5-2010

Automated Extraction of Structures from Sketches of Biological Specimens

Jamie J. Schirf
Computer Sciences, jschirf@cse.unl.edu

AUTOMATED EXTRACTION OF STRUCTURES FROM SKETCHES OF BIOLOGICAL SPECIMENS

by

Jamie Joseph Schirf

A THESIS

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfilment of Requirements

For the Degree of Master of Science

Major: Computer Science

Under the Supervision of Ashok Samal and Scott Gardner

Lincoln, Nebraska

May, 2010

AUTOMATED EXTRACTION OF STRUCTURES FROM SKETCHES OF BIOLOGICAL SPECIMENS

Jamie Joseph Schirf, M.S.

University of Nebraska, 2010

Advisers: Ashok Samal and Scott Gardner

The goal of this study was to develop automated techniques to extract biological structures from sketches of biological specimens. This will form the basis for a searchable database of information about the specimens. Having such a database enables researchers to efficiently search for specimens with particular qualities or identify unknown specimens.

After some preprocessing of the images, the important internal organs of the specimen are extracted using image analysis techniques. The shape, size, and organization of the organs are used to categorize and then to reorganize them in the image. Results using a large database of sketches of trematodes, in important class of parasites, show that we can extract the internal organs accurately.

Using the database, it will be possible to build an intelligent computer based image retrieval (CBIR) system around it. As more data are added, this will prove to be of significance to researchers, practitioners, educators and indeed anyone researching such specimens.

ACKNOWLEDGEMENTS

I would like to thank my advisers Ashok Samal and Scott Gardner for their constant patience and guidance. I would also like to thank Berthe Choueiry for serving on my committee. Finally, I would like to thank my family for their steadfast encouragement.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the modern age we have unprecedented access to information. We have vast libraries filled with documents, museums packed with specimens, and archives filled with records [7]. And yet, amazing as having this information is, what we can now do with it is even more so. With the advent of modern technology and its plentiful storage, we are no longer bound to the use of physical media. The importance of hard copy is passing. Nowadays, it is not uncommon for entire books to be published online, never going to the presses. The advantages of this are numerous, from being environmentally sound to economically superior. However, what of the knowledge we've already acquired? Much of our knowledge is still locked in the physical realm, on paper or as preserved slides. Not only is it harder to study in this state, it's also vulnerable to loss. For example, a natural disaster striking a museum could result in the loss of priceless information. When such information is lost, it is lost forever. By converting this information to digital form, such loss becomes far less dramatic.

1.1 Motivation for Digitization

Naturally, the preservation of information might be the most important reason for the digitization of knowledge, but it's only one benefit. Digital information is also easier to study. It can be searched automatically, as with a search engine. Likewise, it can be downloaded to almost anywhere in the world virtually instantaneously. A less obvious, but still crucial benefit, is that digital information can be interpreted by a computer. This capability allows us to make use of Data Mining for processing vast amounts of data, searching for interesting patterns of information. In this way, new information can be discovered in these "old" data.

1.1.1 Preservation of Knowledge

The presentation and preservation of information is the whole purpose of museums and libraries. Modern technology is absolutely essential for this goal.

Many museums are currently looking at ways to convert their collections into digital form [11]. Naturally, not all types of collections lend themselves to convenient digitization, but many do. For instance, biological specimens preserved on microscope slides are a likely candidate, given that what the user is interested in is almost certainly the image of the specimen.

There is also a strong need for the digitization of things kept on paper. These might include textbooks, newspapers, periodicals, historic documents maps, schematics, and other such collections. The blanket term for this type of information analysis is Document Image Analysis, a sub-field of Computer Vision.

Text recognition (or character recognition) is when a text is first scanned as an image. Letters are then extracted from it through sophisticated algorithms. Once this recognition is done, the text can be copied from, searched, or edited like any

other digital document.

Far more complicated than text recognition (at least within Western character sets) is the automated interpretation of information from graphics or photographs. This domain is the main thrust of the field of Computer Vision. One important application it is often used for is recognition of faces or fingerprints, though it is also very useful in interpreting images from a wide variety of domains. However, techniques for Computer Vision tend to be very domain specific. The techniques for automatically reading a piece of sheet music would be very different from those to identify bad parts on an assembly line or to interpret satellite imagery.

1.1.2 Data Mining

The final benefit of digitization comes from the fact that digital material can be read and interpreted by computer programs. This potentially allows utilization of Data Mining to analyze the data.

In the case of many collections, images or records can easily number in the millions or more. In these cases, traditional manual means of searching the data for meaningful information is impossible. In these cases, we turn to the practices of Data Mining to search for interesting patterns we might ordinarily overlook.

Through applying Data Mining to large amounts of data, interesting patterns can be found. Unknown correlations or relationships can be discovered. Hidden trends can be unearthed. This information can be presented for human consumption in the form of graphs, diagrams, or decision trees. There is much potential for finding new information in "old" data.

1.2 Motivation for a CBIR System

Information isn't helpful without some means of indexing it and searching through it, and image based information is no exception.

1.2.1 Searchability

A digital collection can be searched through with computerized speed, perhaps for an entry with a particular character string. Of course, it's possible, with good planning and organization, to do far better than such simple searches. Much in the way that search engines like Yahoo or Google allow users to search for sites, a great search would allow users the freedom to quickly search for specific information as well as find new things of which they weren't aware.

It's also possible to construct such systems for searching through images. Some systems annotate images and allow text-based searches. However, a more elegant system is that of CBIR or Content Based Image Retrieval. This allows us to do a search through images based on the images own properties, rather than artificial labels that have been attached to them. Popular CBIR methods include "Query By Sketch", wherein the user supplies a primitive sketch of their desired image, and "Query By Example", where an actual image similar to those desired is supplied.

1.2.2 Sharable

This sort of system lends itself particularly well to sharable web applications. Consider the case of the Harold W. Manter Laboratory (HWML). As an institution that loans out slides of parasite specimens, digitization is particularly important for HWML. HWML has millions of specimens in its collection. These are mailed all over the world for research. However, there is always risk involved in this of specimens being

lost or damaged [6]. Additionally, physically sending specimens around the world is both expensive and time consuming.

This is clearly an excellent role for a CBIR system. With that, users are able to search online for the images of the specimens they wish to view. This way, there's no expensive postage, no risk, and no wait. It's beneficial to everyone involved.

1.3 Research Contributions

The contributions of this research fall into two areas:

- Algorithms to extract structures from biological specimen sketches
- A scheme for classifying structures for search

As a step towards a CBIR system, this paper demonstrates algorithms for automatically extracting structures from sketches of biological specimens, taken from the Handbook of Trematodes of North America North of Mexico by Stewart C. Schell [9]. The handbook and sketches are meant to allow users to identify unknown specimens. This is a fine role for a CBIR system. For their value for identification, the structures chosen are testes and ovaries. See Figure 1.3.1 for example sketches.

Each of the structures have their own particular methods for extraction. For instance, detection of the testes within a sketch centers around finding white blobs that contain "S" shaped squiggles within them. Detection of ovaries often involves searching for black areas with white patches.

These can then be characterized in a manner appropriate for placing in a database. For example, if the detection process reveals that a sketch has 4 testes, that information can be filed away for later reference. This information would then lend itself to searching.

In the interest of facilitating search, this thesis also designates certain qualitative classes for the internal structures. For instance, instead of recording that a specimen has 25 testes, it would simply record that it has over 20. This makes sense since often in nature the actual number isn't fixed. In the same way, rather than recording firm statistical measures about shape, such as the roundness, the shape of the structures are divided into predetermined classes.

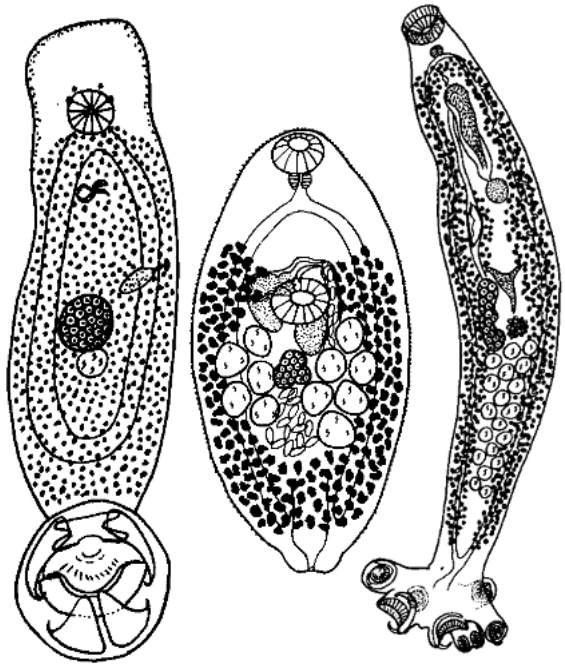


Figure 1.3.1: Example sketches from Schell's Handbook scanned at 300dpi

1.4 Roadmap

Chapter 2 looks at some of the related work with regards to CBIR and document image analysis. Chapter 3 gives a survey of the characteristics of the specimens' structures. Chapter 4 shows how the sketches are extracted and prepared for analysis. Chapters 5 and 6 present the algorithms for detecting the ovaries and testes,

respectively, along with their accuracy. Chapter 7 provides the conclusion and describes avenues for further research.

Chapter 2

Related Work

This thesis makes use of a variety of techniques and theories, particularly from the areas of Document Image Analysis and Computer Based Image Retrieval (CBIR). This chapter gives the reader a brief survey of these areas.

2.1 Document Image Analysis

Document Image Analysis (DIA) is a specialized subset within the field of Computer Vision. Whereas computer vision is concerned with a wide variety of images, the focus of DIA is on processing images of paper documents. This roughly divides into two primary areas: textual processing and graphical processing [10]. Document image analysis is essential for taking hard-copy documents and extracting meaningful information from them without human intervention. Without it, a scanned document would simply remain a static image, unless someone were to perform the tedious task of manually rewriting a digital version of the document.

2.1.1 Textual Processing

Textual processing is concerned with the automated interpretation of the text on the page. Optical character recognition is a central task in text processing. This is the process of taking the graphical image of a character and actually inferring what that character is. From there, whole sentences, paragraphs, and pages can be determined.

A companion operation to this is document lay-out analysis. This is where the components of the document are found [15]. For example, this could be used to determine things like what paragraphs a page has, or what text belongs in an excerpted block, or if something is an embedded graphic. Together with optical character recognition, it enables an understanding of the text of a document.

Depending on how much control was exercised over the scanning process, it may also be important to determine and correct the skew of a page. The skew of a page is defined as "the orientation angle of its lines of text" [17]. This is caused by the page being rotated instead of precisely aligned with the scanner. Depending on the document lay-out analysis techniques used, this may need to be corrected before proceeding.

While interesting, and possibly useful as a companion to this thesis, herein we eschew the textual processing step, and focus on the graphical processing step of DIA.

2.1.2 Graphical Processing

Graphical processing deals more with the interpretation of images within a document. These can be simple elements like lines or curves, or things like diagrams or charts. The majority of the current research into this focuses on processing mechanical drawings, such as schematics. For instance, there's a large demand for moving

old paper-based designs into modern Computer aided drawing (CAD) systems [18].

Processing of maps is also an active area of research. There are still plenty of useful paper maps that could be placed into cartographic databases. Many important subtasks in this field exist such as road extraction, building localization, and symbol identification [19].

