2013 Hawaii University International Conferences
Education & Technology
Math & Engineering Technology
June 10th to June 12th
Ala Moana Hotel, Honolulu, Hawaii

CUMULATIVE DISTANCE HISTOGRAMS AND THEIR APPLICATION TO THE IDENTIFICATION OF MELANOMA

JACK STANGL

CHERI SHAKIBAN

University of St. Thomas

CUMULATIVE DISTANCE HISTOGRAMS AND THEIR APPLICATION TO THE IDENTIFICATION OF MELANOMA

JACK STANGL, CHERI SHAKIBAN

Abstract. This research focuses on the mathematical detection and analysis of border irregularity in skin lesions. In particular, it utilizes cumulative distance histograms and statistical methods to compare the border of a malignant melanoma sample to the border of a nevus, or common mole. We propose that melanoma possess distinguishable border differences from nevi, often undetectable to the human eye. We utilize mathematical methods to detect and quantize this difference for diagnosis. The following research relies heavily on computer vision, the calculation of histograms, as well as curve fitting and residual analysis. The above will be discussed in detail throughout the following sections.

1. Introduction

Melanoma, the most serious type of skin cancer, develops in the cells that produce melanin - the pigment that gives your skin its color [1]. It is worth noting that this cancerous skin lesion is capable of spreading throughout the body, making it difficult to treat in advanced cases. In addition, visual similarities between melanoma and nevi make diagnosis difficult, and often require the use of an invasive skin biopsy to discern between them. Dermatologists often use the ABCD method to determine the necessity of a skin biopsy. This research focuses on B, or *border irregularity*, a trait presumably unique to melanoma. We believe this trait to be detectable using mathematical analysis, in particular, cumulative distance histograms. Using a high resolution image, we attempt to aid dermatologists in detecting this border irregularity, with the intent of reducing expensive and unnecessary skin biopsies. In addition to increased efficiency and reduced diagnostic costs, this technology could even be implemented on a consumer level.

2. Border Irregularity

Not all melanoma have such drastic border irregularity as shown in [Figure 1]. Likewise, not all nevi possess such a smooth border. The distinction between them is often far less noticeable to the human eye, which motivates the use of computer vision. The process begins with a number of high resolutions photographs. A border detection program, provided by Dan Hoff [2] processes the image [Figure 2]. The program is run in Matlab, and is used to determine the outermost border of the skin lesion.

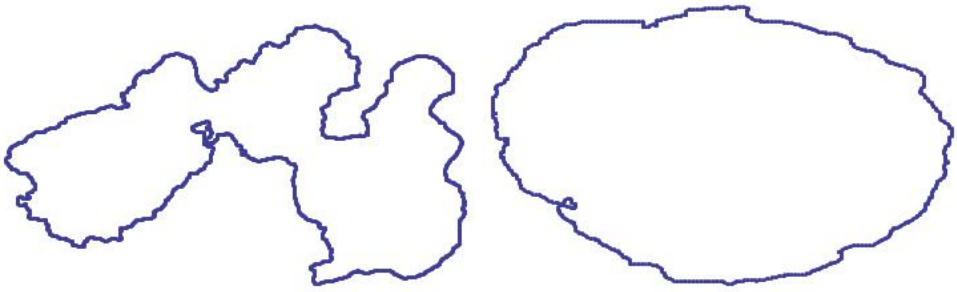

FIGURE 1. Mathematica generated scatter plots of melanoma (left) and nevus (right)

FIGURE 2. Matlab generated images of melanoma from a border detection program, courtesy of Dan Hoff [2].

The output of this program consists of a matrix M of cartesian coordinates (x_i, y_i).

$$M = \begin{bmatrix} x_{(1)} & y_{(1)} \\ x_{(2)} & y_{(2)} \\ \vdots & \vdots \\ x_{(i)} & y_{(i)} \\ \vdots & \vdots \\ x_{(n-1)} & y_{(n-1)} \\ x_{(n)} & y_{(n)} \end{bmatrix}$$

such that $n > 100$ and $M_{(i)}$ represents row i of M. This cartesian data set is then exported as a text file for further calculation with Mathematica.

