Triaging Diagnostically Relevant Regions from Pathology Whole Slides of Breast Cancer: A Texture Based Approach

Mohammad Peikari*, Mehrdad J. Gangeh, Judit Zubovits, Gina Clarke, and Anne L. Martel

***Abstract*—Purpose: Pathologists often look at whole slide images (WSIs) at low magnification to find potentially important regions and then zoom in to higher magnification to perform more sophisticated analysis of the tissue structures. Many automated methods of WSI analysis attempt to preprocess the down-sampled image in order to select salient regions which are then further analyzed by a more computationally intensive step at full magnification. Although it can greatly reduce processing times, this process may lead to small potentially important regions being overlooked at low magnification. We propose a texture analysis technique to ease the processing of H&E stained WSIs by triaging clinically important regions. Method: Image patches randomly selected from the whole tissue area were divided into smaller tiles and Gaussian-like texture filters were applied to them. Texture filter responses from each tile were combined together and statistical measures were derived from their histograms of responses. Bag of visual words pipeline was then employed to combine extracted features from tiles to form one histogram of words per every image patch. A support vector machine classifier was trained using the calculated histograms of words to be able to distinguish between clinically relevant and irrelevant patches. Result: Experimental analysis on 5151 image patches from 10 patient cases (65 tissue slides) indicated that our proposed texture technique out-performed two previously proposed colour and intensity based methods with an area under the ROC curve of 0.87. Conclusion: Texture features can be employed to triage clinically important areas within large WSIs.**

***Index Terms*—Bag of visual words, breast cancer, fast *k*-means, image analysis, machine learning, pathology, texture.**

I. Introduction

THANKS to the advances in digital imaging techniques, whole slide scanners are being employed to digitize whole slide (pathology) images (WSIs) at microscopic resolution. Digitizing pathology images offers many advantages; it enables them to be archived and viewed electronically and that also means that they can be analyzed by computer algorithms. In order to perform the histological assessment of hematoxylin and eosin (H&E) stained WSIs, pathologists start with their visual inspection of tissue at low magnifications to identify regions where tissue structures have changed compared to normal tissues. Regions with unusual hyperplastic nuclei, and/or calcifications are usually an indication for more sophisticated analysis of the area. Therefore, pathologists zoom into those regions of interest to observe the tissue at higher magnified views. However, if some features are not visible at lower magnification then the pathologist has to scan through the entire image at high magnification which may be very time consuming. Therefore, automatically triaging WSIs to find areas with clinically important information could ensure that all relevant regions are further examined by pathologists and irrelevant regions are simply neglected. Furthermore, such automated procedure could also be useful in number of other applications. For instance, automatic region extraction techniques when sampling suitable areas of high tissue activity for gene expression analysis [1], [2], choosing important relevant regions within tissue sections to be punched when doing tissue microarray (TMA) analysis [3], or appropriate region selection for automated grading work flows [4], [5] could benefit from this technique.

The objective of this work is to present an automated method to facilitate and accelerate the analysis of breast digital pathology images by automatically detecting diagnostically important patches of data and discarding the rest. This is done by finding appropriate representations of images, based on their texture variations in different tissue components and employing a suitable machine learning technique to understand the differences between the extracted visual patterns.

A. Related Works

Mining of visual patterns from medical images may have several potential advantages both for research and clinical practice as mentioned in [5]–[7]. Automatically processing pathology images by computer algorithms may be required to increase the throughput and accuracy of data analysis. In addition, these high throughput computations could improve the analysis related to gene profiling [1] and protein expression [8] studies.

Earlier studies have tried to distinguish between different tissue components from immunohistochemical and H&E stained images by looking at small rectangular image segments

Manuscript received June 04, 2015; revised August 05, 2015; accepted August 15, 2015. Date of publication August 20, 2015; date of current version December 29, 2015. This work is funded by the Natural Sciences and Engineering Research Council of Canada (NSERC). *Asterisk indicates corresponding author.*

*M. Peikari is with the Department of Medical Biophysics, University of Toronto, ON, M4N 3M5 Canada (e-mail: mpeikari@sri.utoronto.ca).

M. J. Gangeh is with the Department of Medical Biophysics, University of Toronto, ON, M4N 3M5 Canada.

J. Zubovits is with Faculty of Medicine, University of Toronto, ON, M5G 2M9 Canada.

G. Clarke is with the Department of Physical Sciences, Sunnybrook Research Institute, Toronto, ON, M5G 2M9 Canada.

A. L. Martel is with the Department of Medical Biophysics, University of Toronto, Toronto, ON, M4N 3M5 Canada, and also with the Department of Physical Sciences, Sunnybrook Research Institute, Toronto, ON, M5G 2M9 Canada.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2015.2470529

for representative intensity, colour, and context information [9]–[12]. Some studies have aimed to classify tissue components from TMA images where only selected and specific parts of the specimen are captured [3], [13]. Some others have proposed pixel-wise labeling approaches that are more suitable for small sized datasets and may slow down the systems that are dealing with large sized data such as WSIs [11], [13]. In addition, colour and intensity variation could occur due to different H&E colour staining protocols being used, tissue thickness and/or scanning devices [14]. The problem of colour and/or intensity variation in pathology images may hamper the performance of computer aided detection techniques which is still under investigation to find reliable solutions [15], [16]. We will be considering post-surgical whole-mount images for detecting relevant regions and therefore the aforementioned pixel-wise labeling approaches may not be efficient due to their large size ($> 10^9$ pixels in many cases). In contrast to pixel-wise labeling methods, a patch-based analysis technique is a more desired approach to speed up the processing time for WSIs. In [17], our preliminary results on a limited dataset and a slightly different feature set showed that a patch-based texture model of pathology slides could out-perform a recently proposed patch-based colour model technique in [10]. In addition, proposing a technique that reduces the dependency on colour and/or intensity is beneficial. Finally, generating a training dataset from all over the tissue area that includes the overall heterogeneity of the tissue structures is crucial.

