Applications of Cumulative Histograms in Diagnosing Breast Cancer

Anna Grim

Abstract

We present several algorithms that use invariant cumulative histograms to non-invasively diagnose tumors detected on a mammogram. First, we define three specialized cumulative histograms called the cumulative centroid, kappa, and kappa-s histogram. Then we compute metrics over each cumulative histogram to quantitatively distinguish benign versus malignant tumors. Our methodology has been tested on a dataset of 150 tumors and we include an ROC analysis of our results.

1 Introduction

Non-invasive diagnosis of breast tumors is challenging because benign and malignant tumors detected on a mammogram can be indistinguishable to the human eye. Despite malignant tumors having a more irregularly shaped contour, visual assessment of the tumor by a human is subjective and unreliable for an official diagnosis. Thus, surgical incision and histological examination of the tumor is the standard diagnostic procedure. However, due to the large number of mammograms performed each year, this leads to many unnecessary procedures and increases the risk of false positive diagnosis. We propose a cumulative histogram based methodology that automatically diagnoses a tumor detected on a mammogram. The methodology is based on the observation that benign tumors present on a mammogram with an elliptically shaped contour as seen in Figure 1. In contrast, malignant tumors have finger-like proliferations along the tumor contour called spiculations, which give these tumors an irregularly shaped contour as seen in Figure 2 [5],[13].

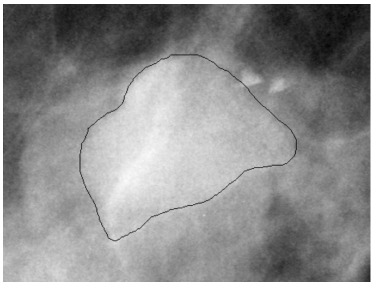

Figure 1: Benign contour

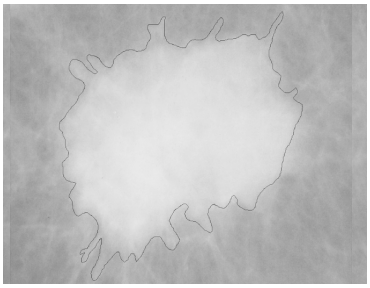

Figure 2: Malignant contour

This paper is an application of the work done by Olver in [3], where he defines a cumulative distance histogram as follows.

Definition 1. *The cumulative distance histogram of a finite set of points* $\{p_1, \ldots, p_n\}$ *is the function* $\Lambda : \mathbb{R}^+ \to \mathbb{N}$ *defined by*

$$\Lambda(r) = \frac{1}{n^2} \#\{(i,j) : d(z_i, z_j) \leq r\},$$

where d *is the Euclidean distance metric.*

The graph of a cumulative distance histogram provides a representation of the geometric shape of a set of points, which is invariant under rigid transformations. For this reason, distance histograms have been used in object-based query along with color and angle histograms [9]. In this application, a distance histogram is computed using distances between the center of mass and points along the contour of an object extracted from an image or video frame. This shape information can then be used to query the content of images and videos. Another application of cumulative histograms is to measure border irregularity in skin lesions. In this application, a cumulative distance histogram is computed over the border of a skin lesion and the shape of the histogram is used to determine a diagnosis of malignant or benign. In this paper, we will extend the work done in [2] and [10] by defining a cumulative histogram called a *cumulative centroid histogram* that uses centroid rather than arbitrary distances in Section 2.1. Then we will define two additional cumulative histograms called the *cumulative kappa histogram* and *cumulative kappa-s histogram* in Section 2.2 that profile a contour' s curvature. In Section 3, we will summarize our results and provide an ROC analysis of these methods.

2 Methodology

2.1 Cumulative Centroid Histogram

Let $P = \{p_1, ..., p_n\} \subset \mathbb{R}^2$ be a finite set of points with centroid p_c such that $p_i = (x_i, y_i)$.