3. Scaling

In the realm of photography, no image is the same. Even when the subject and photographer remain unchanged, it is impossible to retain the same angle, focus, and scale between pictures. Due to this inconsistency, a method of scaling is required to ensure that user bias does not show in the data from the border

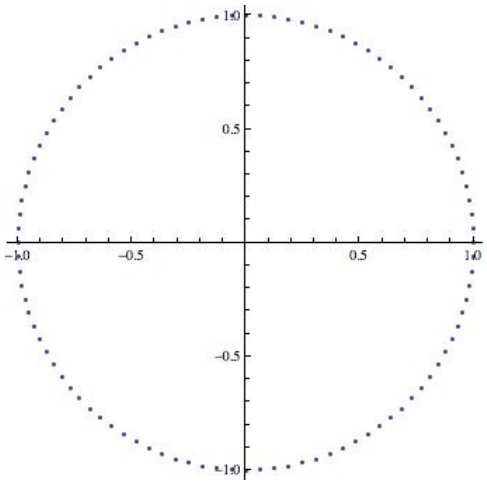

FIGURE 3. Mathematica generated set of data points representing circle $x^2 + y^2 = 1$.

detection program. The most influential bias that can show up in the Invariant Histogram method is from the size of the sample border. As such, each list of data points exported from the border detection program is scaled using the largest distance l from the centroid to a data point. Specifically, each coordinate (x_i, y_i) is multiplied by a scaling factor $s = \frac{1000}{l}$. The largest distance from the centroid to a border data point is then 1000 units.

4. Distance Histograms

A distance histogram is a bar graph formed from a number of sampled distances across a figure. It is an intermediate step to achieving our final goal of creating a *cumulative* distance histogram. To explore the concept of a distance histogram, we step away from melanoma and nevi and look instead at a simpler example. Consider the circle $x^2 + y^2 = 1$. We represent this circle using 100 data points [Figure 3], and store them in M as mentioned previously. A sample distance d_i is calculated by taking $\|M_a - M_b\|$ such that a and b are random integers between 1 and n. This is the distance between two randomly selected points on the circle. Each sampled distance d_i is then stored in $\vec{f_1}$.

$$\vec{f_1} = \begin{bmatrix} d_{(1)} \\ d_{(2)} \\ \vdots \\ d_{(i)} \\ \vdots \\ d_{(m-1)} \\ d_{(m)} \end{bmatrix}$$

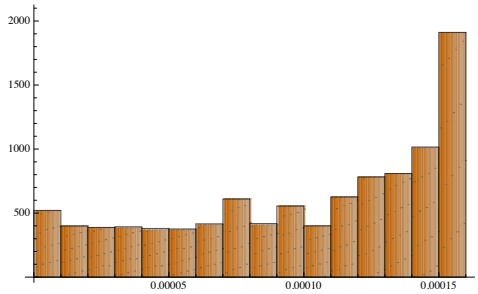

FIGURE 4. Distance histogram h. x axis: distance intervals. y axis: distance count within each interval of length k. In this example, k is chosen by Mathematica to give the most visually pleasing result. We will specify our own value of k in later calculations.

This iterative process continues until $\vec{f_1}$ contains at least $n \cdot 10$ sampled distances. For the circle example, 1000 sampled distances are recommended. This value (number of distances in $\vec{f_1}$) is represented by m. Under a specified interval length k, a distance histogram h can be created from $\vec{f_1}$ [Figure 4]. Distance histogram h is an intermediate, visual step that is not necessary for the creation of a cumulative distance histogram. It is however helpful to understand the difference between the two, which is why both are created.

5. Cumulative Distance Histograms

A cumulative distance histogram is then created from our array of distances $\vec{f_1}$. Similar to a distance histogram, the cumulative distance histogram has bars, or levels. However, it is organized such that each bar builds on the previous one, hence the name *cumulative* distance histogram.

For the calculation of a distance histogram, we allowed Mathematica to choose a visually pleasing value of k. However, for the calculation of a *cumulative* distance histogram, we care less about its visual appearance, and more about its mathematical consistency. Thus, we specify our own value for the interval length k. The magnitude we choose for k depends on the situation. A rough estimate can be found by calculating $k \mid k \leq \frac{1}{100} \cdot \max \vec{f_1}$. In the circle example, $k = 10^{-5}$. Once an interval length k has been selected, an interval count operation is performed (BinCount in Mathematica) to organize the data into q intervals of length k. q and k are related by $q = \lceil \frac{\max \vec{f_2}}{k} \rceil$. The distance count for each interval is stored in $\vec{p}$.

$$\vec{p} = \begin{bmatrix} p_{(1)} \\ p_{(2)} \\ \vdots \\ p_{(i)} \\ \vdots \\ p_{(q-1)} \\ p_{(q)} \end{bmatrix}$$

