12-9-2008

Worm-Web Search: A Content-Based Image Retrieval (CBIR) System for the Parasite Image Collection in the Harold W. Manter Laboratory of Parasitology, University of Nebraska State Mueum

Ramalingamurthy Meduri
University of Nebraska - Lincoln, myrlmurthy@yahoo.com

Ashok Samal
University of Nebraska - Lincoln, samal@cse.unl.edu

Scott Lyell Gardner
University of Nebraska - Lincoln, slg@unl.edu

TEXAS TECH UNIVERSITY
Center *for* Advanced Study
of Museum Science *and* Heritage Management™

MUSEOLOGY

Museum of Texas Tech University Number 11 9 December 2008

WORM-WEB SEARCH: A CONTENT-BASED IMAGE RETRIEVAL (CBIR) SYSTEM FOR THE PARASITE IMAGE COLLECTION IN THE HAROLD W. MANTER LABORATORY OF PARASITOLOGY, UNIVERSITY OF NEBRASKA STATE MUSEUM

RAMALINGAMURTHY MEDURI, ASHOK SAMAL, AND SCOTT L. GARDNER

ABSTRACT

We have developed a prototype web-accessible content-based image retrieval (CBIR) system that allows internet/web-based sharing of biological collections that contain large numbers of images of archived specimens. This system will enable both researchers and educators to access verified, high quality data on biological collections that are available in any museum with digitized holdings. The CBIR system that we are testing can play an important role in understanding global biodiversity because no knowledge of the specific names of specimens need be known before useful information can be extracted from such databases. Our CBIR framework allows users to search image collections using query by content, query by example, and query by sketch. Additionally, we have developed a web based administrative interface to maintain and monitor any image collection. A key idea of our project is to develop a framework that assists educators and researchers to identify biological specimens and to study biodiversity with greater precision and speed. This is a web-based system that is low cost, extensible, flexible, and easily configurable. Our CBIR prototype was implemented and tested using specimens of the family Opecoelidae (trematodes of marine fishes) that are housed in the Harold W. Manter Laboratory of Parasitology Collection at the University of Nebraska. In the future, we plan to field-test the prototype and assess the ability to provide image retrieval from a variety of biological images and with a multitude of image features.

Key words: biological, CBIR, collection, content-based, database, image retrieval, museum, parasite, trematode, web-based

INTRODUCTION

Content-Based Image Retrieval (CBIR) is a process of automatic recovery of desired data from a collection of digital images from one or more computer image databases. The search and retrieval process can be based on either textual, ANSI based descriptions, or actual features depicted in digital photographs such as

color, texture, shape, or structure. Traditional databases and information retrieval systems are mainly based on ANSI or ASCII text matching through queries of structured normalized databases or via searches through embedded metadata in digital images. CBIR systems differ from text-based retrieval of metadata linked images (Mallik 2006; Mallik 2007) in that digital photographs or images are treated as arrays of pixel intensities and, in the sense of database normalization, these combinations of pixel arrays are mostly unstructured. One of the central issues in CBIR is the development of algorithms to extract useful information such as important structures or other morphological features of biological specimens from raw images before any kind of logical selection process of images or photograph is performed (Anonymous 2002). CBIR systems draw many techniques from the fields of image processing and computer vision, and development and optimization of CBIR methods and systems is an active area of research (Mallik 2006; Mallik 2007).

Content-based image retrieval is being used in a wide range of research and development projects, including: understanding the needs and information-seeking patterns of online computer users; the development and identification of flexible and comprehensive methods for image descriptions; data extraction and classification of raw images based on multiple image features; database design and indexing for storing large digital image volumes; and automated species identification with associated optical character recognition and database entry (e.g., HERBIS 2006). Also being pursued is research into content-based queries that use data derived from judgments made by humans. These methods are increasing scope and veracity of retrieval of images with useful data. In addition, flexible and efficient access methods to stored images are being developed following the advent of new computer-human interfaces (Eakins 2002).

The present paper utilized specimens of parasites that are stored in the Harold W. Manter Laboratory of Parasitology (HWML). The HWML contains the second largest collection of parasites from animals in the western hemisphere and includes parasites and their data from studies of systematics and taxonomy, specimens obtained from biodiversity surveys, studies of ecological characteristics of various animal and plant groups, and specimens that document laboratory and ecological experiments in host-parasite biology. Specimens housed in natural history collections are like goldmines of information in that they represent species that exist or once existed on the earth thus providing documentation of a range of complex biological associations with the full range of life histories represented in collections across the earth. The HWML collections span the range of biology, with specimens in the collection documenting both current and past biodiversity and especially the intimate relationships ranging from free-living symbioses to parasitism, which is one of the most dramatic modes of interactions between organisms.

Fundamental to all studies of biodiversity, systematics, and taxonomy are the biological collections that are maintained in museums worldwide. These collections are becoming increasingly important by enabling researchers to use comparative methods to recognize newly evolved or emerging parasitic diseases of humans, domestic animals, and food resources. These collections provide snapshots in time of sylvatic and even domestic biodiversity that are internationally a major concern in this time of dynamic environmental change and anthropogenically mediated environmental degradation. Parasites are now viewed as significant components of biodiversity that must be included in plans for survey and inventory, conservation plans, and other national needs focused on understanding environmental integrity and ecosystem function (Brooks and Hoberg 2000; Hoberg 2002)

One of the problems with identification (ID) of parasites and use of the data that reside in various biological collections worldwide is that it is very difficult for non-experts to identify parasites that may be discovered in a specimen of a host. We view the large collection of preserved parasites in museums worldwide as an untapped resource that can be utilized to enable the development of new or more rapid methods of identification of parasites using both (or either) non-invasive, non-harmful ID approaches, or ID methods dependent on DNA extraction, amplification, sequencing and matching homologous sequences (DNA Barcodes). In the present paper, and using the first approach of non-invasive, non-harm to the specimen, we provide a partial solution to the ID problem

using a model population of parasites from the HWML collections. The development of a content-based image retrieval system would be a boon to both advanced and beginning researchers and may be robust, if programmed to account for natural variation in natural morphological structures in addition to taking into account variation caused by collecting and storing the specimens.

