Validating a Geographical Image Retrieval System

Bin Zhu
Management Information Systems Department, University of Arizona, Tucson, Arizona 85721, E-mail: bzhu@bpa.arizona.edu

Hsinchun Chen
Management Information Systems Department, University of Arizona, Tucson, Arizona 85721, (520) 621-4153, E-mail: hchen@bpa.arizona.edu

This paper summarizes a prototype geographical image retrieval system that demonstrates how to integrate image processing and information analysis techniques to support large-scale content-based image retrieval. By using an image as its interface, the prototype system addresses a troublesome aspect of traditional retrieval models, which require users to have complete knowledge of the low-level features of an image. In addition we describe an experiment to validate the performance of this image retrieval system against that of human subjects in an effort to address the scarcity of research evaluating performance of an algorithm against that of human beings. The results of the experiment indicate that the system could do as well as human subjects in accomplishing the tasks of similarity analysis and image categorization. We also found that under some circumstances texture features of an image are insufficient to represent a geographic image. We believe, however, that our image retrieval system provides a promising approach to integrating image processing techniques and information retrieval algorithms.

1. Introduction

Despite the increasing volume of images in geographical databases, access to them is hindered by the difficulty of retrieving information from large collections. To make an image system accessible requires integration of image processing and information retrieval technologies. Currently, these two research areas are completely separate and rarely interact. To counteract the non-scalability of the traditional approach of textual annotation, effective algorithms for image feature extraction and image representation have been developed in the image processing field. This is evidenced by several recent prototypes such as the Photobook system developed at MIT (Pentland et al., 1994) and by commercial system such as the QBIC system developed by IBM (Flickner et al., 1995). However, incorporation of those algorithms into an effective image retrieval system has not been completely explored.

In this paper, we describe an interactive experiment to validate the performance of a geographical image retrieval system that is a subsystem of our Geographical Knowledge Representation System (GKRS) (Chen et al., 1998). Its goal is to support aerial photograph analysis and retrieval. The system integrates the algorithm for feature extraction created by the Alexandra Digital Library project (Ma & Manjuath, 1996; Manjuath & Ma, 1996) with a classification technique called self-organizing map (SOM) (Kohonen, 1995), which has been successfully applied in textual information retrieval in the Illinois/Arizona digital library project (Chen et al., 1998; Orwig et al. 1997). As a first step, this system provides three functionalities: similarity analysis, region segmentation, and image categorization. Details of the image system are described in section 3.

Most research in image retrieval compares several potential algorithms by running them on data sets from the Brodatz library (Picard & Kabir, 1993). Research has rarely evaluated the performance of an algorithm against that of human subjects. To address this weakness, the objective of our experiment was to compare the performance of our image retrieval system with that of human subjects. Our experiment used 30 human subjects, a remote sensing expert, and 10 different images of the same size. During the experiment, the system and the human subjects accomplished similar tasks. Their results were evaluated by comparison with the results presented by the expert. A description of this experiment and the results are presented in sections 4 and 5, respectively.

2. Image Retrieval System

An image retrieval system requires the incorporation of both image representation and information retrieval techniques. The traditional algorithm for representing an image

is based on its author, date, and content. However, this approach is unable to capture the complete contents of an image and requires manual effort to define and enter the necessary annotation. Another approach, searching images based on their low-level features, has therefore been introduced and become a promising research alternative. In 1983, Bertin identified seven visual variables from which graphics are constructed. These are size, position, color value, texture, color hue, orientation, and shape. However, in the research of image processing, an image is represented by its low-level features like color, texture and shape.

Another important factor that affects an image retrieval system is its retrieval model. Most existing image retrieval systems are based on pattern recognition techniques (Huang et al., 1996). The image retrieval process of these systems assumes that users have complete knowledge of the low-level features of an image to map the pattern they perceive. This assumption, however, is not true under most circumstances. In addition, users may perceive patterns on the same objects differently and, consequently, may map the same objects to different queries. An efficient image retrieval system needs to incorporate an algorithm that will translate high-level image queries provided by users into low-level visual features (Picard & Kabir, 1993).

3. SOM-AIR: An Image Retrieval System Based on Self-Organizing Map

This paper reports our ongoing work in developing SOM-AIR: an aerial photograph image retrieval system based on SOM. Our system retrieves images based on texture pattern. The system's clickable image interface enables users to specify their queries without specific knowledge of texture. The current system supports three functionalities: similarity analysis, region segmentation, and image categorization by creating an image thesaurus (Ramsey et al., 1999).

3.1. Feature Extraction

Automatic feature extraction is a crucial step in creating a scalable image system. Our system focuses on texture feature extraction. A variety of algorithms can be employed to extract low-level features in image retrieval systems. For instance, QBIC calculates the texture features of an image according to its coarseness, contrast, and directionality. Photobook consists of three parts: the Appearance Photobook, the Shape Photobook, and the Texture Photobook. In the Texture Photobook, Wold-based representations are used to extract the texture features of an image. In the prototype system for the Alexandria Digital Library Project, Manjunath & Ma (1996) used Gabor filters to extract texture features of an aerial photo.

The selection of an algorithm for image representation varies with the image type. For instance, an algorithm may work well with medical images but may not be appropriate for geographical images. In our prototype system, since we used aerial photos as input, we also employed Gabor filters as our image representation algorithm. As indicated by Manjunath & Ma (1996), Gabor filters perform well in representing aerial photos. Our experiment, presented in the next subsection, also indicated that Gabor-filter-extracted features and associated similarity measures could map the human perception of aerial photo similarity.

