Automated Metadata for Image Mining[1]

Faleh Alshameri
School of Information Technology
George Mason University

Edward J. Wegman
Center for Computational Statistics
George Mason University

1 Introduction

Datasets on the scale of terabytes preclude any serious effort by individual humans at manually examining and characterizing the data objects. This is particularly true with image data because the normal mode of analysis is by visual observation. A reasonable approach in this framework is to augment the metadata associated with an image by attaching digital objects reflecting the content of the image. The idea is to augment traditional exogenous metadata with endogenous metadata by an automated process. The content could be characterized by geometric structure, e.g. linear or circular features, or by texture features. The basic idea is to launch a program working in the background to collect feature vectors associated with a particular image. The feature vectors are transformed in to digital objects, which are then attached to the metadata associated with image. The investigator can then search on the digital objects with a Boolean search extracting only the images that are most consistent with the desired properties.

Our test bed for image data consisted of 50 gigabytes of image data from NASA's TERRA satellite, the first of the polar orbiting Earth Observing System satellites. The image data, provided to us by LaRC with assistance from staff at JPL, came from the Multiangle Imaging SpectroRadiometer (MISR). The MISR instrument records images at 9 look angles in 4 spectral bands (R,G,B, and NIR) of dimension 128 by 512 [Red is recorded at 512 by 2048]. The data products issued by NASA are categorized by the amount of processing. We used MISR Level 1B2 data, which are NASA's Georectified Radiance Product.

This project suggests a methodology that would allow an Earth scientist to quickly identify candidate images containing one or more features of interest from a massive database that no human could hope to investigate without data mining assistance. In what follows we present some prototypical feature vectors. The methodology we suggest is not limited to these features which are offered only for demonstration porposes. An interactive website can be found at http://scs.gmu.edu/~falshame/home.html.

2 Statistical Feature Methods

As a first attempt to demonstrate this technology we chose features based on the Gray Level Co-occurrence Matrix (GLCM). The GLCM is a common technique in statistical image analysis that is used to estimate image properties related to second-order statistics. GLCM considers the relation between two neighboring pixels in one offset, as the second order texture (Lee et al., 2004). The first pixel is called reference and the second one the neighbor pixel, which we chose to be the one to the east (right) of each reference pixel. GLCM measures the occurrence of one gray tone in a specified spatial linear relationship with another gray tone within the same area. It can reveal certain properties about the spatial distribution of the gray level in the texture image. GLCM is a two dimensional

[1]**Keywords:** GLCM, NDVI, entropy, energy, homogeneity

matrix of joint probabilities $p_{i,j}$, which measures the probability that gray level j follows the gray level i at pixel.

2.1 GLCM Framework

There are several steps to build symmetrical normalized GLCM. These steps as follows:

1- create framework matrix: on this step the matrix will be filled starting from the top left cell to bottom right cell. Pixels along the right edge have no right-hand neighbor (no wrap).

2- Expressing the GLCM as a probability: In this step the GLCM is transformed into a close approximation of a probability table. The probability is measured by applying the normalization equation: $p_{i,j} = \frac{v_{i,j}}{\sum_{i,j=0}^{N-1} v_{i,j}}$, where i and j are the row and column numbers respectively and $v_{i,j}$ is the value in the cell i, j of the image, $p_{i,j}$ is the probability for the cell i, j, and N is the number of rows or columns.

3. Features Implementation

(Haralick et al., 1979) has proposed 14 features that can computed from the GLCM. Some of these features are related to first-order statistical concepts, such as contrast and variance and have clear textural meaning like pixel pair repetition rate and spatial frequencies detection. Other features contain textural information and at the same time are associated with more than one specific textural meaning (Baraldi et al., 1995). On this research we developed a set of features part of it are based on GLCM. Adjacent pairs of pixels (assuming 256 gray levels) are used to create 256 by 256 matrix with all possible pairs of gray levels reflected. Images with similar GLCM are expected to be similar images. In our preliminary research some of the features that were constructed are based on the GLCM, including homogeneity, contrast, dissimilarity, entropy, angular second moment (ASM), and energy. Other features we computed include histogram-based contrast, alternate vegetation index (AVI) (greenness/NIR ratio), and normalized difference vegetation index (NDVI).

4 Comparisons of Feature Vectors

The feature vectors measure different aspects of the images. Figure 4.1 is a parallel coordinate plot of the eight feature vectors for a subsample of the images. The colors are determined by three features within the AVI which appears to discriminate among vegetation (green), water (red), and desert (blue). The high degree of mixing suggests that the eight features are substantially measuring different aspects of the image.

4.1 Comparison between NDVI and AVI To compare the vegetation value between the two vegetation indices, the Alternate Vegetation Index (AVI), and the Normalized Difference Vegetation Index (NDVI), we selected some images (from red band and NIR band for the NDVI, and from green and NIR bands for AVI) and we computed the vegetation values of both indices. The Figure 4.2 show some selected images and the computed histograms for the NDVI and AVI.

