A Study of Shape Distributions for Estimating Histologic Grade

Jasper Z. Zhang*, Sokol Petushi†, William C. Regli*, Fernando U. Garcia† and David E. Breen*
* Drexel University † Drexel University College of Medicine

Abstract— Breast cancers can be histologically categorized (graded) based upon their architectural patterns and cellular types. Inaccurate histologic grading can result in inappropriate treatment for a given patient. Computational analysis of breast cancers offers an operator-independent method for histologic grading that should enhance grading reliability. We present the initial efforts to develop computational technologies that may be used to automatically and objectively estimate the histologic grade of breast cancer tumors. The approach utilizes image processing and shape analysis of imaged histologic sections. Our work is based on the hypothesis that cellular structures found in breast cancer tumors can be transformed into distinct high-dimensional shape distributions using geometric measures from stochastic geometry. The resulting shape distributions define well-populated regions of the associated high-dimensional space. Mapping an unknown breast cancer sample into this high-D space and determining to which region it belongs will allow for the automatic estimation of its histologic grade.

I. INTRODUCTION

B reast cancers can be histologically categorized (graded) based upon their architectural patterns and cellular types. With the advent of new therapies that depend on specific characteristics of tumors and the known lack of intra- and inter-reproducibility of histologic grading [3], new methods that can better identify tumor characteristics and give consistent histologic grading are needed. To cope with these issues, more and more pathologists have turned to automating their work through computational histology. But as research has shown, it is difficult to extract meaningful information from a histologic image [8]. This paper describes a method for acquiring biomedical trait information from segmentations of histologic images, which can then be used to estimate histologic grade for new unknown cases.

Our approach to automated histologic grading is based on image processing and shape analysis of imaged histologic sections. The approach assumes that the spatial structure of the invasive cancer cells found in breast tumors can be transformed into distinct high-dimensional shape distributions using geometric measures from stochastic geometry. Shape distributions [5][6] represent the structure and shape of 2D/3D objects with a probability distribution produced by repeatedly and randomly applying geometric measures. Stochastic geometry focuses on analyzing and quantifying the connections between geometry and probability in order to describe and characterize small-scale structural features, large-scale spatial events, and aggregate statistical geometric properties [1][14].

The overall process that we propose is shown in Figure 1. This process takes in a learning set to define classification groups and an unknown case or set of unknown cases whose histologic grade will be estimated. First, histology specimens are scanned into digital images. The digital images are segmented to produce binary images that represent only the background and the areas of interest (invasive cancer cells in our study). The binary images are transformed into shape distributions [6], represented by histograms. The distributions are then analyzed and used to estimate the histologic grade of the unknown cases. This paper focuses on a subset of the above process, specifically the effectiveness of shape distributions to capture distinguishing features in the spatial structure of invasive cancer cells in breast tumors that may be used to estimate histologic grade.
II. IMAGE ACQUISITION AND SEGMENTATION

The images used in our study include breast cancer samples with histologic grade ranging from one to three, with no healthy specimens. The input test data included ten Grade 1, ten Grade 2 and eleven Grade 3 specimens. All images were stained using the Hematoxylin and Eosin (H&E) process, and scanned at a magnification of 10x to produce images with a 6,000 pixels\(^2\) resolution.

The next stage in the process involved segmenting the raw images to produce binary images. The segmentation stage utilizes a semi-unsupervised technique that combines optimal color-space information with structure-oriented multi-level thresholding and filtering, while incorporating high-level domain knowledge [7]. It produces binary images where white pixels represent invasive cancer cells and black pixels are background. The segmentation process utilizes a graphical user interface (see Figure 2) that allows the user to "train" the system to identify the desired cells.

III. SHAPE DISTRIBUTION GENERATION

The shape distribution generation process takes in binary images that specifies the cancer cell spatial structures and produces a one dimensional shape distribution, represented as a histogram, as its output. A number of geometric measures were implemented to generate the shape distributions. These include: inside radial contact, inside line contact, area, perimeter, area vs. perimeter, curvature, aspect ratio and major axis direction.

A. Inside Radial Contact

Inside radial contact is a geometric measure that has been utilized to represent biological materials [12] and stochastic microgeometry [11]. Applying the inside radial contact measure provides insight into the size distribution of the structures in an image by probing the image with disks. The measure determines the maximum radius of a disk centered at an arbitrary pixel inside a structure that can be fit without intersecting the structure’s boundary.

The algorithm can be efficiently implemented by applying a distance transform to the image [2]. In order to minimize memory usage we utilized a local flood-filling approach to calculating distances. The distance field is then transformed into a shape distribution by rounding the floating point distance values and collecting the resulting integer values into the nearest bucket within the representative histogram. Each bucket value in the shape distribution represents the minimum distance from a pixel inside each segmented region to the boundary of the region. Representative shape distributions are presented in Figure 3, which contains the distributions produced by applying the radial contact measure to the ten Grade 2 histology segmentation images.