A web-browser operated, content-based image retrieval system will provide another method of access to information available in the large collection of archived biological specimens stored for research use not only in the HWML, but the system will be expandable to many other biological image databases. When the system is fully implemented, it will enable researchers with interests in parasites to better understand global parasite biodiversity by increasing the speed at which species can be identified and cataloged. The lack of a robust method of sharing specimen-based data on georeferenced parasites (or other organisms) and their surrogate images over the internet is one of the most frequently cited limitations for sharing parasite specimen information with various institutions that have poor national mail or courier access or are politically insecure or unstable. Physically sending specimens to laboratories located in these areas puts irreplaceable specimens in danger of loss.

In the present paper, we describe a prototype of a CBIR system that can be used to access a subset of the HWML collections. Our database consists of images of various parasite specimens that belong to species of 25 different genera in the trematode family Opecoelidae Ozaki 1925 (Phylum Platyhelminthes: Class Trematoda). We describe the design of the system and the querying methods used in the implementation. This will serve as a model to eventually bring the whole HWML collection online for complete and open access to the specimens and data stored within, and as mentioned, this also can be used as a model for the establishment of CBIR systems in other biological collections such as nematology, entomology, ichthyology, botany, mammalogy, and others.

The design of such a system should be sufficiently flexible to allow scaling up or down to a variety of taxonomic levels; from families with few representative genera and species up to superfamilies and genera with perhaps millions of images. CBIR systems have the potential for furthering specimen-based research and education by improving the precision and speed in identifying specimens using morphology. As the system evolves we will plug in a wide range of data sources and image processing algorithms to support and enhance specimen-based research.

The system was established and test images were generated using specimens from the parasite collection of the Harold W. Manter Laboratory of Parasitology at the University of Nebraska-Lincoln. We were interested in testing such a system because lack of access by researchers to biological specimens and their associated data precludes exploration of questions aimed at trends in biodiversity, pathogen spread, or patterns of co-occurrence of species over geographic space and through time (Peterson et al. 1999; Peterson et al. 2002; Peterson et al. 2004; Peterson 2006; Haverkost et al., submitted). For these reasons, addition of new collections with verified geographic information to databases is deemed a critical priority that will enable research in predictions of changes in diversity or extinction.

Online biological databases enable researchers to access and analyze data in ways never before possible, producing new synthetic results and insights into causes and uses of biological diversity, rates of speciation, and general ecological and evolutionary processes (Peterson et al. 1999; Haverkost et al., submitted). Some excellent examples of facilitating networks of online databases include MaNIS (Wieczorek 2007) and PARASITE (Gardner 2007), which are distributed internet-based programs/databases that serve to link museum collection database sources. The establishment of such internet-based resources is essential for the continued growth of biodiversity research.

DISCUSSION

Efficient design of image databases is a critical aspect in a CBIR system because each image is usually large and a typical system has many images. In addition, each image in a biological collection also contains a significant amount of metadata. Metadata are embedded data associated with images that may provide identifiers for the images on origin, host species, geographic collection locality, image creator, resolution, film type, camera type, time/date stamp, and other information. The database must be organized in a way that facilitates efficient retrieval and accurate search results. Hence traditional database techniques that are focused on tabular data for fast indexing, searching, and retrieval are not suitable for image video databases. For example, it is not possible to use relational (less than or equal to) and mathematical operators (addition or subtraction) when searching content based images. In addition, multimedia data are embedded in multi-dimensional or even non-dimensional formats; a domain that does not directly allow sorting. To circumvent some of the shortcomings in image-based searching and content retrieval, we have combined two approaches for searching through images in an image database: 1) textual or metadata description as the basis for searching, and 2) using features of the specimen for searching the image databases (Berman and Shapiro 1998). Commercial CBIR systems such as QBIC (Query by Image Content) by IBM and VIR developed by Virage Inc. allow retrieval based on color, texture, composition, shape, and structure.

Query Methods

Image querying is a mechanism designed to retrieve desired images from an image database based on a combination of color, texture, or shape features, image types, and metadata associated with an image such as creator, time, and location of image creation. Image queries can be broadly classified into two categories: (a) text-based image retrieval, and (b) feature-based image retrieval.

Text-based Image Retrieval.—Text-based image retrieval systems normally store representations of pictorial documents (such as photographs, prints, paintings, drawings, illustrations, slides, video clips, etc.) in static archival databases, and incorporate multimedia database management systems to store and provide wider access to them (Eakins and Graham 1999). However, these systems typically do not provide CBIR facilities and are based for their functionality on text keywords that have to be added by human indexers to enable retrieval of stored images. Image systems such as iBase, Index+, Digital Catalogue, Fastfoto, FotoWare, Signpost, Cumulus, and Picasa allow storing, retrieving, browsing, viewing, and searching an image collection using numbers, dates, categories, sub-categories, key words, or free text.

Feature-based Image Retrieval.—Image features refer to the characteristics that describe the contents of an image. These include visual features that can be directly extracted from the image. Visual features are typically categorized as color, texture, shape, and structure (Umbaugh 2006). A typical CBIR system (Fig. 1) allows users to formulate queries by submitting an example of the type of image being sought, though some offer alternatives such as selection of color from a palette or an input sketch. The system then identifies stored images whose feature values match those of the query most closely and displays thumbnails of these images on the screen. Some of the more commonly used features for image retrieval are described here.