There are different types of numerical features employed in conjunction with appropriate machine-learning techniques in the literature. The most widely used include one or a combination of: colour, texture, shape, and scale invariant local features. Colour features are based on the raw values or histogram of pixel intensities in different channels of various colour models providing powerful information for object recognition [18], [19]. Texture features are usually based on pixel grey level co-occurrence or run-length matrices [20], [21]; wavelet [22], Fourier, or discrete-cosine transforms; and/or convolving first or second order derivatives of Gaussian-like filters with varying sizes, scales and directions to highlight the variation in intensity levels between neighboring pixels/edges of the image structures [23]. Different texture classification tasks are present in the literature that use various filter banks including: Leung-Malik [23], [24], Schimid [23], [24], and maximum response (MR) [19], [23], [25], [26]. Shape features are based on the morphologic, geometric, and/or topological variation between different pre-segmented binary object contours [7], [20], [27], [28]. The scale invariant feature transform (SIFT) is a method of extracting features based on finding the local maximas (points of interest) by comparing a series of difference of Gaussian (DoG) images and generating a d-dimensional pixel gradient histogram of a window around the points of interest [29]. The SIFT descriptors are shown to be a robust key point descriptor in various image retrieval and matching applications since they are invariant to image transformations, illumination changes and noise. For a more detailed review of the state-of-the-art techniques in analysis of pathology images please refer to [14], [30].

II. Methodology

A. Image Dataset and Data Collection

We have used whole-mount H&E stained digital pathology slides (65 WSIs) from breast lumpectomy surgical specimens of 10 patients. The slides were prepared and scanned in the Biomarker Imaging Research Laboratory at Sunnybrook Research Institute (SRI) using the method described in [31]. Eight patient datasets were scanned at 5X (50 WSIs, 2 μm/pixel) magnification scale and two of them at 10X (15 WSIs, 1 μm/pixel) by a TissueScope scanner (Huron Technologies International Inc., Waterloo, Canada). The following steps were followed to generate a ground-truth training dataset. Adaptive thresholding and morphological opening were used as preprocessing steps on down-sampled images of the slides to remove clearly irrelevant regions, such as large areas of fat and paraffin (Fig. 1). Patches of 512×512 pixels (corresponding to an approximate area of 1 mm^2 for 5X and 0.25 mm^2 for 10X images) were collected from each slide by overlaying a grid of uniformly spaced boxes on tissue regions (Fig. 1). The locations of overlaid boxes were randomized by their starting point on the tissue area. The horizontal and vertical distances between pairs of grid boxes were 1000 and 500 pixels respectively. This ROI selection approach is more time efficient compared to the manual marking of different regions within the slides which may take longer. It also reduces the number of patches the pathologist needs to review and avoids selection bias. A graphical user interface (GUI) was developed in collaboration with a pathologist to capture the biological information within every ROI. The interface allowed the expert pathologist to scroll through the images randomized by their case identification number stored in the database and evaluate the presence of diagnostically important information within every 512×512 pixel patch. For each patch, the pathologist checked off tick boxes corresponding to each tissue type or feature present; the list was extensive and, in addition to diagnostically relevant features such as cancers, atypias, microcalcifications and lymphocytic vascular invasion, the presence of other irrelevant features such as fat, stroma, normal ducts and lobules was also noted. The collaborating pathologist reviewed 5151 image patches (corresponding to 10 patient cases) using the GUI. Fig. 2 shows a subset of this ground-truth set.

B. Proposed Methods

Our proposed automated identification aims to distinguish between different relevant or irrelevant tissue regions based on analyzing the structures each of their components present in pathology slides. The goal is to achieve a high sensitivity of at least 95% while maintaining the highest possible specificity. To eliminate the possibility of overlooking small potentially important regions at low magnification, in our automated method, we chose to examine the tissue structures by subdividing the slides into small square patches at highest possible magnification. We concentrate on texture analysis in order to develop a method with minimal dependence on colour and/or intensity information. Our proposed method proceeds with the following multi-step process: 1) image patches are converted to smaller

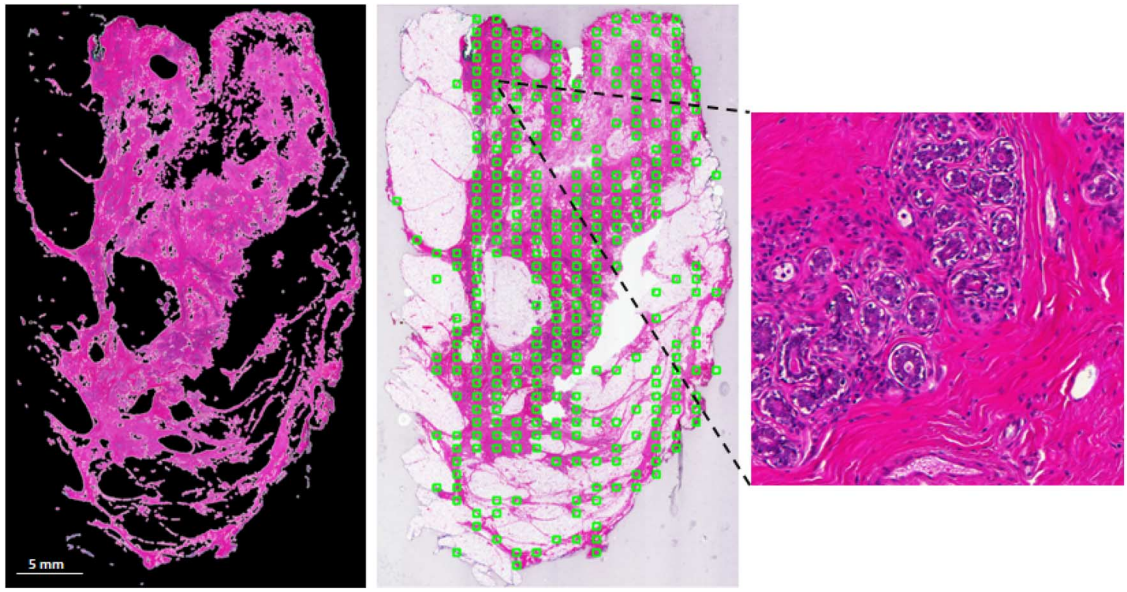

Fig. 1. Adaptive thresholding to remove clearly irrelevant structures before patch selection for data collection process (left), 512 × 512 pixel uniformly spaced box patches (in green) on tissue images (middle), and an example of a 512 × 512 pixel image patch from the WSI (right).

