A Content Based Pattern Analysis System for a Biological Specimen Collection

Joyita Mallik[*], *Ashok Samal*[†], *and Scott L. Gardner*[‡]
University Of Nebraska-Lincoln, NE
[*]*jmallik@cse.unl.edu*, [†]*samal@cse.unl.edu*, [‡]*slg@unl.edu*

Abstract

Over the years many research collections of biological specimen have been developed for research in biological sciences. Number of specimens in some of these collections can be as high as several millions. There is a move to convert these physical specimens into digital images. This research is motivated by the need to develop techniques to mine useful information from these large collections of specimen images. Specific focus of this research is on the collection of parasites in the Harold W. Manter Laboratory (HWML) Parasite Collection, one of the top four parasite collections in the world. These parasites closely resemble in shape and have flexible bodies with rigid extremities. They have only a few specific structural differences. In this paper we present a technique to retrieve specimens based on shape of a given sample. This form of mining based on the shape of the specimen has the potential to discover linkages between specimens not otherwise known.

1. Introduction

Various collections of biological specimen have been developed over the years for research in biological sciences. With the advent of digital imaging technology
and sharply reduced cost of storage media, there has been an attempt to convert such biological specimen collections into digital image databases. Many of these databases contain millions of images. Thus, there is a need to develop automated tools to extract useful information from these databases based on the semantic content of the images.

In this paper, we describe a system for mining information from a large collection of parasite specimen based on the specimen shape. The specimens in an image are characterized by their elongated shapes with a flexible body but rigid extremities. We label these specimens to be Flexible Body with Rigid Extremities (FleBoRE) objects. In this paper, we describe a model for FleBoRE objects, mechanisms to extract them from specimen images and a framework to retrieve objects of similar shape from a specimen image database.

The shape of parasite specimens is of great interest to researchers. Search for a specimen by shape or structure will result in images that may point to unknown linkages. This will help in mining underlying correlations between specimens that have not been discovered before. To the best of our knowledge, there is no shape based search system that is focused on databases of parasite specimens characterized by semi-flexible objects as described here.

2. Related Work

Most of the image data mining applications in literature are very different from our image dataset. [1, 2, 3, 4]. The special attributes of those images and the manner in which they are used are very different from parasite specimen images.

For example, a system for mining data in forensic image databases is discussed in [1]. The challenge in this application is to combine information from various sub-databases such as database of images collected from the crime scene, database of images collected from the suspect, etc. This application makes use of different features such as color, texture, shape, structure and motion alone or in combination.

Another example discusses data mining techniques for digital mammography [2]. In this system, some existing features (type of tissue, position of breast) are combined with extracted features such as statistical parameters like mean, variance, skewness and kurtosis, computed over smaller window of the original image. The apriori algorithm was then applied to mine association rules.

2.1. Similarity Computation

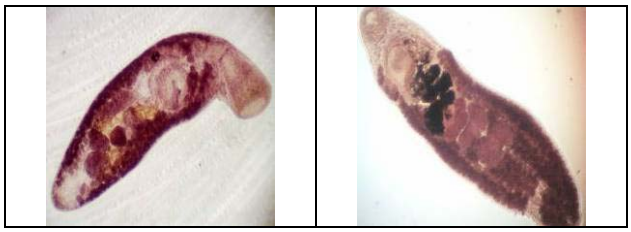
One of the critical components of an image retrieval

0-7695-3019-2/07 $25.00 © 2007 IEEE
DOI 10.1109/ICDMW.2007.69

system is the function that computes the similarity between the query specification (image, shape, sketch, or structure) and the corresponding representation of the images in the database. This similarity function is computed *n* times, where *n* is the number of images in the database. The calculation of similarity is often based on an *image distance measure,* which depends on the features that are used to calculate it. Color, texture, and shape features are common in commercial and experimental image retrieval systems [5, 6].

