MONTANA STATE UNIVERSITY

Extraction of data from digital images of microorganisms
by Paul Andrew Shope

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in
Computer Science
Montana State University
© Copyright by Paul Andrew Shope (1993)

Abstract:
A method is proposed for the extraction of data from digital images of microorganisms. Digital images of microorganisms have certain characteristics which make information extraction difficult. These characteristics are identified, and four steps are suggested to effectively overcome the obstacles inherent in these images. The steps are image enhancement, microbe feature extraction, microbe feature representation, and microbe recognition and enumeration. Examples of the techniques which comprise these steps are shown and each technique's effectiveness is described.

EXTRACTION OF DATA FROM DIGITAL IMAGES

OF MICROORGANISMS

by

Paul Andrew Shope

A thesis submitted in partial fulfillment
of the requirements for the degree

of

Master of Science

in

Computer Science

MONTANA STATE UNIVERSITY
Bozeman, Montana

April 1993

ii

APPROVAL

of a thesis submitted by

Paul Andrew Shope

This thesis has been read by each member of the thesis committee and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the College of Graduate Studies.

4/22/93
Date

Chairperson, Graduate Committee

Approved for the Major Department

4/26/93
Date
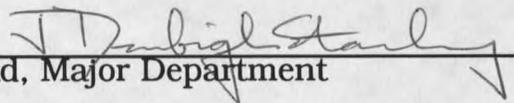
Head, Major Department

Approved for the College of Graduate Studies

4/30/93
Date

Graduate Dean

STATEMENT OF PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a master's degree at Montana State University, I agree that the Library shall make it available to borrowers under rules of the Library.

If I have indicated my intention to copyright this thesis by including a copyright notice page, copying is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for permission for extended quotation from or reproduction of this thesis in whole or in parts may be granted only by the copyright holder.

Signature _____

Date _4/27/93___

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

A method is proposed for the extraction of data from digital images of microorganisms. Digital images of microorganisms have certain characteristics which make information extraction difficult. These characteristics are identified, and four steps are suggested to effectively overcome the obstacles inherent in these images. The steps are image enhancement, microbe feature extraction, microbe feature representation, and microbe recognition and enumeration. Examples of the techniques which comprise these steps are shown and each technique's effectiveness is described.

CHAPTER 1

INTRODUCTION

Image Processing is a field which has grown rapidly over the last decade. Digital image processing can be defined as the procedures used to manipulate digital images to:

1) have improved pictorial information for human interpretation; [Gonzalez]

2) extract data autonomously.

There have been many advances in image processing techniques and new techniques are being devised all the time. Yet, even with these advances, little is known about how to apply these techniques to many specific problems. Almost all problems require more than one image processing step in order to attain the desired result. This is due to the complex nature of imaging problems. Therefore, the formulation of systematic approaches is vital to solve these problems efficiently. These formulations could be compared and blended into one image processing strategy which would more closely resemble the human visual system. It is toward this goal that this paper tries to better define one specific image processing problem and the corresponding systems.

The problem which will be addressed involves collecting information from digital images of microorganisms. This type of information is useful for researchers in microbiology, environmental engineering, health, and many other related fields. The problem is to develop procedures for manipulating images and extracting information in a structured and robust manner, so that they can be applied to many applications. This means that the process must not only be effective for solving a given imaging problem, but also consistent over a range of problems involving microorganisms.

The goal is to extract data from images. There are two categories of data that need to be collected for this problem. These are 1) quantitative data and 2) qualitative data. These two categories are not independent of each other, but are simply used as a means to entirely describe the data being extracted. Under the category of quantitative data fall such things as the width, height, number, and position of microbes in the image. Shape, texture and classification of microbes are qualitative in nature. These two data categories comprise the goal of image processing as applied to images of microorganisms.

