Virtual Microscopy and Grid-enabled Decision Support for Large Scale Analysis of Imaged Pathology Specimens

Lin Yang, *Member, IEEE,* Wenjin Chen, Peter Meer, *Senior Member, IEEE,* Gratian Salaru, Lauri A. Goodell, Viktors Berstis, *Senior Member, IEEE,* and David J. Foran, *Member, IEEE*

***Abstract*—Breast cancer accounts for about 30% of all cancers and 15assisted analysis hold promise for classifying subtypes of disease and improving prognostic accuracy. We introduce a Grid-enabled decision support system for performing automatic analysis of imaged breast tissue microarrays. To date, we have processed more than 100,000 digitized specimens (1200×1200 pixels each) on IBMs World Community Grid (WCG). As part of the Help Defeat Cancer (HDC) project, we have analyzed the data returned from WCG along with retrospective patient clinical profiles for a subset of 3744 breast tissue samples and the results are reported in this paper. Texture based features were extracted from the digitized images and isometric feature mapping (ISOMAP) was applied to achieve nonlinear dimension reduction. Iterative prototyping and testing were performed to classify several major subtypes of breast cancer. Overall the most reliable approach was gentle AdaBoost using an eight node classification and regression tree (CART) as the weak learner. Using the proposed algorithm, a binary classification accuracy of 89% and the multi-class accuracy of 80% were achieved. Throughout the course of the experiments only 30% of the dataset was used for training.**

***Index Terms*—Tissue Microarray, Texton, Grid Computing, AdaBoost**

I. Introduction

BREAST cancer is one of the leading cancers for women. It is the second most common cause of cancer death in white, black, Asian/Pacific Islander and American Indian/Alaskan native women [1], [2]. Early detection and improved therapy planning are crucial for increasing the survival rates of cancer patients.

Tissue microarray (TMA) technology makes it possible to extract small cylinders of tissue from pathology specimens and arrange them on a recipient paraffin block such that hundreds can be assessed simultaneously [3], [4]. Although TMA technology is still evolving, the underlying methods have already been tested extensively and validated for use in several key areas of cancer research. Recently, several leading research groups participated in efficacy studies in which they compared the accuracy of TMA-based analysis with assessments, which had been rendered using traditional whole tissue sections or cDNA microarrays. These findings were reported for a range of disorders including breast cancer [5], [6], prostate cancer [7] and gastric cancer [8]. It is now generally accepted that two to four samples taken from different regions of each donor tissue block provides enough information to allow reliable evaluation of the specimen.

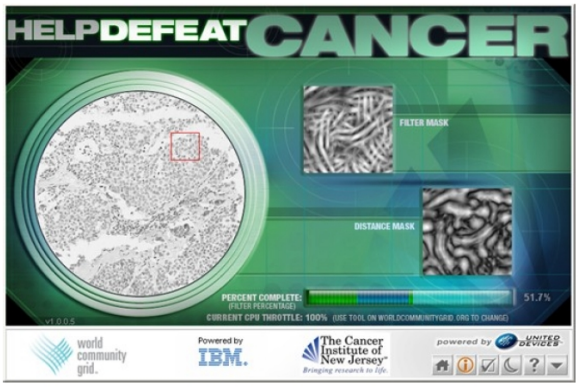


Fig. 1. A screenshot of Help Defeat Cancer (HDC) clinet running on IBM world community grid (WCG).

One of the advantages of TMA arrays is that they allow for amplification of limited tissue resources by providing the means for producing large numbers of small core biopsies, rather than generating one single specimen section. Using TMA technology, a carefully planned array can be constructed such that a 20 years survival analysis can be performed on a cohort of 600 or more patients using only $100-200$ microliters of antibody. Another major advantage of the TMA technique is that each constituent disc within a given array is treated in an identical manner in terms of incubation times, temperatures and washing conditions. Currently, the primary methods used to evaluate tissue arrays involve interactive review of specimens which are subjectively evaluated and scored. An alternate, but less utilized approach is to sequentially digitize each specimen for subsequent semi-quantitative assessment. Both strategies ultimately involve interactive evaluation of TMA samples, which is a slow, tedious process which is prone to error. Reducing the amount of time and effort to process TMA could potentially lead to acceleration of the pace of cancer research.

L. Yang is with the department of Electrical and Computer Engineering, Rutgers University, Piscataway, NJ, 08854 and The Cancer Institute of New Jersey, New Brunswick, NJ, 08903. W. Chen is with The Cancer Institue of New Jersey, New Brunswick, NJ, 08903. P. Meer is with the department of Electrical and Computer Engineering, Rutgers University, Piscataway, NJ, 08854. G. Salaru and L. A. Goodell are with the Department of Pathology, UMDNJ-Robert Wood Johnson Hospital, New Brunswick, NJ, 08903. V. Berstis is with IBM research, Austin, TX, 73301. D. J. Foran is with the Center for Biomedical Imaging & Informatics, UMDNJ-Robert Wood Johnson Medical School, Piscataway, NJ, 08854 and The Cancer Institute of New Jersey, New Brunswick, NJ, 08903.

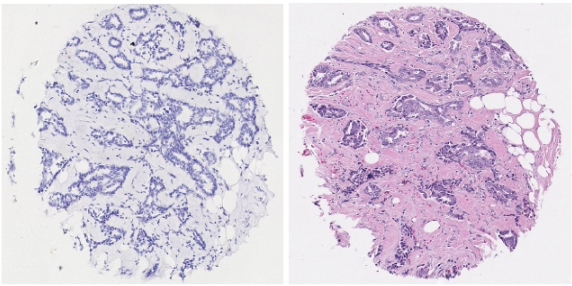


Fig. 2. Two staining examples. On the left is a hematoxylin stained tissue and on the right is hematoxylin & eosin stained tissue.