Our system divides one image into small tiles, each of which has 128×128 pixels. Tiles are the smallest units in our system and are represented by their Gabor features (Ma & Manjuath, 1996). Our system constructs a feature vector of length 60 to represent each tile and stores all the feature vectors in a feature database.

3.2. Define Similarity Measure

Taking a user's query and finding the most similar tiles is the process of similarity analysis. Users expect an image retrieval system to return a set of images that match their queries. We chose to use Euclidean distance in the texture feature space as the similarity measure, because it is commonly employed in existing image systems and information retrieval systems. Our system considers two image tiles to be similar if they have relatively small values of Euclidean distance. As a result, when a user specifies a texture pattern (e.g., orchard), the system calculates the Euclidean distances between this tile and all the other tiles. The system then sorts all the other tiles in ascending order according to their distances and returns the top 10 tiles as the ones most similar to the one referenced. The images to which these chosen tiles belong can also be displayed on the screen. Figure 1 presents the interface of this functionality.

3.3. Segmentation

The goal of region image segmentation is to distinguish "objects" with large geographic features (e.g., airports, dams) and to support more specific queries such as "Find images that have airport." Our system partitions one image into nearly 6,000 tiles and accomplishes the task of region segmentation by grouping adjacent similar tiles. The whole process is described in Manjuanth & Ma (1996).

Based on the feature vectors of its tile elements, our system calculates a feature vector for each of the regions created. Users can specify their queries in terms of regions and the system can retrieve similar regions and the corresponding images. Figure 2 is an example of image segmentation.

3.4. Image Categorization

An efficient way to index and retrieve extracted features is crucial to the scalability of an image system. We used Kohonen's Self-Organizing Map (SOM) (Kohonen, 1995) to categorize the texture features extracted and visualize the categories created. Users can browse the map created and find the desired texture. After clicking on the category of

FIG. 1. A user can browse the whole image provided by the interface and click on a tile of interest. The system then pops up another window called the result window. This window is titled in "Node View" and displays tiles similar to the clicked tile. At the same time, in order to indicate locations of those tiles in the original image, the system highlights the clicked tile in red, and the other similar tiles in blue in both the original image and the result window

interest, all the tiles belonging to this category and corresponding images can be browsed.

SOM is defined as a mapping from an input space onto a two-dimensional array of output nodes, each of which is connected with an input vector via variable scalar weights. In our system, an image is divided into tiles, and each tile is represented by its feature vector. We use the extracted feature vectors as inputs of the SOM. The SOM algorithm we use in our image system is summarized below:

(1) Initialize the weights connecting the input vector and the output nodes at a small random value and initialize the neighborhood size.
(2) For each input vector, calculate the Euclidean distance between the input vector and output nodes and find the "winning" output node with the minimum Euclidean distance to the input vector.
(3) Adjust the vector weights of both the "winning" nodes and their defined neighborhood nodes with a learning rate.
(4) Repeat 1, 2, and 3 for every input vector.
(5) Repeat 1, 2, 3, and 4 for several iterations, using a decreasing neighborhood size and a decreasing learning rate.
(6) Assign every tile to an output node that has the minimum Euclidean distance to the input vector; label each node with the tile having the shortest Euclidean distance from it.
(7) Merge adjacent output nodes that have the same label to form a category.

Thus the SOM map created can be considered to be a graphical categorization of the images. Figure 3 presents the interface of the image categorization.

4. Experiment Design

Our geographical image system gives rise to several research questions:

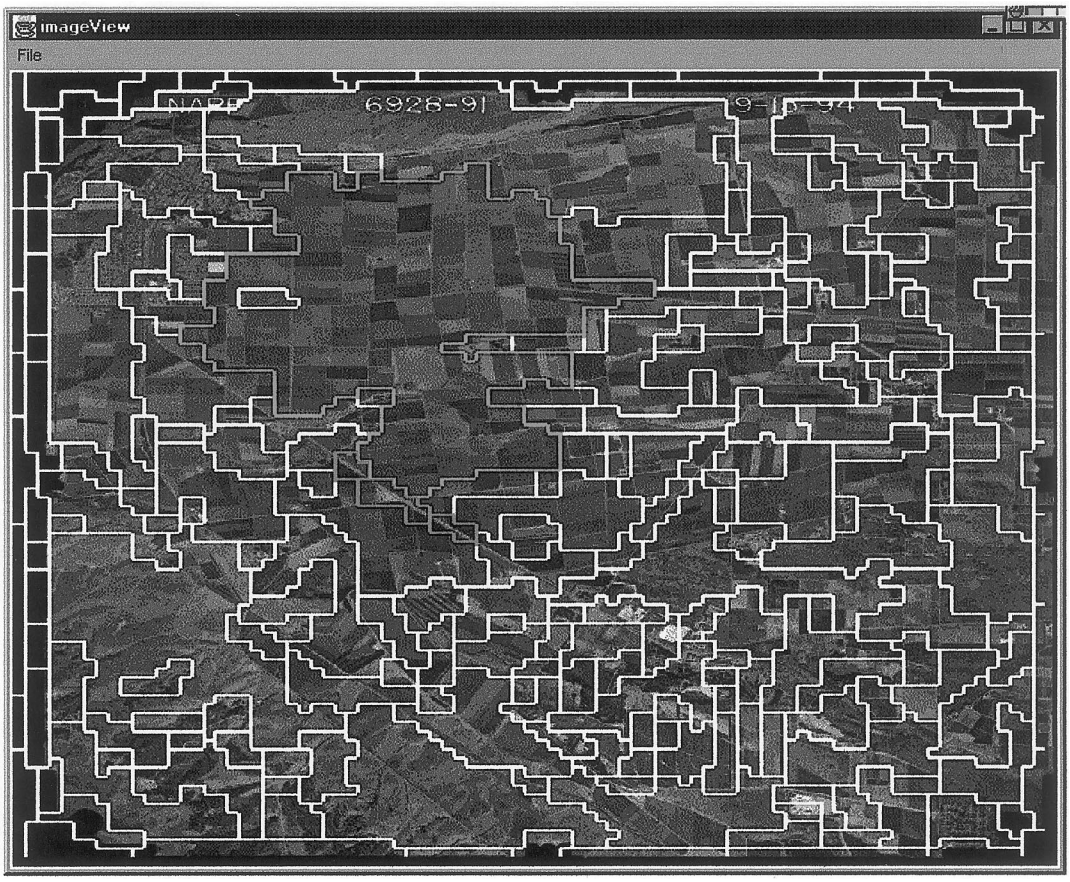

FIG. 2. The whole image in this figure contains 6400 × 6400 tiles. The system forms the regions by drawing lines that conform to the boundaries between tiles. One region represents a large geographical feature. For instance, the region highlighted in red is farm land, while the regions highlighted in green represents roads. Users can click on a region of interest, such as roads or vegetation, and the system will respond by displaying all similar regions.

- To what extent do extracted features and corresponding similarity measures map a human's perceived similarity?
- Does the segmentation method distinguish objects that could be distinguished by a normal person?
- Does applying SOM to the feature vectors produce better results than manual indexing?

In order to address these questions, we designed and conducted a three-part experiment. To address the quality of the similarity analysis of the system, the first part involved comparing the similar tiles retrieved by the system with similar tiles chosen by human subjects. We next compared segmentation results of the system with those of human subjects. The third part of the experiment addressed the effectiveness of the adopted SOM in image categorization. Each of these steps involved 10 subjects and 10 different images. Every image consisted of 12 × 16 (192) tiles. Thirty human subjects were involved in our experiment; most of them were graduate and undergraduate students in the Department of Management Information Systems at the University of Arizona (UA). The others were graduate students from other UA departments. In each comparison, both the system and human subjects performed the same tasks. An expert who had had three years of experience in analyzing remote sensing pictures evaluated both sets of results.

4.1. The First Part: Similarity Analysis

We chose 10 different images. For every image, we randomly selected six tiles as reference tiles. We assigned every subject one image and its corresponding reference tiles. For each reference tile, we asked the subject to evaluate every tile on the image and assign a score, from 0 to 10, based on its similarity to the reference tile. We suggested that subjects judge similarity according to the content of tiles and assign a score based on their own visual perceptions of coverage, layout, and texture orientation. We designed and implemented an interface for this part of the experiment, which is displayed in Figure 4.

The interface consisted of four components, the Image Panel, the Reference Panel, the Results Panel, and the Control Panel. The entire process of the first part of the experiment was as follows. A subject chose a reference tile by clicking on one of the tiles on the Reference Panel. This tile became the current referenced tile and was highlighted in red. After selecting the reference tile, the subject could choose one or more image tiles from the image by clicking on image tiles on the Image Panel. The selected image tiles were highlighted in red and were displayed in the Results Panel. By clicking on the scroll bar on the Control Panel, the subject could assign to the selected image tiles a score based