It is beneficial to understand the hindrances or obstacles which may present themselves during the data extraction process. The images may vary in many different ways; microorganisms are often different sizes and shapes. This makes finding one effective process more difficult. Images

can be captured at any number of different magnifications. Thus, a single variety of microorganism can appear at almost any size, depending on the scale of the image. Image collection factors also add to the inconsistent condition of images. In many cases, images contain one or more of the following image collection problems:

1. Resolution

If the image resolution is not great enough, important detail will be lost, making data retrieval more difficult. Increasing resolution through magnification may be necessary if certain types of data are to be collected.

2. Lighting

Areas with too much or too little light may fade out or hide detail, whereas areas with uneven lighting may hinder feature extraction. Images should be collected with as even lighting and as much contrast as possible.

3. Blur

Blur can be caused by a combination of the two problems listed above (Resolution and Lighting) or by movement of some part of the image collection mechanism. This movement might be the

microscope, the slide, or the microorganisms themselves. These parts of the image collection mechanism should be kept as uniform as possible to enable the greatest amount of information to be transferred to the image.

4. Noise

Noise is extraneous image data which has polluted the image collection process at some point. For example, noise can be introduced by the image capturing camera, or in the process of digitization.

5. Experimental Conditions

Experimental Conditions, such as warping of the media to be imaged and varying surface characteristics, cause unwanted visual effects to occur in the image.

These are the elements (size, shape, scale, and collection factors) which make images vary greatly. If the effect of these elements can be reduced, either before or after processing, the problem of extracting data becomes increasingly simple.

There are other obstacles besides image variation which can complicate the solution. One of these involves the use of processing

resources. Processing time and space will often need to be minimized. The method used to solve this problem needs to be independent of processing time and space available as much as possible. Having a certain limitation of processing time or space should not restrict one from applying this approach. The last obstacle which should be mentioned is that of human interaction. The interaction of humans at various points in the process of data extraction can be very useful. For instance, a human might better be able to adjust a certain parameter of an imaging function so that separation of cells is the most apparent. This interaction may not be available or desired, but also could be necessary in some cases. Thus, the process for obtaining microorganism data should not depend on, nor rule out the use of human interaction.

The last step in defining the problem, is to analyze the tools one has available to build a solution. Even though tools are definitely part of the solution, they are also part of the problem. The main tool used to implement this process will be a serial-based algorithmic language (in this case the C language). This makes certain aspects of image processing more difficult since visual data seems better suited for parallel processing. However, the approach which will be presented in this paper will hopefully allow for any implementation of the steps involved, whether they be serial-based, parallel-based, or some combination of the two.

The problem can be summarized as consisting of three components.

These are:

1) The Goal

- Extract the desired quantitative and qualitative data from digital images.

2) The Obstacles

- Varying image data including characteristics of microorganisms(size, shape, etc.), scale of the images, and image collection factors;
- Utilizing processing resources efficiently;
- Allowing for but not depending on human interaction.

3) The Tool

- A serial-based algorithmic programming language.

Given these basic components, an approach can be formulated. The methods developed in this approach are required to be both systematic and flexible. In order to be systematic the approach must attack the problem in a logical manner, and have ways of overcoming each obstacle. To do this, the desired output should be well defined. The image processing steps will depend on what kind of data is to be extracted. For example, the steps for finding the positions of microorganisms will differ from the steps for enumerating the microorganisms. Secondly, the problem must be broken down into smaller more manageable components. Each step should correspond to a certain sub-goal, each being fairly simple in nature. Lastly, the steps involved in the process must make measured strides toward the goal without the loss of significant data. One step in an image processing

problem might be to find the edges of the objects in an image. If this function is performed carelessly, valuable feature information, which may be needed in a upcoming step, is lost.

The second requirement is flexibility. This means that the process should allow for different types of input data and not be effective only on specific images. This requirement should not be difficult to meet if the range of images is chosen in a reasonable manner. If the input space is too large, the approach will lose its effectiveness.

