RanPEC: Random Projections with Ensemble Clustering for Segmentation of Tumor Areas in Breast Histology Images

Adnan Mujahid Khan[1]
amkhan@dcs.warwick.ac.uk
Hesham El-Daly[2]
Hesham.El-Daly@uhcw.nhs.uk
Nasir Rajpoot[1]
nasir@dcs.warwick.ac.uk

[1] Department of Computer Science
University of Warwick
Coventry, UK

[2] University Hospital
Coventry and Warwickshire,
UK

Abstract

Segmentation of areas containing tumor cells in breast histology images is a key task for computer-assisted grading of breast tissue slides. In this paper, we present a fast, unsupervised, and data-independent framework for dimensionality reduction and clustering of high-dimensional data points which we term as Random Projections with Ensemble Clustering (RanPEC). We apply the proposed framework to pixel level classification of tumor vs. non-tumor regions in breast histology images and show that ensemble clustering of random projections of high-dimensional textural feature vectors onto merely 5 dimensions achieves up to 10% higher pixel-level classification accuracy than another state-of-the-art information theoretic method which is both data-dependent and supervised.

1 Introduction

Breast cancer is the leading cancer in women in terms of incidence both in the developed and the developing world. According to the World Health Organization (WHO), the incidence of breast cancer is increasing in the developed world due to increased life expectancy and other factors. Histological grading of breast cancer relies on microscopic examination of Hematoxylin & Eosin (H&E) stained slides and includes: assessment of mitotic count in the most mitotic area, tubule/acinar formation and degree of nuclear pleomorphism over the whole tumor. Recent studies have shown that the proliferation (mitotic) rate provides useful information on prognosis of certain subtypes of breast cancer. This is a highly subjective process by its very nature and consequently leads to inter- and even intra- observer variability. With digital slide scanning technologies becoming ubiquitous in pathology labs around the world, image analysis promises to introduce more objectivity to the practice of histopathology and facilitate its entry into the digital era [5].

Segmentation of areas containing tumor cells in breast histopathology images is a key task for computer-assisted grading of breast tissue slides. Good segmentation of tumor regions can not only highlight areas of the slides consisting of tumor cells, it can also assist in

determining extent of tissue malignancy.Though some algorithms for segmentation of nuclei, quantitative evaluation of nuclear pleomorphism, and grading of lymphocytic infiltration in breast histology images have been proposed in the literature in recent years (see, for example, [2, 5]), tumor segmentation in breast histology images has received relatively less attention.

In this paper, we address the problem of segmentation of tumor regions in a breast histology image using a features based classification approach. The proposed algorithm employs a library of textural features (consisting of just over 200 features), representing each image pixel as a point in a high-dimensional feature space. Due to the so-called *curse of dimensionality*, the high-dimensional feature space becomes computationally intractable and may even contain irrelevant and redundant features which may hinder in achieving high classification accuracy. Recent feature selection and ranking methods, such as the commonly used minimum redundancy maximum relevance (mRMR) [10], employ information theoretic measures to reduce dimensionality of the problem and have demonstrated success in several problem domains. However, major limitations of such approaches include data dependence and the requirement for training the feature selection in a supervised manner. We show that these limitations can be overcome via the proposed Random Projections with Ensemble Clustering (RanPEC) without compromising the segmentation accuracy down to pixel level.

The remainder of this paper is organized as follows. Given a set of features computed for each image pixel, we present a general framework in Section 2 which employs orthogonal random projections with ensemble clustering for assigning a label to each of the image pixels. Section 3 gives some details of the segmentation algorithm, in particular how a library of texture features is computed. Comparative results and discussion are presented in Section 4. The paper concludes with a summary of our results and some directions for future work.

2 Random Projections with Ensemble Clustering

Let $X = \{\mathbf{x}(i,j) \mid (i,j) \in \Omega\}$ denote the set of d-dimensional feature vectors for all pixels in an image $I(i,j), \forall (i,j) \in \Omega$, where Ω denotes the set of all legitimate pixel coordinates for I and $\mathbf{x} \in \mathbb{R}^d$. Suppose now that we reduce the dimensionality of all such vectors to a low-dimensional space $\mathbb{R}^r$ using a linear mapping Φ as follows: $\mathbf{y} = \Phi\mathbf{x}$, where $\mathbf{y} \in \mathbb{R}^r$ and $r << d$ and Φ is a $r \times d$ matrix containing random entries. According to the Johnson-Lindenstrauss Lemma [6], the above mapping can be used to reduce dimensionality of the feature space while approximately preserving the Euclidean distances between pairs of points in the higher d-dimensional space.

