STATISTICAL MINING ON IMAGE DATABASE USING SUPERVISED CLASSIFICATION

A.Daniel¹
Assistant Professor,
School of Computing Sciences and Engineering,
Galgotias University,
Delhi NCR, Uttar Pradesh, India
danielarockiam@gmail.com

U.Hari Haran²
Information Technology
Galgotia College of Engineering and Technology
Greater Noida, India
udhayakumarhariharan@gmail.com

Abstract— we present a protocol for the evaluation of a content-based remote sensing image information mining system. The knowledge driven information mining system (KIM) which easily enables users the access to remote sensing image archives via internet communication, the analysis of information details by an intelligent man-machine interface (MMI) The query performance of content based image retrieval system mainly depends on the datasets stored in the Clustered Database. We analyze the complexity of image data. In order to provide users fast access to the content of Clustered Database, the system is composed of two main modules. The first includes computationally intensive algorithms for off-line data ingestion in the database image. The next module consists of a graphical man–machine interface that manages the information fusion for interactive interpretation and the image information mining functions. Bayesian classification determines the accuracy of interactive training. In this interactive system user effort, characteristics of the internet design, guidance provided by the system, and duration of a user session are critical aspects, which are observed and measured.

Keywords— Image information mining, Knowledge Information Mining (KIM), man-machine interface (MMI)

I. INTRODUCTION

Remote sensing images have a large scope of applications ranging from weather and climate studies via urban and suburban land usage analysis, finding of deforestation, and study of marine environment up to engineering and geological applications. Image clustering and categorization is a means for high-level description of image content. The aim of the system is to find a mapping of the remote sensed image into classes (clusters) such that the set of classes provide essentially the same information about the image archive as the entire collection of image from database. The final class provide a summarization and visualization of the image content that can be used for different tasks related to enhancement of image from clustered database for the management of images. Image clustering enables the implementation of efficient retrieval algorithms and the creation of a user-friendly interface to the database. Statistical mining is a problem that is getting more and more attention in retrieval of sensed images. Many of the research works had been undertaken to design efficient image retrieval techniques from the image or multimedia databases. Mining is applied to search for similar scenes in remotely sensed images, although image database would require a pre-processing in order to convert it to a categorical map by using a supervised classification based on support vector machine.

State of the art of data mining and advanced database technology helped in retrieval of useful hidden information from large image database [1,22,23,24,25,26,27,28]. The classifier ranks the unlabeled pixels and automatically chooses those that are considered the most valuable for its improvement [3]. The QBIC tools for exploration and mining of images would significantly enhance their value of database [6], [1]. In [1], [2,7,8,9,10], [5], the remote sensing image information corresponding to spectral characteristics is identified by supervised classification based on support vector machines. The support vector machine (SVM) was considered as the core learner and it contains two different regularization schemes: 1) the inclusion of relational operators between tasks. 2) The pairwise Euclidean distance of the predictors. These methods fall on modifications of the kernel used in the
standard SVM. Gabor wavelet coefficients are used to extract the spatial information in SVM [1]. In [4,11,12,13,14,15], it is illustrated that the preprocessing aid in accurate geometric connection and it is mandatory for land and agriculture applications.

In the proposed research which aims at establishing a knowledge driven information mining system (KIM), which easily enables the users to access the remote sensing image archives via internet communication and analyze the information details by an intelligent man-machine interface (MMI). Query-by-image-content (QBIC) tools are in demand in geospatial community because they enable exploration and mining of the rapidly increasing database of remotely sensed images. The query performance of content based image retrieval system mainly depends on the datasets stored in the Clustered Database. This method works on the principle of query on image content, input as a reference Scene and its output is a similarity map indicating a degree of alienness between a location on the map and the reference[16,17,18,19,20,21,22].