Color: Color represents the chromatic attributes of the image as they are captured. Different ranges of geometric color models based on hue, saturation, and lightness (HSV) or the primary colors (RGB) are used for discriminating between images. However, images retrieved based on color may represent totally different objects portrayed within the images. In addition, color depends on the lighting conditions and, hence is not the ideal basis for retrieval of images taken under different imaging conditions. Common color features include (a) average color, obtained by computing average values of the color bands, (b) color regions, a natural method for adding spatial information by dividing the images into fixed sub-images or zones, (c) color histograms, which represent the frequency of different colors in the image, (d) color sets consisting of the most prominent colors in the image, and (e) color correlograms, or the representation of spatial correlation of color pair changes with distance.

Texture: Texture is a property of distinguishing images based on uniformity, density, coarseness, roughness, regularity, intensity, periodicity, directionality, and randomness of image surfaces. Statistical methods treat each texture as a multidimensional feature vector. The main methods that are commonly used for texture features are (a) pixel neighborhood, where intensity values of current pixel properties are compared with estimated probabilities of its neighboring pixels, (b) co-occurrence matrix, that models the repeated occurrence of certain gray-level configuration in a texture, (c) Tamura texture representation, based on human visual perceptions of coarseness, contrast, directionality, line-likeness, regularity, and roughness, (d) Markov random field, a stochastic surface that is statistically similar to original texture, and (e) various wavelet and Gabor transforms (Shapiro and Stockman 2001; Umbaugh 2006).

Shape: The shape of an object can be expressed through a set of features such as the area, local elements of its boundary, and characteristic points. Two shapes can be compared by matching the computed geometric transformation that can then be used to match one with the other. Shape representations can be categorized into (a) boundary-based, where shape is represented as its boundary in the image, and (b) region-based, where the object is represented by the information on the whole area of the image. Common boundary-based shape features include (a) chain code, where a boundary is traced in either direction and code is assigned to each boundary pixel depending on the direction of the next boundary pixel, (b) Fourier descriptors, and (c) wavelet descriptors that represent local properties of a boundary. Common region-based shape features include (a) heuristic region descriptors such as area, Euler's number, circularity, eccentricity, elongations, rectangularity, and the orientation of major axes, (b) moment invariants, and (c) the principal components or eigenvector representations.

Structure: Structure of an object is defined by the spatial layout of its salient components. Spatial relations such as relative location, adjacency, overlapping, and containment are used for discriminating different images in image databases. Image segmentation is utilized to capture the spatial relationships among objects based on object centroid or minimum bounding rectangles. A symbolic image is a logical representation of the original image where objects are labeled with unique symbol names. Then, the symbolic image is described with a suitable structure representation that should mediate the spatial knowledge embedded in the images as closely as possible. The three most common structural representations are (a) 2D strings, where symbolic image is represented by a 2D string according to its projection along horizontal and vertical axes, (b) different types of trees, e.g. the R-tree and its successive variants that provide a dynamic data structure for multidimensional rectangles, and (c)

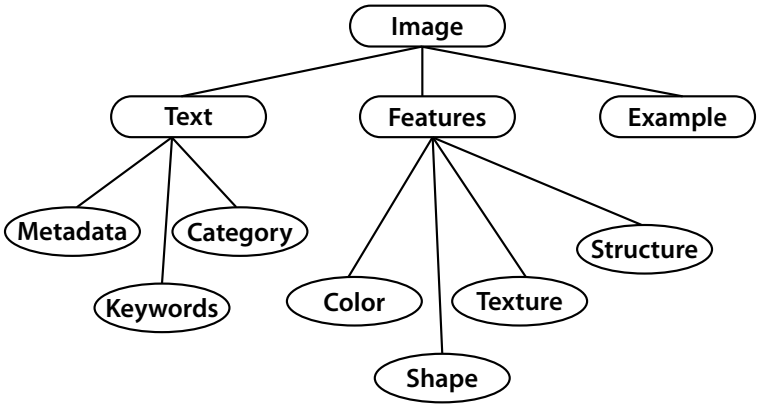

Figure 1. A typical Content Based Image Retrieval (CBIR) system shown in diagrammatic fashion.

graphs, in which spatial relationships among objects in an image are represented as edges of a weighted graph called spatial-orientation graph. Spatial similarity is then determined by the number and extent to which the edges of the spatial-orientation graph of the database image conform to the corresponding edges of the graph of the query image.

Indexing Methods

Text-based retrieval systems typically use indexing methods such as hash tables for faster retrieval. In image retrieval systems, however, visual information must be indexed to improve accuracy, consistency, and speed of searching. Most common indexing methods include indexing of string attributes and indexing by visual attributes. In a string-based scheme, string attributes of images are expressed in the form of keywords, strings, or scripts. Hashing tables and signature files are the most common methods deployed in this type of indexing. A signature is a code that represents a feature that can be inserted into a hash table. These are used as filters for excluding images not relevant for the search criteria. In visual attribute based schemes, attributes (e.g. shape, color, texture, and structure) are modeled in a multidimensional metric feature space known as point access methods (PAM). The performance of PAM depends on the number of features used to represent the properties and the distance measures. Index structures can be static or dynamic. Static structures do not allow addition or removal of items, whereas dynamic structures allow changes in the database without the need to rebuild the complete index. In visual information retrieval, weights are used to indicate the relative relevance of visual features. These weights are often adjusted at run time. Index structures should therefore incorporate similarity computation so that clusters of similar images can be created dynamically (Geradts 2002). Other methods for indexing are based on k-d-trees, R-trees, and SS-trees. SS-trees are dynamic data structures for feature indexing specifically developed to permit similarity indexing in image databases.

Existing Image Retrieval Systems

In this section, we briefly describe some generic CBIR systems. These systems are implemented in diverse environments and require the retrieval system to be locally installed. Some of these CBIR systems are developed for commercial use and some as part of research and educational activities. All these systems operate on predetermined image databases.