One of the major limitations of using random projections for dimensionality reduction and consequently clustering, however, is that the random matrices generated at different runs can produce variable results. Fern *et al.* [4] tackled this issue by generating a similarity matrix from multiple runs of random projections and then using the similarity matrix to drive hierarchical clustering of the data. However, the computational complexity of this approach can make it intractable for use in a large-scale setting. We propose an ensemble clustering approach to address the issue of variability in the results of clustering low dimensional feature data generated by random projections. Let $\{\mathscr{C}_1, \mathscr{C}_2, \ldots, \mathscr{C}_{n_c}\}$ denote the results of clustering the r-dimensional feature data $\mathbf{y}(i,j)$, for all $(i,j) \in \Omega$. In other words, pixel at location (i,j) has n_c labels, where n_c is the number of runs for ensemble clustering. The random projections with ensemble clustering (RanPEC) algorithm for assigning labels to each pixel is given in Algorithm1.

Algorithm 1 Random Projections with Ensemble Clustering (`RanPEC`)

1: **Input**: $X = \{\mathbf{x}(i,j) \mid (i,j) \in \Omega\}$ (where $\mathbf{x} \in \mathbb{R}^d$) the set of high-dimensional feature vectors for all image pixels, r the dimensionality of the lower-dimensional space, n the number of clusters, and n_c the number of runs for ensemble clusters.
2: **Initialization**: Generate random matrices Φ_k, $k = 1,2,\ldots,n_c$, of the order $r \times d$ with matrix entries drawn at random from a normal distribution of zero mean and unit variance.
3: **Orthogonalization**: Use Gram-Schmidt method of orthogonalization to ensure that all rows of Φ_k are orthogonal to each other and have a unit norm. In other words, ensure that $\Phi_k^T \Phi_k$ is an identity matrix, for all $k = 1,2,\ldots,n_c$.
4: **Random Projections**: Project all the feature vectors into r-dimensional space $Y_k = \{\mathbf{y}_k(i,j)\}$ where $\mathbf{y}_k(i,j) = \Phi_k \mathbf{x}(i,j)$ and $\mathbf{y}_k(i,j) \in \mathbb{R}^r$, for all $k = 1,2,\ldots,n_c$ and $(i,j) \in \Omega$.
5: **Ensemble Clustering**: Generate clustering results $\mathscr{C}_k = \{L_k(i,j)\}$ using a clustering method of your choice on the r-dimensional random projections Y_k, for $k = 1,2,\ldots,n_c$ and for all $(i,j) \in \Omega$. Use majority votes in the clustering results to decide the label $L(i,j)$ for image pixel at (i,j) coordinates.
6: **return** $L(i,j)$ for all $(i,j) \in \Omega$.

3 The Segmentation Framework

The RanPEC algorithm described above operates on the set of feature vectors X. In this section, we describe how an input image is pre-processed before computation of feature vectors and application of RanPEC on the feature vectors, followed finally by a set of post-processing operation. An overview of the segmentation framework is shown in Figure 1 with the help of a block diagram. Below we provide a brief description of each of the building blocks, without going into details due to space restrictions.

3.1 Pre-processing

Stain color constancy is one of biggest challenges of H & E staining based on light microscopy. Several factors such as thickness of the tissue section, dye concentration, stain timings, stain reactivity result in variable stain color intensity and contrast. Our pre-processing pipeline consists of stain normalization, background estimation, and edge adaptive smoothing. We used Magee *et al.*'s approach to stain normalization [8]. The background removal was achieved by masking areas containing mostly white pixels. Finally, we converted the stain normalized and background free image into the CIE's La*b* color space and applied anisotropic diffusion to its b* channel in order to remove the inherent camera noise while preserving edges.