The data (no longer distances, now distance *counts*) must be normalized and stored in $\vec{f_2}$. To perform this operation, we need the sum s of the bin counts stored in $\vec{p}$,

$$s = \sum_{i=1}^{q} \vec{p_i} = \{p_1 + p_2 + \cdots + p_i + \cdots + p_{q-1} + p_q\}$$

$$\vec{f_2} = \frac{\vec{p}}{s_2} = \begin{bmatrix} {p_{(1)}}^* \\ {p_{(2)}}^* \\ \vdots \\ {p_{(i)}}^* \\ \vdots \\ {p_{(q-1)}}^* \\ {p_{(q)}}^* \end{bmatrix}$$

where ${p_{(i)}}^*$ is $p_{(i)}$ normalized. From here we form the cumulative distance histogram $C(x)$,

$$C(x) = \sum_{i=1}^{x} {p_{(i)}}^* = \{{p_1}^* + \cdots + {p_i}^* + \cdots + {p_x}^*\}$$

Applied to the previous circle example, the cumulative distance histogram can be seen in [Figure 5]. There are a number of ways to compare cumulative distance histograms. In this paper, we will explore linearity and logistic fitting, however first we apply what we have covered so far to melanoma and nevus samples. To recap, we measured distances across a figure, and performed the intermediate step of creating a distance histogram. We then chose an interval length k, and counted the number of distances that fall into each step. After normalizing the distance counts, we created a cumulative distance histogram $C(x)$ from the data. We now apply this method to melanoma and nevi, and look for contrasting qualities.

6. Application to Melanoma and Nevi

Up until this point, we have only calculated one cumulative distance histogram $C(x)$ for a circle. The following calculation involves 33 melanoma samples and 33 nevi samples provided by MoleMap [3], a dermatology corporation in New Zealand. Cumulative distance histograms for melanoma samples are stored in $R_1(x), R_2(x), ...R_{32}(x), R_{33}(x)$. For nevi samples, $G_1(x), G_2(x), ...G_{32}(x), G_{33}(x)$ are used. We now derive $R_1(x)$ for one melanoma sample. The process is the same for all other samples, both melanoma and nevi.

First, store the data points (x_i, y_i) to matrix M, assign the number of data points to n, and calculate the required number of distance measurements to m.

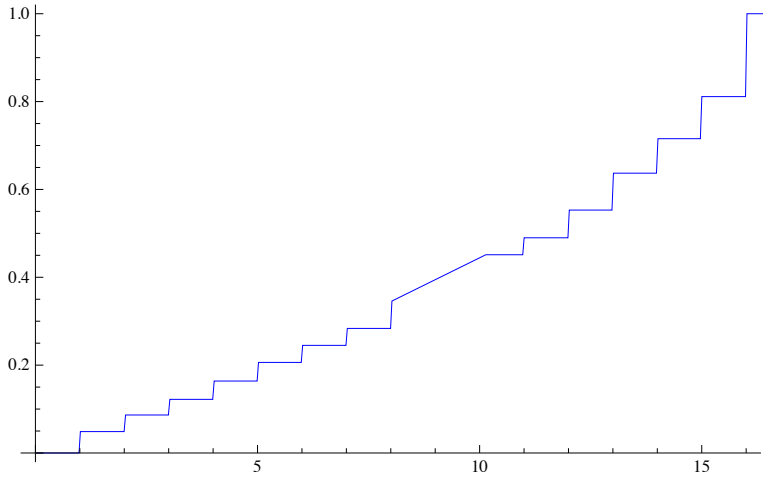

FIGURE 5. Cumulative distance histogram $C(x)$. x axis: distance intervals. y axis: normalized cumulative distance count.

$$M = \begin{bmatrix} 80.1 & -244 \\ 80.1 & -243 \\ \vdots & \vdots \\ x_{(i)} & y_{(i)} \\ \vdots & \vdots \\ 558.9 & -162 \end{bmatrix}$$

$$n = 1615$$
$$m = 10 \cdot n = 16150$$

Calculate the distances d_i and store in $\vec{f_1}$. Previously, we defined each distance d_i as $\|M_a - M_b\|$, where a and b are random integers between 1 and n.

$$\vec{f_1} = \begin{bmatrix} d_{(1)} = 216.821 \\ d_{(2)} = 228.37 \\ \vdots \\ d_{(i)} \\ \vdots \\ d_{(1614)} = 99.1234 \\ d_{(1615)} = 198.903 \end{bmatrix}$$