B. Inside Line Contact

The inside line contact geometric measure probes the structures in an image by finding the longest line passing through two arbitrary points in a segmented region. We have implemented this via an exhaustive line sweeping algorithm. For each line passing through the complete image, we determine which of the segmented regions it intersects. The shape function measures the length of each line segment produced by intersecting the full line with the segmented regions. Each bucket in this shape distribution stores the count of how many resulting line segments have a length specified by the bucket.

Computationally, the measure calculates lines from every boundary pixel of the image to all other boundary pixels, ensuring that the start and end pixels are not on the same boundary edge. We also avoid processing duplicate lines, since a line starting from pixel A and going to pixel B will produce the same result as one from pixel B to pixel A.

C. Area, Perimeter and Area vs. Perimeter

The area measure is computationally the simplest of all metrics. It finds the area, in pixels, of all the separate white regions in the segmented images. Applying the area metric produces a distribution of the region sizes in the given segmented image.

The perimeter measure counts the boundary pixels of each
separate white region in a segmented image. A boundary pixel is a white pixel with at least one neighboring black pixel. During processing as each boundary pixel is identified, the perimeter pixel count for the region is incremented. Once the perimeter of a region is calculated the associated bin count in the perimeter histogram is incremented.

Area vs. perimeter is a measure that combines the previous two measures. This metric captures the surface to volume ratio of the segmented regions, and quantifies the irregularity of the cell shapes. It is calculated by dividing the area by the perimeter for each separate region.

D. Curvature

The curvature measure is similar to the area vs. perimeter measure in that it attempts to capture the irregularity of the individual cells. It does this by calculating a surface roughness metric at the perimeter pixels of the segmented regions. We utilize the level set curvature formulation [13] in Eq. 1 to compute curvature values at the perimeter pixels in a blurred version of the binary segmentation.

$$
\nabla \cdot \left( \frac{\nabla \phi}{|\nabla \phi|} \right) = \nabla_x \nabla_y \phi - 2 \nabla_x \nabla_y \phi_x \phi_y + \phi_x \phi_y^2 \\
(\phi_x^2 + \phi_y^2)^{3/2}
$$

(Eq. 1)

The measure is implemented by first generating a blurred gray scale image with a convolution. The binary image is passed through a Gaussian kernel one segmented region at a time. The kernel width is two pixels with a sigma of three. Once we obtain the Gaussian blurred binary image for a specific region, we calculate Eq. 1 at the perimeter pixels using the intensity values of the blurred image. The absolute value of these curvature values are used to increment the appropriate bins in the curvature histogram.

E. Aspect Ratio and Major Axis Direction

Aspect ratio is one of two measures based on eigenvalues and eigenvectors derived from the individual segmented regions. Aspect ratio is the ratio between the smallest and largest eigenvalues computed for a region, and quantifies the elongation of the cells. We divide the smallest eigenvalue by the largest to produce consistent results in the range of 0.0 and 1.0 for each separate region. The aspect ratios for each region are scaled (to be described in the next section) and used to increment the counts in the appropriate bins.

The major axis direction measure (which we have labeled eigenvector) computes the distribution of cell shape directions within an image. It takes each segmented region and calculates the dot product, which is related to the angle, between the region’s major axis and the average major axis of all regions within an image. The absolute value of the resulting dot product is scaled and then used to increment the appropriate bin in the “eigenvector” histogram.

IV. DATA PROCESSING

A. Pre-processing

A number of filters were applied to the segmented images before they were transformed into shape distributions. Segmented regions with certain properties were removed from the segmentations because they were deemed to be invalid cellular regions. Segmented regions smaller than 64 pixels and larger than 1500 pixels were filtered out before shape distribution processing. The reason for the area filter was to ensure that we are not introducing noise or segmentation errors into the shape analysis. We determined that regions smaller than 64 pixels in area are too small to be a cell, but instead are byproducts of over segmentation. The regions that are larger than 1500 pixels in area tend to be several cellular structures clumped together, an artifact of under segmentation. Another cause of large regions within the segmentation can be caused by tubular formations or other non-cellular structures within the tumor.

Regions with an aspect ratio below 1:6 were filtered from the segmentations. The aspect ratio filter was used to remove any structure that is too “skinny”. It was deemed that cells could not have an aspect ratio less than this. Any structure longer than 1:6 is most likely some kind of non-cell contaminant introduced during the grossing and slide mounting process.