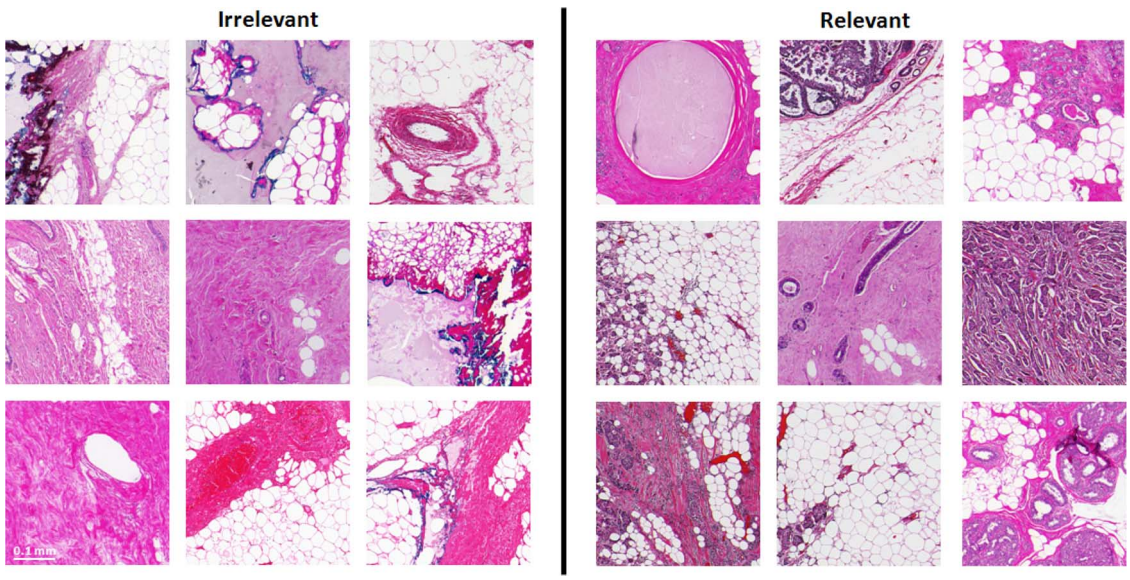

Fig. 2. A subset of ROIs used as ground-truth in this study with their labels. The problem of colour and intensity variation is visible in some of the patches presented in this demonstration. It is clear that the clinically relevant information may have covered different portions of the patches in the dataset.

tiles and texture features are extracted from them forming feature vectors, 2) bag of visual words (BoW) method is employed to regroup the features extracted from each tile and represent individual patches with one histogram of words, and 3) image patches are assigned with the correct class labels using a support vector machine (SVM) classifier. Details of each step are explained in the subsequent subsections.

1) Texture Feature Extraction: We concentrate on texture features in order to minimize colour and intensity dependencies in the analysis of pathology images. The diagnostically relevant regions usually contain higher variability, in their texture due to aggregation of nuclei and/or lymphocytes. To capture this variability we propose the following two texture feature calculation techniques and compare them with other state-of-the-art methods:

Raw pixel representation: In order to isolate the fine textures in pathology images, patches are divided into smaller non-overlapping tiles and converted to grey scale. The raw values of the grey level pixels of every 2D tile are normalized (to have mean $\mu = 0$, and standard deviation $\delta = 1$), vectorized, and used to generate a multidimensional feature vector.

Filter bank representation: Image patches are divided into smaller non-overlapping tiles and are converted from RGB to Lab colourspace. The hue and saturation channels are discarded and the luminance is normalized (to have mean $\mu = 0$, and standard deviation $\delta = 1$) and considered for texture analysis. This produces a tile image with high contrast between the nuclei and the background tissue, see Fig. 3. Texture features are directly calculated from each tile and normalized.

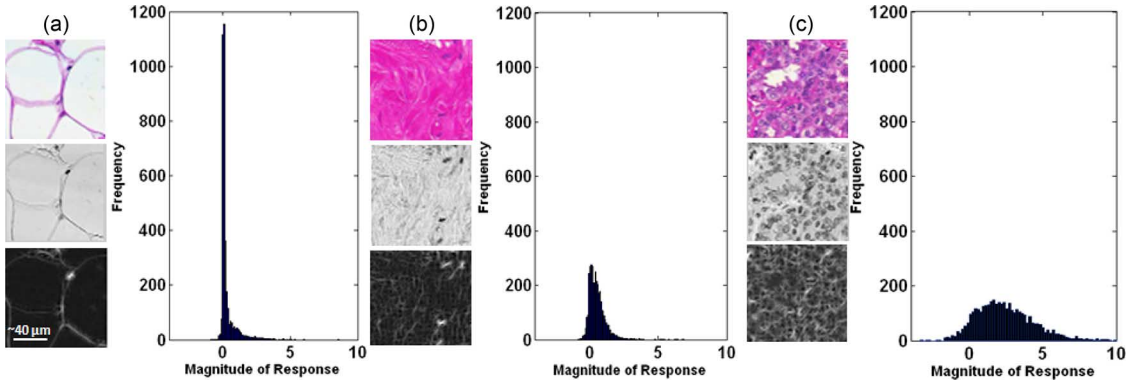

Fig. 3. Patches of three different tissue types (in RGB) corresponding to (a) fat, (b) stroma, and (c) epithelial cells -top- along with their normalized luminance channel images (after converting them from RGB to Lab) -middle-, and maximal filter responses after convolving Gaussian-like filters at all directions of one scale ($(\sigma_x, \sigma_y) = (1, 3)$) -bottom-, and plot showing their histogram of filter response magnitudes.