Choose interval length k, and determine number of intervals $q \mid q = \lceil \frac{\max \vec{f_2}}{k} \rceil$.

$$k \leq \tfrac{1}{100} \cdot \max \vec{f_2}$$

$$k \leq \tfrac{1}{100} \cdot 2.262 \cdot 10^{-4}$$

$$k \approx 10^{-6} \leq 2.262 \cdot 10^{-6}$$

$$q = \lceil \tfrac{\max \vec{f_2}}{k} = \tfrac{2.262 \cdot 10^{-4}}{10^{-6}} \rceil = 227$$

Find the distance count for each interval p_i.

$$\vec{p} = \begin{bmatrix} p_{(1)} = 28 \\ p_{(2)} = 36 \\ \vdots \\ p_{(i)} \\ \vdots \\ p_{(236)} = 3 \\ p_{(237)} = 3 \end{bmatrix}$$

Find the sum of the distance counts, normalize them, and store in $\vec{f_2}$.

$$s = \sum_{i=1}^{q=237} \vec{p_i} = 10000$$

$$\vec{f_3} = \frac{\vec{p}}{s_2 = 10000} = \begin{bmatrix} {p_{(1)}}^* = {}^7/_{2500} \\ {p_{(2)}}^* = {}^9/_{2500} \\ \vdots \\ {p_{(i)}}^* \\ \vdots \\ {p_{(q-1)}}^* = {}^3/_{10000} \\ {p_{(q)}}^* = {}^3/_{10000} \end{bmatrix}$$

Calculate $R_1(x)$.

$$R_1(x) = \sum_{i=1}^{x} {p_{(i)}}^*$$

Repeat the process for $R_2(x), R_3(x) ... R_{32}(x), R_{33}(x)$, and $G_2(x), G_3(x) ... G_{32}(x), G_{33}(x)$. The results can be seen in [Figure 6]. The presence of 66 results imposed on one graph can be hard to work with visually. In an attempt to mathematically discern a difference between the melanoma and nevi samples, the linearity of each cumulative distance histogram was analyzed. [Figure 7] is a measure of the residuals for each sample and its fitted linear regression.

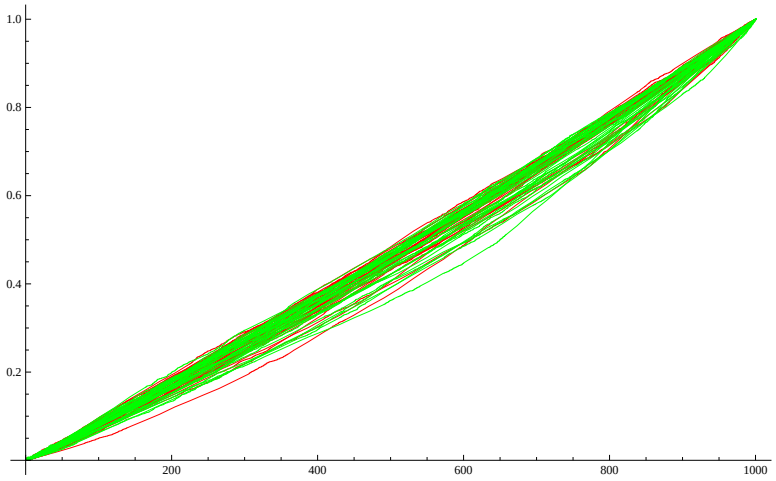

FIGURE 6. Cumulative distance histograms for melanoma $R_1(x), R_2(x), ...R_{32}(x), R_{33}(x)$ and nevi $G_1(x), G_2(x), ...G_{32}(x), G_{33}(x)$. Melanoma samples are plotted in red, nevi in green. x axis: distance intervals. y axis: normalized cumulative distance count.