B. Post-processing

In order to create histograms with consistent bin value ranges the geometric measurement values for the shape distributions from the filtered segmentations were multiplied before the binning process. It was decided that a reasonable number of bins for the shape distribution histograms would be between 100 and 350, with the exception of inside radial contact since scaling this measure would not provide an increase in accuracy. The scaling of geometric measure values was especially important for metrics with values between 0.0 and 1.0, e.g. aspect ratio. The multipliers applied to the geometric measures outputs are listed in Table 1.
Additionally, since the bin counts for the distributions could range from 2 or 3 to tens of millions, the individual bin values were converted to a logarithmic scale before numerical comparisons were performed. The resulting, limited value range produced more stable computations during later stages of data processing.

Some of the geometric measures, e.g. curvature, produced sporadic, noisy, unreliable values in the high output ranges. We removed the noisy tails from these shape distributions in a way that standardized the bin ranges for each geometric measure’s set of shape distributions. The maximum bin location for all shape distributions produced by a particular measure is defined to be the highest bin location with the first zero bin count. We apply the same process to the head of the histogram to eliminate all sparse heads. This guarantees that every distribution produced from the same measure will start and end at the same bin locations.

V. ANALYSIS USING INFORMATION RETRIEVAL

Once the shape distributions were generated, they were analyzed to evaluate their effectiveness for capturing distinguishing spatial features in the histology specimens and correlating them to the specimen’s histologic grade. The key to judging the effectiveness of the shape distribution approach and the individual measures is to determine if they can correctly identify the histologic grade of a specimen by comparing its shape distribution to distributions of specimens of known grades.

A. Sub-region windows

Our initial observations led us to believe that only certain sub-regions of a histogram were needed to produce acceptable comparison results. We therefore performed computational studies that involved sliding a window through the bins of the shape distributions produced by a particular geometric measure. This process sweeps through the bin ranges with a window of a given size by incrementing the starting location of the window one bin location at a time. The window sizes range from 50% to 100% of the total bin range and were incremented by 10%.

For each window and for each set of shape distributions produced by a particular measure we conducted classification tests. In these tests a specimen’s shape distribution would be compared to the distributions produced by the same geometric measure. Via a voting process, the specimen would be classified. Finding the sub-region that best characterized a specimen’s histologic grade would involve determining the sub-region that could be used to most effectively identify the actual grade of the known sample.

B. K-nearest-neighbor queries

The K-nearest-neighbor (KNN) query [4] was utilized to estimate the histologic grade of a test specimen when comparing shape distributions. In the KNN query the shape distribution of an “unknown” specimen is compared against all other specimens and a dissimilarity metric is computed for each comparison. The K nearest distributions, i.e. those with the lowest dissimilarity measures, are identified. The distance between two distributions is computed with the Earth Mover’s Distance [10], and the final estimation is produced by polling the K nearest specimens of known grade. The grade that polls the highest is assigned to the tested specimen.

![Classification outcomes. TP – True Positive. FP – False Positive. FN – False Negative.](image)

We chose 4, 7, and 10 as our K values for the classification process. Using these values guarantees no three-way ties, with only two-way ties being possible. We did not use K values greater than 10 because we only have ten cases for most grades.

The dissimilarity metric, Earth Mover’s Distance (EMD), views one distribution as piles of dirt, while the other distribution as a set of holes. The distance between the distributions is defined as the minimum work needed to move the dirt of one distribution into the holes of the other distribution. This is effectively the transportation problem, and may be solved efficiently via linear programming. One of the EMD’s valuable properties is its ability to compare distributions of varying lengths.

C. Precision, Recall and F-measure

After performing KNN queries on all distributions we evaluated the effectiveness of shape distributions to accurately identify the histologic grades of the specimens for each window using the information retrieval metrics of precision, recall and F-measure [9].

Precision and recall are defined in relation to the possible outcomes of a classification operation. As seen in Figure 4, when attempting to classify a specimen, e.g. of Grade 1, three outcomes are possible. If a Grade 1 specimen is correctly assigned to the Grade 1 class this assignment is a True Positive (TP). If a Grade 1 specimen is classified to the incorrect grade, this is a False Negative (FN). If some other grade is classified as a Grade 1, this is a False Positive (FP).