3. A Shape Model for the Parasites

The specimens in our database are images of parasites that are preserved, stained, and mounted on slides. Depending on how soon and how well after acquiring the specimen it is preserved, its body may shrink, elongate, or bend. In addition, the specimens of parasites in this study have no rigid body parts. This gives rise to the flexibility of the body for different stored and preserved specimens. The parasite specimens in our database are characterized by elongated structures that have well formed shapes at the two ends (anterior and posterior). However the middle part of the specimen is fairly flexible. Figure 1 illustrates this variability of the specimen. Thus, it is logical to organize the shape of these objects as three parts: rigid anterior or head, rigid posterior or tail and a flexible middle part or the body. We describe this model in detail in the next section.



(a) (b)

Figure 1. Examples illustrating curling (a) and stretching (b)

The FleBoRE Model: The shape of a parasite specimen object can be represented by three parts: (a) head, (b) tail and (c) body or the trunk. The head and tail of the specimen can be modeled by singly truncated ellipse segments while the body can be viewed as a doubly truncated ellipse. Figure 2 shows the shape model for a typical specimen.

Formally, we define a specimen, S, as:

$S \equiv <S_h, S_b, S_t>$

where S_h, S_b, and S_t are the head, body and the tail of the specimen, respectively. The head of the specimen, S_h, is represented as a singly truncated ellipse as follows:

$$S_h = < a_h, b_h, c_h>$$

where a_h, and b_h are the lengths of the semi-major axis and semi-minor axis, respectively, of the head ellipse and c_h is the distance between the center of the ellipse and the truncated end of the head (See Figure 2).

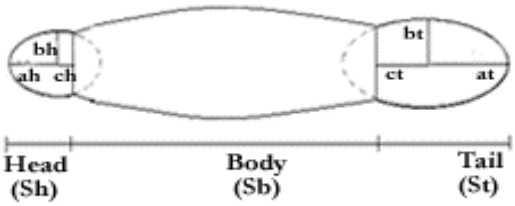


Figure 2. Schematic of the FleBoRE shape model.

Similarly, the tail part of the specimen, S_t, is represented as a singly truncated ellipse as follows:

$$S_t = < a_t, b_t, c_t>$$

where a_t and b_t are the lengths of the semi-major axis, semi-minor axis of the tail ellipse, respectively and c_t is the distance between the center of the ellipse and the truncated end of the tail (See Figure 2).

The body of the specimen can be flexible and therefore is not represented by any parameters in the shape model. Using this FleBoRE model any specimen can be represented as a set of 6 parameters:

$$S \equiv< a_h, b_h, c_h, a_t, b_t, c_t>$$

Since we are interested in finding underlying patterns between objects with similar shapes, we use this model to represent the shape of a parasite specimen and try to find similar shaped specimens by querying based on this model.

Query by Shape Problem: The problem of query by shape can be defined as: Given a database of FleBoRE objects representing a specimen collection, C, a query image of a specimen, S_q, a similarity function, ψ, and a threshold parameter, s_{min}, find a set of images, $\{S_i\} \subseteq C$ such that $\psi(S_i, S_q) > s_{min}$.

Approach: We first use the feature extraction process to derive the shape parameters for the FleBoRE objects. This step is performed for all the images in the database as well as for the query image. We then use a similarity function to compute the

degree of likeness between two FleBoRE objects. If the degree of similarity is above a pre-determined threshold, the corresponding images in the database are retrieved.

We assume that there is only specimen in an image that is wholly contained. Furthermore, the color of the specimen is ignored since it is artificially induced during the staining process.

4. Extraction of Shape Parameters

The images are first converted to gray scale. The specimen is then isolated in the image and the other parts of the image are ignored during the rest of the process. The boundary of the specimen is determined and is used to compute shape parameters. The steps are explained in details below.

Object Separation: The method of object separation is tuned for the types of images (specimen samples) that we have and is not a generic technique. We first use the Canny edge detector as the first step in obtaining boundary of the specimen. The internal structure of the specimen results in significant intensity changes and the edge image produces a very dense image for the interior of the specimen. To handle these problems we use morphological operations.

