Multi-resolution, pattern-based segmentation of very large raster datasets

J. Jasiewicz[1,2], J. Niesterowicz[1], T. F. Stepinski[1]

[1]Space Infromatics Lab, University of Cincinnati, 401 Braunstein Hall, Cincinnati, OH 45221-0131, US
Email: {niestejk; stepintz}@mail.uc.edu

[2]Instititute of Geoecology and Geoinformation, Adam Mickiewicz University, Dziegielowa 27, Poznan, Poland
Email: jarekj@amu.edu.pl

Abstract

We present an algorithm which efficiently segments very large categorical rasters based on patterns of their categories. It operates on a grid of motifels – square blocks of raster cells representing a local pattern. Our algorithm is based on the seeded region growing principle but it uses a novel grid topology and seeds stack with individual thresholds. It has a single free parameter – the spatial scale of a pattern. Algorithm was proven to be robust on land cover data, topographic landforms data, and high resolution color-quantized RGB images. We present a multi-scaled segmentation of NLCD2011 as an example. Potential applications of the new algorithm include ecology, geomorphology, pedology, forestry and agriculture, and urban studies.

1. Introduction

Segmentation is the process of partitioning a raster dataset into multiple homogeneous segments. The goal of segmentation is to spatially generalize a raster so it provides more insight and is easier to analyze. The bulk of the existing work (Zhang *et al*. 2008) has focused on segmentation of images of relatively small scenes. However, segmentation of datasets that originated from remote sensing and cover large, continental or even global-scale areas, are also important, but existing segmentation algorithms are ineffective for such large datasets and the custom algorithms are lacking. Examples of such datasets are the National Land Cover Dataset (NLCD), which covers the conterminous US (CONUS) with the resolution of 30 m, or the SRTM-based DEM, which has a world-wide extent at 90 m resolution. Segmentation of NLCD could yield landscape types – precursors to ecoregions, and segmentation of world-wide DEM could delineate physiographic regions.

In this paper we describe a segmentation algorithm especially designed for very large rasters. Specificities of such datastes are as follows. (1) They are the mosaics of multiple datasets, thus it is better to segment a secondary product of uniform quality (for example, a land cover) rather than a montage of primary data of variable quality (for example, a montage of Landsat scenes). (2) They are large; for example, the NLCD consists of ~8 billion cells and has the size of ~16 GB. (3) The goal of the segmentation is to identify regions characterized by patterns which are homogeneous on the scale that is large in comparison to the resolution of the raster, since the need for pattern-based segmentation.

To deal with a large size of the input the proposed algorithm is based on the concept of Complex Object-Based Image Analysis (COBIA) (Vatsavai 2013, Stepinski *et al*. 2015). In COBIA the raster is divided arbitrarily into a regular grid of local blocks of cells at minimal

computational cost. These blocks of cells are the basic units of analysis; we will refer to them as *motifels* – the smallest processing elements containing local motifs (patterns) of raster variable. The size of the motifel sets the scale of patterns to be segmented. Using COBIA reduces the size of the problem by orders of magnitudes as elementary grid elements change from cells to motifels. To simplify a description of motifels we only consider categorical rasters. This is less restrictive than it may appear as one major application, land cover, is already a categorical raster, and the second, DEM, can be easily converted into categorical raster using geomorphons algorithm (Jasiewicz and Stepinski, 2013). With categorical input we represent patterns within motifels by a category co-occurrence histogram and we use the Jensen-Shannon divergence (JSD) to measure a degree of dissimilarity between motifels.

The proposed segmentation algorithm works within the GeoPAT toolbox (Jasiewicz *et al*. 2015) – a collection of GRASS-GIS modules for pattern-based geoprocessing. It is based on the seeded region growing concept, but, in addition of being applied to motifels instead of pixels, it has two original innovations: (a) a novel *brick*-topology grid, and (b) a novel method for ordering seeds into a stack in order of increasing local inhomogeneity.

2. Pattern-based segmentation algorithm

The algorithm consists of five steps: 1) building brick-grid; 2) analyzing local homogeneity; 3) ordering seeds; 4) small areas removal (optional); 5) homogeneity enhancement (optional).

The result of segmentation depends on the topology of the grid. Four possible topologies are shown in Fig.1. The square, 4-connected grid is a standard in segmentation because the square, 8-connected grid may lead to segments that permeate each other. However, directions of the growth of segments are overly limited by the 4-connectivity. The hexagon grid would be ideal but keeping track of hexagonal motifels of square cells is computationally too expensive. Our solution is to use 6-connected square grid with brick topology. In the first step, co-occurrence histograms are calculated and stored for each motifel in the brick grid.