II. METHODOLOGY

QBIC system is designed to query a category-valued remote sensing image database \( G \) consisting of pixels \( g_i = [x_i, y_i, c_i] \) \( i = 1, \ldots, N \), where \( x_i \) and \( y_i \) are the \( i \)th pixel spatial coordinates and \( c_i \) is the categorical class label. \( N \) is very large, on the order of \( 10^9 \)–\( 10^{10} \) for continental- or global scale databases. The pixel’s class labels are given by \( C = c_1, \ldots, c_K \), where \( K \) is typically low, on the order of ten. The system works on the principle of query by example. Let \( A \subset G \) be a reference scene—a small subset of the overall database. Images may represent a particular pattern of Remote sensing image database classes in a study area—a locality selected for one reason or another as interested or representative. The purpose of the system is to calculate a degree of alienness between a reference scene and all other scenes \( \{A_i \subset G, i = 1, \ldots, M\} \) in the database. The number \( M \) is determined by a particular division of \( G \) into a set of individual scenes. Thus, the data set of the Remote sensing Image Database, the system will identify all images having compositions and patterns of image classes similar to those present in the reference scene.

A. Image Pattern Signature

Map as an array of pixel values corresponds poorly to contextual understanding of the map. Original pixel values are used to calculate more compact mathematical description of a map. Degree of pattern similarity is accessed by this method. Map is much simpler than an image, we design a pattern signature using only two features—distributions of colors and sizes of patches. The connected region of single color is called as patches. We refer to such region as patch rather than 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 the value of pixel, patch is homogeneous.

An image is segmented into patches using a standard connected component algorithm [9] which also calculates their sizes. Each pixel in the image is characterized by two features: its color and the size of the patch to which it belongs. An image pattern signature is a probability distribution of 2-D variable (color and patch size). Categorical value is identified by color, The patch size is a continuous variable and needs to be discretized for calculation of pattern signature.

A scene signature is a 2-D distribution of scene pixels with respect to color and patch size. Images show an importance of using both colors and patch sizes in defining a scene signature.

B. Image Pattern Similarity

The work [11], [12]–[14] on measuring similarity between categorical rasters is limited and focuses on the issue of change detection rather than on QBIC. In particular, past work on NLCD2006 Dataset it is difficult to complete an search of an image. In such context, similarity function is designed to measure a degree of departure from one scene to another in a pixel-by-pixel method. That function will assign a small value of similarity to the two images which are identical but rotated with respect to each other. Thus, it cannot be applied in the context of QBIC where scenes need to be assigned high values of similarity regardless of their rotation, translation, or pattern deformation, as long as they exhibit the same overall style or motif of spatial pattern. The novelty of our proposed similarity function is its design geared toward recognition of spatial patterns that are similar in the sense of overall style or motif rather than in the sense of specific spatial correspondence. To the best of our knowledge, the only previous work addressing the issue of pattern similarity in the sense of pattern style was conducted in the context of landscape ecology rather than in the context of remote sensing. Sets of landscape indices were used [10] to compare the maps; however, no single-valued measure of similarity was developed.

Our Image pattern similarity measure is based on the notion of mutual information. The probability distribution function (pdf) of variable \( Y \) is just a 2-D histogram constructed from all pixels. Second, we define another random variable \( X \) that assigns equal probabilities to selecting one of the two possible images. The pdf of variable \( X \) is \( P_X(xk) = 1/2 \) for both \( x_1 = A \) and \( x_2 = B \). This random variable formalizes a simple act of randomly choosing one of the two scenes. Finally, we define a joint pdf \( P_YX(yij, xk) \) which assigns a probability to choosing a scene \( k \) and drawing a pair \( yij \) from this scene. We use informational entropy to characterize a probability distribution. Joint entropy \( H(YX) \) is the entropy of joint pdf \( P_YX(yij, xk) \), and specific conditional entropy \( H(Y | X = xk) \) is the entropy calculated only from the histogram of scene \( xk \). Finally, conditional entropy \( H(Y | X) \) is an average of the two specific conditional entropies \( H(Y | X = x1) \) and \( H(Y | X = x2) \) or an average of entropies calculated from histograms restricted to individual scenes. Mutual information \( I(YX) \), given by \( I(YX) = H(Y) - H(Y | X) \), measures an average reduction of unpredictability of \( Y \) if the specific scene
is set. Because \( H(Y) > H(X) = 1 \), the range of \( R(X, Y) \) is between zero and one as the upper bound of mutual information is given by \( \min[H(X), H(Y)] \). The value of \( R(Y, X) = 0 \) if both scenes have identical 2-D histograms, and the value of \( R(Y, X) \rightarrow 1 \) if the scenes are dominated by distinct single colors. Thus, \( r(A, B) = I(X, Y) \) is a convenient measure of “distance” between two scenes A and B, if, by the distance, we understand the increasing difference in patterns of colors and patch sizes in the two images. Conversely, \( \text{sim}(A, B) = 1 - I(X, Y) \) is a convenient measure of similarity between the two images from the pattern point of view. The range of \( \text{sim}(A, B) \) is between 0 and 1, we express similarity in terms of percentage.

