Computer Vision and Image Understanding 147 (2016) 3–22

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

A variance-based Bayesian framework for improving Land-Cover classification through wide-area learning from large geographic regions

Tommy Chang*, Bharath Comandur, Johnny Park, Avinash C. Kak

Department of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, 47907, USA

ARTICLE INFO

Article history:
Received 10 July 2015
Revised 2 April 2016
Accepted 2 April 2016
Available online 6 April 2016

ABSTRACT

Common to much work on land-cover classification in multispectral imagery is the use of single satellite images for training the classifiers for the different land types. Unfortunately, more often than not, decision boundaries derived in this manner do not extrapolate well from one image to another. This happens for several reasons, most having to do with the fact that different satellite images correspond to different view angles on the earth's surface, different sun angles, different seasons, and so on.

In this paper, we get around these limitations of the current state-of-the-art by first proposing a new integrated representation for all of the images, overlapping and non-overlapping, that cover a large geographic ROI (Region of Interest). In addition to helping understand the data variability in the images, this representation also makes it possible to create the ground truth that can be used for ROI-based wide-area learning of the classifiers. We use this integrated representation in a new Bayesian framework for data classification that is characterized by: (1) learning of the decision boundaries from a sampling of all the satellite data available for an entire geographic ROI; (2) probabilistic modeling of within-class and between-class variations, as opposed to the more traditional probabilistic modeling of the "feature vectors" extracted from the measurement data; and (3) using variance-based ML (maximum-likelihood) and MAP (maximum a posteriori) classifiers whose decision boundary calculations incorporate all of the multi-view data for a geographic point if that point is selected for learning and testing.

We show results with the new classification framework for an ROI in Chile whose size is roughly 10,000 square kilometers. This ROI is covered by 189 satellite images with varying degrees of overlap. We compare the classification performance of the proposed ROI-based framework with the results obtained by extrapolating the decision boundaries learned from a single image to the entire ROI. Using a 10-fold cross-validation test, we demonstrate significant increases in the classification accuracy for five of the six land-cover classes. In addition, we show that our variance based Bayesian classifier outperforms a traditional Support Vector Machine (SVM) based approach to classification for four out of six classes.

1. Introduction

Our interest in pixel-level classification of satellite images[1] is driven by the role that such classifications can play in solving problems related to the geolocalization of everyday photographs and videos of outdoor scenes. Using spatial relationships between different land-types - say between arable land, a pond, and a nearby road - to establish correspondences between the "objects" seen in a photograph and those extracted from the satellite images can significantly reduce the candidate locations for where the photograph was taken. For obvious reasons, reliable pixel level classification of satellite data is a necessary prerequisite to the development of such solutions to geolocalization problems.

While there has been much work during the last several decades on the classification of multispectral data in satellite images Anderson (1976); Unsalan (2003), as such this work cannot directly be used for solving geolocalization problems in general.

* Corresponding author. Tel.: +17654913680.
E-mail addresses: chang177@purdue.edu (T. Chang), bcomandu@purdue.edu (B. Comandur), jpark@purdue.edu (J. Park), kak@purdue.edu (A.C. Kak).

[1] In reality, pixel-level classification of a satellite image amounts to carrying out land-type classification of an array of points in that portion of the earth's surface that is viewed in the image. Before classification, the multispectral data in a satellite image goes through a processing step called orthorectification that "maps" the array of pixels in an image to an array of latitude/longitude (lat/long for short) coordinates.

http://dx.doi.org/10.1016/j.cviu.2016.04.001
1077-3142/

Before we explain the reasons for why that is the case, note that the traditional approaches to satellite data classification use standard statistical pattern recognition techniques, with the more recent contributions also using decision trees, random forests, neural networks, support vector machines, and so on DeFries and Chan (2000); Duro et al. (2012); Huang et al. (2002); Marchisio et al. (2010).

One thing that is common to practically all these traditional approaches is that the training and the testing data for constructing a classifier are drawn from the same satellite image. There are very few examples where authors have derived the decision boundaries in one satellite image and shown usable classification results in a different satellite image in a given geographic area.[2] In practically all these cases, the classifiers derived for one satellite image fail to perform adequately on the other satellite images even in the same general geographic region. And, when such results have been shown, it is usually for an adjacent area for which the satellite images were captured at the same time as for the image from which the training data was drawn.