Definition 2. *Two points* $p_i, p_j \in P$ *are collinear with the centroid* p_c *if*

$$det \begin{pmatrix} x_i & y_i & 1 \\ x_j & y_j & 1 \\ x_c & y_c & 1 \end{pmatrix} = 0$$

The use of the determinant in Definition 2 is geometrically motivated because the determinant is the area of the parallelogram determined by the points p_i, p_j, and p_c. When the determinant is zero, then the parallelogram is degenerate and the two points are collinear with the centroid.

Proposition 2.1. *Collinearity is invariant under uniform scaling and defines an equivalence relation over* P.

Proof. If two points p_i and p_j are collinear with p_c, then

$$det \begin{pmatrix} x_i & y_i & 1 \\ x_j & y_j & 1 \\ x_c & y_c & 1 \end{pmatrix} = 0$$

by Definition 2. If we scale the points by some $\lambda \in \mathbb{R}$, then the determinant as calculated in (1) would remain to be zero. To prove the equivalence relation, any point p_i is collinear with itself and the centroid p_c because the determinant as calculated in Definition 1 would not have full rank and consequently have determinant zero, so reflexivity holds. Collinearity preserves symmetry and transitivity because geometrically these points all lie on the same line through the centroid. □

When P is a Jordan curve, then each point is collinear with at least one other point on P. The situation is more complicated when P has a finite number of points. For example, if the distribution of points on P is sparse, then there may be no pairs of collinear points. However, we can prove a condition that guarantees when each point is collinear with another point for a discrete contour.

Proposition 2.2. *If $\#P$ is finite and all points in P are uniformly spaced with respect to angular position from the centroid, then each point on P is collinear with another point on P if and only if the parity of $\#P$ is even for $\#P > 1$.*

Proof. We begin by proving the converse of our claim. Since collinearity is invariant under uniform scaling by Proposition 2.1, then we can contract P to a circle $\tilde{P}$ and preserving every collinear relationship. Since the points in P are uniformly spaced with respect to angular position, then $\tilde{P}$ must be the vertices of a regular polygon. If the parity of $\#P$ is even, then each point is collinear with another by the symmetry of an even sided regular polygon. The forward direction of our claim holds again by the symmetry of a regular polygon. If the parity of $\#P$ is odd, then there does not exist a pair of points that are collinear through the centroid in $\tilde{P}$ because $\tilde{P}$ is an odd sided regular polygon. □

Although Proposition 2.2 provides criteria for when each point is collinear to another on a discrete contour, it is unlikely that an arbitrary discrete contour will satisfy the hypothesis of this proposition. However, we assume that the distribution of points on P is relatively dense, so for purposes it suffices to relax our definition of collinearity by instead using *approximate collinearity.*

Definition 3. *The measure of collinearity φ between two points p_i and p_j is defined to be*

$$\varphi(i,j) = \left| det \begin{pmatrix} \tilde{x}_i & \tilde{y}_i & 1 \\ \tilde{x}_j & \tilde{y}_j & 1 \\ x_c & y_c & 1 \end{pmatrix} \right|,$$

where $\tilde{p}_i$ and $\tilde{p}_j$ denote p_i and p_j rescaled along the line $\overline{p_i p_j}$ such that $\|\tilde{p}_i - p_c\| = \|\tilde{p}_j - p_c\| = 1$.

Definition 4. *Two points p_i and p_j are approximately collinear if $\mu(i,j) = 1$ such that*

$$\mu(i,j) = \begin{cases} 1, \textit{ if } j = argmin_k\{\varphi(i,k) : 1 \leq i,k \leq n\} \\ 0, \textit{ otherwise.} \end{cases}$$

The function in Definition 4 is used to pair each point on P with another point that is nearest to being collinear. This definition of approximate collinearity suffices because we assume that the point distribution on P is relatively dense. However, it may no longer be true that approximate collinearity defines an equivalence relation over P because reflexivity and transitivity may not hold. Now that each point is collinear with another, we can define the cumulative centroid histogram.