In the case of this research, the focus is on sketches of biological specimens. This is a fairly novel area for the use of graphical processing. Almost all of the existing body of research deals with sketches of inorganic constructs with well-defined rules for sketching. These biological sketches do not have such rules or regularity.

From these sketches, meaningful digital information is extracted about their inner structures. Of course, this naturally leads to the problem of needing a system for searching for such information.

2.2 CBIR

The solution to that is CBIR. CBIR is the practice of searching for images based on automatically-derived features, such as shape or texture [3]. For instance, a CBIR system might allow one to query for an image with large round red objects. The key feature is that it is a computer making the interpretation.

This is in direct contrast to systems that rely on human input annotations of images. Unfortunately, this is time consuming, and also falls victim to inconsistent interpretation. Significant studies have been performed demonstrating that different people's interpretation of the content of images can vary widely based on their environment [14].

In many cases, the labeling work can be crowd-sourced, mitigating the time consuming nature of doing it manually. For example, there are many online communities

that allow users to upload images that may then be tagged with annotations. This allows the community to quickly locate desired images. One example of such a site is Safebooru.org [20], which maintains a database of Japanese pop-art.

Research into CBIR currently enjoys a high level of activity in a wide variety of situations. The majority of the current research deals with scientific images, particularly those related to medicine and biology. Given the topic at hand, a brief survey of CBIR's use in biology follows.

2.2.1 For Biological Specimens

CBIR has found a good degree of use for searching through collections of specimen photography, for a wide variety of purposes. For example, at the University of New Orleans, a CBIR system was designed for identifying fish species of the genus Carpiodes [12]. It works by identifying various landmarks on the fish, and using them to compute the similarity of different specimens. With this, a user may use an image of a captured specimen to identify it. This is known as Querying By Example.

Another system was devised by the National Chi-Nan University for finding butterflies [11]. Rather than search with another butterfly photograph, their system provides the user with an applet to sketch an approximation of the butterfly they desire. Then, on the basis of color, shape, and pattern, the most similar images are retrieved. This is known as Querying By Sketch.

Research extends beyond the animal kingdom as well. For instance, research between Ajou University and Korea University has produced a CBIR database for leaves [21]. Their system allows users to sketch out the leaf they are searching for. One particularly interesting feature about it is that it is designed to be accessed wirelessly, opening possibilities for easy queries while out in the field.

Of course, not all systems require the user to provide an example photograph or sketch. A system could also be set up in such a way where the user specifies some qualities they are looking for that have been previously detected in the image database. For instance, one could try to identify a leaf based on metrics like the number of points or the pattern of veins.

Summary

In this chapter we have identified areas of research related to this one. We started with some background on textual and graphical processing. Then we presented a survey of various pieces of CBIR research. In particular, a number of uses of CBIR for biological specimens were highlighted.

Chapter 3

Structural Characteristics

Biological species are characterized by regularity of the physical structures they posses. It is the presence, shape, number, location, and organization of the constituent structures that form the basis of taxonomic classification. In this thesis, the scope is limited to the trematodes found in Schell's Handbook.

For the identification and classification of trematodes, several internal structures are useful. The most important structures include testes, ovaries, vitelline glands, suckers, clamps, and many others. However for this research, work was constrained to testes and ovaries, which play a central role in identifying most species of trematodes.

3.1 Characteristics of Testes

Testes are depicted as white blobs containing dark "S" shaped squiggles. Examples may be seen in Figure 3.1.1. However, these come in different shapes and sizes. Therefore, it is useful to partition the testes into two classes of larger and smaller instances: T1 and T2. It's important to remember that this distinction is made solely to aid the process of identifying them. These classes are mostly, but not

entirely, disjoint. There are cases where the class of a testis is ambiguous.

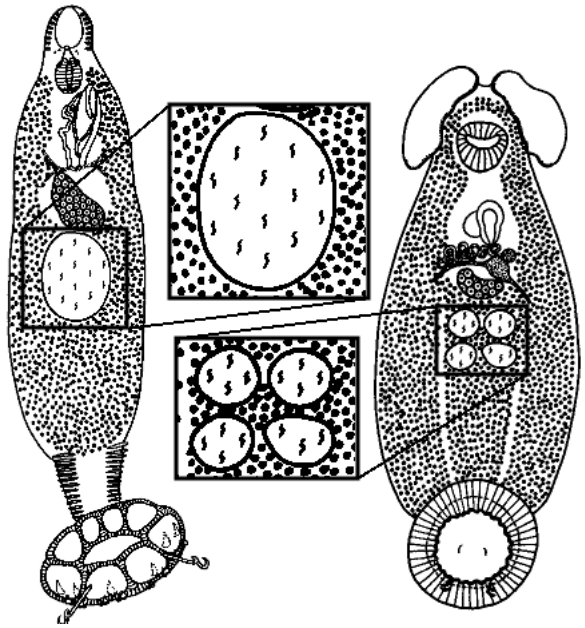
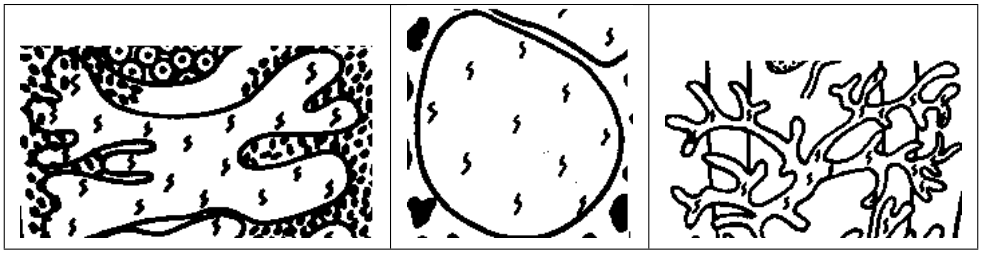


Figure 3.1.1: Sketches with highlighted testes

3.1.1 Larger Testes (T1)

The major deciding factor for whether a testis is of class T1 or class T2 is the size of the testis with respect to the size of the specimen. T1 testes are the larger class, sometimes taking up as much as 18%, or less than 1% of the specimen's surface area. Because the size of the testes correlates with the number of squiggles within, T1 testes almost always have 3 or more squiggles. The shape of T1 testes varies widely. It's most often round, but can also be elongated, tubular, or even branching. These testes often occur in isolation, or in pairs. However, in rare cases, a specimen might have 8 or more. See Figure 3.1.2 for examples of such T1 testes.



a	b	c

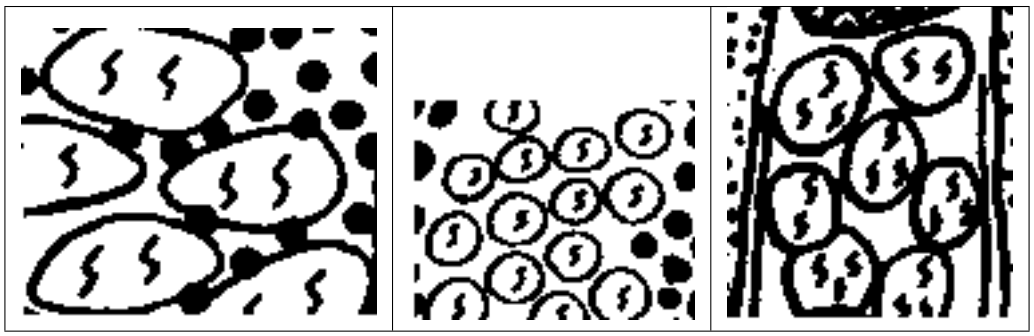
a.) Branching T1 testis with many squiggles.
b.) Round T1 testis with many squiggles.
c.) Branching T1 testis with several squiggles.
All of these testes are large enough to have many squiggles, which makes them T1 testes.

Figure 3.1.2: Examples of T1 testes

3.1.2 Smaller Testes (T2)

What distinguishes T2 testes from T1 testes is again the size with respect to that of the specimen. Their size can extend to as much as 3% of the specimen's area, though it largely hovers around 1.5% or lower. As for squiggles, there are always four or less in class T2 testes. They tend to be fairly round or elliptical, not exhibiting the diversity of shape in T1. However, what sets these apart most from their class T1 counterparts are their surroundings. They are never found alone, but always with other class T2 testes. They always come in populations of at least 4.

In some species, the number can range in the hundreds. Their arrangement differs at times. Sometimes they are densely packed, somewhat like bricks. Other times there is space between them. While class T2 testes must contain at least one squiggle, at times this is placed in contact with the enclosure and becomes hard to detect. Figure 3.1.3 shows some examples of T2 testes. Table 3.1 shows the characteristics of T1 and T2 testes side by side.



a	b	c

a.) Elliptical T2 testes with two squiggles.
b.) Round T2 testes with one squiggle.
c.) Round T2 testes with three squiggles.
All of these testes are small enough to only have a small number of squiggles. And they all occur in groups. That's what makes them T2 testes.

Figure 3.1.3: Examples of T2 testes

Table 3.1: Characteristics of testes compared

	T1	**T2**
Size	.2% to 20%	.2% to 3%
Shape	Varies widely	Round or elliptical
Structure	Contains 3 or more squiggles	Contains 1 to 4 squiggles
Number	1 or 2 in most cases	At least 4, sometimes hundreds
Distribution	Usually in close proximity	Packed together or scattered

3.2 Characteristics of Ovaries

These sketches depict ovaries as a sack-like region filled with round white blobs with a single black spot inside. The image is likely meant to communicate a repository of eggs. At least, this is the ideal case. Unfortunately, oftentimes the white blobs are indistinct, and the ovary is often left as a black region with scattered fragments of white strewn throughout. See Figure 3.2.1 for some example ovaries.

Ovary detection is a matter of pattern detection. Techniques are needed to detect both patterns involving the well-defined round white blobs and the darker ones with white fragments. These two patterns have been called O1 and O2. Even within a

single ovary, both patterns are sometimes present. To find the whole ovary, both techniques would be needed.