Query by Image Content (QBIC).—QBIC was the first commercial application developed by IBM for content-based image retrieval (Flickner et al. 1995). QBIC supports the use of user-provided pictures or user-drawn sketches as the example images. The query is based on one reference image and one feature at a time. Available features include average color, color histogram, color layout, texture, and shape. The visual queries can also be combined with textual keyword predicates.

Virage.—The Virage Image Engine is a commercial CBIR developed at Virage Technologies that supports queries based on color, texture, shape, and structure of the image (Gupta and Jain 1997). These features can be combined in a query, and the user can adjust the weights associated with each feature. Virage is intended as a portable framework for different CBIR applications, and the architecture of the system is designed to support "plug-in" modules for specific needs.

Photobook and FourEyes.—Photobook is an interactive CBIR application developed at MIT (Pentland et al. 1996). It is divided into three separate image descriptions, namely Appearance Photobook (face recognition), Texture Photobook, and Shape Photobook, which can also be combined. This application supports a variety of matching algorithms: Euclidean, Mahalanobis, divergence, vector space angle, histogram, Fourier peak, and wavelet tree distances, as well as any linear combination of these algorithms. The latest version of Photobook also supports matching algorithms via dynamic code loading. Photobook includes FourEyes, an interactive tool for image segmentation and annotation. The user selects some image regions and gives them labels, and FourEyes extrapolates the labels to other regions on the image and in the database.

MARS.—Multimedia Analysis and Retrieval System (MARS) is an interdisciplinary research effort involving multiple research communities at the University of Illinois (Ortega et al. 1999). The main focus of MARS is to develop methods to organize various features into an adaptive retrieval architecture, instead

of finding the "best" representations for any particular image request.

VisualSEEK.—This is a content-based image and video query system developed at the Image and Advanced Television Lab of Columbia University (Smith and Chang 1996). It integrates feature-based image indexing by color with region-based spatial query methods. This enables queries with multiple color regions in the sketch image. Queries may be conducted by sketching a layout of color regions, by providing the URL of a seed image, or by using instances of prior matches.

NETRA.—NETRA is a prototype image retrieval system that is currently being developed at the University of California, Santa Barbara (Ma and Manjunath 1999). NETRA uses color, texture, shape, and spatial location information of segmented image regions to search and retrieve similar regions from the database.

Performance Evaluation of CBIR Systems

Evaluation of retrieval performance is an important aspect in content-based image retrieval. Many of the performance measures are adapted from generic information retrieval systems. The most common evaluation measures used in information retrieval are precision and recall, and are defined as follows:

$$\text{Precision} = \frac{\text{Number of Relevant Documents Retrieved}}{\text{Total Number of Documents Retrieved}}$$

$$\text{Recall} = \frac{\text{Number of Relevant Documents Retrieved}}{\text{Total Number of Relevant Documents in the Collecton}}$$

Other measures used to measure the efficiency include the error rate and retrieval efficiency, defined below:

$$\text{Error Rate} = \frac{\text{Number of Non - Relevant Documents Retrieved}}{\text{Total Number of Documents Retrieved}}$$

$$\text{Retrieval Efficiency} = \frac{\text{Number of Relevant Images Retrieved}}{\text{Total Number of Images Retrieved}}$$

DESIGN OF A CBIR SYSTEM USING THE HWML PARASITOLOGY COLLECTION

The design of our CBIR system integrates the features of both a text-based search engine and image-based search. Our system consists of three main modules: (a) text-based retrieval engine, (b) image-based retrieval engine, and (c) specimen administration and management. Our primary goal was to develop a framework that supports identifying biological specimens with greater ease, precision, and speed. Keeping these in mind, we use a web-based framework in which it will be easy to use a "plug-and-play" approach to incorporating sophisticated image processing techniques, advanced search mechanisms, and accurate evaluation schemes.

Design Overview

As described above, we have divided the system into three central components, each handling a specific task. This modular design accounts for, and allows the incorporation of, new advancements in each module without disturbing the whole system. Each module has well defined user and implementation interfaces. The text-based retrieval module provides an interface in which a user can specify a lookup of a specimen, or series of specimens, using associated image and specimen metadata. Each specimen in the HWML Parasite Collection has a corresponding text-based metadata entry in the database PARASITE (Gardner 2007a). These associated data include the systematic position (taxonomic categories) of the species, specimen, and image of any specimen, geographic location, collector or contributor, host or ecological data, and some image description. The user (searcher) can search on these metadata features individually or in any combination. The image-based retrieval engine currently supports two forms of query, including: 1) query by example, and 2) query by sketch. The administration module provides the interfaces for updating and tracking information from specimens, maintaining images and metadata information associated with specimens, and

monitoring and tracking specimen loans and requests for data. The three components are described in detail in the next three sections (see Gardner 2007, Gardner 2007a). Figure 2 gives a schematic of the complete system.

When a user specifies the query by describing the features in the image (text-based search), we simply fetch the images that match the described features. The image database contains the images as well as their metadata. We also developed a feature database which stores the description of the images in terms of their image features, e.g. color, texture, etc. This is done using a training stage in which all images are analyzed and have their features computed. If the user wants to use a query-by-image approach, they can specify the image either by providing an example or by drawing a sketch of the specimen to which a match is desired. The image is then processed to compute its features. The query features are then matched to all the images in the database to obtain the images whose features match most closely with the query.

Text-based Image Retrieval

The text-based image retrieval module supports five basic functionalities to allow the user to search the image databases based on the metadata associated with each specimen. Text-based image retrieval is intended for subject matter experts who can easily link between image content and the metadata descriptors without the need of further data diagnosis.

Search by taxonomy.—This feature allows a user to search the image database based on family name, subfamily name, genus, parasite type, its scientific name, subspecies, host scientific name, host taxon, host age, host gender, host subspecies, host location, and infection type.