Definition 5. *The centroid function of a finite set of points $P \subset \mathbb{R}^2$ is the function $\eta : \mathbb{R}^+ \to \mathbb{N}$ defined as*

$$\eta(r) = \#\{(i,j) : \|p_i - p_j\| = r \textit{ and } \mu(i,j) = 1\}.$$

Definition 6. *The cumulative centroid histogram of a finite set of points $P \subset \mathbb{R}^2$ is the function $\Lambda_c : \mathbb{R}^+ \to [0,1]$ defined to be*

$$\Lambda_c(r) = \frac{1}{n} \sum_{s<r} \eta(s).$$

Since we assume $\#P$ to be finite, then the support of $\eta(r)$ must also be finite which implies that only a finite number of nonzero terms are summed in the definition of $\Lambda_c(r)$. In Section 1, we defined the cumulative distance histogram $\Lambda(r)$ which is constructed using arbitrary rather than centroid distances. In the application of diagnosing breast tumors using only the contour, we have found that the cumulative centroid histogram is more accurate than using a cumulative distance histogram. To provide some reasoning for this claim, we provide an example of the differing shape between a $\Lambda(r)$ versus $\Lambda_c(r)$ computed over the same set of points.

Example 2.1. *Let $Q = \{q_1, \ldots, q_m\}$ be a finite set of points with even cardinality from a circle of radius R with points uniformly spaced with respect to angular position.*

Since Q satisfies the hypothesis in Proposition 2.2, then for each point on Q there exists another point that is collinear. This guarantees that the cumulative centroid histogram has non-empty support. We have included a plot of $\Lambda(r)$ versus $\Lambda_c(r)$ computed over Q and show the graphs in Figure 3, respectively. The cumulative centroid histogram provides a better shape representation of a circle because all points on a circle are an equal distance from the origin, which is reflected in the histogram and stated as Proposition 2.3. Since benign tumors tend to have circular and elliptically shaped contours, then a cumulative centroid histogram provides a better shape representation than a cumulative distance histogram.

Proposition 2.3. *For the point configuration Q, there exists a unique r^* such that $\Lambda_c(r) = 1$ for $r > r^*$ and $\Lambda_c(r) = 0$ for $r < r^*$.*

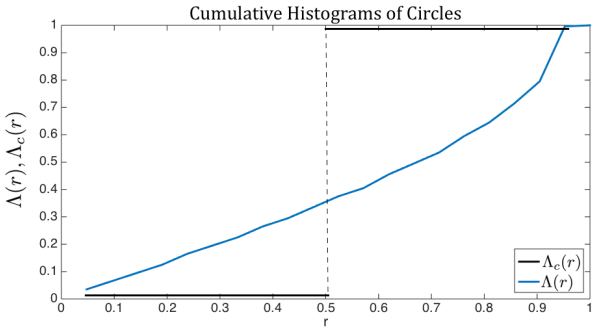

Figure 3: $\Lambda_c(r)$ and $\Lambda(r)$ of a circle

For our second example, we show a comparison of a cumulative distance histogram and cumulative centroid histogram computed over a benign and malignant contour that is shown is Figures 4 and 5, respectively. In these figures, we show an interpolation of $\Lambda_c(r)$ defined over its support. The cumulative centroid histogram provides a better representation of the irregularity in a malignant contour by having more frequent and quick changes in concavity. This shape is distinct from the cumulative centroid histogram of a benign tumor which has a characteristic "s" shape with a single and gradual change in concavity. Our metric Ω for diagnosing tumors will be based on this difference in frequency and intensity of each change in concavity on the cumulative centroid histogram. In short, Ω is defined by finding the maximal value in the approximate second derivative of $\Lambda_c(r)$ near each change in concavity and determining the number of maximal values within a given threshold.