CHAPTER 2

OUTLINE OF APPROACH

The data extraction process consists of a series of steps, where each step consists of certain types of operations which may be performed on the input images. Note that the process may end after any step, depending on the type of data to be extracted. For example, visual clarity problems can normally be solved by applying the techniques in the first step.

A formal definition of a gray scale image is an N x M array where each element in the array contains an integer value which approximates the continuous image f(x,y). [Gonzalez]

$$
\begin{bmatrix}
f(0,0) & f(0,1) & \cdots & f(0, M-1) \\
f(1,0) & f(1,1) & \cdots & f(1, M-1) \\
\cdot & & & \\
\cdot & & & \\
\cdot & & & \\
f(N-1, 0) & f(N-1, 1) & \cdots & f(N-1,M-1)
\end{bmatrix}
$$

This definition of an image will be assumed in all processing examples that follow.

Image enhancement is the first step in the approach. Image enhancement consists of all operations which globally reduce unwanted image information and emphasize the pertinent image data. Image enhancement differs from later steps in that it is more preparation than manipulation. The main goal of enhancement techniques is to process an image so that the result is more suitable for a specific application. [Gonzalez] Specific image enhancement operations include:

A. Histogram Processing and Contrast Adjustment

- These operations affect how light or dark the image appears.

B. Image Thresholding

- Thresholding tries to separate the foreground data (objects) from the background data.

C. Smoothing Filters

- These filters tend to blur the image and reduce the amount of noise in the image.

D. Sharpening Filters

- These filters highlight detail and emphasize object separation from background.

The second step in the approach is feature extraction. The microorganisms in images have certain features which can be differentiated from the surrounding image data by means of image processing operations. The operation used to find these features is:

Edge Detection

    - Edges between objects and background are found. Edges can be defined as places in an image where pixels differ in some specified way (intensity, texture, etc.)

Feature Representation is the next step in the approach. Feature Representation takes the features gathered in the last step and processes them so they are normalized and easier to manipulate. The success of this step is dependent on how effective the feature extraction step was. This step dependency may be found at each level of the approach. The operations involved in this step include:

A. Edge Reconstruction

    - Edges Shapes are built as closed forms so that they may be more easily represented.

B. Shape Number Formulation

- The shape of an object is recorded in a normalized way through the formulation of chain codes and shape numbers.

C. Signature Formulation

- The shape of a candidate microbe is represented by its signature.

The last step in the process is microbe recognition. In this step, the information collected in the previous steps is examined and the objects are classified as microorganisms, undesired objects, or noise data. The operations which make up this step are:

A. Classification of Microbes

- Objects are identified and placed in the appropriate category.

B. Enumeration of Microbes

- Objects of the same classification type are counted.

C. Position Identification

- The x and y coordinates of the microbes are identified.

In summary, the steps which comprise the data-collecting process are:

1. **Image Enhancement;**
2. **Microbe Feature Extraction;**
3. **Microbe Feature Representation;**
4. **Microbe Recognition.**

The work which has been done in these areas will be presented in the next chapters. This will provide a framework by which many of the microbial imaging problems can be solved.

CHAPTER 3

IMAGE ENHANCEMENT

Histogram Processing and Contrast Adjustment

The first component of image enhancement which will be discussed is histogram processing. The histogram gives information about the probability of an occurrence of a certain gray-level. [Gonzalez] Histogram equalization attempts to use these probabilities to produce an enhanced image with groups of highly probable pixels spread out among the available pixel levels. It does this by using a transformation function equal to the cumulative distribution of the original probability density function which is derived from the histogram.