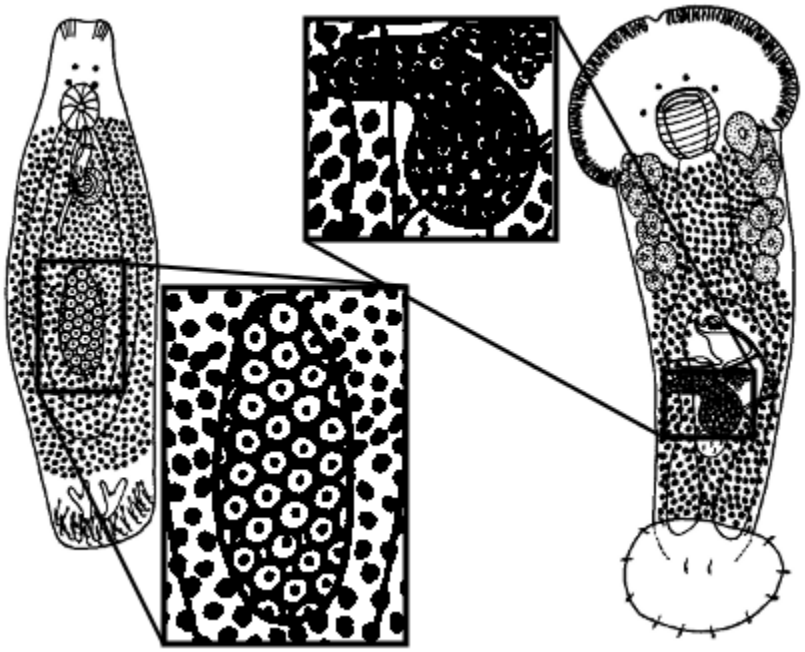


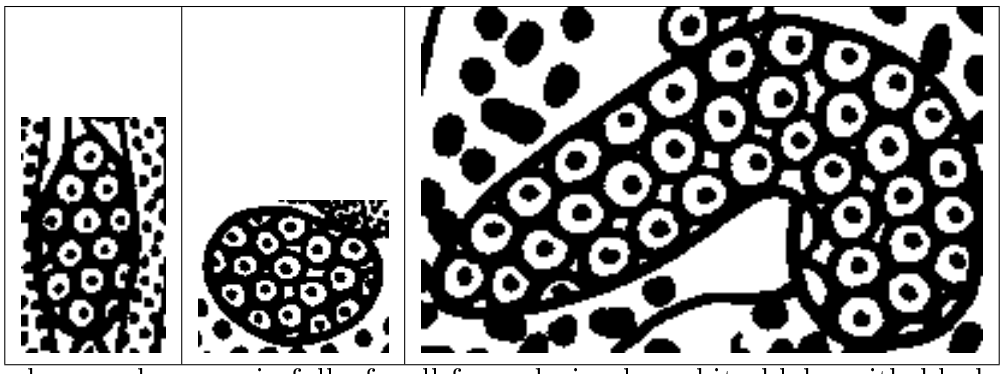
Figure 3.2.1: Sketches with highlighted ovaries

3.2.1 Light Ovary Pattern (O1)

Areas with distinct white blobs with black holes in their centers, are classified as the O1 pattern. The black holes in the center don't touch the sides, which would cause the white blob to mutate to a crescent. Oftentimes, we can see a line wrapping around the outside of the blobs, giving the impression of a sac. Figure 3.2.2 shows ovaries with an abundance of the O1 pattern.

3.2.2 Dark Ovary Pattern (O2)

Dark areas whose white blobs have been mostly broken up and separated are classified as the O2 pattern. These often seem to occur when the ovary's surface area or shape in



Notice how each ovary is full of well-formed circular white blobs with black inner cores. This is the defining characteristic of the O1 pattern.

Figure 3.2.2: Examples of ovaries with the O1 pattern

the sketch didn't allow for the space necessary for the distinct white blobs. Sometimes however, it merely seems to be a stylistic choice. Figure 3.2.3 shows examples of ovaries mostly composed of the O2 pattern.

In the ovaries in the middle and the right, notice the almost complete absence of unbroken white circular blobs. This is the hallmark of the O2 pattern. But the ovary on the left, while mostly O2, also contains a region of several unbroken blobs towards the top. This is an example of how a single ovary can have both patterns.

Figure 3.2.3: Examples of ovaries with the O2 pattern

Summary

In this chapter, we covered the characteristics of the ovaries and testes. Testes are divided into two types, while ovaries have two different patterns. By partitioning

the structures based on their characteristics, algorithms tailored to these specific characteristics can be used. Chapter 5 explores the nature of these algorithms.

Chapter 4

Image Preparation

4.1 Preparation of Sketches

For this thesis, a source of images was needed. That source was Schell's handbook. However, a mere scan of the book was not enough. Operations were needed to extract out the individual images and prepare them for further processing. This chapter covers those operations.

4.1.1 Extraction

Naturally, the first thing to be done was extraction. Any given page might have half a dozen sketches. Figure 4.1.1 shows a sample page. It would probably have been possible to use an automated means of extracting these images, but in the end it was decided to do it manually. There were several reasons for this:

1. Each image had to have its species recorded. This would have required manual action anyway.

2. A good number of the sketches weren't of any use. For example, it was not

uncommon for sketches to only show parts of a specimen. Some sketches are merely specific organs of interest.

One by one, each image was considered. Sketches deemed worthwhile where cut out and had their names recorded. This was somewhat time consuming, but probably no more so than devising an automated schema and cleaning up afterward.

4.1.2 Cleaning

The end product of this was a collection of nearly 700 images. However, the images were still not ready for any sophisticated processing yet. In addition to various bits of noise in the background from the scanning process, each image was accompanied by a scale bar. This bar is useful to researchers for determining the scale of a sketch, but it would only get in the way of computer vision. Also, each image had the name of the species listed below it.

Automated removal of these pieces of noise actually proved to be quite simple. All that needed was to perform a flood-fill operation from the corner, and then perform a connected component operation on everything else. The specimen would certainly be the largest object in the image. Everything else could then be eliminated. Figure 4.1.2 shows an image before and after eliminating the scale bar and text.

The one unforeseen problem with this was that a small number of drawings of specimens had holes in their borders. This meant the flood fill would fill into the specimen, and large parts of it would be lost. There were very few of these, and they were ultimately discarded.

Summary

This chapter looked at the necessary techniques for extracting the sketches from the handbook. After extraction, images had various leftovers removed. The result of this process was a database of images which were then used for structural detection.

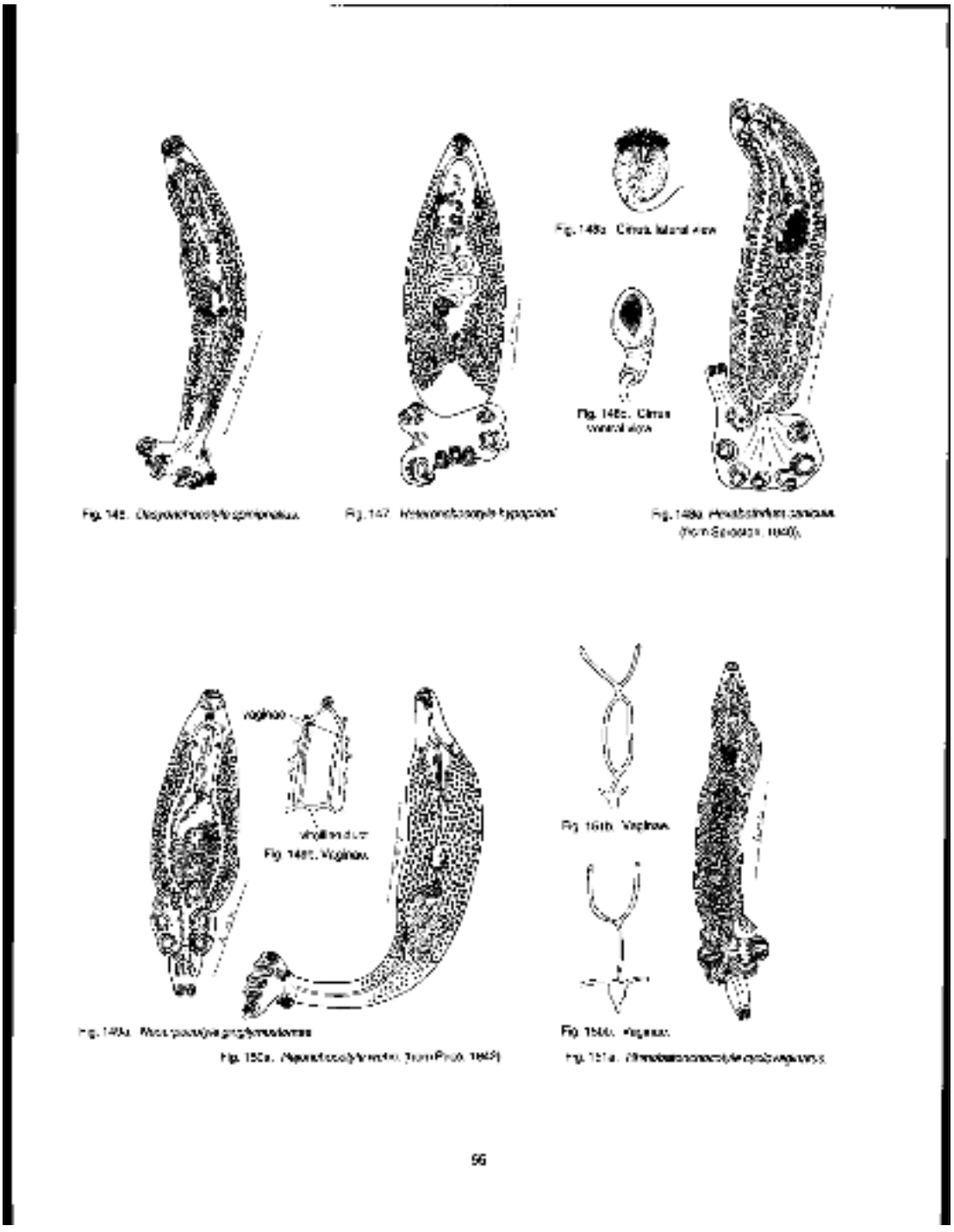


Figure 4.1.1: Sample page

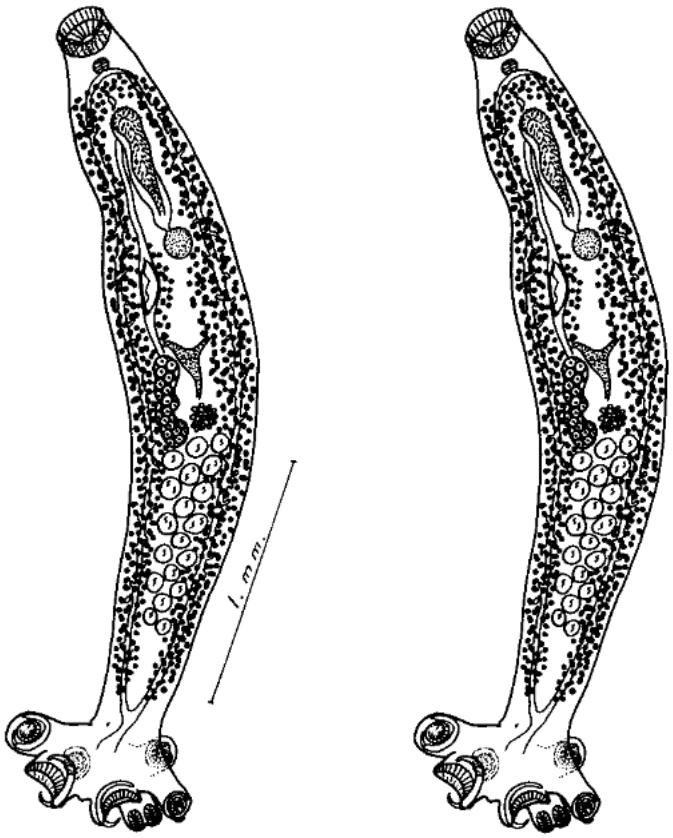


Fig. 146. *Dasyonchocotyle spiniphallus.*

Figure 4.1.2: Image before and after cleaning

Chapter 5

Testes Detection and Classification

In this chapter, we discuss the detection of the testes. This detection is divided out into algorithms for T1 and T2 testes. Finally, we cover the accuracy of the detection.

The common thread between detection of T1 and T2 testes are what have been termed"squiggles". These are the black, "S" shaped blobs that testes have within them. All testes have them, and they aren't used outside of testes, so the squiggles are clearly an important part of detecting testes. The key to detecting testes is detecting the squiggles within them, no matter the type of testes.