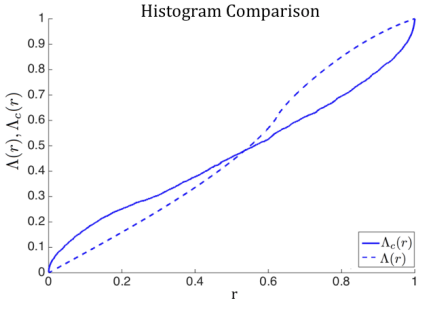

Figure 4: $\Lambda_c(r)$ and $\Lambda(r)$ from a benign contour

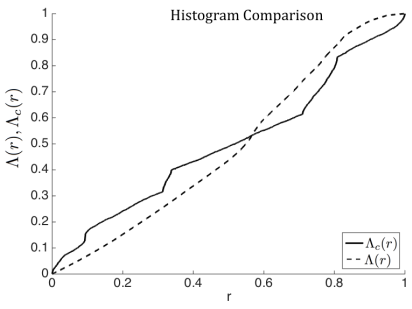

Figure 5: $\Lambda_c(r)$ and $\Lambda(r)$ from a malignant contour

We will begin by computing a second derivative of $\Lambda_c(r)$, but must be careful because $\Lambda_c(r)$ is a step function with finite support. Let $\tilde{\Lambda}(r)$ be a linear interpolation determined by the support of $\Lambda_c(r)$, choose some small $\epsilon > 0$ and define the approximate second derivative to be

$$\Lambda_c''(r) = \left| \frac{1}{\epsilon^2} \big(\tilde{\Lambda}_c(r-\epsilon) - 2\tilde{\Lambda}_c(r) + \tilde{\Lambda}_c(r+\epsilon) \big) \right|.$$

Since $\Lambda_c''(r)$ is a numerical approximation, we define a change in concavity as a point $r^\star$ such that $\Lambda_c''(r^\star) < \delta$ for some small $\delta > 0$. Next, partition the domain [0,1] of $\Lambda_c(r)$ with respect to the points where $\Lambda_c''(r) < \delta$ so that $[0,1] = \bigcup_{i=1}^{n} U_i$ and let $U = \{U_1, \ldots, U_n\}$. We will define the characteristic function ω which will be used in the definition of the metric Ω.

Definition 7. *Let ω be the characteristic function defined over some interval V and value $\lambda \in \mathbb{R}$ such that*

$$\omega(V, \lambda) = \begin{cases} 1, \textit{ if } \lambda \in V \\ 0, \textit{ otherwise.} \end{cases}$$

Definition 8. *Let Ω be defined over some interval V and the set of intervals U determined by the changes of concavity of $\Lambda_c(r)$ such that*

$$\Omega(U, V) = \sum_{i=1}^{n} \omega(V, max\{\Lambda_c''(U_i)\}).$$

The interval V in the definition of Ω allows us to use a grading system to measure the changes in concavity on a cumulative centroid histogram. In practice, we use three-grade system with the intervals $V_1 = (10^{-8}, 10^{-7})$, $V_2 = (10^{-7}, 10^{-6})$, and $V_3 = (10^{-6}, \infty)$. In general, we observe that for benign tumors either $\Omega(U, V_1) = 1$ or $\Omega(U, V_2) = 1$ with $\Omega(U, V_3) = 0$ because they have a single, gradual change in concavity. In contrast, $\Omega(U, V_2)$ and $\Omega(U, V_3)$ are large for malignant contours because there are more frequent and quick changes in concavity.