Although strides have been made towards automating some aspects of the analysis [9], [10], the full promise of TMA has not yet been realized, in part, because of a lack of reliable methods for performing large-scale quantitative comparative analysis. We recently undertook a collaborative project with IBM, the "Help Defeat Cancer" (HDC) [11] project, which enabled us to utilize the massive computational power of the World Community Grid (WCG). A screenshot of one of the thousands of distributed client computers participating in the HDC is shown in Figure 1. The WCG enabled us to demonstrate the feasibility of using spectral and spatial signatures to characterize staining characters of imaged cancer specimens. In this paper, we report a Grid-enabled framework and efficient classification algorithm for high-throughput analysis of digitized breast cancer specimens. We analyzed a subset of the data returned from WCG along with the patients' retrospective clinical profiles to discriminate among benign breast tissues and two other subtypes of breast cancer. More background information can be found on the HDC project though Wikipedia using the keyword "Help Defeat Cancer" [12].

The remainder of the paper is organized as follows: Section 2 introduces background information on the World Community Grid. In Section 3, we introduce the data generation and system framework. Section 4 explains the data analysis methods, including feature extraction, dimension reduction and classification. Section 5A provides comparative experimental results of nine binary classifiers. Section 5B shows the performance of extending the binary gentle AdaBoost algorithm to the multi-class problem. Section 6 concludes the paper.

II. World Community Grid

IBM World Community Grid [13] (WCG) is a philanthropic project which utilizes otherwise unused CPU cycles from personal computers around the world and aggregates the combined computational power. WCG was established to address challenging large scale non-profit research projects which can benefit humanity. It takes advantage of otherwise wasted energy and at the same time creates a virtual supercomputer that by some measures exceeds the capacity of traditional supercomputers. The result is that some otherwise impractical or intractable research projects can be brought to successful completion. Investigators can submit a research proposal for consideration by the WCG project committee. If approved by the advisory board, the project is run at no cost to the research team. Findings are subsequently placed in the public domain. Suitable research areas include, but are not limited to biomedical, climatology, environment, conservation and emergency preparedness.

WCG enabled the most computationally intensive components of the Help Defeat Cancer (HDC) project to run at optimal speed, thereby increasing the accuracy and sensitivity with which expression calculations and pattern recognition procedures were conducted. By harnessing the collective computational power of WCG, we were able to analyze a larger set of cancer tissue specimens than what would be possible using traditional computer resources. This added level of speed and sophistication led to improved capacity to detect subtle changes in measurable parameters, and prognostic clues which are difficult to observe by visual inspection alone.

The research proposal for the HDC project was originally submitted in August 2005. By May, 2006, the research team composed of researchers from The Cancer Institute of New Jersey, Robert Wood Johnson Medical School, Rutgers University and the University of Pennsylvania School of Medicine delivered test programs to the technical support team of WCG for review. The IBM team subsequently performed a thorough security review of the code and modified it for use on the Grid. The changes included footprint reductions, incorporation of robust checkpointing and Grid I/O modifications.

Imaged pathology specimens were generated using a high-throughput whole slide scanner and transferred from laboratories within Robert Wood Johnson Medical School (RWJMS) and The Cancer Institute of New Jersey (CINJ) to the secure Boulder Colorado IBM hosting site where World Community Grid servers reside. As results were computed, they were returned to the servers at RWJMS and CINJ. The total of the transfers approached one terabyte of data. About 2909 years of run-time in the form of slightly more than 5 million work packages were harvested from the personal computers contributed to World Community Grid. This includes an approximate 3 times redundancy of work to ensure that the computations were not in error or tampered with. Because of the fairly large working set memory required for the program, only machines with over 1 GB of RAM were selected to run the project.

III. Data Generation and System Framework

The Tissue microarrays (TMA) used in the HDC project were collected from The Cancer Institute of New Jersey, Yale University, University of Pennsylvania and Imagenex Corporation (San Diego, CA). To date over 300 slides containing cohorts of hundreds of tissue discs each and originating from 45 TMAs were digitized at 40× resolution using a Trestle MedMicro virtual microscopy system. The output images typically contain 1-3 billions of pixels and were stored as a compressed tiled TIFF file sized at 0.5 to 2 Gigabytes. Our registration protocols [14] were applied to the scanned images to identify rows and columns of the tissue arrays. Any tissue cores that suffered from exceedingly pronounced artifacts were excluded from the study. Images of each tissue core were systematically extracted from the archive and packaged as

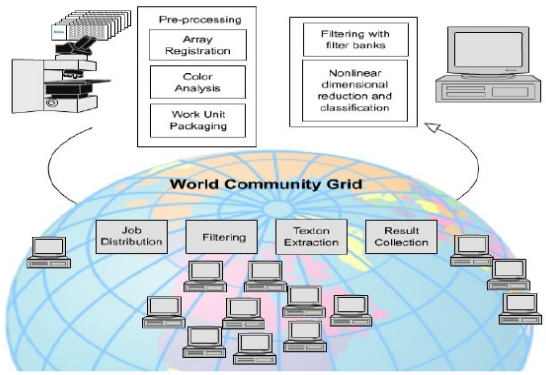


Fig. 3. Two staining examples. On the left is a hematoxylin stained tissue and on the right is hematoxylin & eosin stained tissue.

workunits for the HDC project. The dimension of each image was 1200×1200. The specimens under study had previously been stained with hematoxylin and hematoxylin & eosin. Two staining examples are shown in Figure 2. A texton extraction algorithm was applied on the staining maps of the two dyes which were generated using color decomposition [14]. Each of the resulting staining maps as well as the luminance measure generated from the original color image were uploaded as separate workunits to the WCG. The work-flow and logical units are shown in Figure 3.

IV. Data Analysis

As TMA is being utilized increasingly in cancer research, the development of accurate and efficient method to evaluate TMA specimens remains a major goal. The individual tissue discs comprising a given TMA contain complex, heterogeneous tissue components, which renders most straight forward quantification methods ineffective. Furthermore, as researchers design experiments using different staining techniques which target specific proteins, the methods used for interpreting these specimens must vary accordingly.

In this section, we explain the methods used to generate and analyze the image features for automatic classification of breast tissue specimens. Textures and intensities were used as feature measures to classify the staining profiles of the imaged tissues. Because the feature vectors lie in a high dimensional space, we applied a nonlinear dimension reduction method to decrease the dimensionality. Through iterative experiments we determined that among several different classification algorithms, the gentle AdaBoost classifier provided the best overall performance in the reduced subspace.

