QUERY AND RETRIEVAL OF LAND COVER PATTERNS

Jaroslaw Jasiewicz∗
Adam Mickiewicz University
Geoecology and Geoinformation Institute
Dziegielowa 27, 60-680 Poznan, Poland

Tomasz F. Stepinski†
University of Cincinnati
Department of Geography
Cincinnati, OH 45221-0131, USA

ABSTRACT

Ability to explore and mine land use/land cover (LULC) maps having high resolution and large spatial extent could significantly enhance the value of such datasets. In this paper we present a method for retrieval of alike scenes from large LULC datasets. Alikeness between the two scenes is defined as similarity between their spatial patterns of class labels. Our method works on the principle of query-by-example; the inputs are the LULC map and a reference scene and the output is a similarity map indicating a degree of alikeness between a given location on the map and the reference scene. The similarity measure is described and its applicability to the National Land Cover Dataset 2006 (NLCD2006) is discussed.

Index Terms— Land cover dataset, query-by-example, similarity

1. INTRODUCTION

We have developed a search-engine-like tool for surveying and analyzing land use/land cover (LULC) datasets of continental extent and high spatial resolution. Examples of such datasets include: the National Land Cover Dataset 2006 (NLCD2006) [1] which covers the conterminous United States at the spatial resolution of 30 m/pixel and CORINE [2] which covers Europe at the spatial resolution of 100 m/pixel. At present, such datasets are used predominantly for extraction of “postage-stamp-size” local subsets restricted to a priori defined study areas. Such usage does not take advantage of the full potential of these datasets, which, given a proper computational tool, could be queried for arbitrary patterns of LULC classes. To give an analogy, a present utility status of large LULC dataset is like having an Internet without an advantage of a browser - there is no possibility of utilizing information without prior knowledge of its existence. In this paper we introduce a pattern-based “browser” for exploration of LULC datasets.

The task of query and retrieval of LULC patterns is analogous to the task of content-based image retrieval (CBIR) - an area of intense research presenting multitude of challenges. We are able to develop a successful query and retrieval tool for LULC datasets because they are simpler than natural images and because even a broad search of LULC is useful. A user is rarely interested in searching for specific “objects,” but rather seeks regions having overall similar composition and spatial pattern of classes.

2. METHODS

Our tool is designed to query LULC dataset $\mathcal{G}$ consisting of pixels $g_i = \{x_i, y_i; c_i\} \ i = 1, \ldots, N$, where $x_i$ and $y_i$ are $i$th pixel spatial coordinates and $c_i$ is the LULC class label. Typically, $N$ is very large, of the order of $10^9\cdot10^{10}$ for continental or global scale datasets. The pixel’s categorical label is from the set $S = c_1, \ldots, c_K$, where $K$ is typically low, of the order of 10. The tool works on the principle of query-by-example. Let $\mathcal{A} \subset \mathcal{G}$ be a small subset of the overall LULC dataset. For example, $\mathcal{A}$ may represent a particular pattern of LULC classes in a study area - a locality selected for one reason or another as interested or representative. Although the tool puts no restrictions on the shape of $\mathcal{A}$, rectangular shape is most practical. The purpose of our tool is to identify all $\{\mathcal{A}_{i} \subset \mathcal{G}, i = 1, \ldots, M\}$ which are “similar” to $\mathcal{A}$. Here $M$ is not an a priori given number, but rather is determined by the results of the query. Similarity between the two localities means that both of them have alike composition and spatial patterns of classes. Thus, in the case of NLCD2006 dataset, the tool will identify all localities across the United States having composition and patterns of land cover classes similar to the one present in the locality submitted to the query.

2.1. LULC pattern signature

In a typical LULC map classes are indicated by different colors; for brevity, we refer to the classes themselves as “colors.” Trying to identify similarity or dissimilarity of patterns of colors present in the two rasters directly from pixels values is not feasible. Instead, a more compact mathematical description - LULC pattern signature - is calculated from original pixels values and used to assess the degree of patterns similarity.

∗The author performed this work while at the University of Cincinnati.
†This work was supported by the National Science Foundation under grant IIS-1103684.
For the pattern signature to serve its purpose well it needs to encapsulate the very character of the pattern that analyst would focus on when making visual comparisons. Our pattern signature is based on only two features: colors and sizes of patches. A patch is a connected region of single color. We call it a patch rather than a segment to make a distinction between it and a region of image resulting from image segmentation; a segment is coherent but inhomogeneous with respect to its pixel values whereas a patch is homogeneous. Locality is segmented into patches using connected components algorithm [3]. For each patch we calculate its size. Each pixel in the locality is characterized by two features: its color and the size (area) of the patch to which it belongs. A pattern signature is a probability distribution of two dimensional variable (color, patch size). Note that a color is a categorical variable whereas patch size is a continuous variable and needs to be discretized for calculation of landscape signature. We discretize patch sizes into bins with ranges based on power of two (i.e. 1-2, 2-4, 4-8 etc); the number of bin depends on the maximum size of locality.