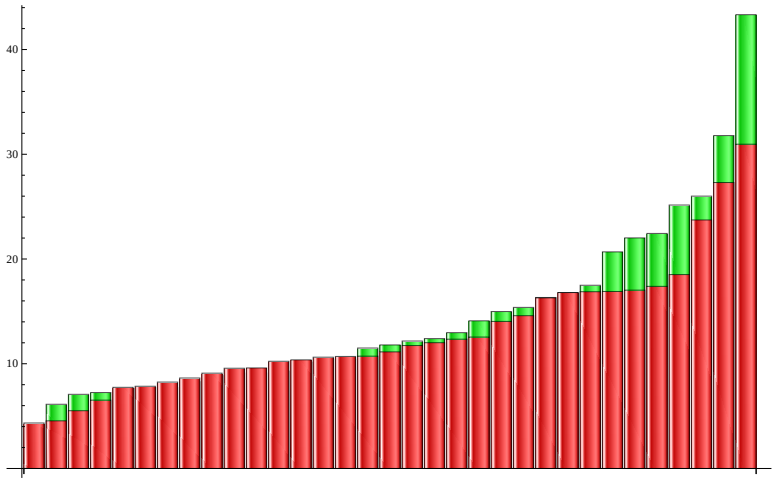

FIGURE 7. Bar graph comparing residuals for linear functions fitted to melanoma and nevi cumulative distance histograms. Melanoma samples are in red, nevi samples in green.

7. Centroid Method

The search for alternative approaches using the Invariant Histogram method led to the involvement of the centroid in the calculation of the cumulative distance histogram. Rather than sampling distances from one randomized point to another, it was discovered that the histogram possessed significantly different features when

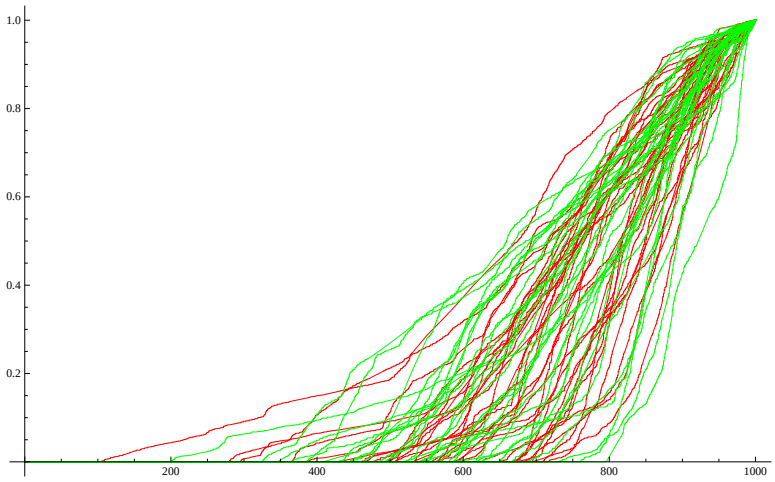

FIGURE 8. Cumulative distance histograms generated using the centroid method. Melanoma samples are in red, nevi samples in green.

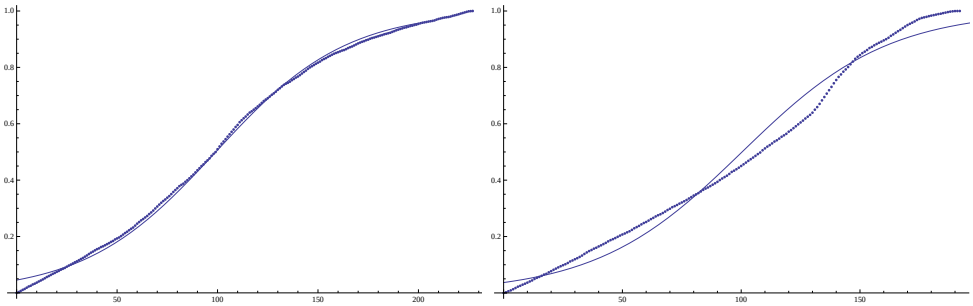

FIGURE 9. Fitted logistic functions for a melanoma sample (left) and nevus sample (right).

using distances measured from the centroid to a randomized point. The 66 results from this approach can be found in [Figure 8].

While comparing melanoma and nevi histograms individually, it was noted that many of the melanoma samples resembled a logistic curve, while many of the nevi samples had opposite concavity at their inflection points [Figure 9]. This difference motivated the use of logistic fitting and residual analysis.

8. Logistic Fitting

We first need a data set (x_i, y_i) to fit a logistic function to. This comes directly from the cumulative step function, and simply involves the exporting of the data set used to plot the steps. Mathematica has a built in logistic fitting command that is ideal for this situation.

Residuals are then calculated for the sample sets. Each bar in [Figure 10] represents the sum of the residuals for each fitted logistic curve and its corresponding cumulative distance histogram.