After that, the single largest connected component in the image corresponds to the specimen. Figure 3(d) shows the image after this step. However, a few small components remain connected in some cases. These "add on"s are easily removed by doing an opening operation. This results in an isolated specimen whose boundary is sharp and accurate.

The boundary pixels of the specimen then are easily determined by a simple binary edge detector which examines the 3×3 neighborhood of each pixel. The final perimeter of the sample image is shown in Figure 3(f). This boundary serves as the starting point for computing the parameters for the FleBoRE model.

FleBoRE Model Matching: Once the boundary of the specimen has been obtained, we match it with the FleBoRE shape model and determine its parameters. Each specimen is then represented by a set of these parameters and stored in a database for matching during the retrieval stage. Model matching is also done for the query image at retrieval time.

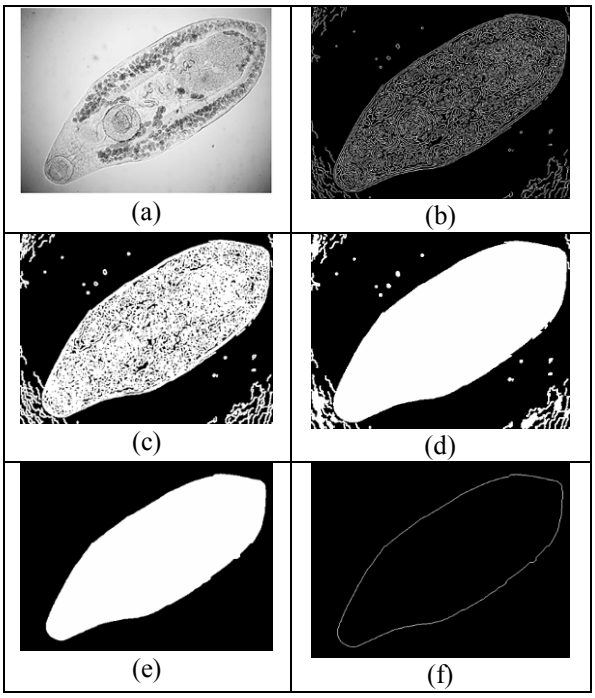


Figure 3. Steps in object separation.

5. Similarity Computation

In our system, the similarity function determines the degree of match between two FleBoRE objects, one from the image database and the other from the query image. We define two similarity functions in this paper. Assume that the two specimens are represented by their parameters as follows:

$$S_q \equiv < a_{qh}, b_{qh}, c_{qh}, a_{qt}, b_{qt}, c_{qt} >$$
$$S_d \equiv < a_{dh}, b_{dh}, c_{dh}, a_{dt}, b_{dt}, c_{dt} >$$

where a_{qh}, b_{qh}, and c_{qh} are the lengths of semi-major, semi-minor axes and the truncation distance for the head part of the query specimen, a_{qt}, b_{qt}, and c_{qt} are the corresponding lengths for the tail part of the query specimen. The parameters a_{dh}, b_{dh}, c_{dh}, a_{dt}, b_{dt}, and c_{dt} are the corresponding parameters for the database image.

Shape Area Similarity: This method of computing similarity calculates the percentage of overlapping boundary between the shapes of two specimens. In order to find out the overlap, we translate the centroids of the two shapes to the origin. The shapes are then rotated so that they are aligned along the same direction. Since the multiple specimens of the same species are not identical in shape or size, we allow some tolerance in the boundary by placing a buffer around it (See Figure 4). We examine the degree of overlap between the query shape and the buffered

database shape. We then reverse the roles and examine the overlap between the database shape and buffered query shape. The similarity between the two shapes is the average of the two measures. Formally,

$$\psi(S_q, S_d) = \frac{Overlap(S_q, S_d) + Overlap(S_d, S_q)}{2}$$

where $Overlap(S_q, S_d)$ is the percentage of the S_q within the buffered shape of S_d.