A. Texton and Features

Figure 2 shows two breast cancer specimens. It can be found that the difference in texture can be used as the discriminative features to separate different types of breast tissues. Traditional texture analysis includes Law's moment [15], cooccurrence matrices [16], run length matrices [17] and autoregressive models [18] et. al.

In recent studies, texture has been represented using texton. Textons are defined as conspicuous repetitive local features

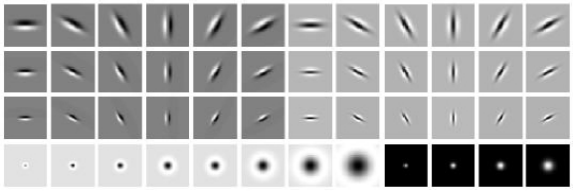


Fig. 4. The LM filter bank used to generate the texture features.

that humans perceive as being discriminative between textures. Unlike many other texture features that describe each texture as a constant relationship – a number, a data vector or a set of model parameters – between each pixel and its surroundings, the concept of a texton supports the existence of numerous distinct textual components in each texture. Therefore, it has advantages in describing textures that have high-level components. Texton based texture analysis has been widely used in many fields of texture related research, including classification [19], [20], [21], segmentation [22] and synthesis [23].

Based on texton theory, we set out to establish a large reference library which could be used as the fundamental vocabulary for distinguishing between cancer and benign tissues. This is referred to as the "bag of visual words" model and has been widely used in recent object recognition literature [22], [24], [25]. In our approach each work unit was first filtered with a texton filter bank. Subsequently, the cluster modes were extracted from the resulting filter responses to generate a universal reference library. The filtering responses collected across all imaged discs can be considered as typical words that describe the underlying histology and staining pattern of the specimens. Thus far, over 100,000 imaged tissue discs have been processed on the Grid.

In our experiments, four different types of filter banks were compared.

1) Gabor filter bank: The basic even-symmetric Gabor filter bank is a set of 2D Gaussian function with variances σ_x and σ_y which are modulated by a complex sinusoid. The sinusoid has center frequencies u and v along x and y-axes, respectively.
2) The Leung-Malik (LM) Filter Bank [26]: The LM filter bank are a set of first and second derivatives of 2D Gaussian function at six orientation and three scales, coupled with eight Laplacian of Gaussian (LoG) and four Gaussian function.
3) The Schmid filter bank [27] : The Schmid filter bank is composed of 13 orientation invariant filters. It is best suited for orientation insensitive texture segmentation.
4) The Maximum Response (MR) Filter Banks [20]: The MR filter bank is quite similar to the LM filter bank. However, in order to achieve the orientation invariance, only the maximum response is chosen as the feature for each scale of the first and second derivatives of the Gaussian. The LoG and Gaussian are chosen as another two features. All combined the dimension of the feature space is eight.

Systematic analysis did not show significant differences among these filter banks in performance and ultimately de-

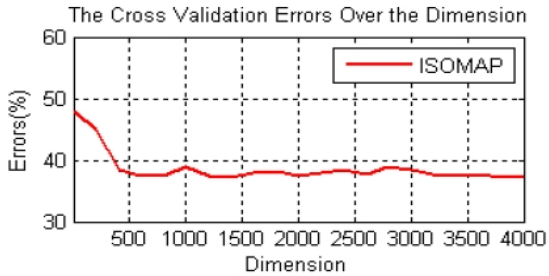


Fig. 5. The five-fold cross validation error over the dimensionality using ISOMAP for nonlinear dimension reduction.

cided to utilize the 49×49 LM filter bank to compute the filter responses. The feature vector is composed of eight LoG filter responses with $\sigma = 1, \sqrt{2}, 2, 2\sqrt{2}, 3, 3\sqrt{2}, 6, 6\sqrt{2}$, four Gaussian filtering responses with $\sigma = 1, \sqrt{2}, 2, 2\sqrt{2}$ and the bar and edge filtering response within six different directions, $\theta = 0, \pi/6, \pi/3, \pi/2, 2\pi/3, 5\pi/6$, $\sigma = 1, \sqrt{2}$, 2. In total, each image pixel was represented by a 48 dimensional feature vector. Figure 4 shows the LM filtering bank in our system.

The image filtering response generated using the collective computation power of the World Community Grid were gathered together and clustered using K-means, where K was set to 4000 in our experiments. The cluster centers, called textons, were used to generate the texton library. The appearance of each breast tissue image was modeled by a compact quantized description called texton histograms. Texton histograms are created by assigning each pixel filter response in the image to its closest texton in the generated texton library, which was calculated using

$$h(i) = \sum_{j \in I} count(T(j) = i) \tag{1}$$

where I denotes breast tissue image, i is the ith element of the texton dictionary, $T(j)$ returns the texton assigned to pixel j. In this way, each breast tissue image was modeled as a texture modes distribution, the texton histogram. Each image was mapped to one point in the high dimension space R^d, where $d = K = 4000$ is the number of textons.

B. Dimension Reduction and Classification

After quantizing the filter response into texton histograms, each image was represented by a 4000 dimension vector. Generally, in such a high dimensional space, one has to consider the "curse of dimensionality" [28, pp. 170]. In this paper, a nonlinear dimension reduction method, the isometric feature mapping (ISOMAP) [29], was applied to find the embedded dimensionality of the original feature space.

1) Nonlinear Dimension Reduction: Although the dimensionality of the input features was quite high, the features have usually exhibited much less degrees of freedom. Given a set of feature vectors $Z = \{\mathbf{z}_1, ...\mathbf{z}_i, ...\mathbf{z}_n\}$ where $\mathbf{z}_i \in R^d$, there exists a nonlinear mapping T which represents $\mathbf{z}_i$ in the low dimension as

$$\mathbf{z}_i = T(\mathbf{x}_i) + \mathbf{u}_i \qquad i = 1, 2, ...n \tag{2}$$

where $\mathbf{u}_i \in R^d$ is the sampling noise and and $\mathbf{x}_i \in R^q$ denotes the representation of the original $\mathbf{z}_i$ in the low-dimensional subspace, where q represents the dimensionality of the reduced subspace.