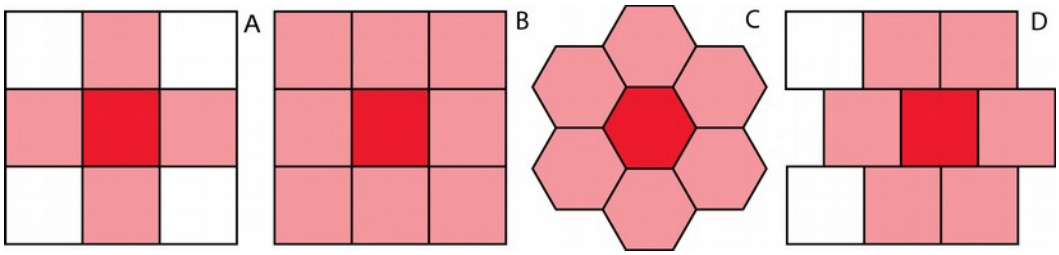

Figure 1: Different grid topologies: (A) square, 4-connected grid; (B) square, 8-connected grid; (C) hexagon, 6-connected grid, (D) brick-like square, 6-connected grid.

The results of the segmentation depend on the selection of the seeds and the order they are grown into segments. Our idea is that priority should be given to growing segments from motifels located in the most homogeneous local neighborhoods. For a focus motifel we calculate a JSDs between it and motifels in its 2-motifel radius neighborhood (18 motifels in the brick topology). Using Fisher's Linear Discriminant we identify a subset of the neighborhood consisting of motifels most similar to the focus motifel. The proxy for local inhomogeneity is an average JSDs between motifels in this subset and the focus motifel. All motifels in the entire grid are ordered into a stack of increasing values of their local inhomogeneities.

Segments are grown from the top of the stack. Starting from a single motifel, all motifels forming a perimeter of a growing segment are checked for accruement. A motifel with

the smallest value of an average linkage (AV - mean dissimilarity to all motifels in the segment) is added to the segment and the process is repeated until the value of the minimum AV is larger than the linkage threshold for this seed. The threshold is equal to the seed's inhomogeneity value plus the standard deviation of JSDs used to calculate the seed's inhomogeneity value. All motifels in the newly formed segment are removed from the stack and the next segment is grown from the remaining top ranked seed. After segments growing ends the segmentation may be optionally enhanced by removal of small segments and possible swapping of motifels at the segments boundaries to increase an overall homogeneity of segments.

3. Results

The algorithm was tested on the entire NLCD, the 30m CONUS DEM classified to topographic landforms, and a high resolution color-quantized RGB image (Niesterowicz *et al*., this conference). Here, we present the results of segmenting the NLCD2011. The algorithm has only a single parameter – the size of the motifel. We segmented the NLCD at the scales from 128 cells (~4km) to 4096 cells (~123km) each time increasing the scale by the factor of two. Fig.2 shows segmentations at four selected levels. Most segments have values of inhomogeneity below 0.1 (the range is between 0 and 1). Limited number of small segments has inhomogeneity values of ~0.3. The segmentation on the scale of 2048 cell resembles most closely the level-IV ecoregions division, but there are significant differences between the two partitioning because much more information than just the land cover pattern is used to delineate ecoregions. The computational time depends on the size of motifel. Segmentation on the scale of 128 cells took 187s on 8-core I7 computer, while segmentation on the scale of 1024 cells took less than 2s.

4. Conclusions and future work

The algorithm is fast, stable, and convenient to use (the only important parameter is the scale of pattern). Note that full version NLCD is used at each scale so less information is lost in comparision with other method of resolution degradation. Using different datasets other than those we have tested may require changing the motifel representation and dissimilarity function. It will become soon available as a part of the new version of GeoPAT. Applications include ecology, geomorphology, pedology, forestry and agriculture, and urban studies. Next step is to extend the algorithm to perform segmentation using multiple inputs thus becoming directly relevant to delineation of ecoregions.

Acknowlegements

This work was supported by the University of Cincinnati Space Exploration Institute, by Grant NNX15AJ47G from NASA, and by the National Science Center (NCN) grant DEC-2012/07/B/ST6/01206.

References

Jasiewicz, J., Netzel, P. & Stepinski, T., 2015. GeoPAT: A toolbox for pattern-based information retrieval from large geospatial databases. *Computers & Geosciences*, 80, pp.62–73.

Jasiewicz, J. and Stepinski, T.F., 2013. Geomorphons—a pattern recognition approach to classification and mapping of landforms. *Geomorphology*, 182, pp.147-156.