Search by geographic location.—This feature allows a user to search the image database based on country/region from which the specimen was collected, state/province, oceanic fishing region (where applicable), geographical location, and latitude and longitude coordinates.

Search by collector or contributor.—This feature allows a user to search the image database based on the information about the person who collected the specimen or contributed it to the HWML parasite collections.

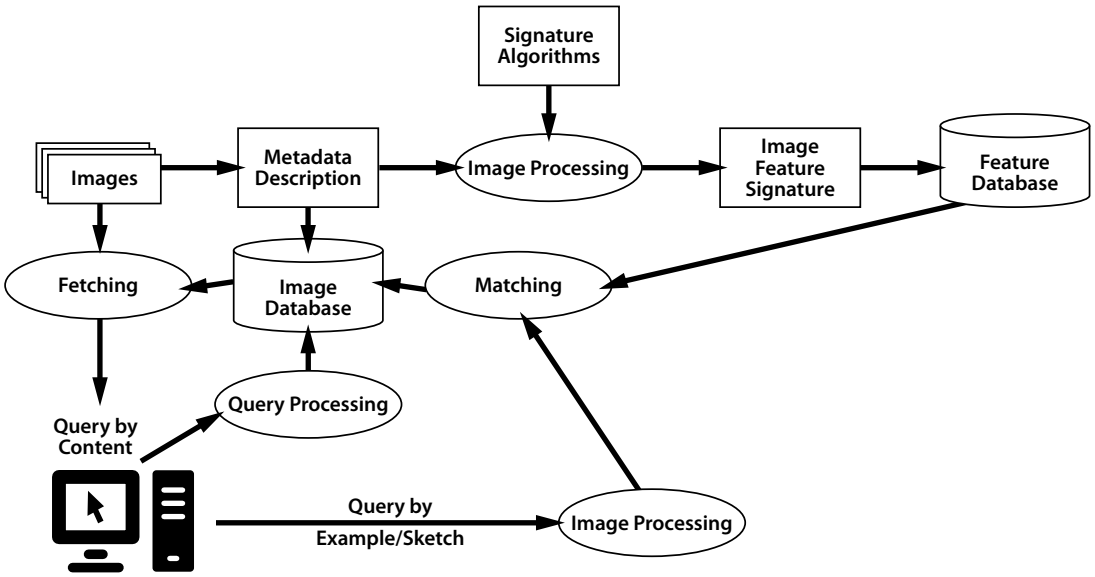

Figure 2. Schematic diagram of the complete CBIR system as developed in the Harold W. Manter Laboratory of Parasitology, University of Nebraska-Lincoln.

Search by image.—This feature allows a user to search the image database based on the information particular to image resolution, image collection number, and storage format.

Meta search.—This is an advanced search option which allows a user to search the image database based on either one or any combination of above image metadata descriptors.

Each of the functionalities in turn provides three different search options to the end user, as follows:

Quick Search: This is a fuzzy search based on keyword. The keyword does not have to exactly match for the image to be retrieved.

Basic Search: This is the central image search feature that is based on key features in each category with options to multi-select/de-select metadata descriptors.

Advanced Search: This option is a combination of quick and basic search options where users can input any keyword or multi-select/de-select metadata descriptors. This search is recommended for advanced users and may require some sophistication on the part of the searcher.

See Figure 3 for a summary of main modules and steps in the text-based image retrieval module.

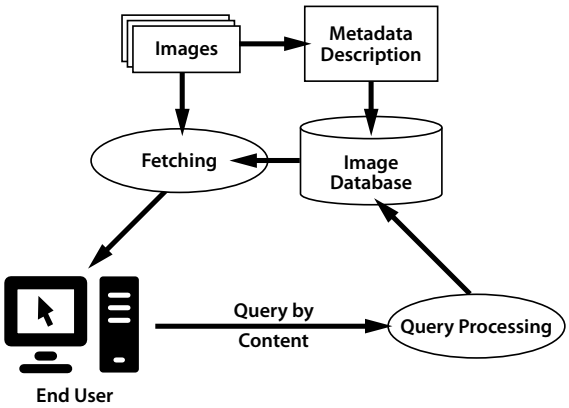

Figure 3. Text-based retrieval module. Summary of steps and processes involved.

Image-based Retrieval

The image-based retrieval module consists of two interfaces for allowing a user to search the image databases using either a sample specimen image or by sketching the shape of the specimen.

Query by example.—This mechanism allows a user to submit an image to begin the search procedure. The goal is to identify all images in the database that are similar to this query image. The query image may be chosen from the image database itself or from a locally stored specimen image collection. For the images stored in the HWML image database, shape information is already available without the need for further segmentation and feature extraction. If the query image is not in the image database, then it is first processed to derive its features before the matching is initiated.

Query by sketch.—This method lets a user draw a sketch of the shape of the specimen and uses it to retrieve images of specimen that match it.

Success in answering queries at this level requires some sophistication on the part of the user. Complex reasoning and often subjective judgment may be required to make the link between image content, and the abstract concepts it is required to illustrate. In addition, sophisticated image processing algorithms to extract shape, structure, color, and texture image features will enhance the systems usability and precision levels significantly. Steps in this module are shown in Figure 4.

Administration

The administration module supports functionalities to allow system administrators and collection managers to enter, maintain, and update information (including loan requests) about specimens. The main modules are listed below.

User administration.—This module allows a system administrator to add/edit/delete the profile of users who can access the system.

Image administration.—This module allows system administrators or staff to add/edit/delete new or existing specimen information along with images and the metadata.