3.2 Extraction of Textural Features

We collected a library of frequency domain textural features for each pixel in the image. These consisted of Gabor energy, phase gradients, orientation pyramid, and full wavelet packet features. We used the spatial filter for two-dimensional Gabor function [3] with orientation separation of 30° (i.e., 0°, 30°, 60°, 90°, 120°, and 150°) and 14 scales, resulting in 84 Gabor channel images. Energy of each filter's response at a pixel location was then

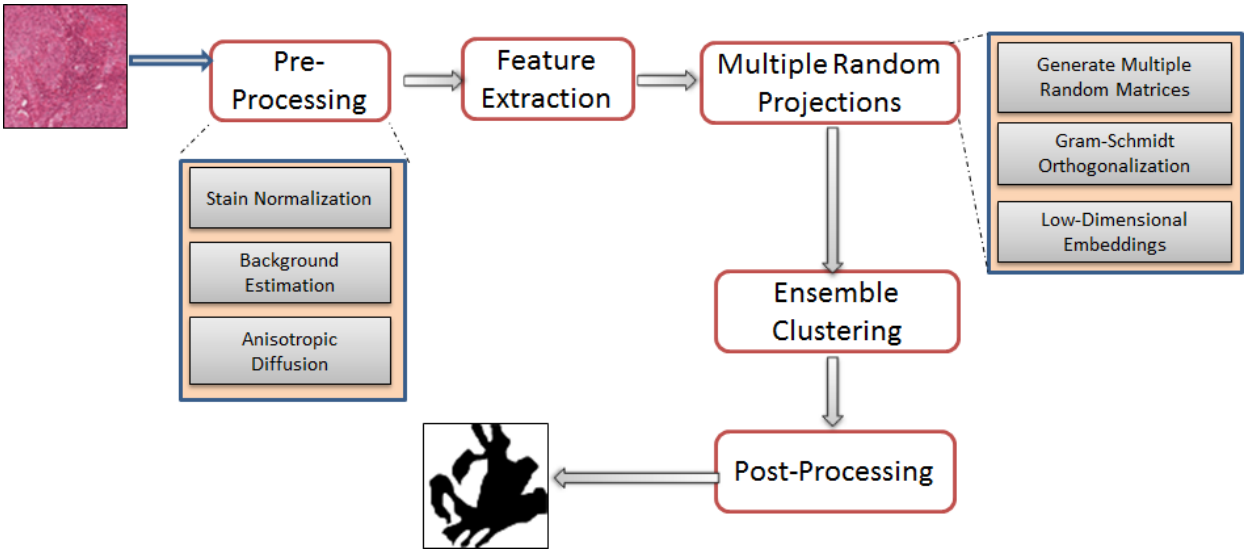


Figure 1: Overview of the proposed tumor segmentation framework.

used as a feature for that filter. Phase information can be used as an important cue in modeling the textural properties of a region. In [9], Murtaza *et al*. used *local frequency* estimates in log-Gabor domain [7] over a range of scales and orientations to yield a signature which uniquely characterizes the texture of a village in satellite images. We computed phase gradient features at 3 scales and 16 orientations to compute 48 filter responses over a window of 15×15 pixels. Next, we used the 3rd level orientation pyramid (OP) features proposed by Wilson & Spann [1, 12], resulting in 21 features. Finally, a set of 64 3-level full wavelet packet features [11] is computed to cater for fine resolution spatial frequency contents in the two texture classes (i.e., tumor and non-tumor). These four sets of features and two proximity features were then concatenated forming a 219-dimensional feature vector per pixel.

3.3 Feature Ranking for Dimensionality Reduction

Feature Ranking (FR) is a family of techniques in which a subset of *relevant* features is used to build a robust learning model that aims to achieve equal, if not better, accuracy of representing high dimensional structures. By removing irrelevant and redundant features from the data, we can improve both the accuracy of learning models and performance in terms of computational resources. Peng *et al.* [10] proposed maximum Relevance Minimum Redundancy (mRMR) feature selection method which employs mutual information to rank features. We compare the performance of mRMR feature selection with the proposed random projections with ensemble clustering (as described in Section 2).