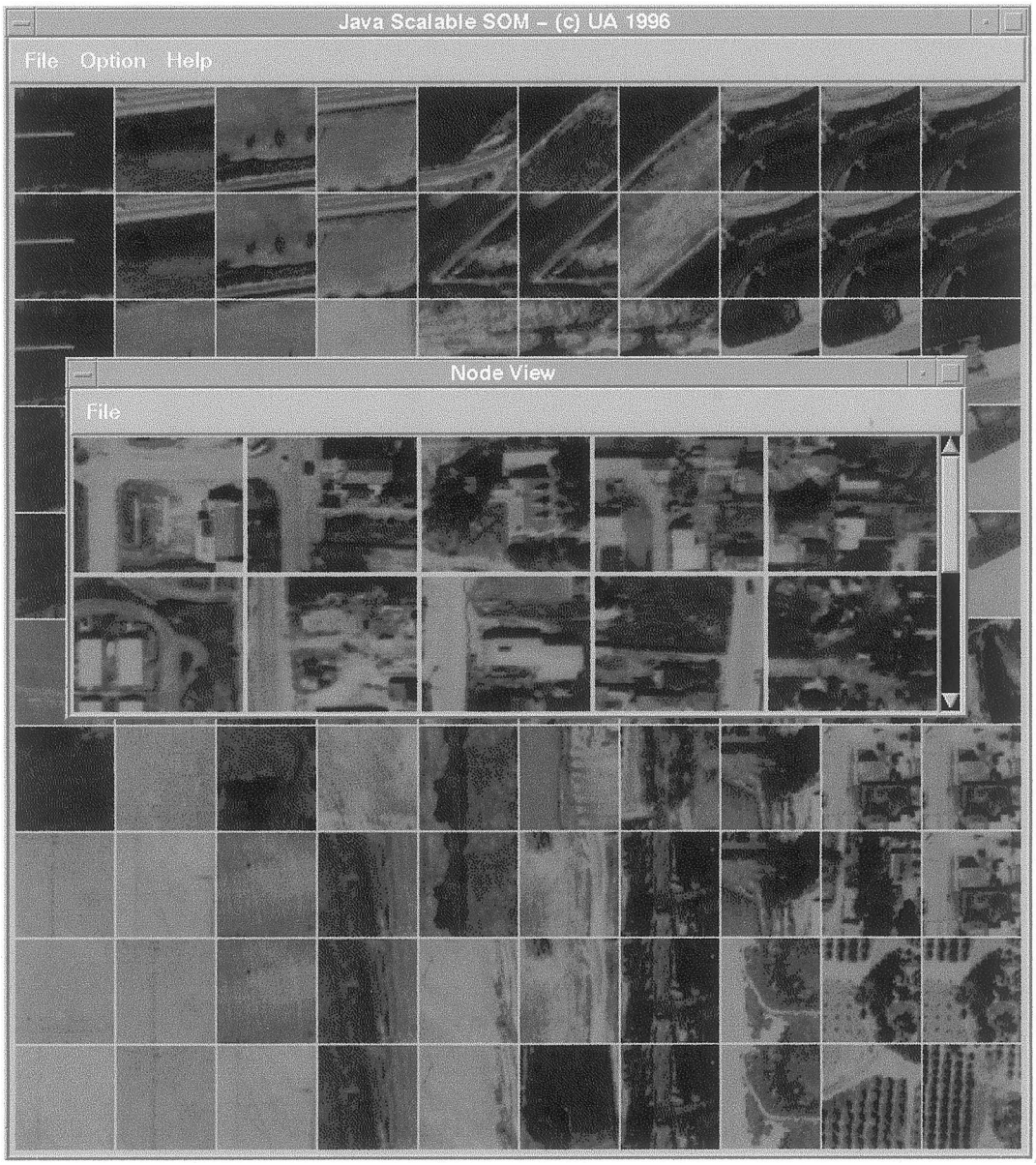

FIG. 3. This is the two-dimensional display of a SOM with 100 (10 × 10) output nodes, each of which is represented by its representative tile. Output nodes with the same representative tile are considered to be in the same category. When a user clicks on a representative tile, the system pops up another window titled "Node View" and displays all the tiles belonging to that category in this window. Since we use the image tile itself as the label of each category, users can easily find a desired category and browse all the image tiles of interest by clicking on the representative tile.

on the similarity between the current reference tile and the tiles selected. He then could click on the "ok" button on the Control Panel to finish this assignment. As a result, the tiles on the Results Panel were removed and the corresponding tiles on the Image Panel were highlighted in green.

4.2. The Second Part: Region Segmentation

A different group of 10 subjects participated in the second part of the experiment. We used the same images as in the first part and asked the subjects to draw lines around areas in the image that they considered similar. For instance, if an area was predominately orchard, it was to be enclosed as one area, while an area of buildings was to be enclosed as a different area. The only restriction was that in drawing the boundaries of the regions subjects had to use tile boundary lines. We established this restriction in order to make the results comparable to those of computers. We present the interface for this part of the experiment in Figure 5.

4.3. The Third Part: Image Categorization

We continued to use the same set of images in the third part of the experiment, but a different group of 10 human subjects participated. Every human subject worked with one image that was categorized by using the SOM algorithm.

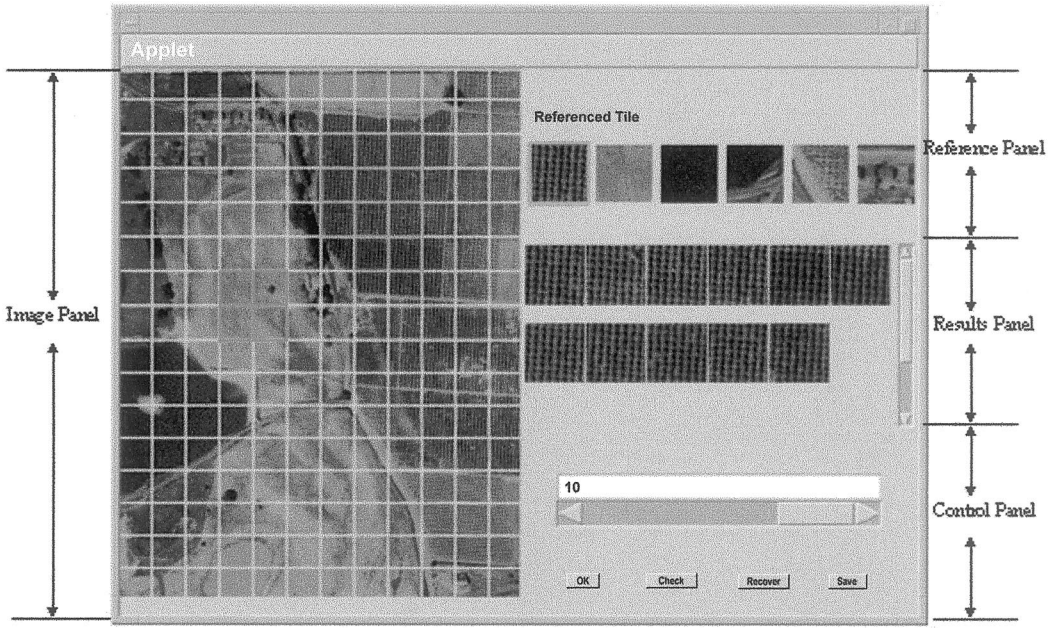

FIG. 4. The interface consists of 4 components, the Image Panel, the Reference Panel, the Results Panel, and the Control Panel. The Image Panel displays an image of 12 × 16 (192) tiles, while the reference panel presents the 6 reference tiles and shows the current reference tile highlighted in red. Tiles on the Image Panel being inspected are also highlighted in red and are displayed on the Results Panel. In this figure, the subject is examining 11 image tiles and there are 4 tiles on the Image Panel that have been assigned scores corresponding to the current reference tile. The buttons "check" and "recover" are used to check how many tiles have been assigned a score corresponding to current referenced tile. The "ok" button is used to finalize the assigning of a score to the image tile under inspection and the "save" button is used to save the results.