Precision measures how many correct queries are performed for each histologic grade with respect to all the queries that map to the specific histologic grade. Precision, as in Eq. 2, is therefore computed by dividing the number of all correct queries (true positives, TP) by the total number of queries mapped to the grade being evaluated (TP + false positives, FP).
Precision $\left( C_j \right) = \frac{TP}{TP + FP}$  \quad (Eq. 2)

Recall, as in Eq. 3, is computed by dividing the number of all correct queries (true positives, TP) by the total number of specimens of the grade being evaluated (TP + false negatives, FN).

$$Recall \left( C_j \right) = \frac{TP}{TP + FN}$$  \quad (Eq. 3)

The F-measure, in Eq. 4, is a metric that combines both precision and recall in a harmonic mean to provide an overall classification performance measure [9].

$$F(C_j) = \frac{2 \times precision \times recall}{precision + recall}$$  \quad (Eq. 4)

The process of evaluating each window for each geometric measure consisted of computing precision, recall and F-measure for the associated classification queries. The results from the evaluation of the best windows for each grade and shape function are presented in Figures 5, 6, and 7. These figures include the best F-measure results for the sliding window process for each geometric measure and each K value used for KNN. Tables 2, 3 and 4 present the details of the best windows for each grade. The curvature metric with a 60% window between bins 103 and 207 provided the best classification of Grade 1 specimens. The area vs. perimeter measure with a 60% window between bins 43 and 82 provided the best classification of Grade 2 specimens. The inside radial contact measure with a 60% window between bins 8 and 20 provided the best classification of Grade 3 specimens.

These peak sub-regions can be visually verified in Figures 8, 9 and 10. These figures present the average of all the shape distributions for Grades 1, 2 and 3 for the curvature, area vs. perimeter and inside radial contact measures. The

---

**Table 2**

<table>
<thead>
<tr>
<th>Metric</th>
<th>Window</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Curvature</td>
<td>Window 103 to 207 (60%)</td>
<td>8/11</td>
<td>8/10</td>
<td>0.7619</td>
</tr>
<tr>
<td>Perimeter</td>
<td>Window 17 to 162 (60%)</td>
<td>6/7</td>
<td>6/10</td>
<td>0.7059</td>
</tr>
<tr>
<td>Aspect Ratio</td>
<td>Window 11 to 51 (50%)</td>
<td>7/10</td>
<td>7/10</td>
<td>0.7</td>
</tr>
</tbody>
</table>

**Table 3**

<table>
<thead>
<tr>
<th>Metric</th>
<th>Window</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area / Perimeter</td>
<td>Window 43 to 82 (60%)</td>
<td>6/6</td>
<td>6/10</td>
<td>0.75</td>
</tr>
<tr>
<td>Perimeter</td>
<td>Window 31 to 152 (50%)</td>
<td>6/8</td>
<td>6/10</td>
<td>0.6667</td>
</tr>
<tr>
<td>Eigenvector</td>
<td>Window 33 to 63 (50%)</td>
<td>7/12</td>
<td>7/10</td>
<td>0.6364</td>
</tr>
</tbody>
</table>

**Table 4**

<table>
<thead>
<tr>
<th>Metric</th>
<th>Window</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inside Radial Contact</td>
<td>Window 8 to 20 (60%)</td>
<td>8/12</td>
<td>8/11</td>
<td>0.6095</td>
</tr>
<tr>
<td>Curvature</td>
<td>Window 48 to 152 (60%)</td>
<td>9/15</td>
<td>9/11</td>
<td>0.6923</td>
</tr>
<tr>
<td>Aspect Ratio</td>
<td>Window 33 to 78 (70%)</td>
<td>8/16</td>
<td>8/11</td>
<td>0.5926</td>
</tr>
</tbody>
</table>
boxes represent the optimal sub-regions that our computational evaluation discovered. They highlight the separation between the shape distributions in these windows, which allow for the relatively successful classification of histologic grade.

VI. CONCLUSIONS AND FUTURE WORK

This paper has presented the results of our investigations into shape distributions for the analysis of segmentations of histologic images. We have described an approach to using shape distributions to estimate histologic grade, and have evaluated its effectiveness on a small number of breast cancer tumor segmentations. While our results are not conclusive, because of the small number of test specimens, they are encouraging and indicate the ability of shape distributions to provide discriminating geometric information that may be used to predict biomedical traits via analysis of histologic images.

We have described a preliminary study, therefore many other issues need further research. For example the issues of scale, magnification and resolution require additional investigation. Additional research is needed to determine which physical scale contains the most distinguishing spatial features. More powerful machine learning and clustering techniques may be employed for analysis and classification of the shape distributions. Other edit distances may also produce improved results. Future studies will benefit from more histologic segmentations for analysis. This would provide increased statistical information, and more data will move us closer to a predictive capability.

REFERENCES


Figure 8. The window of the curvature measure with the best Grade 1 classification ability. Plots represent the average of the Grades 1, 2 and 3 curvature shape distributions.

Figure 9. The window of the area vs. perimeter measure with the best Grade 2 classification ability. Plots represent the average of the Grades 1, 2 and 3 area vs. perimeter shape distributions.

Figure 10. The window of the inside radial contact measure with the best Grade 3 classification ability. Plots represent the average of the Grades 1, 2 and 3 inside radial contact shape distributions.