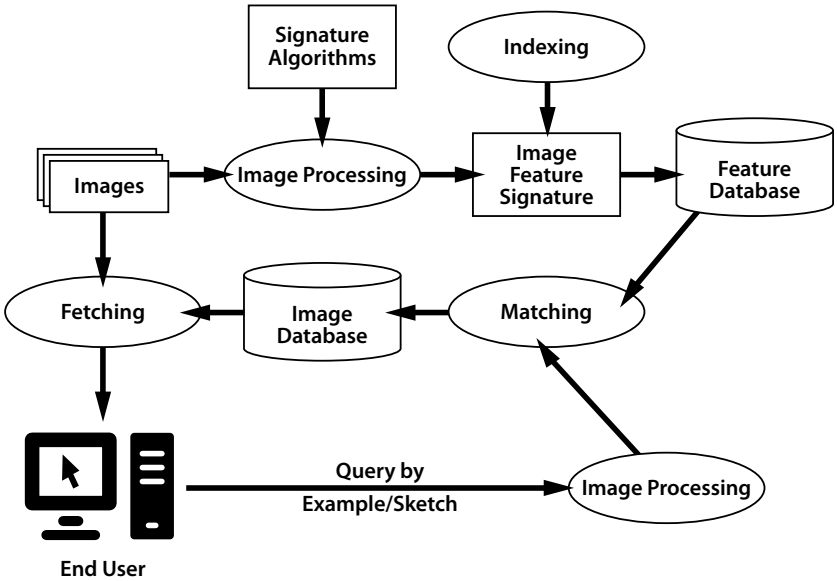

Figure 4. Schematic of image-based retrieval module as implemented in the CBIR system in use in the HWML.

Loan requests.—This module allows system administrators or staff to monitor the loan requests that come through the system from educators and researchers requesting specimens.

Implementation

In this section, we discuss the implementation details of our CBIR system prototype. The image database is stored in a MySQL database. The image-based retrieval module uses a JAVA/JSP/Applet environment that is directly connected to the MySQL database through the Internet. The prototype will serve as the basis for future development and extension to the system.

The main objective for our implementation was modular design and extensibility. The framework developed supports basic elements of both text-based and image-based search techniques. We now briefly discuss the computing environment and other implementation details of the text-based and image-based query processing modules.

To support the modular design objective we chose Java as the programming language to implement the framework. For the database, we chose an open-source implementation MySQL because it is free and readily available. The query environment is implemented to be web-based. We chose Java JSP/Servlets to implement the back-end of the query processing to support flexible query mechanisms, modular extension, and plug-and-play capability.

Image-based query processing has been implemented to support both "query by example" and "query by sketch". Currently, the same back-end is used for both systems but it can be easily changed with our modular design approach.

Figure 5 shows the interface used by the query by example module. After a query image is submitted by the user, the system first performs the feature calculations by using image processing operations. These features are compiled into the image signature. Then, this signature is compared with the signatures of all the images in the specimen database. A list of images whose signatures are the closest are returned to the user for display.

The system also includes a query by sketch, so a user with limited knowledge of the morphological characteristics of a specimen in hand may draw the main features to retrieve a result. To support the query

Figure 5. Image of the interface used in the query by example module of the CBIR system in the HWML (not all results shown).

by sketch feature, we developed an applet that allows the user to draw a sketch of the specimen including its internal structures. The sketch is then used for comparison with the images in the database. Figure 6 shows the sketching module view.

For our prototype program to search for image data, we used the coefficients of the Fast Fourier Transformations (FFT) as the primary features of the image. The FFT coefficients provide a rich summary of the entire image and have been found useful in similar applications. The modular design allows us to add any number of other features to the system as needed. The FFT coefficients of all the images in the database are pre-computed and stored in the database. We use a simple Euclidean measure to compute the distance (inverse of similarity) between two images – the smaller the distance between the signatures of two images, the more similar they are and vice versa.

To evaluate the efficiency of our system, we digitized a set of 610 specimens of trematodes from the family Opecoelidae. The specimens were digitized using a compound microscope and a Pixera™ camera system attached to a PC. Figures 7-8 show sample query images and the top 20 matches (in decreasing order of scores) for each query using the system. As expected, the query image always matches exactly with itself. The other matches usually had a relatively high correlation with the genus or species of trematodes. Many of the matches were different specimens of the same species (of the query image). Precision is often used to determine the efficiency of retrieval. It is defined as:

$$\text{Precision} = \frac{\text{Number of Relevant Images Retrieved}}{\text{Total Number of Images Retrieved}}$$

The average precision of query by example is about 0.25 and that of query by sketch is 0.14.