For every image, the label tiles of created categories were regarded as representative tiles. We used these system-selected representative tiles as suggested categories and asked human subjects to put every tile on the image into a category.

Figure 6 presents the interface for this part of the experiment, which resembles that for the first part of the experiment. A subject could select a current representative tile and have some image tiles brought up on the Results Panel by clicking on the Image Panel. The image tiles selected

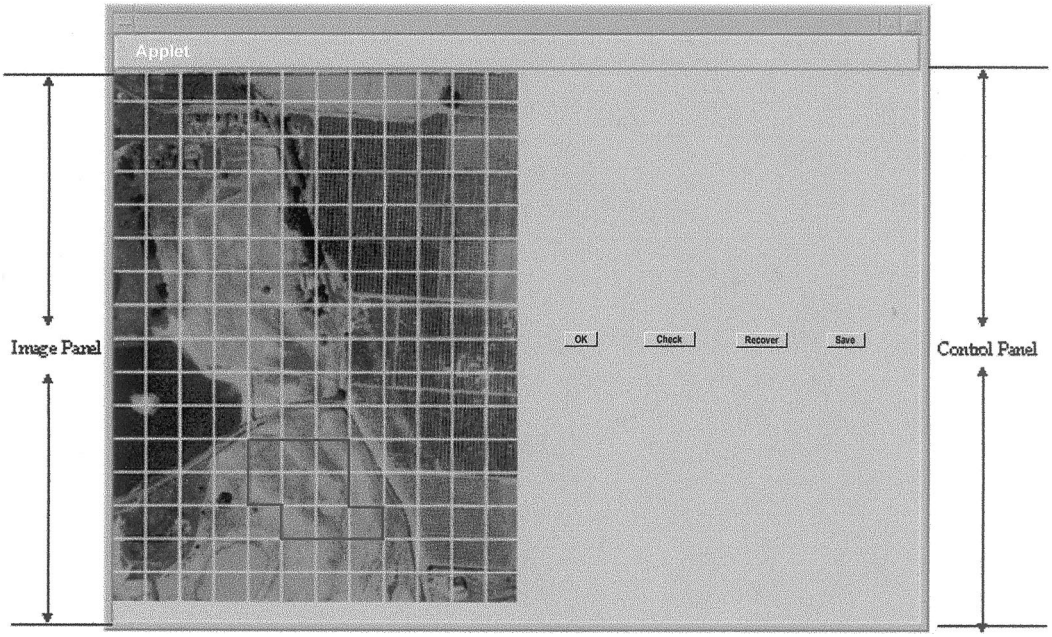

FIG. 5. The region being inspected has a boundary color of blue. The subject can include or exclude an adjacent tile by clicking on it. He can use the "ok" button to finalize his decision about this region. The region with a boundary color of red indicates that this is a created region. The "check," "recover," and "save" buttons on the Control Panel have similar functionality to those in the first part of the experiment.

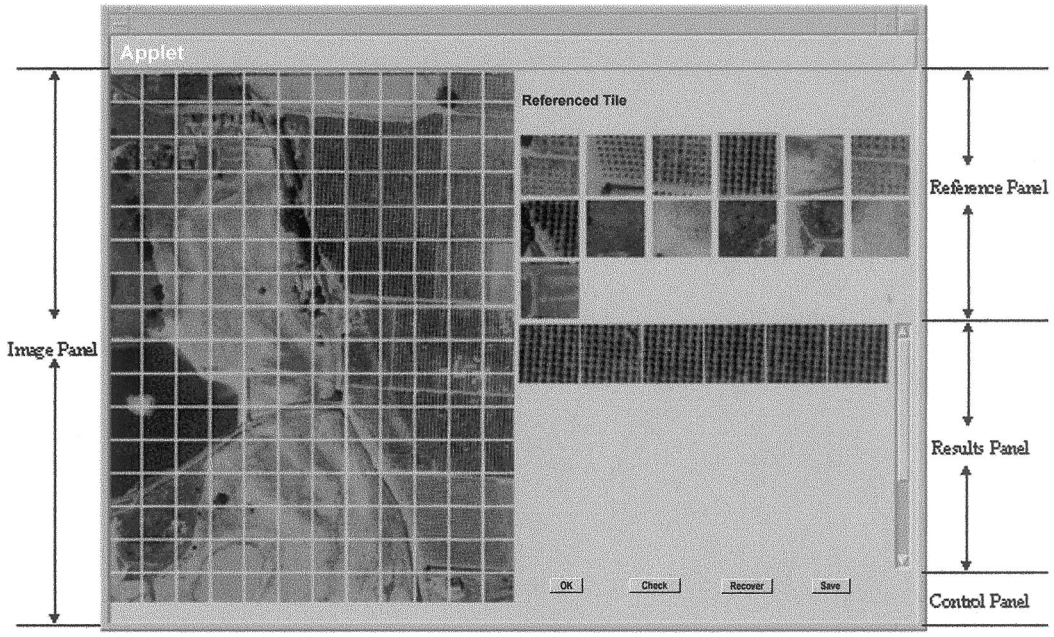

FIG. 6. The interface has 4 components: the Image Panel, the Reference Panel, the Results Panel, and the Control Panel. The Image Panel displays an image, while the Reference Panel presents 13 representative tiles, each of which suggests a category. The current representative tile is highlighted in its own color, while image tiles being inspected are highlighted in the same color and are displayed on the Results Panel.

could then be assigned to the category represented by the current representative tile by clicking the "ok" button on the Control Panel. After such assignment, tiles on the Results Panel were removed and the corresponding tiles on the Image Panel was highlighted in the same color as the current representative tile. The subject could then change the current representative tile and repeat the whole process until all tiles on the Image Panel had been assigned to a category. The "check," "recover," and "save" buttons on the Control Panel functioned as they did in the first part. Subjects took 40–50 minutes to finish this task.

