. However, membrane pixels can have an arbitrary width and not necessarily one pixel-wide as in boundary pixels. In addition, the authors were not concerned with achieving closed contours as usually required in generic segmentation methods.

3. Problem formulation

An edge-based segmentation method can be formulated as follows. Let I be a single-band image of $X \times Y$ resolution such that $I = \{p(x, y) : 1 \le x \le X, 1 \le y \le Y\}$. $p(x, y)$ denotes the pixel value at the Cartesian position (x, y) and $p \in \mathbb{R}$. A function, φ: $p \to [0, 1]$, is deduced such that $\varphi(p(x, y))$ will result in a higher value if (x, y) is a boundary pixel and a lower value otherwise. A segmentation is achieved by an assignment A: $\varphi \to L$ of each pixel in I to a label in $L \in \{0, 1\}$ using φ where 0 indicates a non-boundary pixel whereas 1 indicates a boundary pixel. A is usually defined as:

$$A(x,y) = \begin{cases} 1 & \text{if } \varphi(p(x,y)) \ge 0.5 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

4. Proposed method

The conceptual workflow of the proposed method is shown in Fig. 1. The process starts by separating bands of the remote sensing image. Each band is assigned a committee of boosted CNNs that will individually derive a confidence map. The value at each pixel location in a confidence map indicates the confidence in the existence of a boundary at that particular location. These individual confidence maps are combined using inter-fusion into a fused confidence map. Then, the fused confidence map is intra-fused in order to thin boundaries, connect them, and produce the final segmentation map.

4.1. Method formulation

Let B be a Z-band image such that $B = \{I_z \mid 1 \le z \le Z\}$ where I_z is a single band in the image. Then, each band, I_z, is divided into a set of overlapping patches ($\mathbf{p}(x, y)$) where $\mathbf{p}(x, y) = \{p(x - \lfloor w/2 \rfloor, y - \lfloor w/2 \rfloor), \ldots, p(x + \lfloor w/2 \rfloor, y + \lfloor w/2 \rfloor)\}$ which denotes the $w \times w$ window surrounding the pixel at the Cartesian coordinates (x, y) in a band. Also, w should be an odd number for the window to have a center pixel. This condition is necessary as each CNN classifies the center pixel as a boundary or non-boundary. Also, let $l(x, y)$ be the ground-truth label of a pixel at (x, y) indicating whether it is a boundary or non-boundary. The size of the band is synthetically increased near its boundaries through