2.2 Cumulative Kappa and Kappa-S Histograms

Let $Z = \{z_1, ..., z_n\} \subset \mathbb{R}^2$ be an ordered finite set of points, which we identify as the discrete approximation of a smooth curve $C \subset \mathbb{R}^2$. We begin by calculating the approximate curvature $\tilde{\kappa}$ at each point $z_i \in Z$ by selecting points $z_{i-1}, z_{i+1} \in Z$, forming the triangle illustrated in Figure 10 [10,11]. Let $\triangle$ represent the signed area of the triangle formed by z_{i-1}, z_i, z_{i+1} and s represent the semi-perimeter, so that $\triangle = \pm\sqrt{s(s-a)(s-b)(s-c)}$ with $s = \frac{1}{2}(a+b+c)$ [10]. The approximate curvature at z_i follows as

$$\tilde{\kappa}(z_i) = 4\frac{\triangle}{abc} = \pm 4\frac{\sqrt{s(s-a)(s-b)(s-c)}}{abc} \text{[10]}. \tag{1}$$

To approximate the first derivative of curvature with respect to arc length, $\tilde{\kappa}_s$, take the points $z_{i-2}, z_{i+2} \in Z$ and approximate curvature at z_{i-1}, z_{i+1} using (6). Then the approximate derivative of curvature at z_i is given by

$$\tilde{\kappa}_s(z_i) = \frac{3(\tilde{\kappa}(z_{i+1}) - \tilde{\kappa}(z_{i-1})}{2a + 2b + d + e} \tag{2}$$

with $d = \|z_{i-1} - z_{i-2}\|$ and $e = \|z_{i+2} - z_{i+1}\|$ [11].

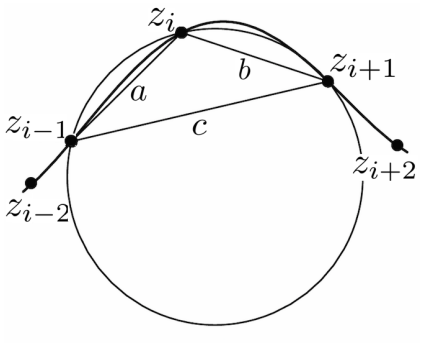

Figure 6: Approximate curvature at z_i

Definition 9. *The curvature histogram of a finite set $Z \subset \mathbb{R}$ is the discrete function*

$$\psi(k) = \#\{z_i : \tilde{\kappa}(z_i) = k\},$$

where $1 \leq i \leq n$ and the derivative of curvature histogram $\psi(k_s)$ is defined in the same manner.

Since curvature is not invariant under uniform scaling, then we must remedy this problem by renormalizing the curvature histogram. In the spirit of Definition 4, we renormalize the curvature histogram into a cumulative kappa histogram.

Definition 10. *The cumulative kappa histogram $\Psi(k)$ of a finite set $Z \subset \mathbb{R}^2$ is the discrete function*

$$\Psi(k) = \frac{1}{n} \sum_{s \leq k} \psi(k)$$

and the cumulative kappa-s histogram $\Psi_s(k)$ is defined in the same manner.

Since curvature and consequently its derivative are defined with respect to distance and distance is invariant under rigid motions, then the cumulative kappa and kappa-s histograms are invariant under rigid transformations of Z. After calculating the cumulative kappa and kappa-s histograms, we observe that the area under the cumulative kappa histogram calculated from a malignant contour is much larger than the area from a benign contour. The contrast is depicted in Figures 7 and 8, where the range and initial derivative of the cumulative kappa

histogram for a malignant contour is much larger. We will define a metric ζ by constructing a step function from Ψ and integrating over this function. The support of Ψ is finite because the tumor contour is discrete, so call this set $R = \{r_1, \ldots, r_n\}$.

Definition 11. *Let $\bar{\Psi}(r)$ be the step function defined as*

$$\bar{\Psi}(r) = \begin{cases} 0 & \text{if } r = 0 \\ \Psi(r_i) & \text{if } r = r_i \\ \Psi(r_{i+1}) & \text{if } r \in (r_i, r_{i+1}). \end{cases}$$

and let $\bar{\Psi}_s$ be the step function defined with respect to Ψ_s

Definition 12. *Let ζ be the measure defined over a finite set of points Z be the function*

$$\begin{aligned} \zeta(Z) &= \int_R \bar{\Psi}(r) \\ &= \sum_{i=1}^{N-1} \bar{\Psi}(r_i)(r_{i+1} - r_i) \end{aligned}$$

and ζ_s be defined similarly with respect to Ψ_s.