5. Experiment Results

Humans' perception of image texture is subjective (Huang et al., 1996); different persons or the same person at a different point of time may have different perception criteria. Therefore the evaluation of the system incorporates the concept of relevance. In this experiment, we considered the objective relevance of the system that indicates the relationship of an image tile to a query specified by a user (Janes, 1994). We relied on an expert to evaluate the performances of the system and the human subjects. By comparing the performance of the system with those of the human subjects, we hoped to determine how useful our system would be in helping a nonexpert user in image retrieval. We used two measures in our evaluation, *precision* and *recall.* Precision represented the relevance of the retrieved information, while recall indicated how much of the relevant information in the database was retrieved (Salton & McGill, 1983). We designed three methods to calculate the precision and recall for the three parts of the experiment.

5.1. Similarity Analysis

Our geographical image system determines the similarity between image tiles based on Euclidean distances in their feature space. We decided on a threshold of Euclidean distance and considered an image tile to be similar to a reference tile when the Euclidean distance between them was shorter than that threshold value. This threshold was determined during the pilot study according to our visual judgment. We also set a threshold value for the scores assigned by human subjects in the first part of the experiment. We believed other images were either similar or not similar to a reference tile. Janes (1991) conducted an experiment to study humans' perception of relevance. During his experiment, in which human subjects were asked to give a "break point" on a continuum of relevance from zero relevance to complete relevance, he found that the break points assigned by human subjects exhibited a wide range. However, the mean value of the break point remained at around 50 on a scale of 100. Therefore, we considered an image tile to be similar to the reference tile if human subjects assigned it a score higher than or equal to 6. Meanwhile, the expert selected all the similar image tiles corresponding to every reference tile. We used the measures of *subject recall* and *subject precision* to evaluate the performance of human subjects and applied *system recall* and *system precision* to represent the performance of our system. These measures were defined as follows:

- *Subject recall* reflected the percentage of total number of tiles similar to those retrieved by the expert that were located by the subjects.
- *System recall* reflected the percentage of the total number of similar tiles retrieved by the expert that were located by the system.
- *Subject precision* reflected the percentage of the number of similar tiles selected by the subjects that were considered by the expert to be similar.
- *System precision* reflected the percentage of the number of tiles retrieved by the system as similar that were considered by the expert to be similar.

When results were compared, the system exhibited no significant difference from human subjects in precision, but it performed less well than human subjects in recall. Ten subjects worked on 10 different images, from each of which we randomly selected 6 tiles as reference tiles to this image. We asked subjects to assign a score to every image tile corresponding to a reference tile based on its similarity to this reference tile and to repeat this process for every reference tile on the interface. Because of limited time and cognitive resources, 9 subjects finished 4 reference tiles and only one subject finished 5 reference tiles. We obtained 41 finished reference tiles at the end of the experiment. Using the previously described threshold, we calculated subject recall, subject precision, system recall, and system precision for each finished reference tile. We thus had 41 evaluations each for subjects and the system. We used Minitab statistical software to perform a one-way analysis on each comparison (Figure 7), using the value of P to indicate the significance and setting the threshold value of P at 10%. At $P > 10\%$ there was no significant difference between the computer results and those of human subjects, while at $P < 10\%$ there was a significant difference. The P value ($P = 0.343$) of precision indicated that the results were not significant, while the results of recall ($P = 0.004$) indicated that our system performed worse in recall than human subjects. *For a given referenced tile, the average number of similar tiles retrieved by human subjects was larger than the average of those retrieved by system.* Although the system did relatively poorly in recall, we do not view this as a major weakness. When people retrieve images from a database, they usually are interested in finding the first N images that are most similar to a query image. Under these circumstances, precision is more important than recall. We found that for a given referenced tile, the average number of similar tiles retrieved by human subjects was 68, while this number for our system was only 28. This fact could explain why the system exhibited relatively poor recall. Therefore, we believe we could easily improve the system's recall by allowing a lower threshold of similarity measures.

An interesting result of this research was the lack of duplication in the top five similar tiles suggested by the system and by human subjects. When the system and human subjects were asked to find the five tiles most similar to a referenced tile, their results were likely to be different. The explanation for this appears to be that the measure of

```
Analysis of Variance of Recall (Similarity)
Source   DF      SS      MS      F      P
Factor    1  0.4316  0.4316   8.80  0.004
Error    81  3.9729  0.0490
Total    82  4.4044
                        Individual 95% CIs For Mean
                        Based on Pooled StDev
Level     N    Mean   StDev -------+---------+---------+---------
System   42  0.6575  0.2418  (-------*--------)
Subject  41  0.8018  0.1985                   (-------*--------)
                               -------+---------+---------+---------
Pooled StDev =  0.2215          0.640     0.720     0.800

Analysis of Variance of Precision (Similarity)
Source   DF      SS      MS      F      P
Factor    1  0.0630  0.0630   0.91  0.343
Error    81  5.6206  0.0694
Total    82  5.6836
                        Individual 95% CIs For Mean
                        Based on Pooled StDev
Level     N    Mean   StDev ----+---------+---------+---------+--
System   42  0.4749  0.2527          (-----------*-----------)
Subject  41  0.4198  0.2740 (-----------*-----------)
                               ----+---------+---------+---------+--
Pooled StDev =  0.2634          0.360     0.420     0.480     0.540
```

FIG. 7. A one-way analysis using Minitab statistical software.