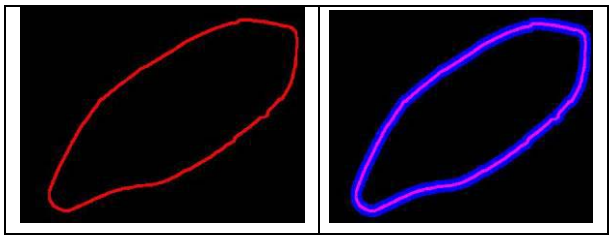


Figure 4. A specimen shape (left) and the buffer around it (right)

Parameter Distance Similarity: In this method, we treat the parameters of each FleBoRE object as a feature vector and compute the distance between them. Since the images of the specimens were taken at different magnification levels during the collection process, we first normalize the ellipse parameters by dividing them by the semi-major axis length. Then one can choose any distance function described in literature [7] e.g. Euclidean distance or Minkowski distance.

Let the ellipse parameters for the head of the query image S_q be $H_q = < a_{hq}, b_{hq}, c_{hq} >$ and that of the database image S_d be $H_d = < a_{hd}, b_{hd}, c_{hd} >$. The normalized heads are then given by:

$$NH_q = \left\langle \frac{a_{hq}}{a_{hq}}, \frac{b_{hq}}{a_{hq}}, \frac{c_{hq}}{a_{hq}} \right\rangle = \left\langle 1, \frac{b_{hq}}{a_{hq}}, \frac{c_{hq}}{a_{hq}} \right\rangle = \left\langle 1, b'_{hq}, c'_{hq} \right\rangle$$

Similarly, the normalized tail is given by:

$$NT_q = \left\langle \frac{a_{tq}}{a_{tq}}, \frac{b_{tq}}{a_{tq}}, \frac{c_{tq}}{a_{tq}} \right\rangle = \left\langle 1, \frac{b_{tq}}{a_{tq}}, \frac{c_{tq}}{a_{tq}} \right\rangle = \left\langle 1, b'_{tq}, c'_{tq} \right\rangle$$

The normalized head and tail of the database specimen are given by:

$$NH_d = \left\langle 1, b'_{hd}, c'_{hd} \right\rangle$$

$$NT_d = \left\langle 1, b'_{td}, c'_{td} \right\rangle .$$

We compute the similarity of the two heads using a Euclidean distance approach as follows.

$$\psi_{head}(S_q, S_d) = \sqrt{(b'_{hd} - b'_{hq})^2 + (c'_{hd} - c'_{hq})^2} .$$

We can compute the similarity between the two tails similarly.

$$\psi_{tail}(S_q, S_d) = \sqrt{(b'_{td} - b'_{tq})^2 + (c'_{td} - c'_{tq})^2} .$$

The overall similarity of the two specimens is a weighted average of the similarity between the head and the tail.

$$\psi(S_q, S_d) = \alpha \times \psi_{head}(S_q, S_d) + \beta \times \psi_{tail}(S_q, S_d)$$

.

where α and β are determined experimentally.

6. Mining Based On Internal Structures

After finding specimens with similar shapes, the next step would be to mine information about its internal organs. Some of the unique defining internal structures of a parasite specimen are shown in Figure 5.

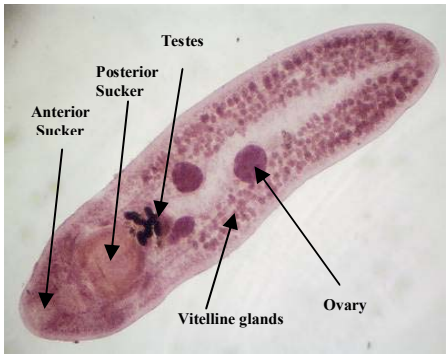


Figure 5. Example of internal structures of parasite specimen.

We briefly present methods to find the location of few such internal structures inside the parasite specimen image.