For instance, an image where the background and microbe gray-levels are similar is shown in Figure 1. Notice that many of upper gray levels are being wasted, that is, they could be used to emphasize differences in lightness and darkness in the image. After equalization, the entire pixel value range is represented (Figure 2). The histograms which correspond to these images are shown in Figures 3 and 4. The process of histogram
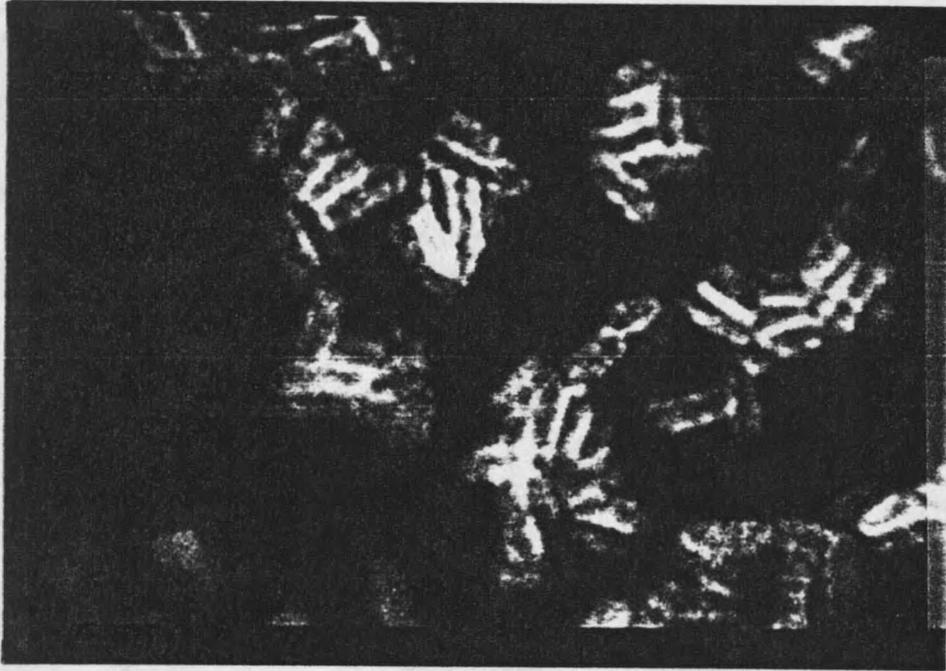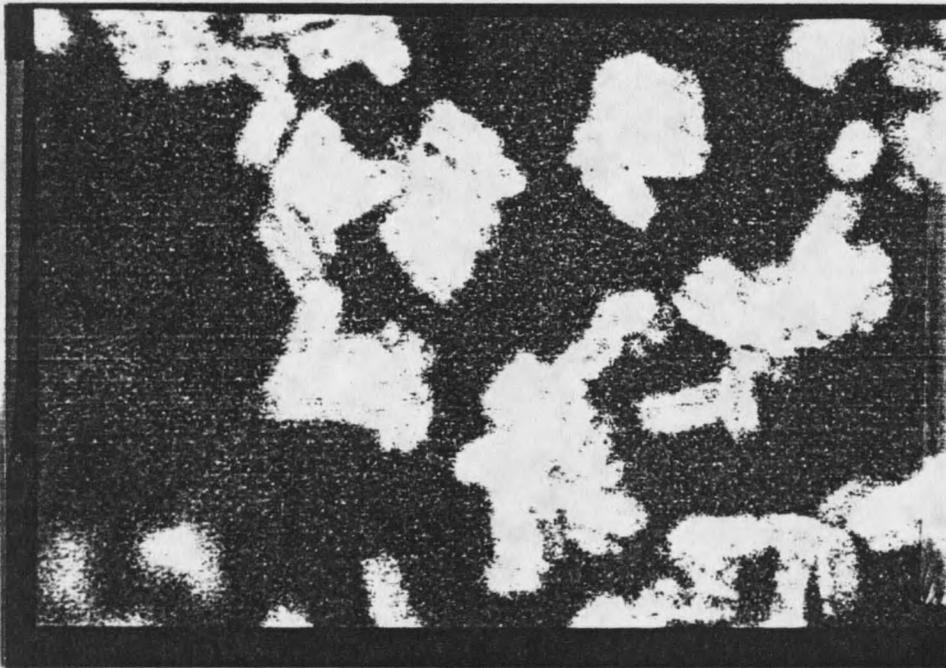
14