Euclidean distance in our system can match the human perception of similarity only to a certain extent; it needs to be combined with other similarity measures to simulate human perception. This finding is consistent with the results of Manjuath & Ma (1996), whose research also indicates that nearest-neighbor searching fails to retrieve some other more relevant patterns.

Both the system and human subjects did well in retrieving tiles with distinguishable texture, but had difficulty in retrieving image tiles that could not be distinguished by texture alone. Both the system and human subjects did well in retrieving image tiles of orchard, which has obviously distinguishable texture (subject recall = 96.9%, system recall = 57.9%, subject precision = 42.5%, and system precision = 100%), but both the human subjects and the system, especially the system, had difficulty in retrieving the pure water tiles (subject recall = 80%, system recall = 37.5%, subject precision = 27%, and system precision = 17.6%). We found that water tiles had the same homogeneous texture as soil or forest tiles. Our system certainly had difficulty in distinguishing these because it represents image in terms of texture features. Human subjects had the same problem. Although they used contrast information in comparing an image tile and its background, their decision was more affected by texture, resulting in their retrieving pure soil tiles as similar to pure water tiles, which had similar gray scales and textures. This indicates that, under some circumstances, texture alone is insufficient in representing image.

Analysis of Variance of Recall (Segmentation)

Source	DF	SS	MS	F	P
Factor	1	5.5284	5.5284	55.73	0.000
Error	3838	380.7327	0.0992		
Total	3839	386.2611			

Individual 95% CIs For Mean
Based on Pooled StDev

Level	N	Mean	StDev	----------+---------+---------+-----
System	1920	0.5248	0.2931	(----*----)
Subject	1920	0.6007	0.3354	(---*----)
				----------+---------+---------+-----
Pooled StDev = 0.3150				0.540 0.570 0.600

Analysis of Variance of Precision (Segmentation)

Source	DF	SS	MS	F	P
Factor	1	17.1459	17.1459	174.20	0.000
Error	3838	377.7571	0.0984		
Total	3839	394.9030			

Individual 95% CIs For Mean
Based on Pooled StDev

Level	N	Mean	StDev	------+---------+---------+---------+
System	1920	0.5352	0.3025	(--*--)
Subject	1920	0.6689	0.3246	(--*--)
				------+---------+---------+---------+
Pooled StDev = 0.3137				0.550 0.600 0.650 0.700

FIG. 8. Minitab's one-way ANOVA test for recall and precision.

5.2. Region Segmentation

In this part of the experiment, the measure of recall and precision were defined as follows:

- *Subject recall* reflected the percentage of total number of image tiles the expert grouped with a tile that had been grouped with that tile by subjects.
- *System recall* reflected the percentage of total number of image tiles the expert grouped with a tile that had been grouped with that tile by the system.
- *Subject precision* represented the percentage of number of image tiles subjects grouped with one tile that had been grouped with that tile by the expert.
- *System precision* represented the percentage of the number of image tiles the system grouped with one tile that had been grouped with that tile by the expert.

The system did worse than humans on both recall and precision. For each tile on an image, we calculated the subject recall, subject precision, system recall, and the system precision. We had obtained 1,920 (192 × 10) evaluations at the end of the experiment. Figures 8 illustrates Minitab's one-way ANOVA test for recall and precision. We continued to use P = 10% as the threshold and we found that the system did worse in both recall (P = 0.00) and in precision (P = 0.00).

Most differences between selections by human subjects and by the system occurred in connection with tiles having undistinguishable texture or more than one types of texture. We found that most of the differences between human subjects' and the system's selections occurred on image tiles that contained more than one land surface type. For instance, if one tile was mainly occupied by orchard but included a small road, human subjects probably grouped it with adjacent orchard tiles, while our system would regard it as another region. Once again, in this part of the experiment, both the subjects and our system had problems similar to those they had in the first part. They both grouped adjacent tiles that had similar texture and gray scale but belonged to different land surface types.

5.3. Image Categorization

In this part of the experiment, the definitions of recall and precision were as follows:

- *Subject recall* represented the percentage of total number of image tiles assigned to a category by the expert that had been assigned to this category by subjects.
- *System recall* represented the percentage of total number of image tiles assigned to a category by the expert that had been assigned to this category by the system.
- *Subject precision* reflected the percentage of total number of image tiles assigned to a category that had been assigned to this category by the expert.
- *System precision* reflected the percentage of total number of image tiles assigned to a category that had been assigned to that category by the expert.

Analysis of Variance of Recall (Categorization)

Source	DF	SS	MS	F	P
Factor	1	0.039	0.039	0.34	0.562
Error	252	29.287	0.116		
Total	253	29.327			

Individual 95% CIs For Mean
Based on Pooled StDev

Level	N	Mean	StDev	--------+---------+---------+--------
System	127	0.4160	0.3335	(----------*----------)
Subject	127	0.3912	0.3482	(----------*----------)
				--------+---------+---------+--------
Pooled StDev = 0.3409				0.360 0.400 0.440

Analysis of Variance of Precision (Categorization)

Source	DF	SS	MS	F	P
Factor	1	0.0069	0.0069	0.09	0.760
Error	252	18.7535	0.0744		
Total	253	18.7604			

Individual 95% CIs For Mean
Based on Pooled StDev

Level	N	Mean	StDev	----+---------+---------+---------+--
System	127	0.3358	0.2336	(----------*----------)
Subject	127	0.3462	0.3071	(----------*----------)
				----+---------+---------+---------+--
Pooled StDev = 0.2728				0.300 0.330 0.360 0.390

FIG. 9. Comparisons of recall and precision.

The system did at least as well as human subjects in image categorization. We calculated the subject recall, subject precision, system recall, and system precision for each category suggested, having obtained 127 evaluations at the end of the experiment. The comparisons of recall and precision are presented in Figure 9. There were no statistically significant differences between human subjects and the system in recall ($P = 0.562$) and precision (0.760), although in general the performance of our system was better.