Unsupervised manifold learning is capable of discovering the degrees of freedom that underlie complex natural observations. We applied ISOMAP to explore the low dimension embedding in the original feature space. In the first step, we determined the neighbors of each point $\mathbf{z}_i$ in the original space R^d and connected the neighbors to form a weighted graph G. The weights were calculated based on the Euclidean distance between each connected pair of points. We then calculated the shortest distance in the graph G, $d_G(i, j)$, between pairs of points of $\mathbf{z}_i$ and $\mathbf{z}_j$. The final step was to apply the standard multiple dimensional scaling (MDS) [30] to the matrix of graph distance $M = \{d_G(i, j)\}$. In this way, the ISOMAP applied a linear MDS on the local patch but preserved the geometric distance globally using the shortest path in the weighted graph G. Cross validation (CV) [31] was applied to evaluate the embedded dimensionality of the original 4000 dimensional feature vector. CV is the statistical method of partitioning samples into subsets

$$CV(\alpha) = \frac{1}{N} \sum_{i=1}^{N} \left| y_i - f^{-k}(\mathbf{x}_i, \alpha) \right| \tag{3}$$

where $\mathbf{x}_i$ is the feature vector in the reduced subspace R^q, $y_i = \{+1, -1\}$, which represents the cancer and benign breast tissue labels. $f^{-k}(\mathbf{x}_i, \alpha)$ is used to denote the classification results using the α-th dimensionality with the k-th partition removed from the training set. In Figure 5 we show the CV errors corresponding to the dimensions of the feature vector. An elbow can be observed when the dimensionality approaches 500, therefore, we choose to reduce the dimension of the original feature vector to 500.

2) Classification: In [32], the k-nearest neighbor (kNN) and classification tree ($C4.5$) were integrated into a Bayesian framework for characterizing breast tissues. However, in our case, each training sample was represented by a feature vector $\mathbf{x}_i$ in the reduced subspace R^q where $q = 500$. This is still a relatively high dimension where the maximal margin classifiers such as support vector machine (SVM) [33] and boosting [34] are better suited. We conducted experiments to compare the performance of four boosting algorithms, the standard AdaBoost, the gentle AdaBoost, the real AdaBoost and LogitBoost with kNN, Bayesian classifier and SVM. The results showed that the maximal margin classifiers [33], [34], such as SVM and boosting, which simultaneously minimize the empirical classification error and maximize the geometric margin, outperformed all the other algorithms. In order to separate two subtypes of breast cancers from the benign, the best binary classifier in our experiments (the gentle AdaBoost) was extended to a multi-class algorithm

The kNN consists of assigning all the features into k most similar cluster centers based on certain similarity measurements. The final label was determined by majority voting from k candidates. The $C4.5$ decision tree is a widely used multiple node tree based classifier, which is built by minimizing the entropy.

Input: Given n features $\mathbf{x}_i$ in R^q and their corresponding labels $y_i = \{-1, 1\}$.
Training:
- Initialize the weights $w_i = 1/n, i = 1, ..., n$. Set $b(\mathbf{x}) = 0$ and the number of nodes $M = 8$ in the CART decision tree.
- For $j = 1...J$
  - Each training sample is assigned its weight w_i. The weighted tree growing algorithm is applied to build the CART decision tree $h_j(\mathbf{x})$.
  - Update using $b(\mathbf{x}) = b(\mathbf{x}) + h_j(\mathbf{x})$.
  - Update the weights $w_i = w_i e^{-y_i h_j(\mathbf{x})}$ and renormalize w_i.
  - Save the j-th CART decision tree $H_j(\mathbf{x})$.

Testing:
- Output the classification: $sign\left[b(\mathbf{x})\right] = sign\left[\sum_{j=1}^{J} H_j(\mathbf{x})\right]$.

Fig. 6. The binary gentle AdaBoost using an eight nodes classification and regression tree (CART) as the weak learner.

Input: Given n features $\mathbf{x}_i$ in R^q and their corresponding labels $y_i = \{1, 2, ..., M\}$, where M represents the number of classes.
Training:
- Using the original n training samples to compose a $n * M$ observation matrix of training samples $\{(\mathbf{x}_i, 1), y_{i1}\}, ..., \{(\mathbf{x}_i, j), y_{ij}\}..., \{(\mathbf{x}_i, M), y_{iM}\}$, where y_{ij} is the $\{-1, +1\}$ response for j-th class of training sample $\mathbf{x}_i$.
- Generate the j-th strong classfier $H_j(\mathbf{x})$ by applying the gentle AdaBoost algorithm in Figure 6 on the j-th row of the observation matrix. Continue this step for each class.

Output:
- Output the final classification result by maximizing $argmax_j H_j(\mathbf{x})$.

Fig. 7. The multi-class gentle AdaBoost using an eight nodes classification and regression tree (CART) as the weak learner.

Let $\mathbf{x} \in X$ represent the low level feature in the reduced subspace R^q, the *Bayesian classifier* is designed to maximize a-posterior (MAP) probability

$$p(C_i|\mathbf{x}) = \frac{p(\mathbf{x}|C_i)p(C_i)}{\sum_{i=1}^{K} p(\mathbf{x}|C_k)p(C_k)} \tag{4}$$

and the Bayesian classifier determines the class C_i by maximizing the posterior probability $p(C_i|\mathbf{x})$.

The *support vector machine* (SVM) was first introduced in [33] for binary classification problem. The strategy is to construct the linear decision boundaries in a large transformed version of the original feature space. The SVM simultaneously minimizes the empirical classification error and maximizes the geometric margins by minimizing the regularization penalty

$$\frac{1}{2}\left\|\mathbf{w}\right\|^2, \text{ subject to } y_i(w_0 + \mathbf{w}^T\mathbf{x}_i) - 1 \geq 0 \tag{5}$$

