Computers & Geosciences 80 (2015) 62–73

Contents lists available at ScienceDirect

Computers & Geosciences

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

GeoPAT: A toolbox for pattern-based information retrieval from large geospatial databases

Jarosław Jasiewicz [a,c,*], Paweł Netzel [b,c], Tomasz Stepinski [c]

[a] Institute of Geoecology and Geoinformation, Adam Mickiewicz University in Poznan, Dziegielowa 27, 60-680 Poznan, Poland
[b] Department of Climatology and Atmosphere Protection, University of Wroclaw, Kosiby 6/8, 51-621 Wrocław, Poland
[c] Space Informatics Lab, Department of Geography, University of Cincinnati, Cincinnati, OH 45221-0131, USA

ARTICLE INFO

Article history:
Received 7 May 2014
Received in revised form
22 December 2014
Accepted 6 April 2015
Available online 8 April 2015

Keywords:
Pattern analysis
Query-by-example
Large geospatial datasets
Similarity
Image classification
GRASS GIS

ABSTRACT

Geospatial Pattern Analysis Toolbox (GeoPAT) is a collection of GRASS GIS modules for carrying out pattern-based geospatial analysis of images and other spatial datasets. The need for pattern-based analysis arises when images/rasters contain rich spatial information either because of their very high resolution or their very large spatial extent. Elementary units of pattern-based analysis are scenes – patches of surface consisting of a complex arrangement of individual pixels (patterns). GeoPAT modules implement popular GIS algorithms, such as query, overlay, and segmentation, to operate on the grid of scenes. To achieve these capabilities GeoPAT includes a library of scene signatures – compact numerical descriptors of patterns, and a library of distance functions – providing numerical means of assessing dissimilarity between scenes. Ancillary GeoPAT modules use these functions to construct a grid of scenes or to assign signatures to individual scenes having regular or irregular geometries. Thus GeoPAT combines knowledge retrieval from patterns with mapping tasks within a single integrated GIS environment. GeoPAT is designed to identify and analyze complex, highly generalized classes in spatial datasets. Examples include distinguishing between different styles of urban settlements using VHR images, delineating different landscape types in land cover maps, and mapping physiographic units from DEM. The concept of pattern-based spatial analysis is explained and the roles of all modules and functions are described. A case study example pertaining to delineation of landscape types in a subregion of NLCD is given. Performance evaluation is included to highlight GeoPAT's applicability to very large datasets. The GeoPAT toolbox is available for download from http://sil.uc.edu/.

1. Introduction

Most spatial datasets in geosciences originate from remote sensing (RS) and are in the form of images. Therefore, there exists a significant body of literature on retrieving information from RS images (Richards, 1999). Image classification – a process of converting an image into a thematic map of semantically meaningful classes – is the most common form of spatial information retrieval from an image (Lu and Weng, 2007). An original approach to image classification utilizes a pixel-based methodology. A pixel is the smallest element of a surface, as depicted in an image, for which a value of a color is stored. A pixel-based classification algorithm assigns class labels to individual pixels. Note that this is fundamentally different from how an analyst interprets an image by perceiving the coherence of colors on multiple scales simultaneously and assigning class labels to multi-pixel tracts on the basis of their textures or patterns. Pixel-based classification algorithms may suffer from poor performance especially if applied to very high resolution (VHR) images, where individual pixels correspond to small elements of real objects and their numerical attributes are not sufficient to recognize the class of an object, or, if applied to very large images where the goal of analysis is to retrieve generalized classes (for example, when the goal is to retrieve landscape types rather than their constituent land cover classes Graesser et al., 2012; Niesterowicz and Stepinski, 2013; Vatsavai, 2013a; Jasiewicz et al., 2014).

Object-Based Image Analysis (OBIA) was developed (Blaschke, 2010; Lang, 2008) to alleviate the problems associated with pixel-based classification. In OBIA image is first segmented to simplify it by grouping pixels into meaningful segments (called "objects") which are homogeneous with respect to pixel-based attributes. In the second step information is retrieved by classifying objects into semantically meaningful classes. OBIA algorithms get closer to the way an analyst interprets an image but they still suffer from a number of shortcomings (Vatsavai, 2013b). First, segmentation

* Corresponding author.
E-mail addresses: jarekj@amu.edu.pl (J. Jasiewicz), pawel.netzel@uni.wroc.pl (P. Netzel), stepintz@ucmail.uc.edu (T. Stepinski).

http://dx.doi.org/10.1016/j.cageo.2015.04.002

itself is a complex and computationally expensive process and there is no single method that performs consistently well (does not under-segment or over-segment portions of an image) on different RS images. Second, because objects are, by definition, homogeneous segments of the surface, OBIA cannot be used to classify an image into highly generalized classes. For example, although OBIA can classify an image into land cover classes (low-level generalization) more accurately than a pixel-based classifier can, it still cannot classify it into landscape types (high-level generalization). In other words, OBIA can utilize information about image texture but not information about spatial patterns.

For the purpose of this paper we define a spatial pattern as a perceptual structure, placement, or arrangement of image objects having a geometric quality. We then define texture as a structure of pixels arranged quasi-randomly and lacking geometric quality. Thus, a single land cover class in a VHR image (for example, a rooftop) is characterized by texture as it appears on image as a quasi-random mosaic of pixels having a range of colors. However, a fragment of a thematic map showing an urban scene consisting of a spatial arrangement of several land cover classes needs to be characterized by its pattern.