Vitelline gland detection: The vitelline glands appear as two bands on either side of the specimen and close to the specimen boundary. The distribution of the glands may or may not be continuous towards the posterior end of the specimen. The thickness of the band formed by the glands, their continuity at the posterior end and their length towards the anterior part of the specimen characterize the specimen.

We first obtain the profile of the pixels from each point in the boundary of the image in the direction perpendicular to the boundary inside the specimen (Figure 6(a)). Since the vitelline glands in the image often appear as dark bands, the intensity profile of the image reflect significant variation of intensity where the glands are present. Figure 6(b) shows an example of detecting the vitelline glands using this approach. Using such a method it is possible to mine similarity patterns of vitelline glands for different specimens.

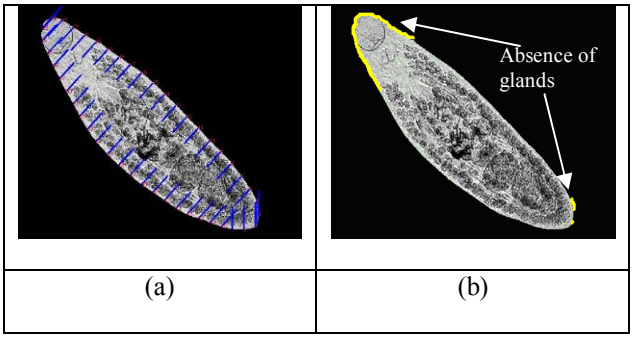


Figure 6. (a) Path for obtaining profile of vitelline glands (b) Detecting the absence of vitelline glands.

Anterior/Posterior Sucker detection: The shape of both the anterior and posterior suckers is circular. Usually the posterior sucker is larger than the anterior sucker. We used Hough transform for finding circles to detect the location of these internal structures. Using the constraints of the size of the specimen, we narrow the range of possible radii for the suckers. Figure 7 shows the result of sucker detection. This method of sucker extraction can help in mining information about the size and relative position of suckers in specimens.

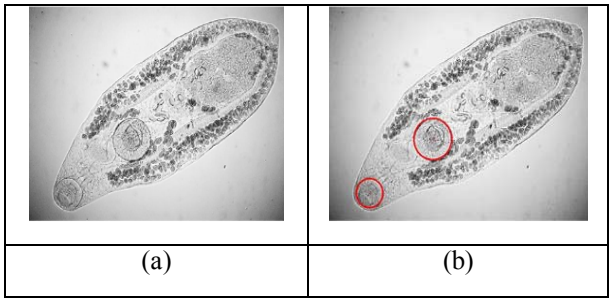


Figure 7. (a) Original image (b) image with detected suckers.

7. Implementation and Results

We implemented a prototype system for a subset of images of the HWML Parasite collection using MATLAB [8].

Specimen Database: As a part of this research we scanned and digitized a small subset of the specimen from the HWML collection. The specimens, typically preserved on slides, are first mounted on 25 mm by 75 mm glass microscope slides. The specimen are captured with a Pixera ™ digital camera at approximately 1.5 mega pixel resolution.

We show the results of sample queries using the two similarity measures described in Section 5. The result of each query gives a similarity score between the input image and every image in the database. We sort the similarity scores in descending order to find the 10 most similar images to the query image.

Retrieval using Shape Area Similarity: Figure 8 shows the results of a sample query. It shows the query image (shown on top) and the retrieved images.

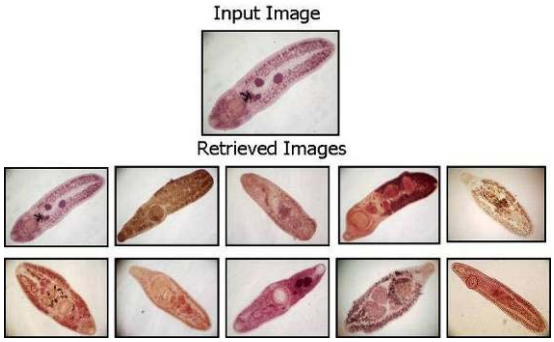


Figure 8. Retrieval using Shape Area Similarity.