When the examples are not linearly separable, the optimization can be modified by adding a penalty for violating the classification constraints. This is called *soft margin SVM* which minimizes

$$\frac{1}{2}\left\|\mathbf{w}\right\|^2 + C\sum_{i=1}^{N}\xi_i, \text{ subject to } y_i(w_0 + \mathbf{w}^T\mathbf{x}_i) - 1 + \xi_i \geq 0 \tag{6}$$

where ξ_i are called slack variables which store the deviation from the margin and C is the soft penalty to balance the training errors and margins. In (5) and (6), $\mathbf{w}$ is the slope of the decision hyperplane and w_0 is the offset. The $\mathbf{x}_i$ denotes the feature vector, and y_i is the ground true labels. We minimize (6) by maximizing the dual problem of (6) which involve a feature mapping $\phi(\mathbf{x})$ through an inner product. The inner product can be evaluated without ever explicitly constructing the feature vectors $\phi(\mathbf{x})$ but through a kernel function $\kappa(\mathbf{x}, \mathbf{x}')$. In our project, we proposed to use a nonlinear Mercer kernel [35] based on χ^2 distance. It was shown that among other choices of distance functions between histograms, χ^2 distance performed the best for the texture similarity measure. The kernel function is defined as

$$\kappa(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{1}{\tau}\chi^2\left(\mathbf{x}, \mathbf{x}'\right)\right) \tag{7}$$

where

$$\chi^2\left(\mathbf{x}, \mathbf{x}'\right) = \frac{1}{2}\sum_{l=1}^{500}\frac{\left(\mathbf{x}(l) - \mathbf{x}'(l)\right)^2}{\mathbf{x}(l) + \mathbf{x}'(l)}. \tag{8}$$

Boosting works by sequentially applying a classification algorithm on a reweighted version of the training data and producing a sequence of weak classifiers $h_j(\mathbf{x})$, $j = 1, 2, ..., W$ where $W = 40$ in our case represents the number of iteration rounds of each boosting algorithm. The strong classifier is assembled from all the weak classifiers $h_j(\mathbf{x})$ to minimize the exponential cost function $exp(-yh_j(\mathbf{x}))$, where y represents the label of the training sample $\mathbf{x}$. In the standard binary AdaBoost classification, the labels were decided by weighted voting to produce the final prediction

$$\widehat{y} = sign\left(\sum_{j=1}^{W}\alpha_j h_j(\mathbf{x})\right) = sign(H(\mathbf{x})) \tag{9}$$

where $H(\mathbf{x})$ is the learned strong classifier. The α_j is the weight of the j-th weak classifier $h_j(\mathbf{x})$ and is computed during training. All the boosting algorithms are designed to minimize an exponential cost function $exp\left(-y\sum_{j=1}^{W}\alpha_j h_j(\mathbf{x})\right)$. If the weak classifier $h_j(\mathbf{x})$ returns a discrete class label $\{-1, +1\}$, the boosting algorithm is called *AdaBoost*. Instead of making a hard decision, if the weak classifier $h_j(\mathbf{x})$ returns a real value prediction like a probability mapped to the interval $[-1, +1]$, it is called *real AdaBoost*. The *gentle AdaBoost* is a modified version of the real AdaBoost algorithm, which applies Newton step rather than exact optimization at each step of minimizing the loss function. The *LogitBoost* is another boosting algorithm which uses Newton steps to fit an additive logistic regress model based on maximum likelihood. The weak classifier we used was an eight node classification and regression tree ($CART$).

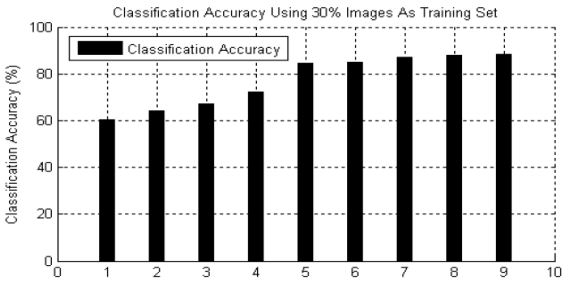


Fig. 8. The binary classification accuracy of nine different classifiers using 30% as the training set. 1) Bayesian classifier. 2) KNN (K = 3). 3) KNN (K = 5). 4) C4.5 decision tree. 5) Support vector machine (SVM). 6) Standard AdaBoost. 7) Real AdaBoost. 8) LogitBoost. 9) Gentle AdaBoost.

We experimentally tested each of the classification algorithm. The gentle AdaBoost using an eight node $CART$ decision tree provided the best results for binary classification problem. Fig. 6 shows the details of the gentle AdaBoost algorithm.

Multi-class experiments were designed to determine the capacity of the system to subclassify different types of cancer. Given a M-class classification problem, where we have N training samples $\{x_1, y_1\}, ...\{x_i, y_i\}, ..., \{x_N, y_N\}$. The $x_i \in R^q$ denotes the i-th feature vector in the reduced subspace and $y_i \in \{1, 2, ..., M\}$ represents the corresponding ground-truth class labels. The target is to find a strong classifier which minimizes a multi-class exponential loss function $\sum_{j=1}^{M} exp(-y_i H_j(\mathbf{x}))$ where $H_j(\mathbf{x})$ is the j-th strong classifier. This is equivalent to run separate boosting algorithms in an one-against-all manner. One-against-all boosting constructs M binary classifier, each of which is used to separate one class from all the others. The j-th strong classifier was trained using boosting with all the training samples satisfying $y_i = j$, $i = 1, 2, ..., N$ as positive and all the others as negative. As the gentle AdaBoost outperformed the other methods in the previous binary classification, we extended it to classify two different subtypes of cancers from benign tissue images in the multi-class experiment. The multi-class gentle AdaBoost algorithm is shown in Fig. 7.

V. Experiments

In these experiments, pathologists were asked to provide independent confirmation of the ground-truth labels of records for the entire mixed set of 3744 digitized breast tissue images. The breast tissue images contained 10 different types, which included normal (NOR), ductal hyperplasia (DH), fibroadenoma (FIB), atypical ductal hyperplasia (ADH), ductal carcinoma in situ (DCIS), lobular carcinoma in situ (LCIS), invasive ductal carcinoma (IDC), invasive lobular carcinoma (ILC), lymph-node-negative metastasis (LNN) and soft tissue metastasis (STM). The goal of the binary classification experiments was to determine the capacity of the algorithms to separate benign from cancer tissue. Based on the discussion with surgical pathologists, NOR, DH, FIB and ADH were grouped as benign breast tissue and the remaining classes were grouped as breast cancer. In the multi-class experiment, DCIS and LCIS were treated as one subgroup of cancer and IDC, ILC, LNN and STM as the other.