Stepinski, T.F., Niesterowicz, J. & Jasiewicz, J., 2015. Pattern-based Regionalization of Large Geospatial Datasets Using Complex Object-based Image Analysis. *Procedia Computer Science*, 51, pp.2168–2177.

Vatsavai, R.R.R., 2013. Object based image classification: state of the art and computational challenges. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data*. pp. 73–80.

Zhang, H., Fritts, J.E. & Goldman, S.A., 2008. Image segmentation evaluation: A survey of unsupervised methods. *Computer Vision and Image Understanding*, 110(2), pp.260–280.

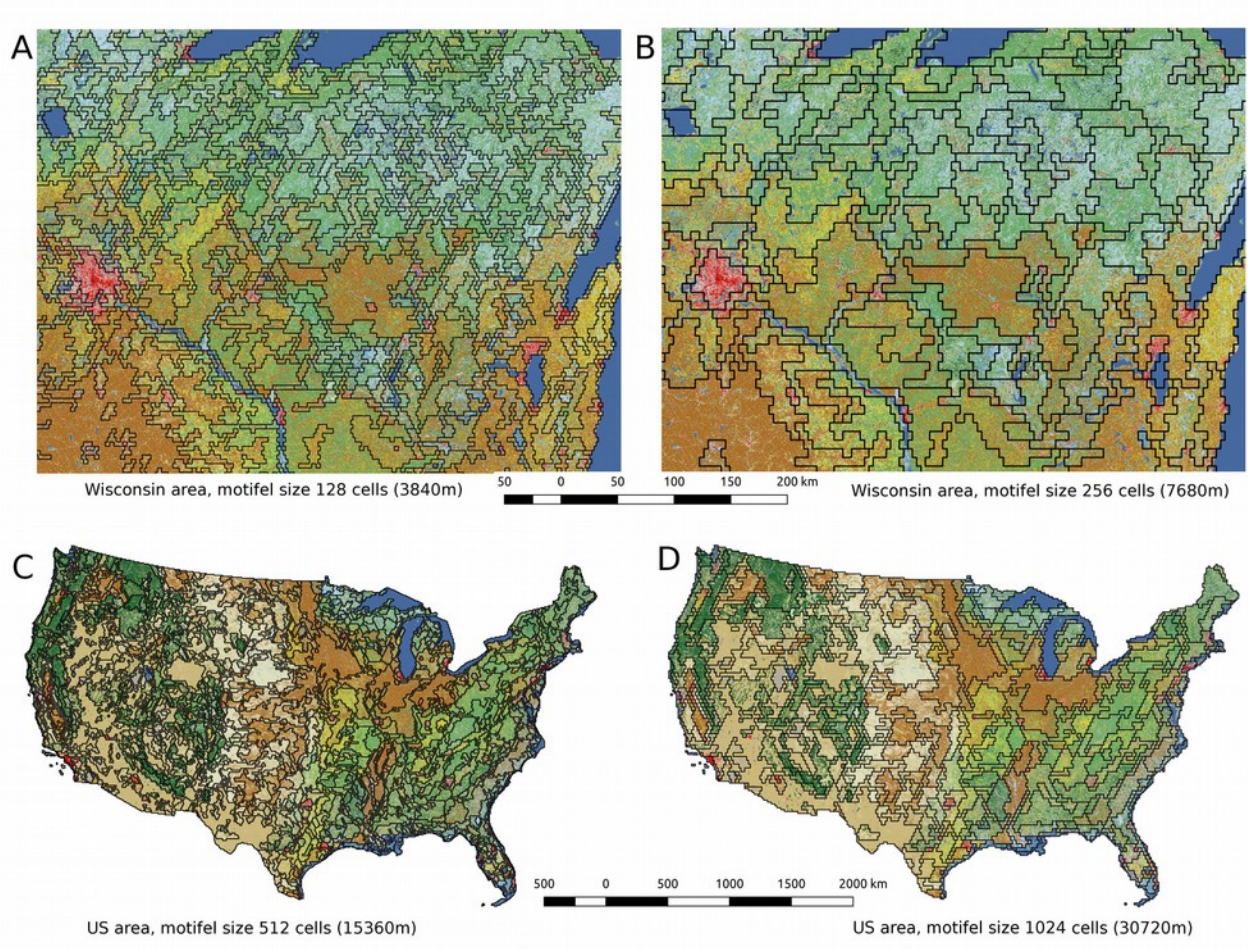

Figure 2: Selected results of segmentation of the NLCD2011. NLCD classes are shown using the standard colors