We chose to employ the root filter set (RFS) bank [32] due to its ability to form rotationally and scale invariant texture features [33]. This filter set has been shown to perform relatively better than similar sets in characterizing random textures of materials [26], [34]. Our initial triaging experiments agree with those presented in the literature. The RSF bank used in this study consists of 38 filters of size $s = 4 \times 4$ pixels: 2 anisotropic edge and bar filters in 6 directions and 3 scales ($(\sigma_x, \sigma_y) = [(1, 3), (2, 6), (4, 12)]$), and 2 rotationally symmetric Gaussian and Laplacian of Gaussian both with $\sigma = 10$ pixels. The filters in the bank are convolved with all of the small image tiles (taken from large patches) making 38 filtered images per image tile, each highlighting textures at different scales and directions.

Since the epithelial and lymphocytic components are relatively small and circular, their texture responses around the edges compared to their surrounding pixels are high in all scales and directions of the filters in the bank. This is shown as a uniformly distributed histogram of filter responses in Fig. 3(c). In contrast, fat cells and stroma have more skewed filter response representations (Fig. 3(a), (b)). Six statistical measures were calculated from every scale of the maximum over the six directions of anisotropic filters. The statistical features were mean, mode, median, standard deviation, kurtosis, and skewness. This compresses the information relating to texture variability contained in the 38 filtered images into just 48 statistical measures, that is $[6(\text{directions}) + 2(\text{rotations})] \times 6$ (statistical features).

2) Visual Content Representation Using Bag of Visual Words: In order to combine the features extracted from every tile of an image patch, we employed the bag of visual words approach. This technique allows for modeling complex image contents without explicitly considering the object models, pattern locations, and their relationships [35]. This method has been previously used to learn discriminative models of biomedical and scenery images and has shown to be robust [23], [24], [35]. Details of this algorithm are shown in Fig. 4. First, image patches are converted into tiles of a specific size depending on the extent of preferred isolation of visual patterns (explained before). Visual features corresponding to textures are retrieved from each tile. For any image tile, visual features are converted into a numerical vector as explained in Section II-B-1. Therefore, the number of feature vectors retrieved from every patch is equal to the number of tiles dividing it.

In this study, feature vectors (extracted from tiles) corresponding to each class (irrelevant/relevant) is separately clustered using a k-means algorithm [34]. Cluster centroids from each class are then combined to generate a feature dictionary of words to which every feature vector can be mapped. Therefore, for each patch, feature vectors from all tiles are mapped to the dictionary elements and a histogram is formed showing the frequency of dictionary elements within a particular patch. We can distinguish among image patches by training a classification algorithm to understand their differences by comparing their histograms of words.

We have implemented a faster version of k-means clustering method since the speed of its standard implementation drops for large number of samples and/or high dimensioned feature spaces. This implementation of k-means is similar to the one proposed in [36] and works by randomly partitioning points from all over the feature space into 'n' exclusive groups and applying standard k-means in parallel on each group. The distances between cluster centres from all groups are used to identify associated cluster centroids and the mean of corresponding centroids are set as new cluster centers of the points in the whole feature space (words in the dictionary). Our implementation of this technique performs at least 10 times faster than the standard k-means. This fast k-means was also employed in the BoW pipeline of the second feature calculation technique (II-B-1-2). This is important to note that the fast k-means clustering method is performed during the training stage only when the dictionary of words is to be learned.

3) Classification of Diagnostically Relevant Regions: An SVM with a radial basis function (RBF) kernel was used to model the feature space and find the best separating margin between the two classes. The RBF kernel has been shown to be a powerful kernel in modeling feature spaces that have nonlinear relationship between class labels and attributes. Furthermore, the RBF kernel has shown to behave like linear and sigmoid kernels at certain parameters [37], [38]. We employed the RBF-SVM implemented in libsvm library [39] and the optimized parameters were found using a confusion matrix calcu-

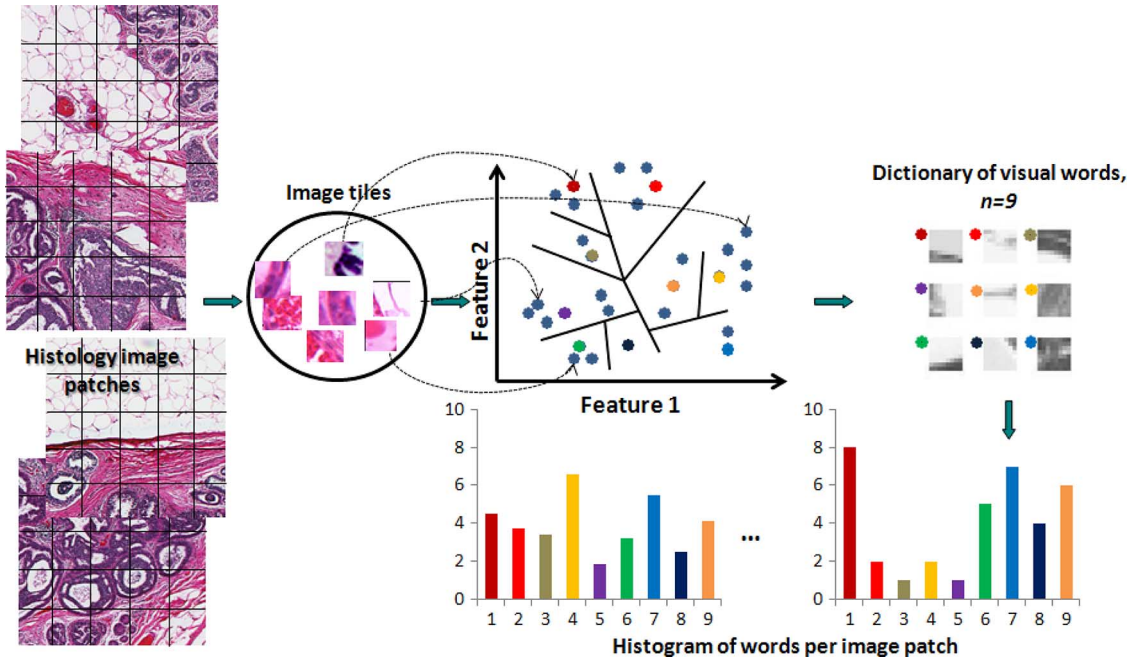

Fig. 4. Bag of visual words content representation pipeline: image patches are split into smaller image tiles. Visual features are extracted from every image tile and mapped onto a multidimensional feature space (shown as 2D for simplicity). Centroids of the clustered feature space (coloured stars) are taken as the visual word dictionary elements. A histogram of words is generated using the feature dictionary for every image patch.

lated after cross-validating on training images. In order to correct for the imbalanced training set, the SVM's weight value of the class with more cases (irrelevant) was set to the ratio of number of cases in relevant and irrelevant classes.