TABLE I
The confusion matrix shows three classes classification accuracy using multi-class gentle adaboost and 30% images as the training set

	Benign	Cancer I	Cancer II
Benign	**84.5**	6.4	7.1
Cancer I	6.8	**81.2**	13.8
Cancer II	8.7	12.4	**79.1**

The mixed set of breast tissue microarrays were digitized using a 40× volume scan on a Trestle/Zeiss MedMicro, whole slide scanner system. We have developed algorithms to automatically delineate the tissue discs comprising the arrays, decompose those discs into their constituent staining maps, and process the images on the World Community Grid. The total number of computers currently participating in the World Community Grid efforts is approximately 250,000 worldwide and growing.

We have analyzed 3744 breast cancer tissues (674 hematoxylin and 3070 hematoxylin and eorin staining) from the total dataset containing 100,000 imaged specimens. Without the Grid, it would require about 210 days of computation to generate the complete texton library using an efficient C++ implementation on a PC with P3 1.5GHz processor and 1G RAM. However, we built the texton library in less than 40 minutes for the breast cancer subset using WCG.

A. Binary Classification of Benign and Caner

In this section, we first present the comparative performance results for four classification methods, kNN, Bayesian, $C4.5$ and SVM, and four types of boosting algorithms. The dataset used in these experiments consisted of 611 benign and 3133 cancer specimens. Each algorithm was tested 10 times using different portions of the training images drawn from random sampling. We select 30% of the images as training and the other 70% was reserved for testing. Figure 8 shows the average classification results. It is clear that the maximal margin classifiers, SVM and boosting, produced comparative good results, while outperforming widely used classifiers such as kNN, Bayesian and $C4.5$ decision tree. The gentle AdaBoost using an eight node CART decision tree provided the best performance. Because the training data was skewed to cancer samples, we obtained higher false positive than false negative. This is indeed preferred and is actually a design criteria for many clinical tests.

B. Multi-class Classification of Benign and Two Subtypes of Cancer

The experimental results are presented for studies in which the original gentle AdaBoost algorithm was modified to accommodate multi-class classification. Based on the direction of the clinical pathologist, we separated six subtypes of cancer tissues into two sub-groups: cancer class I which contains DCIS and LCIS and cancer class II containing IDC, ILC, LNN and STM. The dataset is consisted of 611 benign, 1103

cancer class I and 2030 cancer class II. 30% of the images in each class were randomly selected for training and the remaining 70% was used for testing. The confusion matrix is presented on the right of Table I. Figure 9 shows some correct classification samples and failed cases. The left most three columns are correctly classified samples, and the right most fourth column shows the failed cases. The first row is the benign tissue where the last one is misclassified as cancer class II. The second row represents cancer I while the last tissue image is misclassified as benign. The last row is the cancer II, and the last image is misclassified as cancer I. In Figure 9 we show the large intra-class variances and inter-class similarities which produced the classification errors.

From all these experiments, it was shown that the gentle AdaBoost provided satisfactory results on both binary and multi-class classification of breast tissue images. We obtained an average 89% accuracy in separating benign from cancer tissue and an average accuracy of 80% in classifying two types of breast cancers from benign. In both cases only 30% of the images were used in the training.

VI. Conclusion

We have presented a Grid-enabled framework which utilize texton histograms to perform high throughput analysis of digitized breast cancer specimens. Experimental results have shown that a gentle AdaBoost classifier using an eight node CART decision tree as the weak learner provided the best results. We present the classification results of separating benign from cancer and also two classes of breast cancer. Multi-class classification errors increase significantly as the number of classes increased. In future work, we plan to expand the reference library of texton signatures and develop a robust multi-class classification algorithm to further classify ten different classes of breast cancer. We also plan to expand our studies to include a wide range of cancers, including colon cancer, head & neck, for which we have already generated the texton reference libraries using IBM World Community Grid.

A. Acknowledgement

This research was funded, in part, by grants from the NIH through contract 5RO1EB003587-02 from the National Institute of Biomedical Imaging and Bioengineering and contract W81XWH-06-1-0514 from the Department of Defense. UMDNJ wants to thank and acknowledge IBM for providing free computational power and technical support for this research through World Community Grid. For further information about World Community Grid, please view the IBM link [12], [13]. The authors are also grateful to The Cancer Institute of New Jersey and the Hospital of the University of Pennsylvania for the specimens and support that they have provided for this research.

References

[1] American Cancer Society, *Cancer Facts and Figures 2007*, 2007th ed. American Cancer Society, 2007.

[2] U. S. Cancer Statistics Working Group, "United states cancer statistics: 2003 incidence and mortality (preliminary data)," *National Vital Statistics*, vol. 53, no. 5, 2004.

[3] J. Kononen, L. Bubendorf, A. Kallioniemi, M. Barlund, P. Schraml, S. Leighton, J. Torhorst, M. J. Mihatsch, G. Sauter, and O. P. Kallioniemi, "Tissue microarrays for high-throughput molecular profiling of tumor specimens," *Nature Medicine*, vol. 4, pp. 844–847, 1998.

[4] D. L. Rimm, R. L. Camp, L. A. Charette, J. Costa, D. A. Olsen, and M. Reiss, "Tissue microarray: A new technology for amplification of tissue resources," *Cancer Journal*, vol. 7, pp. 24–31, 2001.

[5] J. Torhorst, C. Bucher, J. Kononen, P. Haas, M. Zuber, and O. R. K. et al, "Tissue microarrays for rapid linking of molecular changes to clinical endpoints," *American Journal of Pathology*, vol. 159, pp. 2249–2256, 2001.

[6] D. Zhang, M. S. Tellz, and T. C. Putti, "Reliability of tissue microarrays in detecting protein expression and gene amplification in breast cancer," *Modern Pathology*, vol. 16, pp. 79–84, 2003.

[7] N. R. Mucci, G. Akdas, S. Manely, and M. A. Rubin, "Neuroendocrine expression in metastatic prostate cancer: Evaluation of high throughput tissue microarrays to detect heterogeneous protein expression," *Human Pathology*, vol. 31, pp. 406–414, 2000.

[8] C. Gulmann, D. Butler, E. Kay, A. Grace, and M. Leader, "Biopsy of a biopsy: Validation of immunoprofiling in gastric cancer biopsy tissue microarrays," *Histopathology*, vol. 42, pp. 70–76, 2003.