Once the squiggles are detected, testes detection is basically reduced to simply searching for white blobs containing squiggles, for both T1 and T2 detection. T2 detection requires some further processing relating to the geographic relationships of the blobs, but in any case, squiggle detection is the essential first step.

5.1 Squiggle Detection

Squiggle detection serves as an interesting problem. The squiggles are tiny black objects, but they certainly are not the only tiny black objects in the image. In

particular, the vitelline glands occupy a similar niche. In Figure 5.1.1, we see several examples of squiggles.

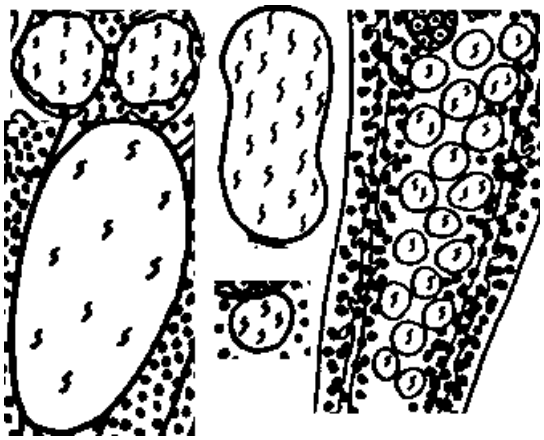


Figure 5.1.1: Testes with various styles of squiggles

5.1.1 Squiggle Characteristics

The first step in detecting the squiggles was collecting statistics about them. From the images, a random sampling of a few dozen was taken. The attributes studied were:

- area – the number of pixels in a squiggle
- aspect ratio – the length divided by the width of the minimum bounding box [13]
- orientation – the angle of the squiggle's major axis with respect to the x axis
- density – the area of the squiggle divided by the area of the minimum bounding box
- eccentricity – the roundness of the squiggle

For each of the above attributes, the range of valid values amongst the sample data was calculated. See Table 5.1.

Table 5.1: Ranges of squiggle attributes

	Area	**Aspect Ratio**	**Orientation**	**Density**	**Eccentricity**
Minimum	25	1.86	75	.64	.87
Maximum	67	3.23	102	.98	.94

It is important to note that there are squiggles that fall outside these measures for one or more attributes. However, these measures provide an excellent baseline for what most squiggles will fall within. While useful, these measures alone were not sufficient to reliably discriminate squiggles. Detection therefore requires an additional tool.

5.1.2 Quadrant Profiling

If aligned vertically, the squiggles always resemble the letter "S", with the line extending from lower left corner to the upper right. This can be thought of in contrast to the number "2", which exhibits the exact opposite layout. Fortunately, squiggles posses axial symmetry, so there is no proper "top" or "bottom".

The method devised for grading conformance to this layout has been named "quadrant profiling". For every potential squiggle, the following is performed:

1. Thin the blob down to its skeleton. Using that, find the two most distant end points.

2. Determine the midpoint between these two points.

3. Translate the blob so that this midpoint lies at the origin. Rotate it so that both end points lie on the X-axis.

4. Count the number of pixels of the blob that lie in each quadrant.

5. The quad score of the blob is the number of pixels in the upper right and lower left quadrants, minus the number of pixels in the lower right and upper left, divided by the total number of pixels. This lies between -1 and 1.

6. To gauge the symmetry, or balance, the absolute value of the difference between the number of pixels in the upper right and lower left quadrants is taken. This is then divided by their sum. This value lies between 0 and 1.

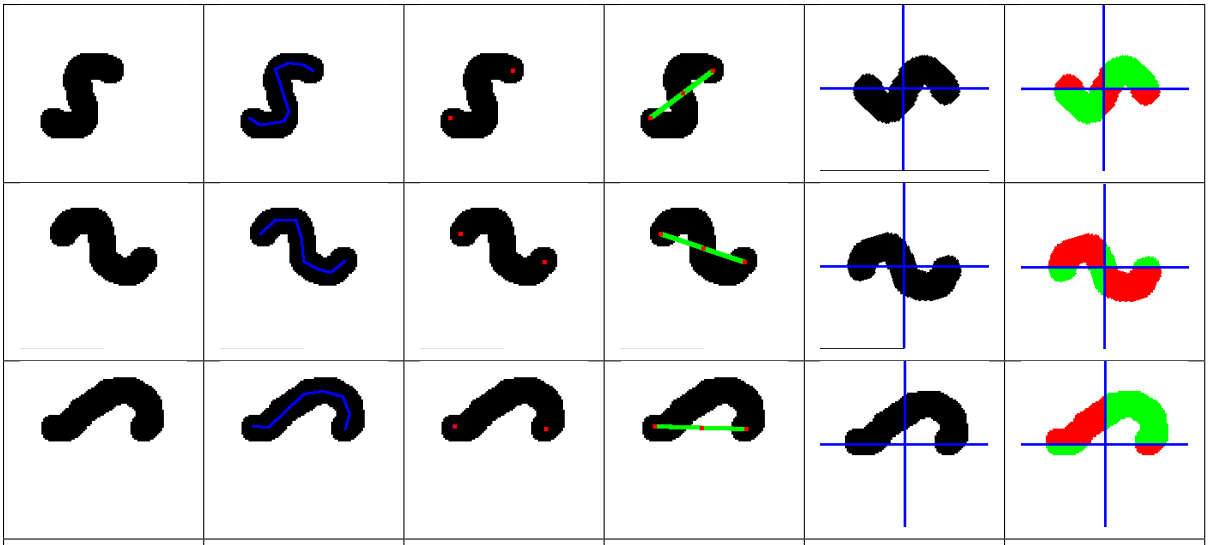
The quad score gives an idea of how well the blob lays in the desired quadrants. A score of 1 is the ideal, while a score of -1 is the least desired case. The balance score tells us how rotationally symmetric the blob is. A score 0 indicates perfect balance, while a score of 1 indicates total imbalance.

Neither the quad score nor the balance alone is very helpful, but together, they give a reasonable answer as to whether the shape is a good squiggle or not. Consider Figure 5.1.2. The top row shows a squiggle. The middle shows a mirrored squiggle. The bottom row shows a different curved blob. Note how the bottom row's blob enjoys a fair quad score, while the middle row's blob enjoys fair balance. This illustrates how both scores are essential to make a sound judgment.

5.1.3 Detection

Squiggle detection takes place over 3 phases:

- Elimination
  - Blobs that cannot possibly be squiggles are eliminated from further consideration.



a	b	c	d	e	f

a.) Blob.
b.) Skeleton.
c.) Endpoints of skeleton found.
d.) Midpoint of the endpoints found, forming blob's axis.
e.) Blob axis aligned with the X-axis with the midpoint as the origin.
f.) Blob is scored based on position of pixels within the quadrants.

Figure 5.1.2: Quadrant profiling illustrated

- Identification
  - Blobs that strongly resemble squiggles are identified.
- Support
  - Blobs supported by a nearby strong match are identified.

In the Elimination phase, each blob is scored based on a series of tests. For example, if a blob's aspect ratio is between 1.6 and 4.5, a point is added to its score. Note that this is greater than the range outlined above. The goal of this phase is to eliminate blobs that cannot be squiggles, not to necessarily designate anything as a squiggle. Similar tests are performed for all the statistics outlined above. If the score of the blob is not satisfactory, it is eliminated from further consideration.

In Identification phase, squiggles considered to be strong matches are selected. This phase makes use of the quadrant profiling method. If a blob has a quad score greater than .2 and a balance less than .23, it is marked as being strong. This still doesn't mean it is certainly a squiggle, but it has a strong chance of being one.

It is impossible to identify whether a single blob is a squiggle looking only at that one blob. So the Support phase explores the geographic relationships among the squiggle candidates. Each blob has its neighboring blobs determined. The neighborhood of a blob is defined as all blobs within 65 pixels (centroid to centroid). Furthermore, the difference between two neighbor's areas must not exceed 35% of the larger one's area.

By this, we essentially propagate the strength. Any strong blob's neighbors become strong. In this way, any blob with a connection to a strong blob will be found to be strong. Any neighborhood of blobs must have at least one strong squiggle in it or it will be eliminated. Also, since squiggles don't occur alone, any blobs without neighbors, whether strong or not, are eliminated. The end product of this is a map of all the squiggles in the image. Figure 5.1.3 provides an overview of the squiggle detection process in a flow chart.

5.2 Testes Detection

After determining the locations of the squiggles, detection of the testes themselves can progress. In each case, this comes down to looking for white blobs of various characteristics with squiggles in them.

5.2.1 Larger Testes (T1) Detection

T1 detection involves finding the large white blobs. The following qualities are tested:

- Area
- Number of Holes
- Squiggle Tests
  - Number of Squiggles
  - Slant of the Squiggles
  - Area of the Squiggles
  - Squiggle Coverage

Tests showed that T1 testes have an area in the range of .2% of image area to 18%. Figure 5.2.1 shows a graph of T1 areas occupying this range. So the first step is performing a connected component operation on the white image areas, and removing any that fall outside this range.

The next step performed is eliminating any blobs that lack holes, or have an excessive number of holes. Aside from the rare cases of encroaching structures, the only holes present should be squiggles. Any blobs with less than 3 holes or more than 80 are discarded. The result of this is an image of blobs which are possible T1 testes.

Each blob is then tested based on the squiggles found within it, to produce a numerical score indicating the likelihood of it being a testis. One test involves counting the number of squiggles within the blob, and comparing that to the area of the blob. Since the two qualities correlate, a large blob should have a large number of squiggles.

Another test involves looking at the standard deviation of the slant of the squiggles in the blob. Squiggles have the useful property of being largely parallel. So a blob is scored favorably when this standard deviation of the slant is low. Likewise, the same measure is used for the standard deviation of the area of the squiggles.

Another measure used is the coverage of the squiggles. For this, the area of the convex hull of the squiggles in the blob is divided by the area of the blob itself. This tells us how well the squiggles cover the blob. Naturally, a high coverage results in a favorable score. Figure 5.2.2 shows a graphical overview of the entire T1 testes detection process.

5.2.2 Smaller Testes (T2) Detection

T2 Detection takes place through the following phases:

- Identification – Small round blobs are detected.
- Strength – Blobs that contain squiggles are considered strong matches.
- Support – Blobs supported by nearby strong matches are identified.

In the Identification phase, all the small round white blobs are detected. These blobs must have an area between .2% and 3% of the total image area. This gives a map of the possible T2 testes in the image. Figure 5.2.3 shows a graph of T2 testes within this range.

In the Strength phase, the blobs from the Identification phase are evaluated based on whether they contain a squiggle. Since T2 testes sometimes have their squiggle obscured by the blob's boundary, the lack of one or more squiggles in a blob is not a sufficient criterion to eliminate it. Instead, the presence of a squiggle in a blob is considered a mark of strength.

Then comes the Support phase. Very much like in squiggle detection, a neighborhood-based strength propagation is performed. The only difference is the definition of neighborhood. In this case, two blobs are neighbors if they are within 130 pixels of each other (centroid to centroid) and differ in area no more than 15%. As before,

lone blobs, and those without any connection to a strong one are eliminated. Figure 5.2.4 gives a graphical overview of the T2 testes detection process.