Figure 1 • Image Before Equalization

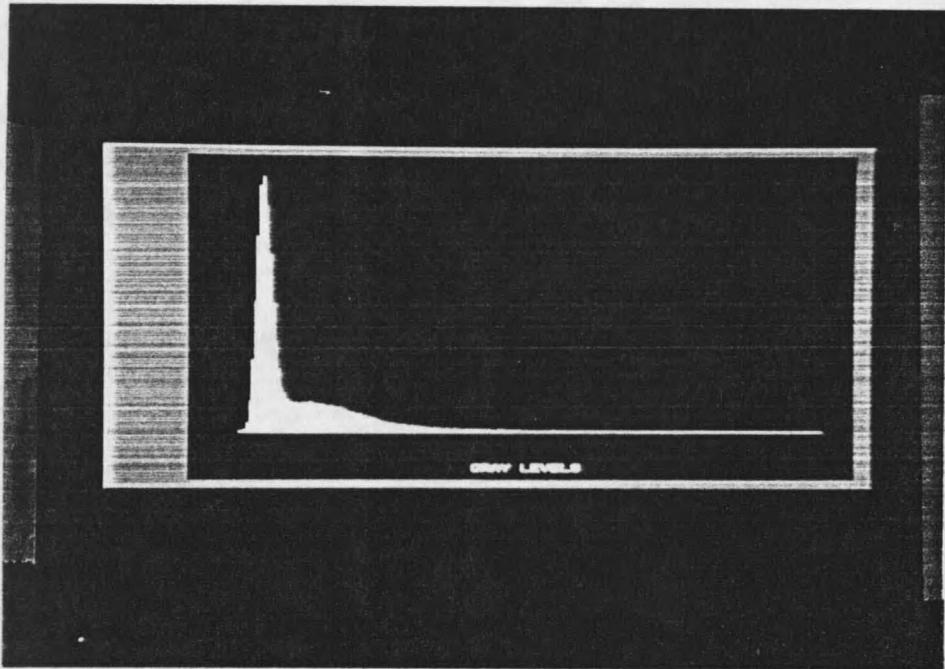**Figure 2 • Image After Equalization**

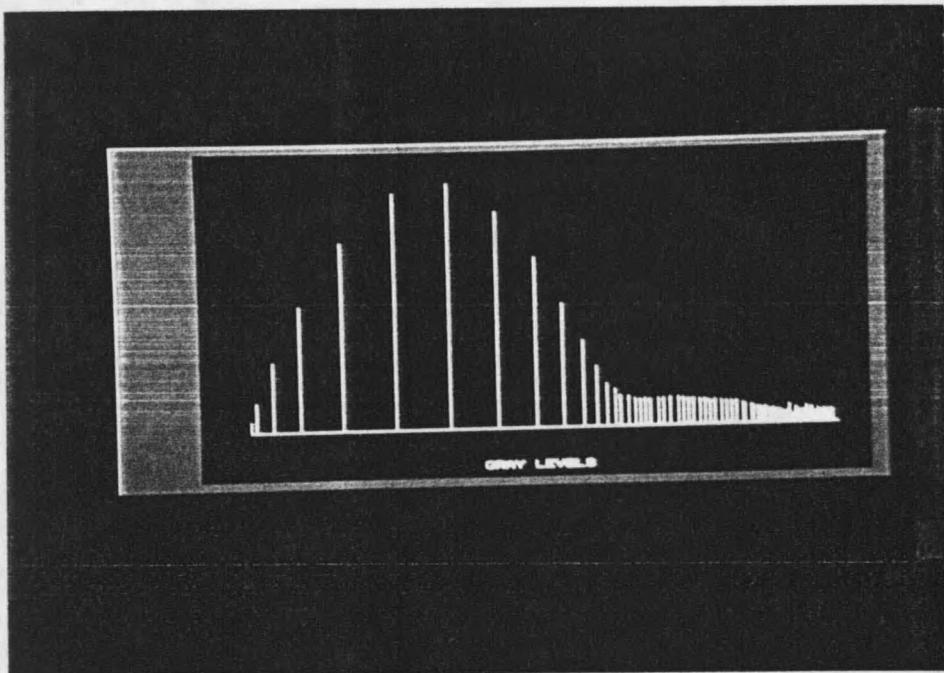**Figure 3 • Histogram of Low Contrast Image**