[9] J. S. Suri and R. M. Rangayyan, *Recent Advances in Breast Imaging, Mammography, and Computer-Aided Diagnosis of Breast Cancer*. SPIE, 2006.

[10] L. Yang, P. Meer, and D. Foran, "Unsupervised segmentation based on robust estimation and color active contour models," *IEEE Trans. on Information Technology in Biomedicine*, vol. 9, pp. 475–486, 2005.

[11] IBM Help Defeat Cancer, "http://pleiad.umdnj.edu/ibm," 2007.

[12] Wikipedia, "http://en.wikipedia.org/wiki/help_defeat_cancer," 2007.

[13] IBM World Community Grid, "http://www.worldcommunitygrid.org," 2007.

[14] W. Chen, M. Reiss, and D. J. Foran, "Unsupervised tissue microarray analysis for cancer research and diagnosis," *IEEE Trans. on Information Technology in Biomedicine*, vol. 8, no. 2, pp. 89–96, 2004.

[15] K. I. Laws, *Texture Image Segmentation*. Ph.D Thesis, University of Southern California, 1980.

[16] R. Haralik, K. Shanmugan, and I. Dinstein, "Texture features for image classification," *IEEE Trans. on System, Man and Cybernetics*, vol. 3, pp. 610–621, 1973.

[17] R. Galloway, "Texture analysis using gray level run lengthes," *Computer Graphics Image Processing*, vol. 4, pp. 172–179, 1975.

[18] J. Mao and A. Jain, "Texture classification and segmentation using multiresolution simultaneous autoregressive models," *Pattern Recognition*, vol. 25, pp. 173–188, 1992.

[19] T. Leung and J. Malik, "Representing and recognizing the visual appereance of materials using three-dimensional textons," *International Journal on Computer Vision*, vol. 43, no. 1, pp. 29–44, 2001.

[20] M. Varma and A. Zisserman, "Classifying images of materials: Achieving viewpoint and illumination independence," *European Conference on Computer Vision*, vol. 3, pp. 255–271, 2002.

[21] O. Cula and K. Dana, "3D texture recognition using bidirectional feature histograms," *International Journal on Computer Vision*, vol. 59, no. 1, pp. 33–60, 2004.

[22] L. Yang, P. Meer, and D. J. Foran, "Multiple class segmentation using a unified framework over mean-shift patches," *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[23] D. Heeger and J. Bergen, "Pyramid-based texture analysis/synthesis," *SIGGRAPH '95: Proc. of the 22nd Annual Conf. on Computer Graphics and Interactive Techniques*, pp. 229–238, 1995.

[24] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 2169–2178, 2006.

[25] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," *European Conference on Computer Vision*, vol. 1, pp. 1–13, 2006.

[26] T. Leung and J. Malik, "Representing and recognizing the visual appearance of materials using three-dimensional textons," *International Journal on Computer Vision*, vol. 43, no. 1, pp. 29–44, 2001.

[27] C. Schmid, "Constructing models for content-based image retrieval," *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 39–45, 2001.

[28] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Wiley-Interscience, 2000.

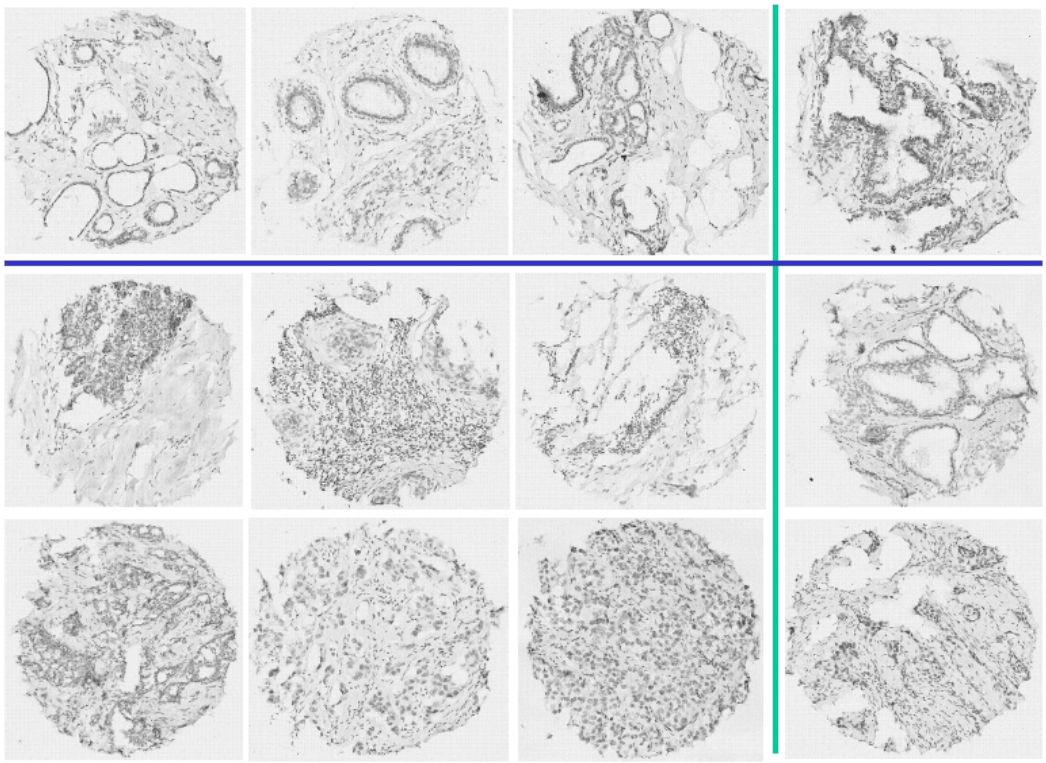


Fig. 9. The multi-class classification results using the gentle AdaBoost. The left three columns are correct classified samples and the right fourth column shows the failed cases. The first row is the benign samples. The second and third rows are the cancer samples.

[29] J. Tenebaum, V. de Silva, and J. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[30] I. Borg and P. Groenen, *Modern Multidimensional Scaling: theory and applications*, 1st ed. Springer-Verlag New York, 1997.