5.3 Detection Accuracy

Table 5.2 is a confusion matrix showing the overall accuracy of the testes detection algorithms on this data set. Below, figures 5.3.1 and 5.3.2 show examples of successful and failed testes detection, respectively.

5.3.1 Larger Testes (T1) Accuracy

T1 detection fairs quite well. It finds roughly 75% of the T1 testes. It suffers from a very small amount of false positives, but not enough to jeopardize the usefulness of the algorithm.

5.3.2 Smaller Testes (T2) Accuracy

On the surface, T2 detection's results are disheartening. It discovers nearly 60% of the T2 testes, but it also suffers from a very large number of false positives. However, these results must be put into context. Less than 1/4 of the images have T2 testes. The other 3/4 or so of the images are the source of a great number of these false positives. In those cases, T1 detection would usually find the testes, so T2's results would likely not be considered.

5.4 Classification of Testes

For the sake of placing the results in a database and allowing users to query for information on the structures in the sketches, it was necessary to qualify the attributes

of the testes into classes. One way would be to record the characteristics of the individual testes. However, in practice this would prove highly redundant. Within a single specimen, most of the testes have the same size and shape. So instead, the classification of testes for a specimen are determined based on their aggregate characteristics. Table 5.3 shows the classification scheme.

The classes of numbers are self explanatory. As for shape, if the compactness (area / convex hull area) is less than 80%, the testes will be considered non-compact. Otherwise, if the roundness is .9 or greater, the testes will be considered round. If neither condition is satisfied, the testes is considered elongated.

Size is in reference to the percentage of the total specimen area. Coverage only comes in to play when there are 4 or more testes. This is a measure of the percentage of the specimen's area that is covered by the convex hull of the testes. For example, if a coverage of 50% were achieved, this would fall into the "spread" class. Figure 5.4.1 shows some example images with their testes classified.

Summary

In this chapter we've covered the algorithms used to detect testes within the image collection. Furthermore, we statistically demonstrated the effectiveness of these algorithms. Finally, we introduced a scheme for classifying the testes into qualitative classes.

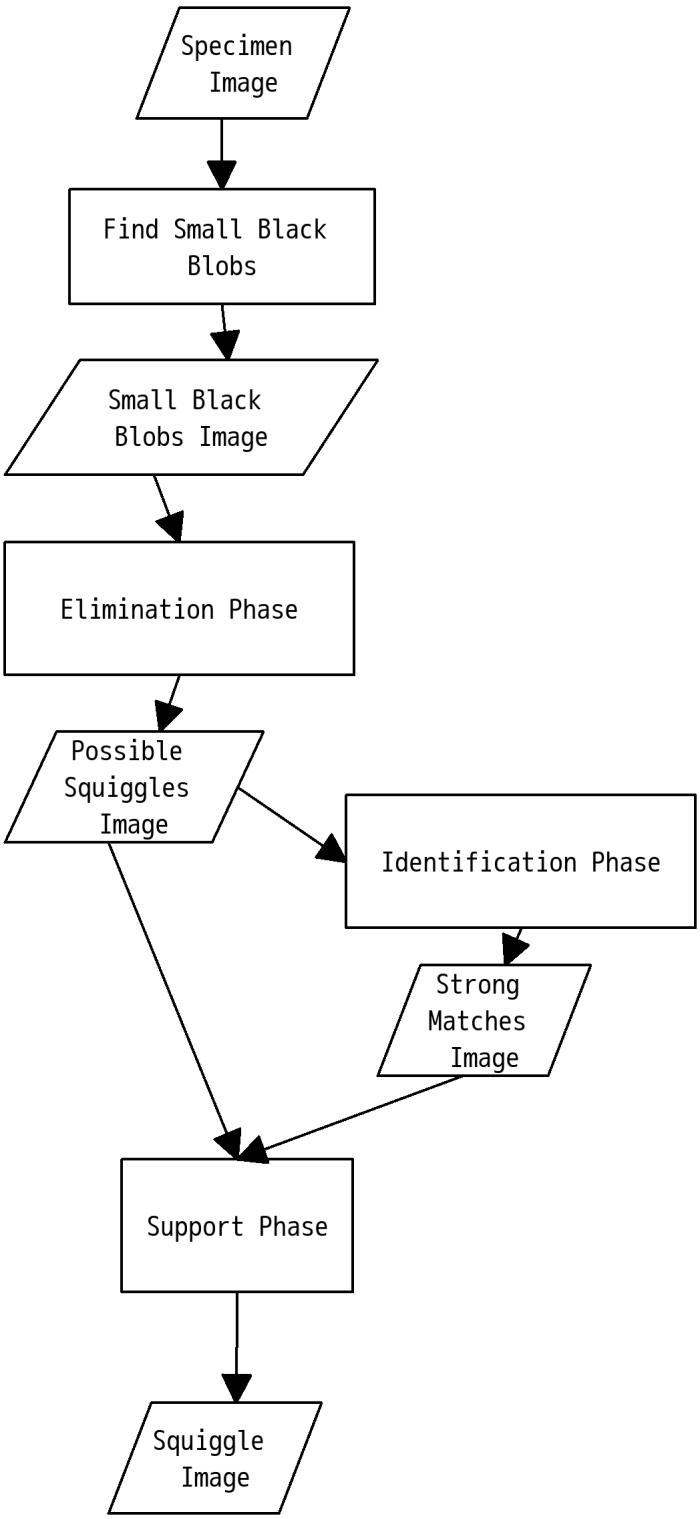
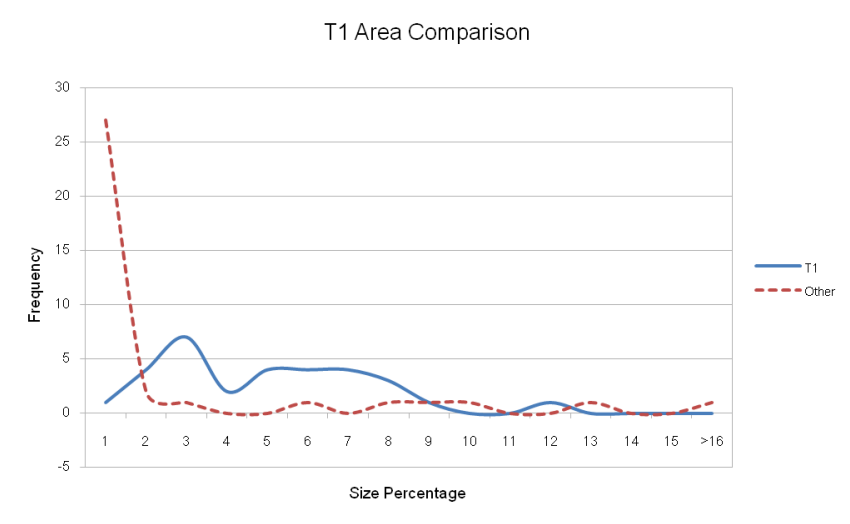
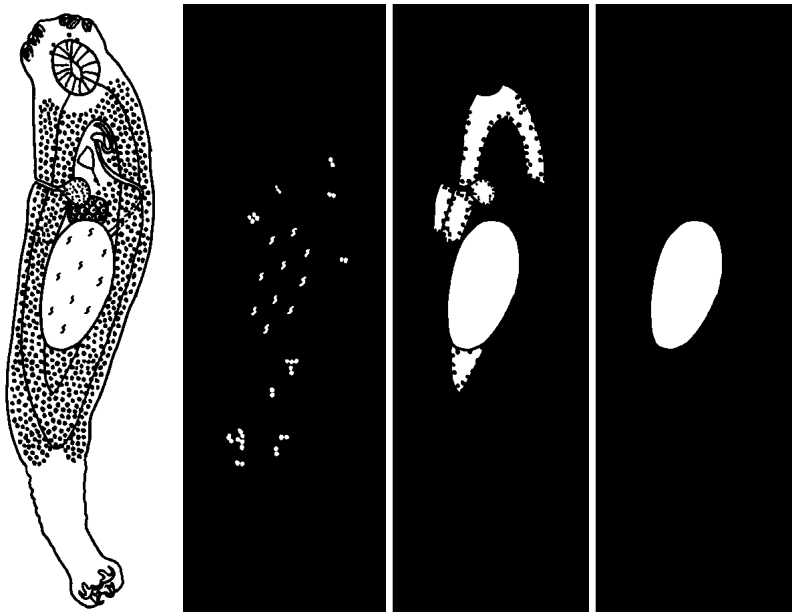


Figure 5.1.3: Flowchart of squiggle detection



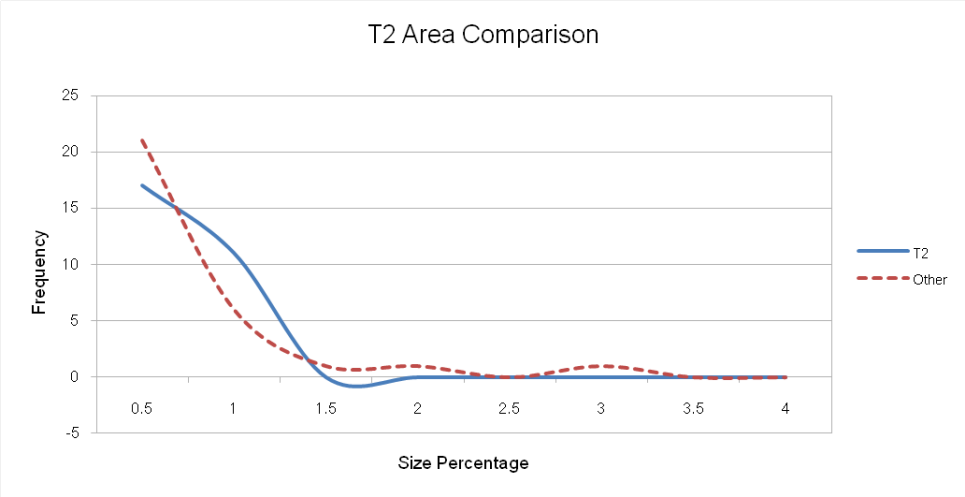
This graph shows a plot of the distribution of areas of T1 testes compared with that of an assortment of random white blobs.

Figure 5.2.1: T1 testes area graph



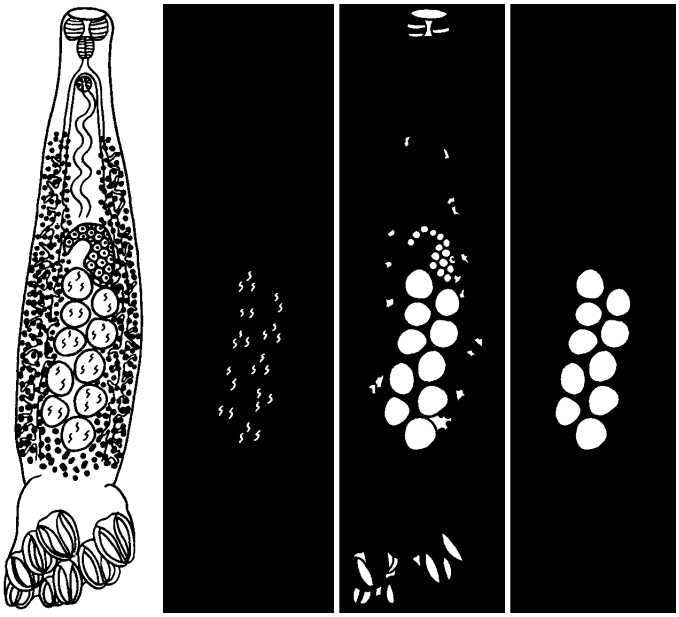
From left to right: the original image, an image showing possible squiggles, an image showing white blobs that may be testes, and the image showing blobs matched with squiggles.