III. Experimental Setup and Results

A. Experimental Design

To assess the performance of our proposed approach, 2 phases of validation were designed. In the first phase, the training dataset with 5X image patches ($n = 2302$, 307 relevant and 1995 irrelevant patches) was divided into 8 distinct groups each consisting of the image patches corresponding to individual patients. An 8-fold subject-wise cross-validation scheme was then used to train and evaluate the classification performances of each group at different tile and dictionary sizes. The dictionary of words were regenerated in every fold of this cross-validation scheme. In each fold, the optimum SVM-RBF parameters were chosen through cross-validating image patches of seven patients in the training set over a range of possible SVM trade-off parameter (C) and the kernel width (γ) values. We defined the optimum parameter set by first identifying all sets that produced a sensitivity of above 95%; from these we then selected the set with maximum specificity.

In the second phase, the median of the 8 optimized SVM parameters corresponding to each fold in the first phase was calculated and used to train an overall SVM model from the training dataset used in phase 1. The overall trained model in this phase was subsequently used to test the classification performance on image patches of two unseen patient cases not used in the first phase ($n = 2849$ patches from 15 slides). Image patches used in this phase were all captured at 10X (pixel size $= 1\ \mu$m) magnification and so were down-sampled to match our previously trained models based on image patches magnified at 5X.

In order to illustrate the performance of the model in a more realistic setting, that is in triaging a whole slide image, the trained model was applied to a whole-mount image scanned at 5X. This image was taken from an unseen new patient dataset that was not included in either phase 1 or 2. The results of the triaging were then compared to the annotations provided by a trained pathologist. The operating point of the optimized models for each combination of SVM kernel parameters was chosen to be the one that had the highest possible specificity for a target sensitivity of at least 95%.

B. Comparison With State-of-the-Art Methods

To compare the performance of the proposed automated methods, we compare it with the following two other state-of-the-art methods: 1) The colour based method proposed in [10]; briefly, the RGB colour information in image patches are unmixed into two channels corresponding to hematoxylin (purple/blue) and eosin (pink/red) stain colours using the method explained in [40]. A 22 dimensional feature descriptor corresponding to 11 uniformly distributed percentiles of the cumulative histogram generated from each of hematoxylin and eosin channels were generated for each image patch and normalized. An RBF-SVM classifier was trained to classify the test image patches into diagnostically relevant or irrelevant regions. 2) A variation of the intensity based method introduced in [12]; briefly, for every image patch, the raw intensity values of each red, green, and blue colour channels are considered to form 3 image intensity histograms which were concatenated to form one 300 dimensional descriptor which was then normalized.

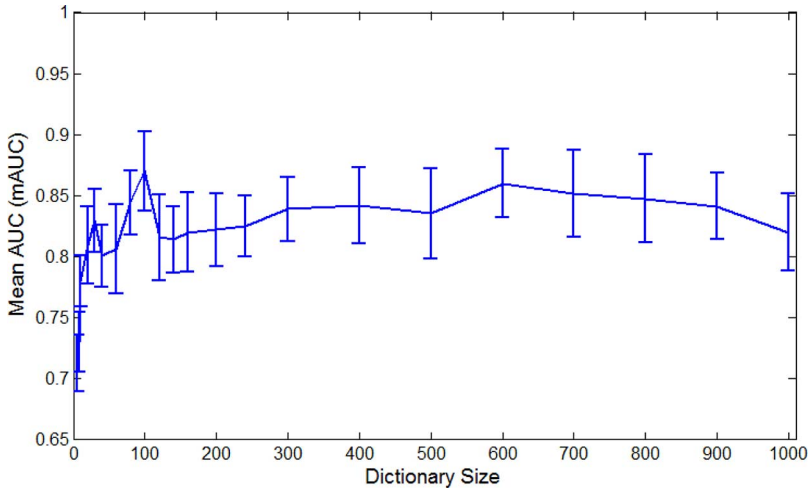

Fig. 5. Effect of dictionary size on the area under the curve of classification using BoW and filter response method on 32×32 tile size. Dictionary size of $s = 100$ words was found to have the best performance while being faster compared to $s = 600$ words on cross-validated 8-patient data ($n = 2302$ patches). Error bars indicate standard error of mean.

TABLE I
RESULTS COMPARING THE CROSS-VALIDATED PERFORMANCES OF THE FOUR METHODS. THE DICTIONARY SIZE FOR METHODS WITH BOW WAS SET TO A CONSTANT VALUE OF 100 ($k = 50$). NUMBER OF IMAGE PATCHES $n = 2302$. ASTERISK ($*$) SHOWS STATISTICALLY SIGNIFICANT DIFFERENCES BETWEEN THE PERFORMANCE OF BOW AND FILTER BANK REPRESENTATION WITH A TILE SIZE OF 32×32 WITH OTHER STATE-OF-THE-ART METHODS USING TWO-TAILED PAIRED WILCOXON SIGNED-RANK TEST

Method	Tile Size	AUC	95% CI	Sensitivity (%)	Specificity (%)
Colour [10]	-	0.76	[0.70, 0.82]*	95	34
Intensity [12]	-	0.75	[0.65, 0.82]*	95	15
BoW & raw-pixel rep.	256	0.63	[0.55, 0.71]*	95	11
	128	0.66	[0.58, 0.74]*	95	14
	64	0.69	[0.60, 0.78]*	95	20
	32	0.73	[0.66, 0.80]*	95	25
	16	0.74	[0.66, 0.81]*	95	27
BoW & filter bank rep.	256	0.72	[0.66, 0.79]*	95	38
	128	0.79	[0.70, 0.88]*	95	45
	64	0.82	[0.76, 0.89]*	95	48
	32	**0.87**	[0.81, 0.92]‡	95	**56**
	16	0.85	[0.79, 0.91]	95	52

*$p < 0.05$ when compared with ‡
AUC= Area Under the ROC Curve
CI= Confidence Interval

An RBF-SVM classifier was trained to classify the test image patches into diagnostically relevant or irrelevant regions.