[31] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," *The 14th International Joint Conference on Artificial Intelligence*, vol. 2, no. 12, pp. 1137–1143, 1995.

[32] A. Oliver, J. Freixenet, R. Marti, and R. Zwiggelaar, "A comparison of breast tissue classification techniques," *Medical Image Computing and Computer-Assisted Intervention*, vol. 4191, pp. 872–879, 2006.

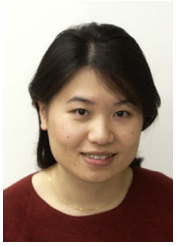
[33] C. Cortes and V. Vapnik, "Support vector networks," *Machine Learning*, vol. 20, pp. 1–25, 1995.

[34] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," *Machine Learning*, pp. 148–156, 1996.

[35] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*, 1st ed. Cambridge University Press, 2004.

Lin Yang received the B.S. degree in Electrical and Communication Engineering from Xian Jiaotong University, Xian, Shaanxi, P.R.China. in 1999 and the M.S. degree in Signal and Information Processing from the Image Processing Center of Xian Jiaotong University, Xian, Shaanxi, P.R.China in 2002. He is currently working toward the Ph.D. degree in the Department of Electrical and Computer Engineering, Rutgers University at New Brunswick, Piscataway, NJ.

His research interests include the development of robust image segmentation and registration, computer vision, statistics pattern recognition, content-based image retrieval and intelligent large-scale multimedia database.



Wenjin Chen received the Bachelor of Medicine degree from Beijing Medical University (now Peking University Health Science Center), Beijing, China in 1997 and the Ph.D. degree in Computational Molecular Biology from the University of Medicine and Dentistry of New Jersey (UMDNJ) and Rutgers University, Piscataway, NJ, in 2005. She is currently serving as Associate Director, Biomedical Imaging at Center for Biomedical Imaging and Informatics, UMDNJ, Piscataway, NJ. Her research interests include tissue microarray analysis, digital microscopy, image pattern recognition and medical informatics.

Peter Meer received the Dipl. Engn. degree from the Bucharest Polytechnic Institute, , Romania in 1971, and the D.Sc. degree from the Technion, Israel Institute of Technology, Haifa, in 1986, both in electrical engineering. From 1971 to 1979 he was with the Computer Research Institute, Cluj, Romania, working on R&D of digital hardware. Between 1986 and 1990 he was Assistant Research Scientist at the Center for Automation Research, University of Maryland at College Park. In 1991 he joined the Department of Electrical and Computer Engineering, Rutgers University, Piscataway, NJ and is currently a Professor. He has held visiting appointments in Japan, Korea, Sweden, Israel and France, and was on the organizing committees of numerous international workshops and conferences. He was an Associate Editor of the *IEEE Transaction on Pattern Analysis and Machine Intelligence* between 1998 and 2002, is a member of the Editorial Board of *Pattern Recognition*, and was a Guest Editor of *Computer Vision and Image Understanding* for a special issue on "Robust Statistical Techniques in Image Understanding". He is coauthor of an award winning paper in *Pattern Recognition* in 1989, the best student paper in the 1999 and the best paper in the 2000 *IEEE Conference on Computer Vision and Pattern Recognition*. His research interest is in application of modern statistical methods to image understanding problems.

Gratian Salaru received the M.D. degree from the University of Medicine and Pharmacy Timisoara, Romania in 1995. After three years of Forensic Pathology at the "Mina Minovici" Institute of Forensic Medicine, Timisoara, Romania, he completed an additional five years of training and graduated from the Pathology residency program at Robert Wood Johnson Medical School/UMDNJ in 2004. He completed a fellowship in Hematopathology at the same institution in 2006. He is board certified in Anatomical and Clinical Pathology and is currently an Assistant Professor of Pathology and Laboratory Medicine at Robert Wood Johnson Medical School/UMDNJ, New Brunswick, NJ. His research interests focus on areas of clinical pathology, rapid HIV testing, digital microscopy, medical informatics and hematopathology.

Lauri Goodell , MD received her Medical degree from UMDNJ- Robert Wood Johnson Medical School in 1991. She is currently Associate Professor and Director of Hematopathology in the Department of Pathology and Laboratory Medicine at UMDNJ-RWJMS as well as Director of the Immunohistochemistry Core Research Lab at the Cancer Institute of New Jersey. Her research interests include pathophysiology of hematopoietic diseases, proteomics, image pattern recognition and telepathology.

Viktors Berstis is a Senior Software Engineer and Master Inventor at the IBM Corporation in Austin Texas. His experience at IBM includes architecting the System/38 - AS/400, developing various compilers, research on high-level automated integrated circuit design, and OS/2. Currently he is the technical lead and scientist for World Community Grid, where he also helps researchers exploit grid computing in their projects. With degrees from the University of Michigan, he is a senior member of the IEEE, and has over 130 US patents. His hobbies include radio controlled airplanes, SCUBA diving, 3D stereoscopic photography, playing the piano, exploiting solar energy and making all sorts of gadgets.

David J. Foran (S89-M91) received the B.S. degree from Rutgers University, New Brunswick, NJ, in 1983 and the Ph.D. degree in biomedical engineering from the University of Medicine and Dentistry of New Jersey (UMDNJ) & Rutgers University, Piscataway, NJ, in 1992. He served as a Physics Instructor at New Jersey Institute of Technology, Newark, NJ, from 1984 to 1985 and worked as a Junior Scientist at Johnson & Johnson Research, Inc., North Brunswick, NJ, from 1986 to 1988. He received one year of post-doctoral training at the Department of Biochemistry at UMDNJ-Robert Wood Johnson Medical School (RWJMS) in 1993. He joined the faculty at RWJMS in 1994 where he is currently an Associate Professor of Pathology & Radiology and the Director of the interdepartmental Center for Biomedical Imaging & Informatics. Dr. Foran also serves as the Associate Director for Research for the university-wide Informatics Institute. He is a member of the graduate faculty in the Program in Computational Molecular Biology and Genetics and he is a Research Associate Professor at the Center for Advanced Information Processing, both at Rutgers University. His research interests include quantitative, biomedical imaging, computer-assisted diagnosis, and medical informatics.