Figure 4 • Histogram of Image After Equalization

equalization may be described in the following way:

1. Obtain the histogram by enumerating each pixel which falls in a given gray-level.

2. Calculate the probability density function based on this histogram. Values will now be between 0 and 1.

3. Calculate the cumulative density function by computing the accumulated probability for each level.

4. Create a transformation function based on this cumulative density function by scaling it to the range of possible pixel values.

5. Apply the transformation function to each pixel in the image.

Histogram equalization is effective for increasing contrast without human interaction. Other methods allow a person to specify a desired histogram modification and have it applied. This would be useful if one knew the gray-level range of features which needed to be emphasized. The easiest way to do this is to:

1. Pick the gray level range to be emphasized;

2. Create a transformation function which adjusts the pixels in this range so they are highly contrasting with the levels which are not to be emphasized;

3. Apply the transformation function.

This process is known as *gray level slicing.* An example of a slicing transformation function and its resulting image are shown in Figures 5 and 6.
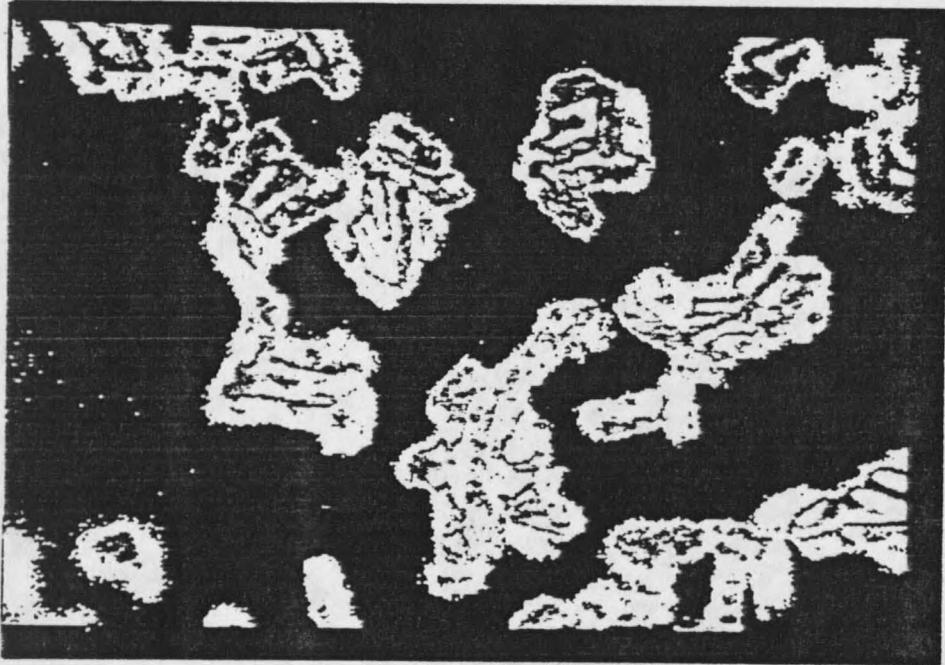
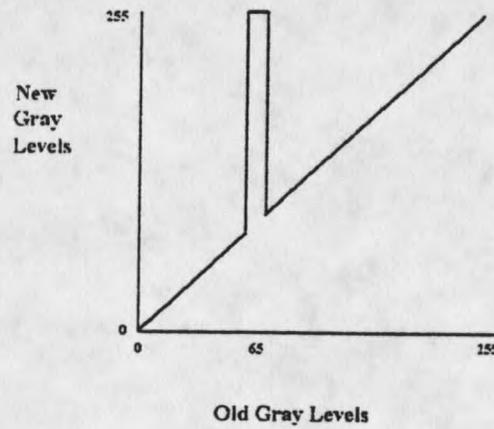**Figure 5 • Gray Level Slicing Image**