From the results we can see that the retrieved specimens have approximately the same shape as that of the query specimen. The retrieved images have a similar aspect ratio as the query image.

Retrieval using Parameter Distance Similarity: In this method of retrieval we compute the distance between the parameters of the FleBoRE objects as an indicator for their similarity.

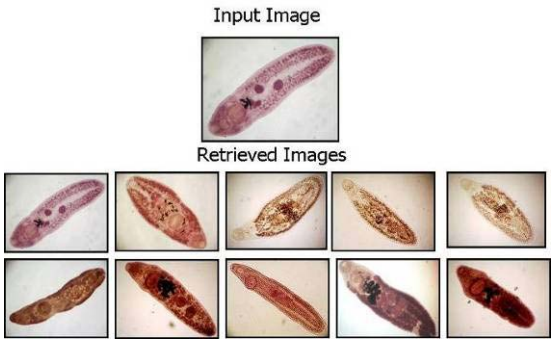


Figure 9. Retrieval using Parameters Based Similarity.

The results in Figure 9 show that we were able to retrieve specimens that are similar in shape to the given specimen.

8. Conclusion

In this research we have developed the framework for a system that can retrieve images similar to a given image for a complex biological specimen collection. The system can potentially help find unknown

relationships between different parasite specimens. The database consists of images of parasites which are characterized by rigid shape at the ends and a flexible body or trunk. We have developed a shape model for these shapes, called FleBoRE objects. We have developed several methods to compute the similarity between two FleBoRE objects. Such measures are important when mining information from image data as it gives a basis for comparison.

We have also developed automated methods to extract the shape and structure of the FleBoRE objects from specimen images. The system has been tested with a collection of parasite images from the Harold Manter Laboratory for Parasitology [9].

Future Directions: The research presented in this paper can be extended in many directions. The extraction of internal features of specimen is an important step in querying the database to find specimens with similar internal structures. Similarity functions need to be developed for comparing structures of two specimens. One can combine the two methods of shape and structure based querying presented here to find if specimens of different families that might have similarity in shape and structure.

9. References

[1] Zeno J. Geradts, Jurrien Bijhold. Data Mining in Forensic Image Databases Proc. SPIE Vol. 4709, p. 92-101, Investigative Image Processing II, 2002.

[2] M. Antonie and O. Zaane and A. Coman, Application of Data Mining Techniques for Medical Image Classification, *Proc. 2nd Int. Workshop Multimedia Data Mining,* 2001.

[3] Ji Zhang, Wynne Hsu, Mong-Li Lee, An Information-driven Framework for Image Mining, *Proc. of 12th International Conference on Database and Expert Systems Applications,* 2001.

[4] Peter Eklund and Jane You and Peter Deer, Mining Remote Sensing Image Data: An Integration of Fuzzy Set Theory and Image Understanding Techniques for Environmental Change Detection, *Proc. Of SPIE 2000: Knowledge and Data Discovery,* 66-67, 2000.

[5] L. G. Shapiro and G. C. Stockman, *Computer Vision,* Prentice Hall, Upper Saddle River, NJ, 2001.

[6] V. N. Gudivada and V. V. Raghavan, "Content based image retrieval systems," *IEEE Computer*, Vol. 28, No. 9, pp. 18-22, September 1995.

[7] S. Umbaugh, Computer Imaging: Digital Image Analysis and Processing, Taylor & Francis, New York, 2005.

[8] The MathWorks - MATLAB and Simulink for Technical Computing, http://www.mathworks.com/. Last accessed on December 1, 2005.

[9] S.L. Gardner. *The parasite collection search page in the Manter Laboratory of Parasitology*. http://manter.unl.edu/hwml/. Last accessed on May 22, 2006.