In general, the value of ζ is large when Z corresponds to a malignant contour in comparison to a benign contour. This pattern is a caused by spiculation, which skews the distribution and increases the variance of curvature and derivative of curvature values on a malignant contour. The distribution is skewed towards zero because there are significantly more points where either $\kappa = 0$ and $\kappa_s = 0$, which results a steep initial slope in Ψ and Ψ_s. The variance of Ψ and Ψ_s is large due to the irregular shape of a malignant contour. Therefore, a large value in $\zeta(Z)$ indicates that Z is more likely to correspond to a malignant contour.

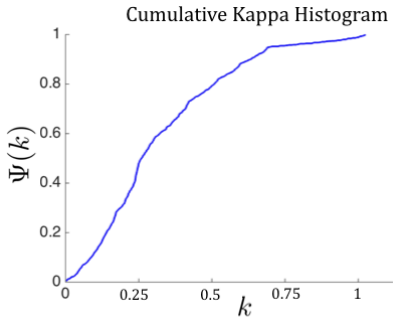

Figure 7: $\Psi(k)$ of a benign contour

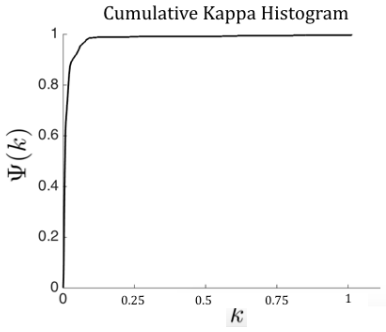

Figure 8: $\Psi(k)$ of a malignant contour

3 Results

3.1 Data Set

The data set contains 78 benign and 78 malignant mammograms diagnosed by expert radiologists. Atypical tumors comprise approximately 10% of the data set with seven spiculated benign and nine circumscribed malignant tumors. The mammograms were downloaded from the University of South Florida Digital Database for Screening Mammography and the Mammographic Image Analysis Society [12,13]. Each mammogram is between 512×512 and 1024×1024 pixels and was taken with either a Lumysis or Howtek scanner. The database provided an official diagnosis and delineation of each tumor contour drawn by radiologists. After downloading the mammograms, each image is individually discretized into a set of approximately 500 (x,y) points using active contour segmentation [14,15].

Figure 9: Benign tumor contours

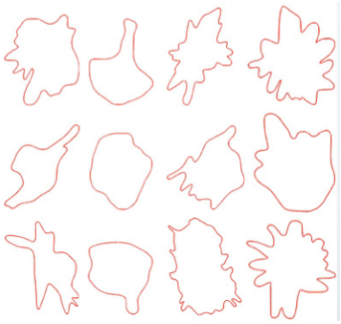

Figure 10: Malignant tumor contours

3.2 ROC Analysis

We used the metrics defined in Sections 2.1 and 2.2 to define two distinct decision trees to diagnose a tumor as benign of malignant. The first decision tree is defined with respect to three-grade system described at the end of Section 2.1 and the second decision tree is defined with respect to metrics ζ and ζ_S. Next, the calculate a receiver operating characteristic (ROC) curve from the decision trees, which is a plot of the true positive rate against the false positive rate. The area under the ROC curve indicates the accuracy of our methodology to correctly diagnose benign and malignant tumors. In Figure 14, the sensitivity and specificity refer to the true positive and false positive rate, respectively. The measure is an objective assessment of the accuracy of our algorithms and objectively compares our methodology against existing automated algorithms. We have defined a correct diagnosis as identifying typical malignant, atypical malignant, and atypical benign tumors as malignant and identifying typical be-

nign tumors as benign. Since atypical benign tumors closely resemble malignant tumors, a biopsy should be clinically tested as a precautionary measure.