C. Results

In initial cross-validated experiments on the training set using the two proposed methods, classification performance was better for tile size $= 32 \times 32$ pixels compared to other sizes. To investigate the effect of dictionary size on the classification performance, we chose this tile size to be constant and systematically changed the dictionary size. The number of words selected for the dictionary size was between 6 to 1000. We found that the performance of BoW and filter bank representation method improved dramatically up to a dictionary size of 30 and continued improving smoothly until the dictionary size reached 100 words (Fig. 5). The performance slightly drops after dictionary size of 100 but improves back until it reaches a size of 600. Differences in classification performance between dictionary sizes of 100 and 600 was insignificant but the time required to learn them was significant. Therefore, $s = 100$ was chosen to be the optimum dictionary size.

On the other hand, to investigate the effect of tile size for all methods, the dictionary size was set to a constant value of 100 and tile sizes were varied from 16×16 to 256×256. Table I summarizes the classification performance over 8-patient dataset for variable tile sizes and dictionary size of 100. In order to do a fair comparison between performances of the three compared methods, the operating point of all classification models was set to be at 95% sensitivity. Significant differences were found among cross-validated AUCs of all compared methods in Table I using paired non-parametric Friedman's test and between AUCs of individual experiments and that of the best performance (tile size $= 32 \times 32$ pixels) using a two-tailed paired Wilcoxon signed-rank test with a statistical power of above 84% (using the G*Power statistical

TABLE II
RESULTS COMPARING THE PERFORMANCES OF THE FOUR METHODS ON A TOTALLY UNSEEN TEST SET OF 2 PATIENTS USING BEST PERFORMING CLASSIFICATION MODELS IN TABLE I. NUMBER OF IMAGE PATCHES $n = 2849$

Method	Tile Size	Accuracy (%)	Sensitivity (%)	Specificity (%)
Colour [10]	-	24	92	14
Intensity [12]	-	27	91	10
BoW & raw-pixel rep.	16	17	99	9
BoW & filter bank rep.	32	**65**	**94**	**62**

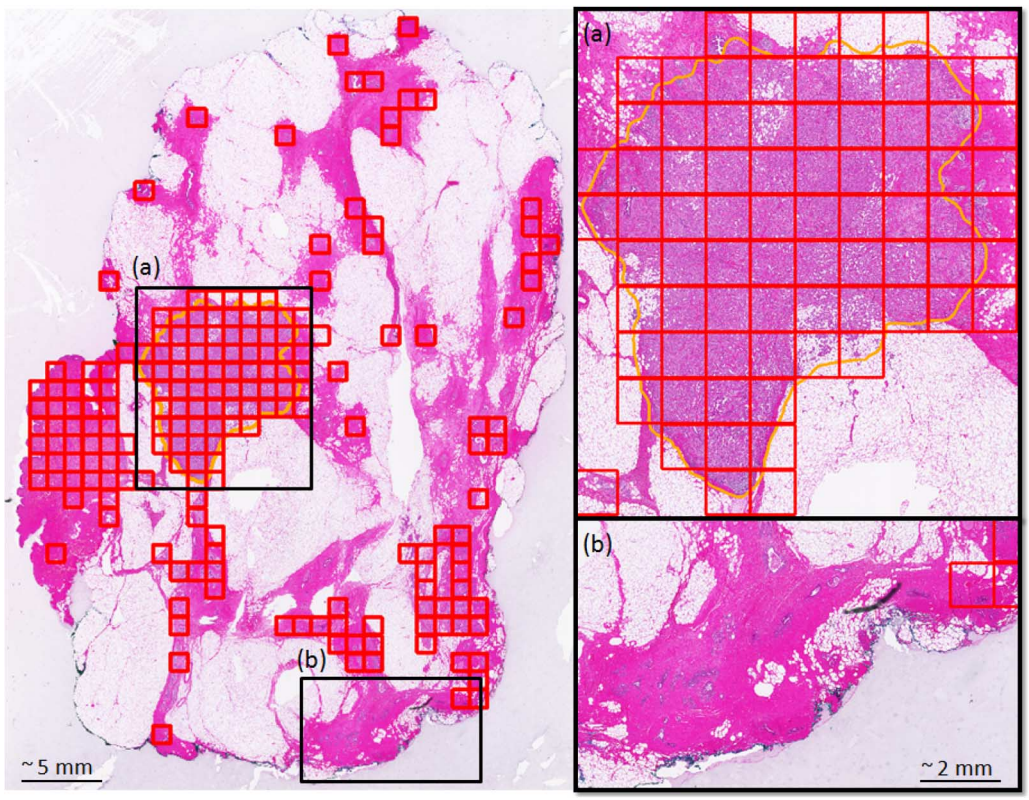

Fig. 6. Result of running our final detection model on an unseen whole-mount tissue slide which shows that most of the irrelevant tissue structure is discarded. (a) detection results (red boxes) matches that of a previous manually annotated clinically important region (yellow contour) by a trained pathologist. (b) correctly discarded region corresponding to normal ductal structure.

software [41]). We found that the tile size had the largest effect on the classification performances. As the tile size reduced down to 32 × 32 the performance improved. The classification performance of a model trained on 8 patient cases and tested on an unseen set of 2 patients data (explained in previous section) is shown in Table II. Results of this analysis indicated that the model generated using our proposed approach is still more reliable in classifying unseen test sets compared to other methods. Finally, this model was also applied to an unseen whole-mount tissue slide which had been annotated manually by a trained pathologist; our model was able to detect all contours identified as invasive or non-invasive cancer while correctly discarded about 81% of the image area (Fig. 6). This means that after applying our triaging technique only about 19% of the image area had to be manually inspected for the image slide presented in Fig. 6.