4 Experimental Results and Discussion

Our experimental dataset consisted of digitized images of breast cancer biopsy slides with paraffin embedded sections stained with Hematoxylin and Eosin (H&E) scanned at 40× using an Aperio ScanScope slide scanner. A set of fourteen images was hand segmented by an expert pathologist. We generate all experimental results taking the pathologist's markings as ground truth (GT). All the images are pre-processed in a similar manner, with stain nor-

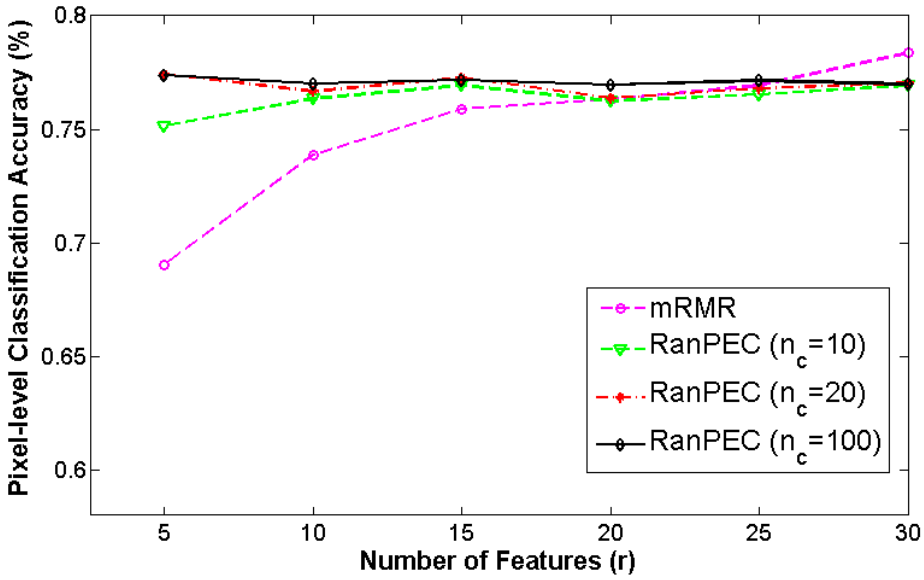


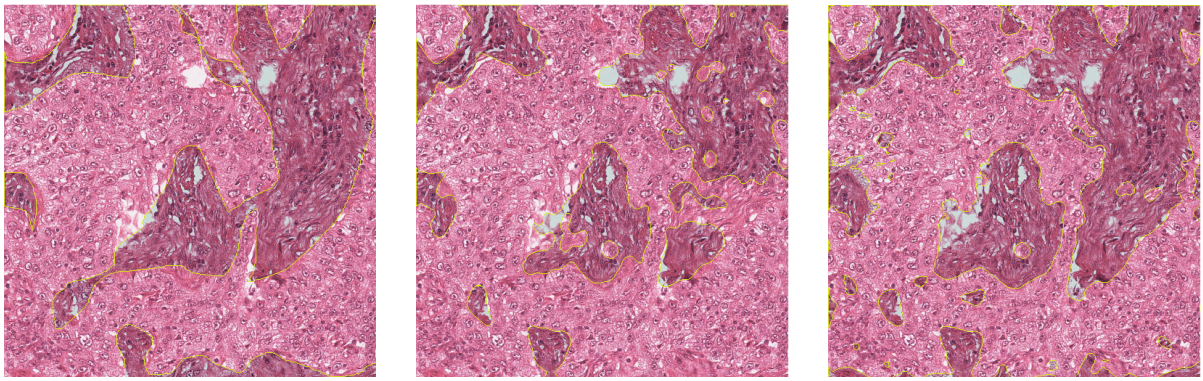
Figure 2: Comparative results of pixel-level classification accuracy (%) versus dimensionality of the feature space for mRMR and RanPEC with n_c=10, 20, and 100.

malization carried out as described in Section 3.1. Background removal is then performed to remove the artifacts caused by staining, fixation and tissue fat. This provides robustness in the subsequent steps of pipeline. As described in Section 3.2, a total of 219 textural and proximity features are calculated for each pixel of the input image I.