### III. PROBLEM DEFINITION

Many systems have been designed for image retrieval process. In which the visual image features like texture, color, shape are derived automatically to describe image. These visual features are also used for indexing purpose which is easier to user to construct. For target image search, user presents an example image as a query and system returns images which is similar to extracted features. The today’s system which is already available re-retrieved previously checked images in current iteration and face the following disadvantages:

**A. Local Maximum Trap Problem:**

The guarantee of getting target image is none.

**B. Slow Convergence:**

In this the number of iterations is increases to get target image. In this proposed system will retrieve images from image database that are closest to query image and return the target image. Here supervised classification technique is used to minimize the search area. And from that search space it will start the searching by considering minimum bounding box. It will help us to find target image with a smaller number of iteration and minimum time complexity.

### IV. SEARCH METHODS

The performance of search methods is finding out by taking comparison of the algorithm with another algorithm. Search method can be divided into three main categories:

**A. Target Search**

From image database the user can find exact image which is desired to him. In target search, the process of image searching should be continuously repeated till user didn’t get target image or expected result are not found. Until images similar to target image are not derived; target search should not be stopping their process.

The target search is applied to image database when user wants to derive image that already known to him. If we want to know the similar image is already used (registered) or not, at that times target search method is applicable to find out the information [8].

**B. Category Search**

When user want to find relevant images without aware of time the category search method is used. Retrieval of given semantic class of images is the main goal of category search.

**C. Open-Ended Search**

Without any goal, images are searched from any specific database. In throughout the search process, goal is not remained constant, it may vary many times. Due to this, the starting query and the different results while navigating through the database are completely different from initial query [8]. Due to this reason it is too much difficult to retrieve the desire image with such type of search category.

### V. TARGET SEARCH

An important task of target search method is to find intended target images by achieving fast speed and to escape the local maximum trapping problem [7]. Also target search finds the desired image with minimum time complexity by reducing number of iterations.

Re-retrieving already checked images is one of the deficiencies of existing techniques. Due to this the two major problem local maximum trapping and slow convergence is occurred and we cannot get target image. And here our main goal is to leave the already checked images.

**Naive Random Scan Method**

Until the target image was not found out by user, the NRS method at a time continuously retrieves the k different images. A set of k different unchecked images are retrieved at each iteration. Different iteration contains k different images which are unchecked. After some finite iteration the NRS find out the desired image without suffering maximum trapping. But only for a small database set it is suitable [7].

**Local Neighboring Movement Method**

In LNM the re-retrieval of already checked images are not taken out. When LNM suffer from local maximum trapping, LNM have a best solution that it counts the neighbour points of the given query image and choose the one point which is nearer to the target. And this process is repeated until target images are not found by taking so many iterations. In this way, the LNM solve the local maximum trapping problem [7].

### VI. CONCLUSION

We have proposed a methodology for exploration of large category-valued geospatial data sets using a tool based on the QBIC principle. The functionality of our tool is very intuitive—selects a reference scene, run a search, and examines the resultant similarity map. The search time of entire database for the remote sensed images are minimized. We improve the forms of image signature and image similarity to decrease the semantic gap. Image signature is improved by adding an additional feature (patch shape) to the pair of features (color and patch size) used by its current implementation. After add the shape helped to differentiate
between scenes which have similar compositions of colors and similar patch size distributions. The 2-D (or 3-D) histogram representing an image signature is redistributed based on similarities between various colors and patch sizes. Such redistribution incorporates an additional knowledge into the image signature, thus improving correspondence between the value of similarity as calculated by an algorithm and a degree of similarity as perceived by an analyst.

REFERENCES


7. Sivaram, M. "Odd And Even Point Crossover Based Tabu Ga For Data Fusion In Information Retrieval." (2014).