IV. DISCUSSION AND CONCLUSION

The goal of this study was to automatically detect clinically relevant regions from large post-surgical WSIs of breast cancer and discard the irrelevant ones. From the four compared colour, intensity, and texture approaches (two methods), our proposed texture filter representation features were able to better distinguish between the two image classes with a mean F1-score value of 0.44 (on the cross-validated 8-patient data). Our training set was generated by randomly selecting patches from the whole tissue area. This way the selection bias was reduced and the possibility that the overall tissue heterogeneity is considered in our feature set and classification model was increased.

We found that the tile size has the largest impact on the classification performance of a BoW technique. This maybe because as the tile size reduced, there was a higher chance of having one tissue component (epithelial, stroma, or fatty structures) isolated in any given tile. In addition, we found that the effect of dictionary size on classification performance was secondary to that of the tile size. The reason for performance decreasing after the dictionary reached 100 words may be that by further dividing the feature space, large clusters of feature points are simply divided into further smaller and adjacent groups, increasing the dimensionality of feature space without improving performance.

The combination of raw pixel representation and BoW did not perform as well as the other methods in Table I. This could be because in the implementation of this technique, raw pixel representations were directly considered as features for every tile without deriving statistical measures from them. This could generate unrepresentative feature vectors in the presence of irrelevant information for each training patch which may push the relevant training points towards the irrelevant side of the feature space.

In order to further improve the quality of the model generated by our proposed technique, texture features could be derived from overlapping tiles. This may provide more training data and hence will provide more points for the dictionary learning stage which is essential for the histogram of words generation. However, the drawback of such a policy would be an undesired increase in the run time of the dictionary learning process.

We believe that the same proposed automated triaging technique presented in this paper could be helpful in a number of other applications where the goal is to find areas of high tissue activity (with densely populated epithelial cells) in a timely manner. These applications may include automated region of interest detection for TMA analysis and cancer grading work flows. The patch-wise processing of WSIs may improve the speed of classification for large datasets compared to pixel-wise labeling methods since one label is assigned to a patch of 512 $\times$ 512 pixels at once. Our patch-wise labeling method was able to process a WSI (17,000 $\times$ 23,000 pixels, Fig. 6) with close to 900 patches on the tissue part (after removing the background by thresholding) in 27 minutes (on a 64-bit Intel (R) Xeon (R) CPU E3-1241 v3 at 3.50 GHz machine) which makes it suitable for preprocessing applications only. Implementing this technique using parallel or CUDA programming may further speed up the run-time.

For future directions of this research, it is important to classify pathology whole-mount images into regions corresponding to different invasive/pre-invasive structures such as UDH, ADH, DCIS, and/or IDC. In order to achieve this goal, whole-mount slides could first be triaged (using the proposed technique) into clinically relevant/irrelevant regions discarding as much irrelevant regions as possible, and considering relevant regions for further processing only.

Acknowledgment

The authors would like to thank Kela Liu of the Biomarker Research Laboratory, Sunnybrook Research Institute for providing the annotated whole-mount image slide used for illustration of our method in Fig. 6.

References

[1] Y. Yuan *et al.*, "Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling," *Sci. Translat. Med.*, vol. 4, no. 157, p. 157ra143, Oct. 2012.

[2] F. Viti *et al.*, "Semi-automatic identification of punching areas for tissue microarray building: The tubular breast cancer pilot study," *BMC Bioinformat.*, vol. 11, no. 1, p. 566, Jan. 2010.

[3] B. Karaçali and A. Tözeren, "Automated detection of regions of interest for tissue microarray experiments: An image texture analysis," *BMC Med. Imag.*, vol. 7, p. 2, Jan. 2007.

[4] J.-R. Dalle, W. K. Leow, D. Racoceanu, A. E. Tutac, and T. C. Putti, "Automatic breast cancer grading of histopathological images," in *Proc. Annu. Int. Conf. IEEE EMBS Conf.*, Jan. 2008, vol. 2008, pp. 3052–3055.

[5] T. J. Fuchs and J. M. Buhmann, "Computational pathology: Challenges and promises for tissue analysis," *Comput. Med. Imag. Graph.*, vol. 35, no. 7–8, pp. 515–350, 2011.

[6] A. Madabhushi, "Digital pathology image analysis: Opportunities and challenges," *Imag. Med.*, vol. 1, no. 1, pp. 7–10, Oct. 2009.

[7] A. Madabhushi, S. Agner, A. Basavanhally, S. Doyle, and G. Lee, "Computer-aided prognosis: Predicting patient and disease outcome via quantitative fusion of multi-scale, multi-modal data," *Comput. Med. Imag. Graph.*, vol. 35, no. 7–8, pp. 506–514, 2011.

[8] H. R. Ali *et al.*, "Astronomical algorithms for automated analysis of tissue protein expression in breast cancer," *Br. J. Cancer*, vol. 108, no. 3, pp. 602–612, Feb. 2013.

[9] S. J. McKenna, T. Amaral, S. Akbar, L. Jordan, and A. Thompson, "Immunohistochemical analysis of breast tissue microarray images using contextual classifiers," *J. Pathol. Informat.*, vol. 4, p. S13, Jan. 2013.

[10] C. Bahlmann *et al.*, "Automated detection of diagnostically relevant regions in H&E stained digital pathology slides," *Proc. SPIE Med. Imag.*, p. 831504, Feb. 2012.

[11] A. M. Khan, H. El-Daly, E. Simmons, and N. M. Rajpoot, "HyMaP: A hybrid magnitude-phase approach to unsupervised segmentation of tumor areas in breast cancer histology images," *J. Pathol. Informat.*, vol. 4, p. S1, Jan. 2013.