The ROC values of the cumulative centroid histogram and kappa, kappa-s histogram methodologies are 0.983 and 0.966, respectively. In the automated diagnosis literature, we have found our algorithms to be equally or more accurate than existing methodologies. Rangayyah and Nguyen used the 1D and 2D ruler box counting fractal dimension to obtain an ROC curve values ranging from 0.83-0.89 [8]. In addition, they also developed algorithms using compactness, fractional concavity, spiculation index, and Fourier-descriptor-based factor, which obtained ROC curve values ranging from 0.77-0.93 [17]. Another study by Chen, Chung, and Hun used fractal features in an image processing texture analysis using fractals, where they obtained an ROC curve value of 0.88 [4].

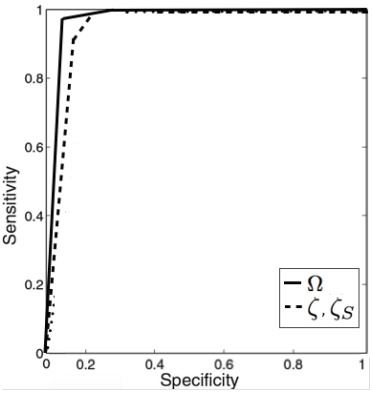

Figure 11: ROC Analysis

4 Conclusion

The results obtained in this research study show that cumulative histograms can be used to diagnose tumors detected on a mammogram. Cumulative histograms can be used to diagnose tumors by accentuating differences in the shape of benign and malignant tumor contours. Some future work may be to apply this methodology to distinguishing between moles and melanomas, whose contours contrast by the degree of irregularity.

5 Acknowledgments

This work was funded by a CSUMS grant number DMS0802959 from the National Science Foundation in collaboration with the University of St. Thomas.

References

[1] Boutin, M., Numerically invariant signature curves, *International Journal of Computer Vision* **40** (2014).

[2] Brinkman, D., and Olver, P.J., Invariant Histograms, *American Mathemtatics Monthly* **119** (2012), 4-24.

[3] Calabi, E., Olver, P., Shakiban, C., Tannenbaum, A., and Haker, S., Differential and numerically invariant signature curves applied to object recognition, *Int. J. Computer Vision* **26** (1998), 107-135.

[4] Chen, D., Classification of breast ultrasound images using fractal features, *Journal of Clinical Imaging* **29** (2005), 235-245.

[5] DeBerardinis, R., The biology of cancer: metabolic reprogramming fuels cell growth and proliferation, *Cell Metabolism* **7.1** (2008), 11-20.

[6] Lankton, S., Hybrid geodesic region-based curve evolutions for image segmentation, *International Society for Optics and Photonics* (2007), 65104U-1.

[7] Lankton, S., and Tannenbaum, A., Localizing region-based active contours, *Image Processing, IEEE Transactions* **17.11** (2008), 2029-2039.

[8] Rangayyan, R. and Nguyen, T., Fractal Analysis of Contours of Breast Masses in Mammograms,*Journal of Digital Imaging* **20.3**, 223-237.

[9] Saykol, E., Gudukbay, U., and Ulusoy, G., A Histogram-based Approach for Object-based Query-by-Shape-and-Color in Image and Video Databases, *Image and Vision Computing* **23** (2005), 1170-1180.

[10] Shakiban, C., and Stangl, J., Cumulative Distance Histograms and their Application to the Identification of Melanoma, preprint.

[11] Suckling, J., (1994): The Mammographic Image Analysis Society Digital Mammogram Database Exerpta Medica. International Congress Series 1069.

[12] University of South Florida. University of South Florida Digital Mammography Home Page. http://marathon.csee.usf.edu/Mammography/Database.html.

[13] Vogelstein, B., and Kinzler, K., The multistep nature of cancer, *Trends in Genetics* **9.4** (1993), 138-141.