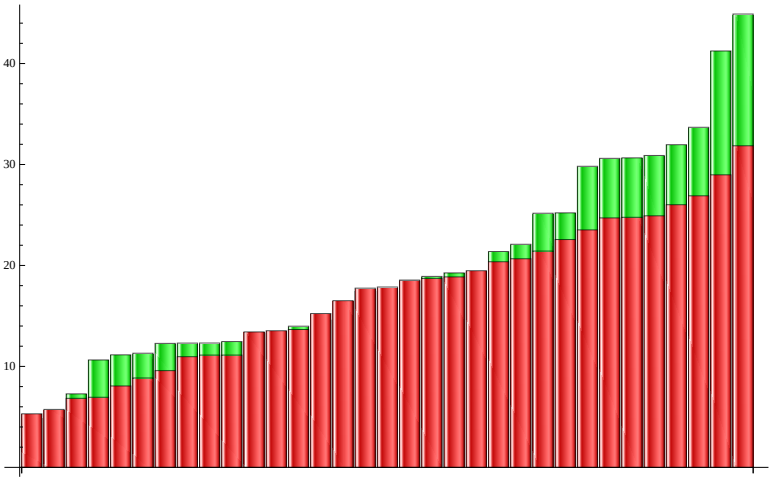

FIGURE 10. Bar graph comparing residuals for logistic functions fitted to melanoma and nevi cumulative distance histograms using the centroid method. Melanoma samples are in red, nevi samples in green.

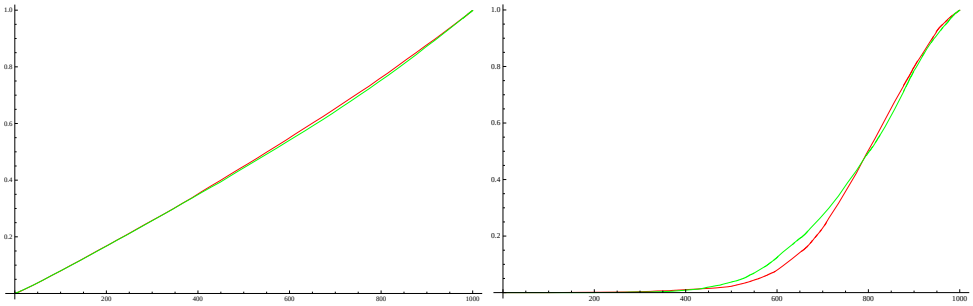

FIGURE 11. A comparison of average cumulative distance histograms using the standard method of point to point sampling (left) and using the centroid method of point to center sampling (right). Melanoma samples are in red, nevi samples in green.

9. Average Cumulative Distance Histograms

The last avenue explored using this method involves the calculation of an *average* cumulative distance histogram for the melanoma set and for the nevi set. This avenue was explored with the intent to compare each melanoma and nevus sample individually to the average cumulative distance histograms. It could then be determined whether each sample best fits the average melanoma histogram or the average nevus histogram. A comparison of the average cumulative distance histograms can be found in [Figure 11].

The average cumulative distance histograms were found by averaging the bin counts on each interval. Visually, this is the averaging of the vertical jumps, or y values, in the cumulative distance histograms. While comparing individual samples

with the average cumulative distance histograms, it was determined that 24 of the 33 melanoma samples more closely resemble the average melanoma histogram, while only 14 of the 33 nevi samples more closely resemble the nevi average histogram. Using the centroid method, 21 of the 33 melanoma samples more closely resemble the melanoma average histogram, while 16 of the 33 nevi sample more closely resemble the nevi average histogram.

10. Future Research

The results thus far have not been promising for this method, and due to the small sample size of only 33 melanoma and 33 nevi, we are hesitant to do any further analysis with this method. In the future, we hope to obtain a much larger sample size. Once this occurs, we will be able to obtain much broader results for the invariant histogram method, and more easily be able to identify areas of the processes that need improvement. In addition, other unrelated methods such as signature curves and ellipse fitting are being explored in the hopes that ventures into geometry and curvature can uncover unique, contrasting qualities between melanoma and nevi.

11. Acknowledgements

We extend our thanks to MoleMap for providing the sample images used in this research. We would also like to thank Dr. Cheri Shakiban for her guidance throughout this project, Dan Hoff for permitting the use of his border detection program, the National Science Foundation (CSUMS grant #260077), and the University of St. Thomas Center for Applied Mathematics for supporting this research.

References

[1] Mayo Clinic. "Melanoma." 2 June 2010. Web. 21 September 2011.
$< http://www.mayoclinic.com/health/melanoma/DS00439 >$
[2] Dan Hoff, Graduate Student, University of Minnesota. Full reference not yet available.
[3] MoleMap Dermatology, New Zealand. Full reference not yet available.