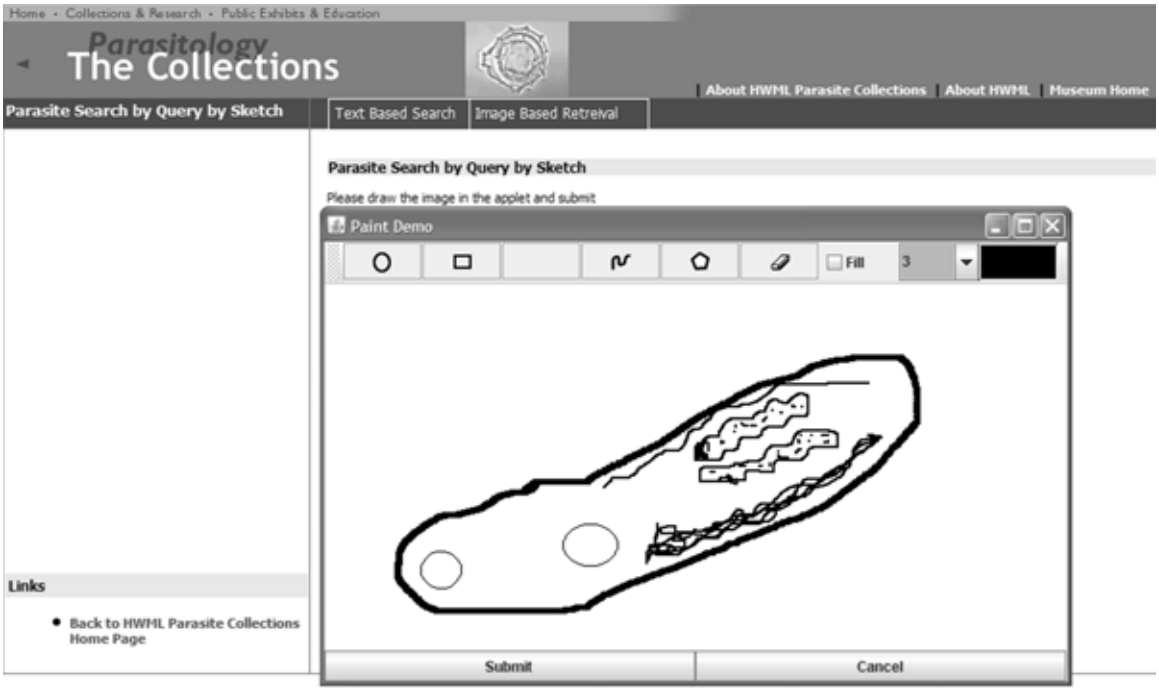

Figure 6. Image of the interface and example of the query by sketch module.

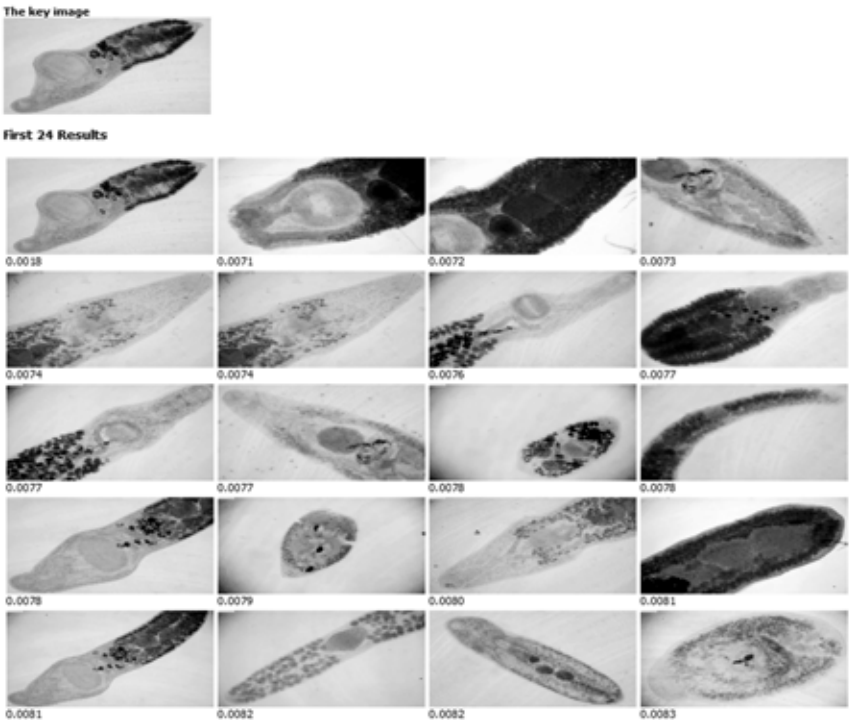

Figure 7. Summary of query images for an example image (top) and resulting matches in decreasing order of scores (not all results shown). See: http://hwml.unl.edu.

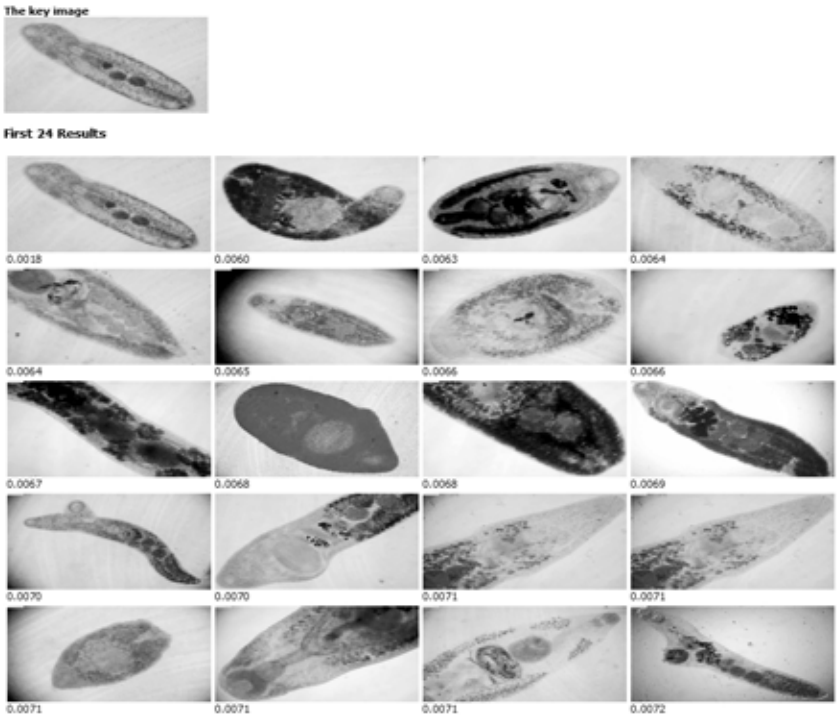

Figure 8. Summary of query images for an example image (top) and resulting matches in decreasing order of scores (not all results shown). See: http://hwml.unl.edu.

CONCLUSIONS AND FUTURE DIRECTIONS

Content-based image retrieval can be a powerful tool that can make a significant impact on the way various kinds of specimen-based research is conducted. Coupled with a web-based interface, a CBIR system will allow any researcher or student anywhere in the world with internet access to query large databases of biological specimens at any time. We have described the motivation, design, and implementation of a prototype CBIR system for the parasite collections in the Harold W. Manter Laboratory of Parasitology. The goal of our CBIR system was to achieve a flexible and extensible framework that would allow quick identification of biological specimens with greater precision and speed for sharing specimen information. The prototype supports access to image databases using both a text-based interface and an image-based query mechanism. Initial evaluation shows that the system is effective and efficient. Future work on a system such as this would include more sophisticated querying mechanisms and advanced image analysis techniques to collect additional descriptive features of specimens being studied.