Figure 5.2.2: The T1 testes detection process



This graph shows a plot of the distribution of areas of T2 testes compared with that of an assortment of random white blobs.

Figure 5.2.3: T2 testes area graph



From left to right: the original image, an image showing possible squiggles, an image showing white blobs that may be testes, and the image showing blobs matched with squiggles.

Figure 5.2.4: The T2 testes detection process

Table 5.2: T1 and T2 testes detection results confusion matrix

Predicted	Actual		
	T1	T2	Non-Testes
T1	739	113	61
T2	62	1,285	3,342
Non Testes	261	898	

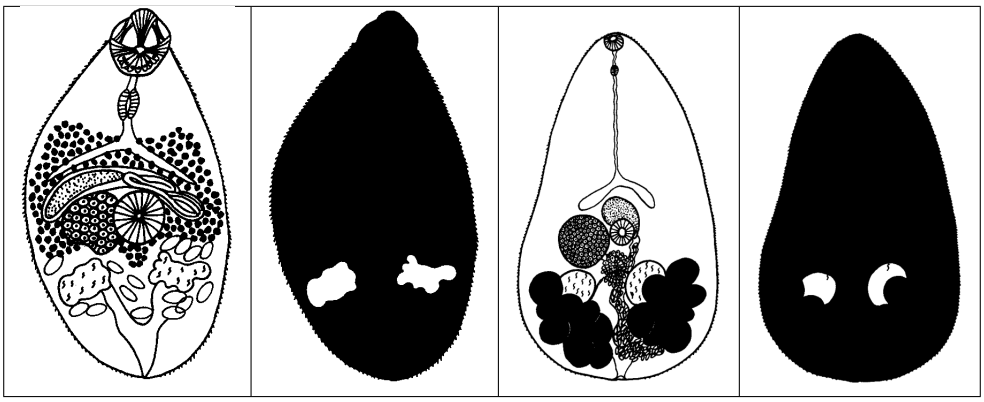
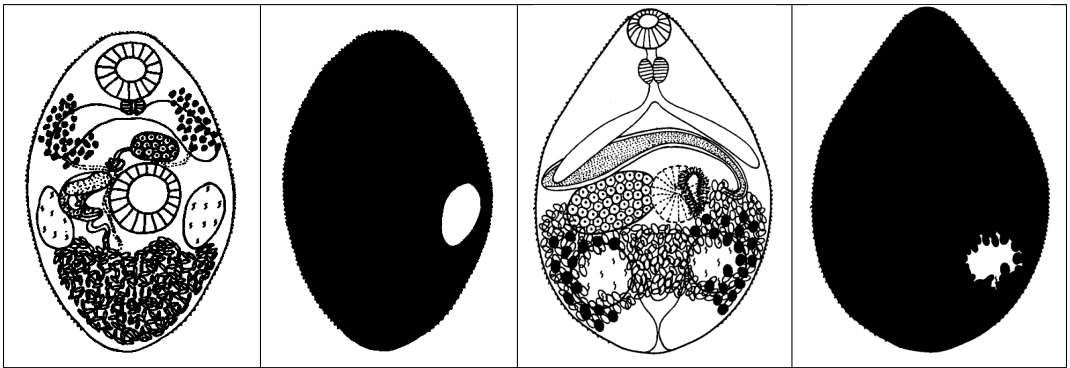


Figure 5.3.1: Successful testes detection



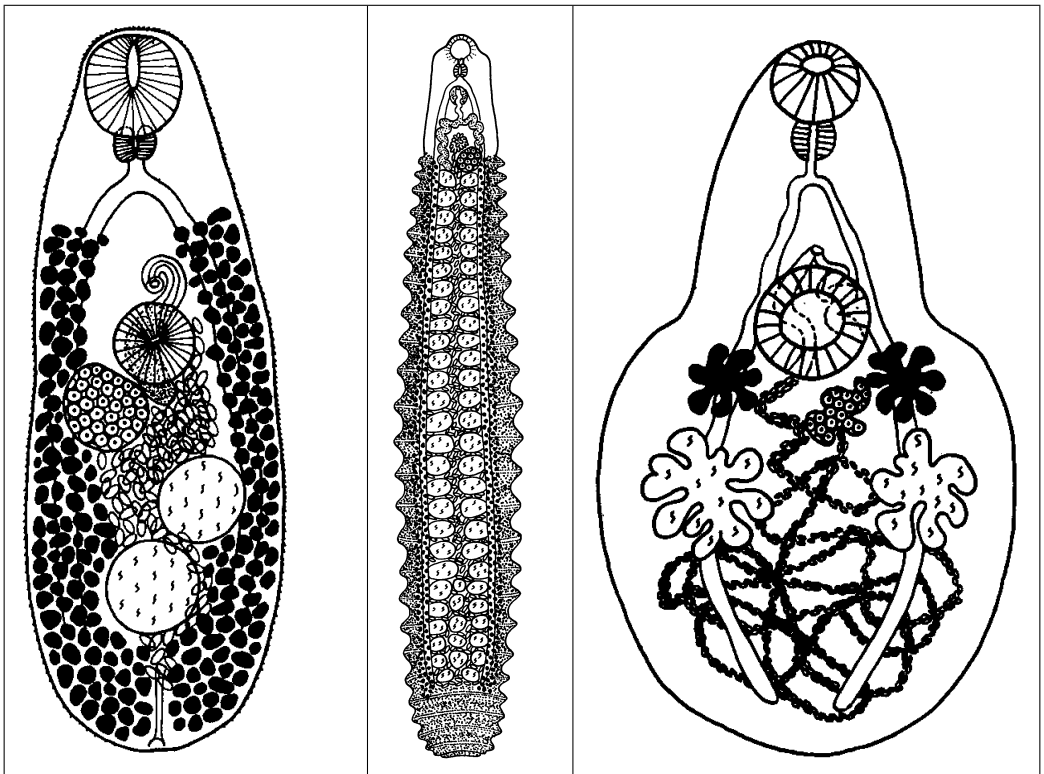
A B

A.) Left testis missed due to poorly represented squiggles.
B.) Left testis missed due to encroaching structures cancelling out some of the squiggles.

Figure 5.3.2: Failed testes detection

Table 5.3: Testes attributes and their classes

Number	**Shape**	**Size**	**Coverage**
0	round	tiny (below 1.5%)	compact (below 25%)
1	elongated	small (1.5% - 2.99%)	spread (25% - 59.99%)
2	non-compact	medium (3% - 4.99%)	ubiquitous (60%+)
4		large (5%+)	
5 - 10			
11 - 20			
20+			



a **b** **c**

	Number	**Shape**	**Size**	**Coverage**
a	2	round	medium	compact
b	20+	round	tiny	ubiquitous
c	2	non-compact	medium	compact

Figure 5.4.1: Sample testes classifications

Chapter 6

Ovary Detection and Classification

In this chapter, we describe algorithms for the detection of ovaries and their performance based on the sketch database. We have developed separate algorithms for the two patterns, as described in Section 2. Ovaries consisting of the O1 pattern are lighter, with distinct internal white balls. In contrast, ovaries consisting of the pattern O2 are darker, with internal white marks scattered throughout. Some ovaries are comprised of both patterns. See figure 6.0.1 for an example of each pattern.

Figure 6.0.1: Example O1 and O2 patterned ovaries

6.1 Detection of Ovaries

Ovary detection revolves around detection of the patterns that comprise them. Because any given ovary might consist of both patterns, reliable ovary detection requires both patterns.

6.1.1 Light Ovary Pattern (O1) Detection

O1 detection hinges on finding the small white balls that make up the body of the ovary. In turn, each of these contains a single black core. In this way, it is actually very much like the testes detection's blobs and squiggles from Chapter 5. See figure 6.1.1 for a graphical annotation of the ball and core.

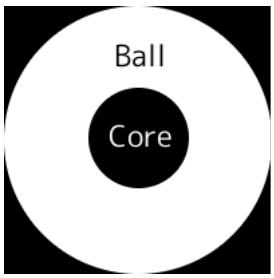


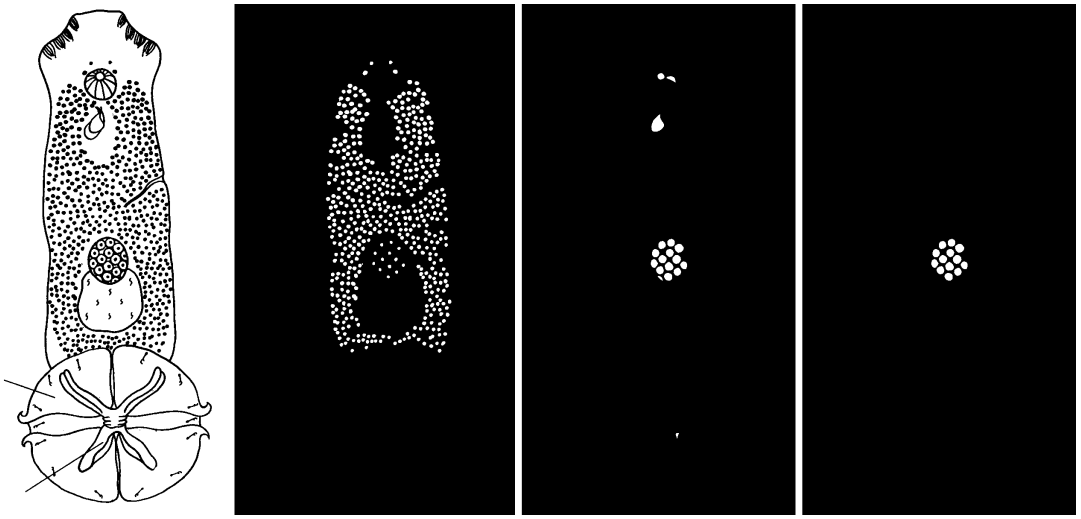
Figure 6.1.1: Annotated ball and core

First a mask of the white balls is detected. Then a mask of the black cores is detected. Finally, these are used together to eliminate balls that aren't likely to belong to an ovary. See figure 6.1.2 for an illustration of the process. Below is the pseudo-code for this algorithm:

```
function ball_mask(image)
for each blob in image

    if (blob_holes != 1) then

        discard blob from image;

        if (blob area > maxBallArea) or (blob area < minBallArea)
          or (blob roundness < minBallRoundness) then

           discard blob from image;

// Any remaining blobs are considered balls
return image;
```

```
function core_mask(image)
for each black blob in image

    if (blob_holes != 0) then

        discard blob from image;

        if (blob area > maxCoreArea) or (blob area < minCoreArea)
          or (blob roundness < minCoreRoundness) then

           discard blob from image;

    // Any remaining blobs are considered cores
    return image;
```

```
function o1_mask(image)
ballimage = ball_mask(image);
    coreimage = core_mask(image);
    for each blob in ballimage
        if the number of cores intersecting != 1 then
            discard blob from image;
    // Add some sort of clustering code here to join the
    // remaining blobs into a single ovary.
    return image;
```



From left to right: the original image, an image showing possible cores, an image showing possible balls, and the image showing balls matched with cores.