Figure 6 • Gray Level Slicing Transformation Function

A more advanced technique for emphasizing certain gray-levels is histogram specification. In this process:

1. The levels of the original image are equalized as described earlier.

2. A new histogram which has a desirable shape is constructed interactively.

3. The transformation function to equalize an image based on this new histogram is computed.

4. The inverse of this transformation function is applied to the image equalized in step 1 to create an image with the desired histogram.

This process is useful if equalization does not contrast effectively or if there are localities in the histogram which need contrast enhancement.

Thresholding

The techniques described above are all examples of contrast adjustment. Each of them involved taking the original pixel value and assigning it a new value which will hopefully emphasize certain desired characteristics. Taking this idea to the extreme, one could simply divide the gray levels into two or more groups and assign each pixel to one of the groups. This is the concept behind *thresholding.* Thresholding distinguishes pixels that have higher gray values from pixels that have lower gray values. [Haralick] This process is very useful in microbial image processing since it is often the case that there are only two or three major

gray-level ranges in these images. The background is often in one range (and usually the largest), and the microbes are in another range. Thus, if these two ranges can be effectively separated, further processing will be much easier. The difficult part of thresholding an image is picking the point or points which will be used to divide up the image. This can be done interactively by viewing thresholded results using various threshold values and choosing the best. This technique can be effective since a threshold which seems to best emphasize the microbes without losing valuable edge information can be chosen. If human interaction is not available, a threshold can be automatically chosen by algorithmic methods. One such method involves minimizing the within-group variance. [Otsu] If the histogram is bimodal, the histogram thresholding problem is to determine a best threshold separating the two modes of the histogram from each other. [Haralick] One method is to pick a threshold for which the weighted sum of group variances is minimized, where the weights are the probabilities of the respective groups.

This technique is described in the following steps:

1. At each possible threshold value, sequentially compute the within-group variance based on the equation:

$$\sigma^2{}_w(t) = q_1(t)\,\sigma^2{}_1(t) + q_2(t)\,\sigma^2{}_2(t)$$

where   t = the threshold value being considered;

q$_1$(t) = the probability for the group with values less than or equal to t;

q$_2$(t)= the probability for the group with values greater than t;

σ$_2$w(t)= the weighted sum of group variances;

σ$^2_1$(t)= the variance for the group with values less than or equal to t;

σ$^2_2$(t)= the variance for the group with values greater than t.

[Otsu]

2. Choose the threshold value which produces the smallest within-group variance value;

3. Apply a transformation function based on this threshold.

Another way of computing this thresholding value is to minimize the *Kullback Information Distance* [Kittler, Illingsworth]. The procedure is similar to minimizing the within-group variance except the following equation is used:

$$H = \frac{1 + \log(2\pi)}{2} - q_1 \log(q_1) - q_2 \log(q_2) + \frac{1}{2}\left(q_1 \log(\sigma_1^2) + q_2 \log(\sigma_2^2)\right)$$

where q and σ have the same meaning as in the within-group variance equation and H is the variable to be minimized.

Both of these techniques can be effective in choosing a threshold if the modes which comprise the image are Gaussian in nature, and are not disrupted by interfering distributions. Examples of microbial images thresholded using these two techniques are shown in Figures 8 and 9.