The case for classifying an image or image-like spatial dataset, for example a Digital Elevation Model (DEM), on the basis of spatial patterns arises in multiple disciplines where a high level of generalization is desired. In RS, with VHR images containing rich spatial information, the use of a pattern-based classification method makes it possible to distinguish between different urban landscapes, for example, between informal settlements, industrial/commercial structures, and formal residential settlements (Graesser et al., 2012; Vatsavai, 2013a). In landscape ecology, it makes it possible to distinguish between different landscape types (Niesterowicz and Stepinski, 2013; Cardille and Lambois, 2009) as well as between different types of forest structures (Long et al., 2010), and in geomorphology it makes it possible to identify and delineate physiographic units (Jasiewicz et al., 2014).

It is only recently that methodologies for pattern-based information retrieval from images and other raster datasets have been proposed. Vatsavai (2013a) proposed a multi-instance learning (MIL) scheme as a means for the pattern-based classification of images. In this method, an image is divided into regular grid of local blocks of pixels. The data (a set of all multi-dimensional attribute vectors from each pixel) in each block is modeled using a multivariate Gaussian distribution. The distance (dissimilarity) between any two blocks, and thus between the two patterns contained in these blocks, is calculated as the probabilistic distance between their modeled Gaussian distributions using the Kullback–Leibler (KL) divergence. Using supervised learning based on the MIL scheme Vatsavai (2013a) and Graesser et al. (2012) classified RS images of several cities into formal and informal neighborhoods.

Independently, we have proposed a general approach for pattern-based information retrieval from all types of geospatial datasets (Jasiewicz and Stepinski, 2013a; Stepinski et al., 2014). For our method to be broadly applicable and computationally efficient it uses an input (image, DEM etc.) that has been preprocessed using a pixel-based classification and thus already converted into a categorical format. This categorical raster is divided into a regular grid of local blocks of pixels. Because the data is categorical, each block can be compactly represented by a histogram of categories or other attributes derived from these categories. We have successfully applied this methodology to search for and classify land-cover patterns in the National Land Cover Dataset (NLCD) (Jasiewicz and Stepinski, 2013a). We have also used it for an assessment of land cover change over the entire United States using the NLCD (Netzel and Stepinski, 2015), and for the identification and delineation of physiographic units using DEM data (Jasiewicz et al., 2014).

The concept of pattern-based information retrieval from geospatial datasets is at the beginning of its developmental cycle. For this concept to mature much more work is needed, including application to many different datasets in multiple contexts. In this paper we present the Geospatial Pattern Analysis Toolbox (GeoPAT) – a collection of GRASS GIS modules that integrate the various tools necessary for experimenting with pattern-based information retrieval from geospatial data. GeoPAT is intended as a convenient platform for experimentation with the pattern-based analysis of rasters including rasters having giga-cell and larger sizes. It integrates into the GIS system procedures for pattern description, pattern similarity, and the search and retrieval of similar patterns. These concepts were originally developed for working with natural images in the context of Content-Based Image Retrieval (CBIR) systems (Datta et al., 2008) but are now utilized by GeoPAT for the purpose of geospatial analytics. Such integration allows a user to perform the standard GIS tasks of mapping, map overlay, and segmentation on a grid of pattern-bearing blocks of pixels in a way which is already familiar (from performing similar tasks on standard images). In other words, GeoPAT extends the standard GIS system by adding a new type of attribute – the pattern signature – and a new type of data query – a query-by-pattern-similarity (QBPS). This significantly lowers the cost of entry into experimenting with pattern-based information retrieval, helps to accelerate further development of this concept, and makes possible the assessment of its utility in various domains.

GeoPAT modules are written in ANSI C and are designed to work within the GRASS GIS 7 (GRASS Development Team, 2012) environment. Embedding GeoPAT in GRASS has a number of advantages: (1) GRASS is an open source software available for major computing platforms, (2) GRASS is especially well-suited to work with large datasets, and (3) incorporating a toolbox into an already existing, well-established environment allows for an integrated computational pipeline that provides convenience and boosts efficiency (Körting et al., 2013). GeoPAT is an actively developed solution. The core of the toolbox consists of the seven modules that compute pattern signatures and perform the GIS tasks of comparing, searching, overlaying, and segmenting the rasters on the basis of similarity between local patterns. These modules provide the basic infrastructure for pattern-based information retrieval and are not expected to be modified by a user. In addition, two libraries provide a selection of functions for extracting pattern signatures and for calculation of similarity/distance between two patterns, respectively. As there are no standard means of representing spatial patterns and calculating a measure of similarity between them, we expect users to add to those libraries as they experiment with different datasets.

The rest of this paper is organized as follows: Section 2 presents an overview of our toolbox architecture. Section 3 describes the most important functions in the shared libraries and Section 4 describes the seven core geoprocessing modules. A case study (Section 5) presents an example on how GeoPAT modules can be utilized to perform regionalization of land cover patterns into landscape types using either unsupervised or supervised approaches. Section 6 gives an assessment of the computational performance of the GeoPAT modules and Section 7 contains our discussion and conclusions.

2. Software architecture

As an introduction to GeoPAT we first give an illustration of the basic idea behind the pattern-based analysis of geospatial data. For this we use a DEM with 30 m resolution. The left panel in Fig. 1 shows a hillshade rendition of a 2000 $\times$2000 cell DEM (we