Figure 6.1.2: The O1 ovary detection process

6.1.2 Dark Ovary Pattern (O2) Detection

O2 Detection actually represents a significant departure from the methods previously explored. O2 patterns have no real consistent internal structures. Likewise, they

are rarely isolated from surrounding structures, so simple connectivity based searches yield little benefit.

The search method begins by searching for dark black areas within the image. O2 patterned regions often contain small white pieces, fragments of the balls present in the O1 pattern. These are dealt with via a series of small morphological openings, aimed to break up any white pieces with a narrow height or width.

The next step is separating the dark black regions from their surroundings. This required morphological closing. Finding the proper size for the closing's structuring element was tricky though, given the differences in scale from sketch to sketch. A handy tool for estimating the scale of the image is the average width of the border of the specimen. This is done by simply sampling the width at a number of points. The highest and lowest outlier values are discarded, and the mean of the remainder is calculated.

The optimal structuring element proves to be a disk shape, with a radius of 2.5 times the average border width. By closing with this, black blobs are reliably separated from their surroundings.

Ovaries are not, however, the only dark black blobs. Often, some of the larger vitelline glands also form dark black areas. However, these often lack the white pieces within. So the next detection step is to find small white objects in dark neighborhoods. The neighborhood of pixel is considered to extend roughly 3 times the width of the specimen's border. Any white pixels whose neighborhoods are sufficiently dark are then considered to be potential ovary components.

Finally, using this new image, any segments from the dark region image that don't have a white object nearby are trimmed off. In this way, many blobs that result from the fusion of an ovary and its surroundings can still be trimmed down to free the ovary. Figure 6.1.3 provides a flowchart illustrating the entire process while figure

6.1.4 gives a graphical overview.

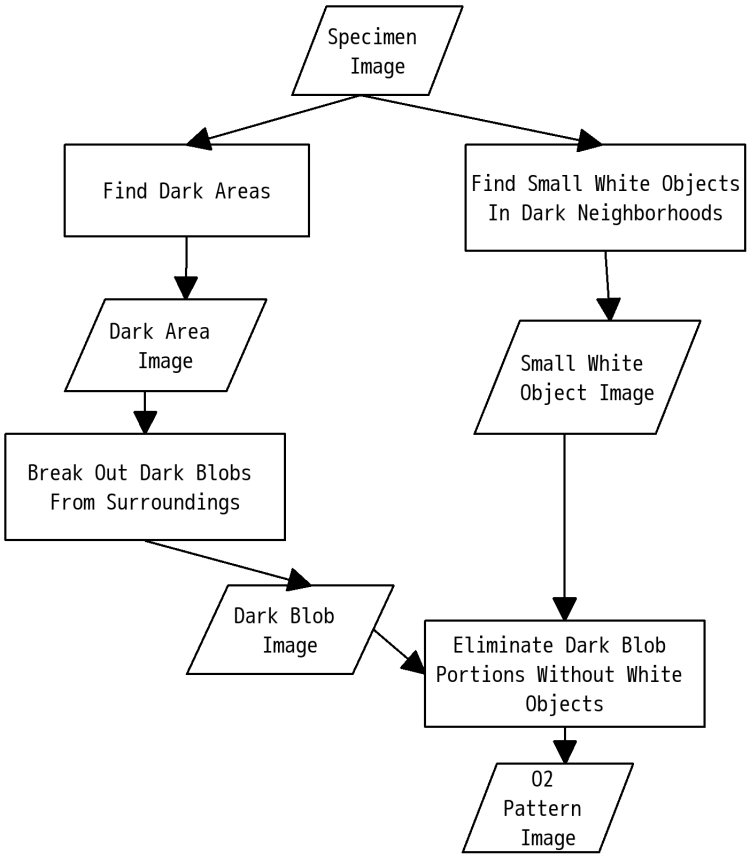
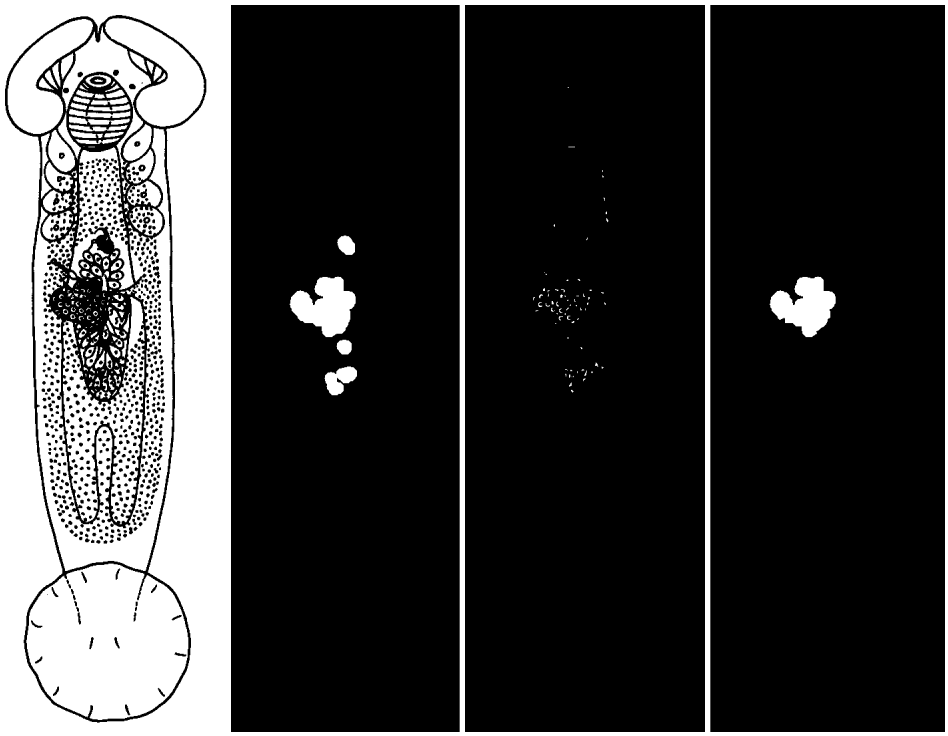


Figure 6.1.3: Flowchart of O2 pattern detection

6.1.3 Combining Pattern Detection Results

The masks of O1 and O2 patterns are combined through a logical OR operation. The final step is, of multiple blobs, discriminating which blob, if any, is the ovary. It proves advantageous to work under the false assumption that every specimen has a single ovary. At first glance, it would seem foolish to make an assumption that is known to be incorrect. However, when one considers that the vast majority of the specimens do in fact have an ovary, this proves to be a good assumption. The exceedingly rare cases of incorrectly indicating an ovary are more than made up for



From left to right: the original image, an image showing dark areas, an image showing small white signals, and the image showing dark areas matched with white signals.

Figure 6.1.4: The O2 ovary detection process

by avoiding frequent failures to indicate an ovary in images that do have one. And it would be quite possible that those images wouldn't have any dark blobs anyway.

So the question becomes: of the remaining blobs, which is most likely to be an ovary? Based on experimental observations, it was discovered that simply selecting the largest blob proves to be quite accurate. Other measures, such as concavity or roundness could have proved useful, but among the ovaries, there are many with very odd shapes that wouldn't score well at all by such metrics.

6.2 Implementation and Results

For the O1 algorithm, the parameter values used are listed in table 6.1. These thresholds were determined by observing a sampling of the sketches and taking measurements of these parameters.

Table 6.1: Variable values for class O1 ovary recognition

Variable	**Setting**
maxBallArea	.23% of Image Area
minBallArea	.005% of Image Area
minBallRoundness	.6
maxCoreArea	.2% of Image Area
minCoreArea	.0045% of Image Area
minCoreRoundness	.8

For testing of the accuracy of the individual pattern detection algorithms, each algorithm was run on a subset of sketches featuring ovaries with that particular pattern. An ovary is considered to have been detected if over 50% of its surface area is detected. It is considered a false positive if something entirely different is selected, or if the detected region outside the ovary exceeds 30% of the ovary's area. See table 6.2 for these results.

By this metric, O1 detection fares very well finding 95% of the ovaries in the sketches when given a collection. O2 pattern detection fares moderately well, finding 80%, while also suffering from some modest false positives. This accuracy is especially surprising given the less rigid definition of the O2 pattern.

Table 6.2: O1 and O2 ovary detection results

Class	Tested	Found	**Missed**	**False Positives**	**Accuracy**
O1	64	61	3	0	95%
O2	32	29	7	8	80%

Table 6.3 shows the results of full ovary detection over the entire data set. De-

tection was very good. Missed ovaries were minimal. False positives were somewhat higher, but not so much as to spoil the usefulness of the algorithm. Figures 6.2.1 and 6.2.2 show some examples of successful and failed ovary detection, respectively.

Table 6.3: Ovary detection results confusion matrix

Predicted	Actual	
	Ovaries	Non-Ovaries
Ovaries	541	80
Non-Ovaries	10	

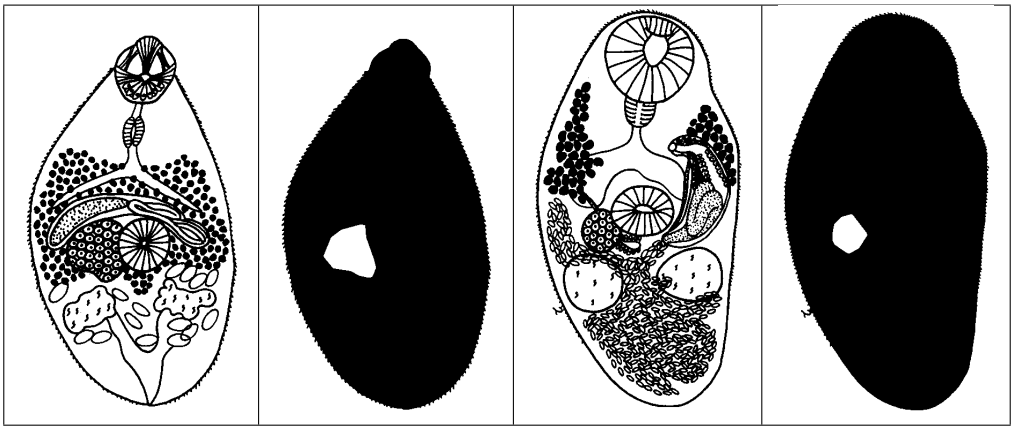
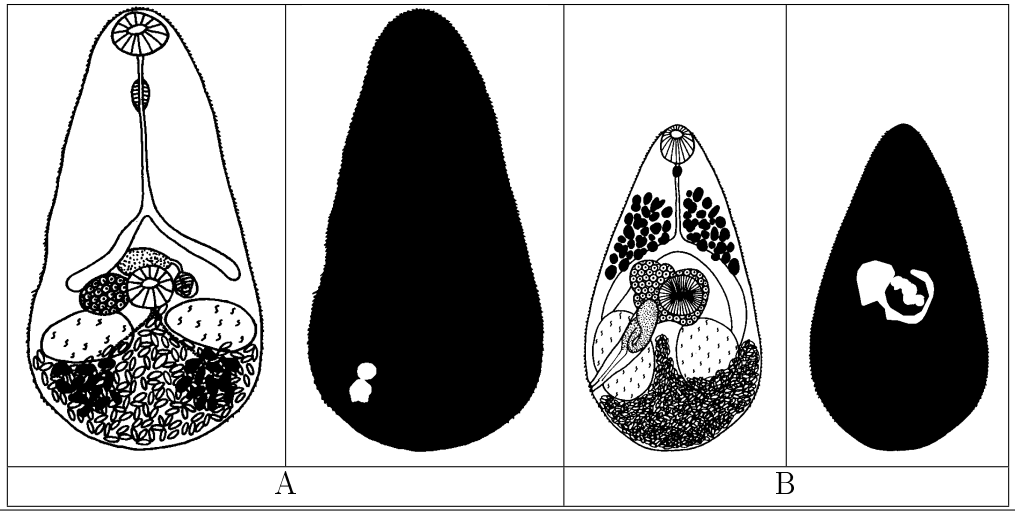


Figure 6.2.1: Successful ovary detection

6.3 Classification of Ovaries

The results must then be qualified for placement in the database. The key features recorded from the ovaries are their shape, size, and number. In the case of the number, since the specimens in the sketch database only have at most a single ovary, this becomes more of a Boolean statement as to whether a given specimen has an ovary. Table 6.4 shows these attributes.