The system was inclined to produce too many categories, and representative tiles sometimes were not really representative. Most of the subjects complained that there were too many suggested categories and that some of the representative tiles were similar to each other. This is probably due to the small size of the input data set to the SOM. As a matter of fact, when we created our system, we used hundreds of thousands of image tiles as the input data of our adopted SOM. A set of 192 tiles was too small for the SOM algorithm. However, it was impossible to increase the image size in the experiment. During our pilot studies, we found that even though the interface relieved human subjects of the need to manage tiles in the image, subjects still could not accomplish task when there were too many tiles due to limited cognitive resources. We found that when the number tiles of exceeded 200, most subjects either gave up on the task or assigned tiles to categories randomly because of cognitive overloading. We were pleased that, based on such a limited input size, our system could perform at a level comparable with that of human subjects.

6. Conclusion

When performing the task of similarity analysis, the system did as well as human subjects in precision, but did worse in recall. While completing region segmentation, the system did worse than human subjects. Our experiment found that the system did at least as well as human subjects in completing the task of image categorization. As a result, we believe our system could do as well as humans in image analysis and categorization. This is especially true for surface types with distinguishable texture. We believe our system addresses two issues in the image retrieval field. It successfully integrates image processing techniques such as the Gabor filter with information analysis algorithms like the SOM. The system created is scalable and has performance comparable to that of a normal person. On the other hand, our system addresses a troublesome aspect of traditional retrieval models, which require users to have complete knowledge of the low-level features of an image. Our system enables users to specify their queries by clicking on images and translates their high-level queries into low-level features.

Our system still has some weaknesses. For example, it represents images by only one low-level feature, texture, which by itself is not sufficient to represent images. Our future work involves improving the performance of the system by combining texture with other low-level image features (e.g., color and shape), improving segmentation techniques, and finding more representative similarity measures. We are also expecting to integrate this geographical image system with other GKRS systems, such as satellite data subsystems and textual information subsystems.

Acknowledgements

This project was mainly supported by:

- NSF/ARPA/NASA Digital Library Initiative. IRI94-113301. 1994–1998 (T. Smith, M. Goodchild, et al., "The Alexandria Project: Towards a Distributed Digital Library with Comprehensive Services for Images and Spatially-Referenced Information")
- NSF/ARPA/NASA Digital Library Initiative. 1996–1998 (H. Chen and T. Smith, "Supplement to Alexandria DLI Project: A Semantic Interoperability Experiment for Spatially-Oriented Multimedia Data")
- NSF/ARPA/NASA Digital Library Initiative. IRI-9411381. 1994–1998 (B. Schatz, H. Chen, et al., "Building the Interspace: Digital Library Infrastructure for a University Engineering Community")

References

Bertin, J. (1983). Semiology of graphics, William J. Berg (Trans.), Madison: University of Wisconsin Press.

Chen, H., Simith, T.R., Larsgaard, M.L., Hill, L.L., & Ramsey M. (1998). A geographic knowledge representation system (GKRS) for multimedia geospatial retrieval and analysis. International Journal of Digital Libraries, 1, 132–135

Chen, H., Houston, A.L., Sewell, R.R., & Schatz, B.R. (1998). Internet browsing and searching: User evaluations of category map and concept space techniques. Journal of the American Society for Information Science, 49 (7), 582–603.

Flickner, M., Sawhney, H., Niblack, W., & Ashley J. (1995). Query by image and video content: The QBIC system. IEEE Computer, 28 (9), 23–33.

Huang, T., Mehrotra, S., & Ramchandran, K. (1996). Multimedia analysis and retrieval system (MARS) project, Data Processing Clinic.

Janes, J.W. (1991). The binary nature of continuous relevance judgments: A study of user's perceptions. Journal of the American Society for Information Science, 42 (10), 754–756.

Janes, J.W. (1994). Other people's judgements: A comparison of users' and others' judgments of document relevance, topicality, and utility. Journal of the American Society for Information Science, 45 (3), 160–171.

Kohonen, T. (1995). Self-organized maps, Springer-Verlag, Berlin Heidelberg, chapter 3.

Ma, W.Y., & Manjuath, B.S. (1998). A pattern thesaurus for browsing large aerial photographs. Journal of the American Society for Information Science, Vol. 49 (7), 633–648.

Manjuath, B.S., & Ma, W.Y. (1996). Texture features for browsing and retrieval of image data. IEEE Transactions on Pattern Analysis and Machine Intelligence, special issue on digital libraries, 18 (8), 837–842.

Orwig, R.E., Chen, H., & Nunamaker, J.F. (1997). A graphical, self-organizing approach to classifying electronic meeting output. Journal of the American Society for Information Science, 48 (2), 157–170.

Pentland, A., Picard, R.W, & Sclaroff, S. (1994). Photobook: Tools for content based manipulation of image databases. Proceedings of the International Society for Optical Engineering, vol. 2185 (pp. 34–47).

Picard, R.W., & Kabir, T. (1993). Finding similar patterns in large image databases. In Proc. IEEE Intl. Conf. on Acoust., Speech, and Signal Processing (ICASSP '93), Minneapolis, MN, vol. 5, (pp. 161–164).

Ramsey, M, Chen, H., Zhu, B., & Schatz, B. (1999). R. A collection of visual thesauri for browsing large collections of geographic images. Journal of the American Society for Information Science, Vol. 50, No. 9, 826–834.

Salton, G., & McGill, M.J. (1983) Introduction to modern information retrieval. McGraw Hill Computer Science Series.