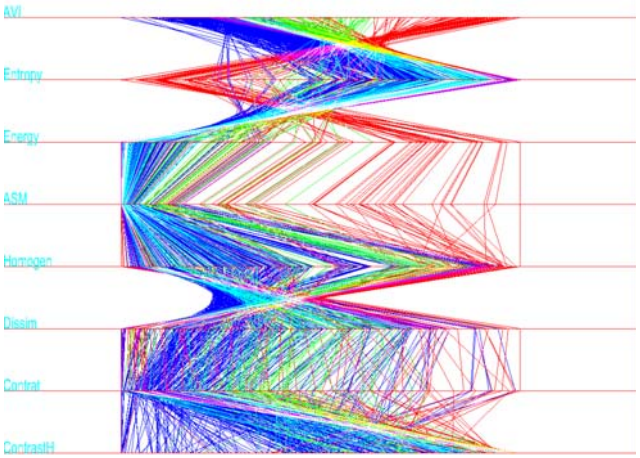

Figure 4.1 Parallel coordinate plot of the 8 features.

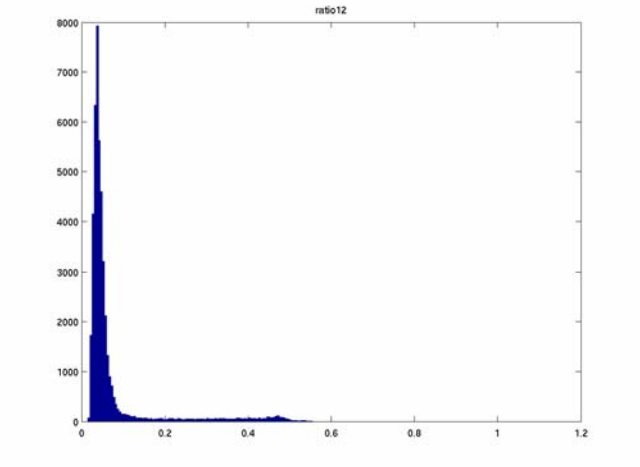

Figure 4.2a AVI= 0.072

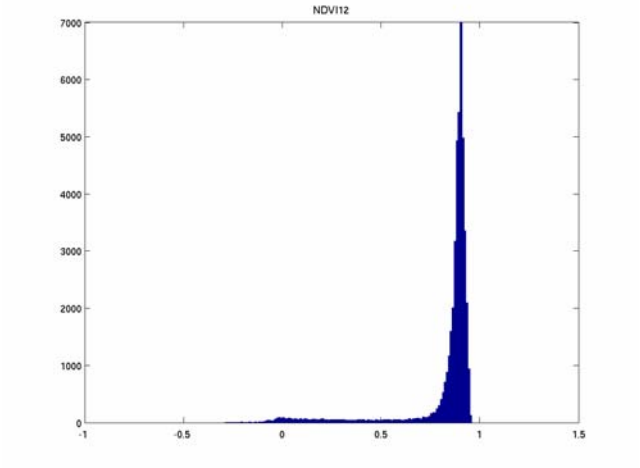

Figure 4.2b NDVI

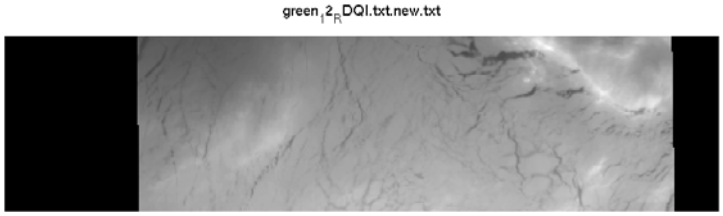

Figure 4.2c Shows the image in the green band for the histogram in Figure 4.2a.

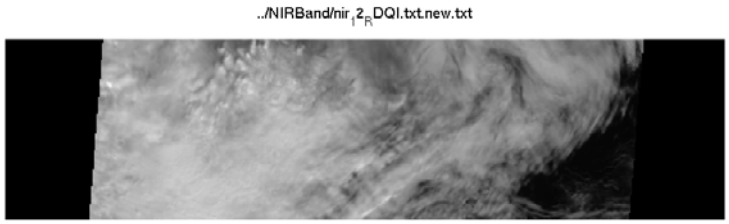

Figure 4.2d Shows the image in the near infrared band for the histograms in Figure 4.2a,b

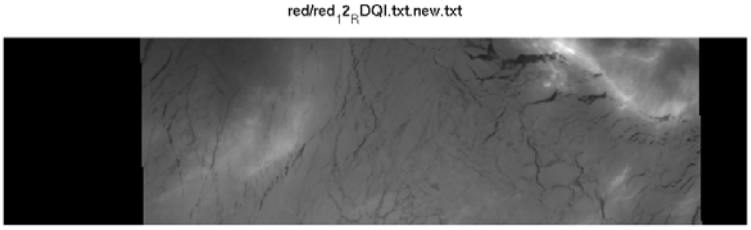

Figure 4.2e Shows the image in the red band for the histograms in Figure 4.2b

References

Baraldi, A., and Parmiggiani, F. (1995), "An Investigation of the Textural Characteristics Associated with Gray Level Co-occurrence Matrix Statistical Parameters," *IEEE Transaction on Geoscience and Remote Sensing*, 33(2), 293-304.

Haralick, R. M., (1979) "Statistical and structural approaches to texture," *Proceedings of the IEEE*, 67(5), 786-804.

Lee, K., Jeon, S., and Kwon, B. (2004) "Urban Feature Characterization using High-Resolution Satellite Imagery: Texture Analysis Approach," Map Asia Conference, Bejing, China.