LITERATURE CITED

Anonymous. 2002. "Content-Based Image retrieval (CBIR) of biomedical Images." A report to the Board of Scientific Counselors, September 26-27. Communications Engineering Branch, Lister Hill National Center for Biomedical Communications, National Library of Medicine.

Berman, A. P., and L. G. Shapiro. 1998. A flexible image database system for content-based retrieval. 17th International Conference on Pattern Recognition.

Brooks, D. R., and E. P. Hoberg. 2000. Triage for the biosphere: The need and rationale for taxonomic inventories and phylogenetic studies of parasites. Comparative Parasitology 67:1-25.

Eakins, P. J. 2002. Towards intelligent image retrieval. Pattern Recognition 35:3-14.

Eakins, J. P., and M. E. Graham. 1999. Content-based Image Retrieval: A report to the JISC Technology Applications Programme. Institute for Image Data Research, University of Northumbria at Newcastle.

Flickner, M., H. W. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. 1995. Query by image and video content: the QBIC system. IEEE Computer 28:23-32.

Gardner, S. L. 2007. "Parasite: The relational database in the Harold W. Manter Laboratory of Parasitology." (http://hwml.unl.edu). University of Nebraska-Lincoln.

Gardner, S. L. 2007a. An experimental on-line CBIR system in the Manter Laboratory of Parasitology. (http://yamaguti.unl.edu:8080/MSProject1/Welcome.jsp). University of Nebraska-Lincoln.

Geradts, Z. 2002. Content-Based Information Retrieval from Forensic Image Databases. Ph.D. Thesis, University of Utrecht, The Netherlands.

Gupta, A., and R. Jain. 1997. Visual information retrieval. Communications of the ACM 40:71-79.

Haverkost, T., S. L. Gardner, and A. T. Peterson. (Submitted). Shared niches reflect common history in host-parasite assemblages. Journal of Parasitology.

HERBIS. 2006. The Erudite Recorded Botanical Information Synthesizer. (www.herbis.org)

Hoberg, E. P. 2002. Foundations for an integrative parasitology: collections, archives, and biodiversity informatics. Comparative Parasitology 69:124-131.

Ma, W. Y., and B. S. Manjunath. 1999. NeTra: A toolbox for navigating large image databases. Multimedia Systems 7:184-198.

Mallik, J. 2006. A content based image retrieval system for a biological specimen collection. M.S. Thesis, Department of Computer Science and Engineering, University of Nebraska-Lincoln.

Mallik, J., A. Samal, and S. L. Gardner. 2007. A Content Based Pattern Analysis System for a Biological Specimen Collection. Seventh IEEE International Conference on Data Mining, Workshop on Mining and Management of Biological Data.

Ortega, M., Y. Rui, K. Cbakrabarti, K. Porkaew, S. Mehrotra, and T. S. Huang. 1998. Supporting ranked Boolean similarity queries in MARS. IEEE Transactions on Knowledge and Data Engineering 10:905-925.

Pentland, A., R. W. Picard, and S. Sclaroff. 1996. Photobook: Content-based manipulation of image databases. International Journal of Computer Vision 18:233-254.

Peterson, A. T. 2006. Ecological niche modeling and spatial patterns of disease transmission. Emerging Infectious Diseases 12:1822-1826.

Peterson, A. T., J. Soberon, and V. Sanchez-Cordero. 1999. Conservation of ecological niches in evolutionary time. Science 285:1265-1267.

Peterson, A. T., V. Sanchez-Cordero, B. Beard, and J. M. Ramsey. 2002. Ecological niche modeling and potential reservoirs for Chagas disease, Mexico. Emerging Infectious Diseases 8:662-667.

Peterson, A. T., J. T. Bauer, and J. N. Mills. 2004. Ecological and geographic distribution of filovirus disease. Emerging Infectious Diseases 10:40-47.

Santini, S., and R. Jain. 1998. Beyond query by example. Proceedings of 6th ACM International Multimedia Conference.

Shapiro, L., and G. Stockman. 2001. Computer Vision. Prentice Hall, New York.

Smith, J. R., and S. F. Chang. 1996. VisualSEEk a fully automated content-based image query system. ACM International Conference on Multimedia, pp. 87-98. Boston, Massachusetts.

Umbaugh, S. 2006. Computer imaging: Digital image analysis and processing. Taylor & Francis, New York.

Wieczorek, J. 2007. MaNIS: Mammal Networked Information System. (http://manis.org) University of California, Berkeley.

Addresses of authors:

RAMALINGAMURTHY MEDURI

Department of Computer Science & Engineering
University of Nebraska-Lincoln
Lincoln, NE 68588-0115
myrlmurthy@yahoo.com

ASHOK SAMAL

Department of Computer Science & Engineering
University of Nebraska-Lincoln
Lincoln, NE 68588-0115
samal@cse.unl.edu

SCOTT L. GARDNER

Harold W. Manter Laboratory of Parasitology
W-529 Nebraska Hall
University of Nebraska State Museum and
School of Biological Sciences
University of Nebraska-Lincoln
Lincoln, NE 68588-0514
slg@unl.edu

Publications of the Museum of Texas Tech University

Institutional subscriptions are available through the Museum of Texas Tech University, attn: NSRL Publications Secretary, Box 43191, Lubbock, TX 79409-3191. Individuals may also purchase separate numbers of Museology directly from the Museum of Texas Tech University.

ISSN 0169-0237

Museum of Texas Tech University, Lubbock, TX 79409-3191