While these traditional approaches may suffice for solving, say, land-cover resource management and forecasting problems, they come up short when solving problems related to the geolocalization of photographs and videos. The reason has to do with the fact that the data in satellite images is affected by the change of seasons, the off-nadir/elevation angle associated with a satellite view, the sun angle, and so on. The logic of the algorithms that one might use for geolocalizing a photograph becomes simpler if the information extracted from the satellite images is invariant to all of the aforementioned changes to the maximum extent possible. The easiest way to achieve such invariance is to use ALL of the satellite data that may be available for a given geographic region. When decision boundaries in a feature space are based on all of the data - meaning data recorded at different times of the year, with different off-nadir/elevation angles, with different sun angles, etc. - any discriminations one is able to make in that feature space are likely to possess the desired properties of invariance. Said another way, *our classifier would be able to make distinctions between the variances associated with the objects that look more or less the same around the year and the variances associated with the objects that change significantly with, say, seasons.*

An additional aspect related to using satellite imagery for solving the problems of geolocalization is that a photograph (or a video) may have been recorded anywhere in a large geographic region of interest whose area may far exceed what is typically associated with a single satellite image. This requires that the land-cover classifications be carried out for the entire ROI using all of the satellite data available for the region.

For reasons stated above, this paper addresses the problem of land-cover classification from a larger geographical perspective than has traditionally been the case in the past. We want to be able to classify all of the data in an ROI (Region of Interest) that can be much larger than the area covered by a typical single satellite image. Consider, for example, the Chile ROI shown in Fig. 1. This ROI, of size 10,000 km^2 is covered by a total of 189 satellite images in the WorldView2 dataset. Our goal is to see if it is possible to create decision boundaries from all of this data taken together so that the overall classification rate for the entire ROI shown in Fig. 1 would be at a usable level of accuracy.

Obviously, before we can design a classifier at the level of an ROI, we must first come to grips with the data variability over the ROI. Understanding data variability at the scale of a large ROI presents its own challenges and can be thought of as a “Big Data” problem. The challenges are created by the typical fast-response and dynamic-storage needs of any human-interactive computer system that must work with very large variable-sized datasets.[3] We have addressed these challenges by developing a special software tool (named PIMSIR for “Purdue Integrated MultiSpectral Image Representation” tool) that is custom designed to achieve the following:

- Rapid visualization of all of the data in an ROI
- Rapid visualization of the geographic area overlaps between the satellite images. Understanding the overlaps is important because any probabilistic modelling of the data at any given geographic *point* is predicated on how much data is available at that point through overlapping satellite views.[4]
- Rapid visualization of the variability of the spectral signatures[5] both spatially and across the views.

This tool is run on a cloud-based cluster of five physical computing nodes, each with up to 48 cores and 256 GB of RAM, that are connected with a 10 Gb network switch. The system is supported by a network storage server with 24 TB of storage.

Even after understanding the extent of data overlap and variability, there remains the big issue of what classification strategy to use for the data. Machine learning now gives us tens of choices for classifiers and it’s not always clear at the outset as to which choice would work the best for a given problem. In this paper, based on our analysis of the data overlap and of the view-to-view variability that we have seen, we chose to design ML (maximum-likelihood) and MAP (maximum a posteriori) classifiers by probabilistic modeling of NOT the spectral signatures themselves, but of the variability in the spectral signatures. We show our ROI based results obtained with this Bayesian classifier, and for comparison, also with the more commonly used SVM classifiers in Section 8 of this paper. As the reader will see, this comparison justifies our intuition regarding the superiority of variance-based Bayesian classification vis-a-vis the more traditional approaches (as exemplified by SVM-based classification).

Moreover, even after a choice is made regarding the classification strategy, there remains the complex problem of how to generate on an ROI basis the positive and negative examples for the different land-cover types for training and testing a classifier. Obviously, it would be very challenging for a human to scan through an entire ROI and manually select such examples. What is needed is a human-in-the-loop random sampling strategy that has the power to yield positive and negative samples that adequately represent ROI based distributions for the different land-cover types. In our classifier training protocol, we use a random sampler based on the Metropolis-Hastings algorithm to select a small number of ROI subregions for presentation to a human and it is for the human to decide whether or not to use that subregion for generating the positive examples for a given land-cover label. (What if a randomly selected subregion is mostly over water while the human is seeking positive examples of high vegetation?) After the human has ac-

[2] We draw the reader’s attention to Wilkinson’s survey Wilkinson (2005) whose major conclusions are just as valid today as they were when the survey was published in 2005. This survey covered 574 experiments in satellite image classification as reported in 138 publications over a period of 15 years. Wilkinson concluded that despite the large body of published research, virtually no progress had been made in satellite image classification over the time period covered by the survey. One of the reasons he highlighted for this lack of progress was the common practice of the researchers drawing their training and testing data sets from the same satellite image.

[3] By very large, we mean datasets that are hundreds of gigabytes in size.

[4] In general, probabilistic modelling is with respect to spatial distribution of the observed data. However, in order to address view-to-view data variability issues, one can also talk about probabilistic modeling with respect to the viewing dimension.

[5] We use the term spectral signature to refer to the 4 or 8 spectral band values at each pixel.