Fig. 1 shows an example of three localities (from the NLCD2006 dataset) and their pattern signatures. We have chosen the localities with very similar composition of colors (dominated by green - standing for deciduous forest, and yellow - standing for pasture). Thus, a similarity measure based on composition of colors alone would indicate that all three localities are equally similar to each other. However, most analysts would assess that localities B and C are more similar to each other than either of them is to the locality A. This is because an analyst will take patterns of colors into consideration. Using our pattern signature and similarity measure as defined in the next subsection, we indeed obtain a similarity relationships between A, B, and C as suggested by visual appraisal.

2.2. LULC pattern similarity

Our LULC pattern similarity measure is based on the notion of mutual information. Consider any two localities A and B represented by their respective pattern signatures. First, we define a random variable Y that assigns probabilities of getting a specific color-membership and patch size-membership while randomly drawing a pixel from a combination of A and B. The probability distribution function (PDF) of variable Y is just a 2D histogram constructed from all pixels in both localities. Second, we define another random variable, X, that assigns equal probabilities to selecting one of the two possible localities A and B. The PDF of variable X is $P_X(x_j) = 1/2$ for both $x_1 = A$ and $x_2 = B$. This random variable formalizes a simple act of randomly choosing one of the two localities. Finally, we define a joined PDF, $P_{YX}(y_i, x_j)$ which assigns a probability to choosing a locality $j$ and drawing a duplet (color, patch size) indexed $i = \{i_c, i_{sz}\}$ from locality $j$.

We use informational entropy [4] to characterize a probability distribution $P(Z)$ of any given random variable Z. Jointed entropy $H(Y, X)$ is the entropy of joined PDF, $P_{YX}(y_i, x_j)$ and specific conditional entropy $H(Y | X = x_i)$ is the entropy calculated only from the histogram of locality $x_i$. Finally, conditional entropy $H(Y | X)$ is an average of
Mutual information, $I(Y, X)$, given by $I(Y, X) = H(Y) - H(Y|X)$, measures an average reduction of unpredictability of $Y$ if the specific locality is set. The value of $I(Y, X) = 0$ if both localities have identical histograms, and the value of $I(Y, X) -> 1$ if the rasters are dominated by distinct single colors. Thus, $r(A, B) = I(X, Y)$ is a convenient measure of “distance” between two localities $A$ and $B$, if by the distance we understand the increasing difference in patterns of LULC classes. Conversely, $sim(A, B) = 1 - I(X, Y)$ is a convenient measure of similarity between the two localities from the LULC patterns point of view.

The novelty of our similarity measure is that it does not directly compare histograms of the two landscapes (it does not rely on mutual information between PDFs characterizing respective landscapes). A direct comparison of indexed rasters using mutual information has been used previously in the context of identifying change between two images or maps; such measure is sensitive to specific spatial correspondence of pixels. In our context we are not interested in a degree of departure from one raster to another, but rather in a “generic” similarity between their patterns. Thus, we want the rasters to be similar even if they are rotated, translated, or deformed with respect to each other, as long as they exhibit an overall pattern similarity.

### 2.3. Search

We utilize an overlapping moving window approach to calculate a similarity between a query LULC pattern and all possible localities in the dataset. Present implementation uses window’s size of 500x500 pixels and shift of 100 pixels. All calculated similarities (even small ones) are stored for visually effective presentation of the results of the query in the form of similarity map. Similarity map shows the entire spatial extent of the dataset with a color gradient showing degree of similarity between a LULC pattern that is queried and the local pattern. Top search results (localities having LULC patterns most similar to the query) are retrieved automatically or on demand by using the similarity map.

### 3. RESULTS

Fig 2 demonstrates the workings of our query and retrieval tool using the NLCD2006 dataset (panel A) and a specific example of a query (panel B). The query is a tiny piece of NLCD2006 showing a pattern of land cover classes characteristic of a particular irrigation technique used in desert agriculture. The tool searches all other places in the U.S. where such (or highly similar) pattern is found. A query results in a generation of a U.S. map (panel C) showing degree of similarity to the queried LULC pattern. Top search results (areas having most similar patterns) are retrieved (panel D) automatically or on demand from the similarity map. Using our query a domain expert is able to put his study area in a
broader geospatial context by gaining information about existence and spatial locations of similar areas within the bounds of NLCD2006 dataset.

4. CONCLUSIONS

This work provides a search tool for exploration of large LULC datasets. The ability to search large volumes of data for similar patterns is taken for granted in many fields, but has not been available for geospatial datasets like the LULC. Standard GIS tools enable searches for a single LULC class, but not for a pattern of classes. As presently implemented, the pattern search tool takes about 1.5 hours (using a computer with a single 4 cores 3GHz Intel processor and 16GB of memory) on the NLCD2006 dataset to complete the search and produce results as shown in Fig.2. Most of this time is spent of data reading; by storing only pattern signatures instead of the dataset the search time can be reduced by two orders of magnitude. Such modification would make practical to implement the tool as a web-based application.

5. REFERENCES