Shape for ovaries is figured exactly as it is with testes. If the compactness (area / convex hull area) is less than 80%, the ovary will be considered non-compact.



A.) Ovary missed in favor of another dark area with pieces of white signal.
B.) Ovary is detected, but a significant amount of the surroundings are included.

Figure 6.2.2: Failed ovary detection

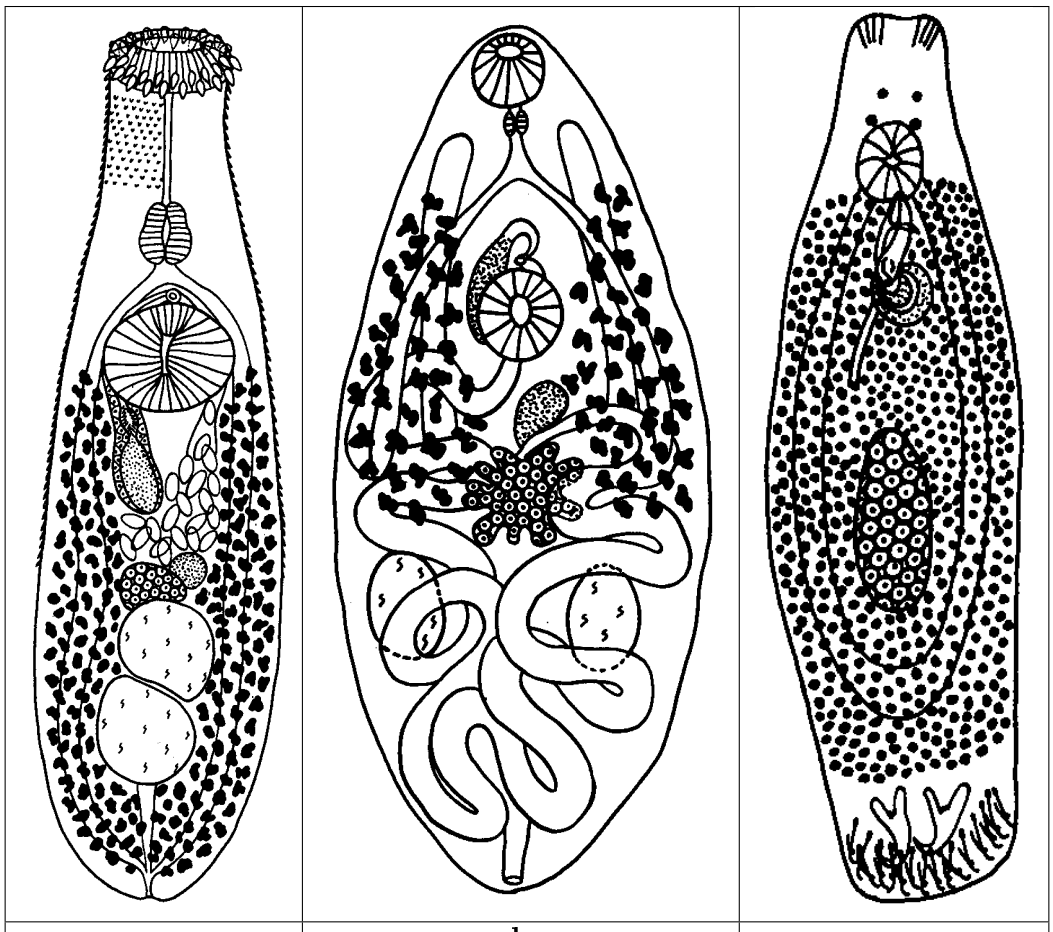
Table 6.4: Ovary attributes and their classes

Number	**Shape**	**Size**
0	round	tiny (below 1.5%)
1	elongated	small (1.5% - 2.49%)
	non-compact	medium (2.5% - 3.99%)
		large (4%+)

Otherwise, if the roundness is .9 or greater, the ovary will be considered round. If neither condition is satisfied, the ovary is considered elongated. Size is in reference to the percentage of the total specimen area. Based on this, size is qualified into one of four bins. Figure 6.3.1 shows some examples with their ovaries classified under this scheme.

Summary

This chapter has demonstrated algorithms for the detection of ovaries in the sketches. Furthermore, it evaluated the accuracy of these algorithms. Finally, it demonstrated



a	b	c

	Number	Shape	Size
a	1	round	small
b	1	non-compact	medium
c	1	elongated	large

Figure 6.3.1: Sample ovary classifications

a classification scheme for the ovaries.

Chapter 7

Summary and Future Work

7.1 Summary

Through this research, I have developed algorithms for automatically deriving structural information about sketches of biological specimens. Algorithms were developed to handle both ovaries and testes. While these algorithms are very specialized for the given collection, they do demonstrate a number of abstract methods that should prove useful for processing other types of sketches.

7.2 Future Work

As with any piece of research, there are plenty of avenues available for further exploration, some of which are outlined here.

7.2.1 Algorithm Tuning via AI

One thing that was not explored in devising and tuning the algorithms was the use of AI. All tuning was done manually through trial and error. Given the testes detection

algorithms' heavy use of parameters and somewhat lackluster accuracy, those would be excellent candidates for further tuning via neural networks or a similar approach.

7.2.2 CBIR System

The fruit of this work emerges in the form of a platform for users to query for specimens based on their structural data, and an initial population of data to put move it out of theory and into practice. A likely next step is building a CBIR system to allow easy queries of the structural data.

Such a system could search the database for specimens whose features resemble those a user specifies. In this way, users can identify unknown specimens, or failing in that, at least narrow down where they fall in the phylogenetic tree.

7.2.3 Additional Applications

There are plenty of areas for expanding this research to other structures and other datasets.

7.2.3.1 Detecting Additional Structures

Ovaries and testes only represent a few of the many structures present within these specimens. There are plenty of other internal structures that prove useful for specimen identification, such as the suckers, vitelline glands, and many others. However, due to the nature of the sketches, each type of structure demands its own techniques for detection. Due to time constraints, the scope of this thesis had to be limited to ovaries and testes.

7.2.3.2 Other Collections

There are many, many more collections of biological sketches to be digitized, even among trematodes. This represents just one small, but important collection. It would be interesting to see how well the detection algorithms from this research can be applied to other such sketches.

Almost all of the existing graphical processing work currently deals in non-biological illustrations, such as maps or mechanical drawings. The techniques in this research should serve as an important step towards advancing graphical processing of sketches in biological fields.

7.3 Conclusion

In conclusion, in addition to the field of parasitology, this research will prove quite useful for a wide range of applications of digitization. Many disciplines could benefit from importing older graphical data from hard copy into digital, searchable forms. Surely any discipline that predates the advent of cheap computer access will have large bodies of data waiting to be made electronically available.

Bibliography

[1] R.E. Patiño-Escarcina, J.A.F. Costa, Content Based Image Retrieval using a Descriptors Hierarchy, in: Proceedings of the 7th International Conference on Hybrid Intelligent Systems, 2007.

[2] S.Liang, Z.Sun, B.Li, Sketch Retrieval Based on Spatial Relations, in: International Conference on Computer Graphics, Imaging and Vision: New Trends, 2005.

[3] M. Borowski, L. Brocker, S. Heisterkamp, J. Loffler, Structuring the Visual Content of Digital Libraries Using CBIR Systems, in: Fourth International Conference on Information Visualisation (IV'00), 2000.

[4] M. Cloonan, S. Sanett, The Preservation of Digital Content, in: Libraries and the Academy Vol 5 No. 2, 2005.

[5] What is digital preservation?, http://www.oclc.org/news/events/presentations/2001/preservation/chapman.htm. Last accessed on October 26, 2008.

[6] A Short History of the Manter Laboratory, http://www.museum.unl.edu/research/parasitology/about.html. Last accessed on October 26, 2008.

[7] J. Han and M. Kamber, Data Mining Concepts and Techniques, Morgan Kaufmann, San Francisco, CA, 2001.

[8] S. Umbaugh, Computer Imaging Digital Image Analysis and Processing, CRC Press, Boca Raton, Florida, 2005.

[9] S. Schell, Handbook of Trematodes of North America North of Mexico, University Press of Idaho, Moscow, Idaho, 1985.

[10] L. O'Gorman and R. Kasturi, Document Image Analysis, IEEE Computer Society Press, Los Alamitos, CA 1995.

[11] J. Hong and H. Chen and J. Hsiang, A Digital Museum of Taiwanese Butterflies, in: Proceedings of the fifth ACM conference on Digital Libraries, 2000.

[12] Y.Chen and H. Bart and F. Teng, A Content-Based Image Retrieval System for Fish Taxonomy, in: Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval, 2005.

[13] D. Chaudhuri and A. Samal, A Simple Method For Fitting of Bounding Rectangle to Closed Regions, in: Pattern Recognition Vol 40 Issue 7, 2007.

[14] Matthew MacDonald, Your Brain: The Missing Manual, O'Reilly Media, Inc., Sebastopol, CA 2008.

[15] Lawrence O'Gorman, The Document Spectrum for Page Layout Analysis, in: IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 15, No. 11, Nov. 1993, pp 1162-1173.

[16] T. Pavlidis and J. Zhou, Page Segmentation and Classification, in: CVGIP: Graphical Models and Image Processing, Vol. 54, No. 6, Nov. 1992, pp. 484-496.

[17] H.S. Baird, The Skew Angle of Printed Documents, in: Proc. Conf. of the Society of Photographic Scientists and Engineers, 1987, pp. 14-21.

[18] S. H. Joseph and T. P. Pridmore, Knowledge-Directed Interpretation of Mechanical Engineering Drawings, in: IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 14, No. 9, Sept. 1992, pp. 928 - 940.

[19] A. Bessaid, H. Bechar, M. K. Fellah. Image analysis and pattern recognition as tools in map interpretation. Electronic Journal "Technical Acoustics", http://www.ejta.org, 2003, 15.

[20] Safebooru, http://safebooru.org/. Last accessed on May 8, 2010.

[21] S. Kim, Y. Tak, Y. Nam and E. Hwang. mCLOVER: mobile Content-based Leaf Image Retreival System. in: Proceedings of the 13th annual ACM international conference on Multimedia pp. 215 - 216.