[12] S. Petushi, C. Katsinis, C. Coward, F. Garcia, and A. Tozeren, "Automated identification of microstructures on histology slides," in *Proc. 2nd IEEE Int. Symp. Biomed. Imag., Macro to Nano*, 2004, vol. 2, pp. 424–427.

[13] N. Linder *et al.*, "Identification of tumor epithelium and stroma in tissue microarrays using texture analysis," *Diagnostic Pathol.*, vol. 7, no. 1, p. 22, Jan. 2012.

[14] M. Veta, J. P. W. Pluim, P. J. van Diest, and M. A. Viergever, "Breast cancer histopathology image analysis, A review," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 5, pp. 1400–1411, May 2014.

[15] Y. Yagi, "Color standardization and optimization in whole slide imaging," *Diagnostic Pathol.*, vol. 6, p. S15, Jan. 2011.

[16] P. Bautista, N. Hashimoto, and Y. Yagi, "Color standardization in whole slide imaging using a color calibration slide," *J. Pathol. Informat.*, vol. 5, p. 4, Jan. 2014.

[17] M. Peikari, J. Zubovits, G. Clarcke, and A. Martel, "A texture based approach to automated detection of diagnostically relevant regions in," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent. Workhop Breast Image Anal.*, 2013.

[18] T. Gevers and A. W. M. Smeulders, "Color-based object recognition," *Pattern Recognit.*, vol. 32, pp. 453–464, 1999.

[19] J. Li and N. M. Allinson, "A comprehensive review of current local features for computer vision," *Neurocomputing*, vol. 71, no. 10–12, pp. 1771–1787, Jun. 2008.

[20] K. Fatima, A. Arooj, and H. Majeed, "A new texture and shape based technique for improving meningioma classification," *Microscopy Res. Tech.*, vol. 77, no. 11, pp. 862–873, Nov. 2014.

[21] O. S. Al-Kadi, "Texture measures combination for improved meningioma classification of histopathological images," *Pattern Recognit.*, vol. 43, no. 6, pp. 2043–2053, Jun. 2010.

[22] H. Qureshi, O. Sertel, N. Rajpoot, R. Wilson, and M. Gurcan, "Adaptive discriminant wavelet packet transform and local binary patterns for meningioma subtype classification," in *Proc. MICCAI*, 2008.

[23] O. Tuzel, L. Yang, P. Meer, and D. J. Foran, "Classification of hematologic malignancies using texton signatures," *Pattern Anal. Appl.*, vol. 10, no. 4, pp. 277–290, Oct. 2007.

[24] M. Varma and A. Zisserman, "A statistical approach to material classification using image patch exemplars," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 2032–2047, Nov. 2009.

[25] Y. Tang, Y.-Q. Zhang, N. V. Chawla, and S. Krasser, "SVMs modeling for highly imbalanced classification," *IEEE Trans. Systems, Man, Cybern. Part B, Cybern.*, vol. 39, no. 1, pp. 281–288, Feb. 2009.

[26] Z. Guo, Z. Zhang, X. Li, Q. Li, and J. You, "Texture classification by texton: Statistical versus binary," *PloS One*, vol. 9, no. 2, p. e88073, Jan. 2014.

[27] S. Kothari, J. H. Phan, A. N. Young, and M. D. Wang, "Histological image classification using biologically interpretable shape based features," *BMC Med. Imag.*, vol. 13, no. 1, p. 9, Jan. 2013.

[28] L. He, L. R. Long, S. Antani, and G. R. Thoma, "Histology image analysis for carcinoma detection and grading," *Comput. Methods Programs Biomed.*, vol. 107, no. 3, pp. 538–556, Sep. 2012.

[29] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
[30] M. N. Gurcan *et al.*, "Histopathological image analysis: A review," *IEEE Rev. Biomed. Eng.*, vol. 2, no. 1, pp. 147–171, Jan. 2009.
[31] G. M. Clarke *et al.*, "Increasing specimen coverage using digital whole-mount breast pathology: Implementation, clinical feasibility and application in research," *Comput. Med. Imag. Graph.*, vol. 35, no. 7–8, pp. 531–541, 2011.
[32] J.-M. Geusebroek, A. W. M. Smeulders, and J. van de Weijer, "Fast anisotropic Gauss filtering," *IEEE Trans. Image Process.*, vol. 12, no. 1, pp. 938–943, Jan. 2003.
[33] J.-k. Kamarainen, V. Kyrki, and H. Kälviäinen, "Invariance properties of Gabor filter-based features overview and applications," *IEEE Trans. Image Process.*, vol. 15, no. 5, pp. 1088–1099, May 2006.
[34] M. Varma and A. Zisserman, "A statistical approach to texture classification from single images," *Int. J. Comput. Vis.*, vol. 62, pp. 61–81, 2005.
[35] A. Cruz-Roa, J. C. Caicedo, and F. González, "Visual pattern mining in histology image collections using bag of features," *Artif. Intell. Med.*, vol. 52, no. 2, pp. 91–106, Jun. 2011.
[36] D. Sculley, "Web-scale k-means clustering," in *Proc. 19th Int. Conf. World Wide Web*, 2010, pp. 1177–1178.
[37] C.-J. Keerthi and S. Lin, "Asymptotic behaviors of support vector machines with Gaussian kernel," *Neural Comput.*, vol. 15, no. 7, pp. 1667–1689, 2003.
[38] C. Lin and H. T. Lin, "A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods," *Neural Comput.*, 2003.
[39] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, 2011.
[40] D. Ruifrok and A. C. Johnston, "Quantification of histochemical staining by color deconvolution," *Analyt. Quant. Cytol. Histol.*, pp. 291–299, 2001.
[41] F. Faul, E. Erdfelder, A.-G. Lang, and A. Buchner, "G * Power 3: A flexible statistical power analysis program for the social, behavioral, biomedical sciences," *Behav. Res. Methods*, vol. 39, no. 2, pp. 175–191, 2007.