Multiple random projections of these textural features are used to generate multiple clustering results from the low-dimensional representation of features using the standard k-means clustering algorithm. A consensus function is then used to combine the partitions generated as a result of multiple random projections into a single partition. A simple majority function is employed on individual cluster labels to produce a single partition. Five replicates of k-means clustering are performed to get a reasonably consistent partitioning.

In order to produce mRMR ranking [10], portion of a subset of GT is chosen as training images. The choice of training images is critical as some of the images have large stromal area and small tumor area while others have vice versa. We ensure that the final training set has approximately similar representation for stromal and tumor areas. Features from all the test images are reordered and k-means clustering is performed. Post-processing is performed on clustering results obtained using both mRMR and RanPEC to eliminate spurious regions and also to merge closely located clusters into larger clusters, producing relatively smooth segmentation results.

Figure 2 presents a quantitative comparison of RanPEC with mRMR. It can be seen from these results that the application of RanPEC with $n_c = 20$ produces quite stable results for almost all values of feature space dimensionality r. Furthermore, the RanPEC results at $r = 5$ generate nearly 10% higher overall pixelwise classification accuracy than mRMR at $r = 5$. Visual results for RanPEC (at $r = 5$)and mRMR (at $r = 65$) as well as the GT are also shown in Figure 3. These results further show that for relatively small dimensionality of the feature space, RanPEC generates superior results than mRMR.



(*a*) (*b*) (*c*)

Figure 3: Visual results of tumor segmentation in a sample image: (*a*) Original image with ground truth (GT) marked non-tumor areas shown in a slightly darker contrast with blue boundaries; (*b*) Results of segmentation with 65-**dimensional** feature space using mRMR (83% accuracy) and (*c*) using RanPEC with 5-**dimensional** feature space and n_c=20 (90% accuracy).

5 Conclusions

In this paper, we addressed the issue of robustness of clustering results in the context of random projections. We proposed a framework for random projections with ensemble clustering and applied it to the segmentation of tumor areas in breast cancer histology images. We showed that the proposed framework RanPEC preserves the Euclidean distance between points in high-dimensional spaces in a robust manner. For our application of tumor segmentation, reasonably high accuracy was achieved using only 5 dimensional feature space.

6 Acknowledgments

The authors would like to thank Warwick Postgraduate Research Scholarship (WPRS) and Department of Computer Science, University of Warwick for funding the research work of AMK. The images used in this paper are part of MITOS dataset, a dataset setup for ANR French project MICO.

References

[1] C.C.R. Aldasoro and A. Bhalerao. Volumetric texture segmentation by discriminant feature selection and multiresolution classification. *Medical Imaging, IEEE Transactions on*, 26(1):1–14, 2007.

[2] A.N. Basavanhally, S. Ganesan, S. Agner, J.P. Monaco, M.D. Feldman, J.E. Tomaszewski, G. Bhanot, and A. Madabhushi. Computerized image-based detection and grading of lymphocytic infiltration in her2+ breast cancer histopathology. *Biomedical Engineering, IEEE Transactions on*, 57(3):642–653, 2010.

[3] J.G. Daugman. Uncertainty relation for resolution in space, spatial frequency, and

orientation optimized by two-dimensional visual cortical filters. *Optical Society of America, Journal, A: Optics and Image Science*, 2:1160–1169, 1985.

[4] X.Z. Fern and C.E. Brodley. Random projection for high dimensional data clustering: A cluster ensemble approach. In *International Conference on Machine Learning*, volume 20, page 186, 2003.

[5] M.N. Gurcan, L.E. Boucheron, A. Can, A. Madabhushi, N.M. Rajpoot, and B. Yener. Histopathological image analysis: A review. *Biomedical Engineering, IEEE Reviews in*, 2:147–171, 2009.

[6] W.B. Johnson and J. Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics*, 26(189-206):1–1, 1984.

[7] H. Knutsson. *Filtering and reconstruction in image processing*. Linköping University Electronic Press,, 1982.

[8] D. Magee, D. Treanor, P. Chomphuwiset, and P. Quirke. Context aware colour classification in digital microscopy. In *Proc. Medical Image Understanding and Analysis*, pages 1–5. Citeseer, 2010.

[9] K. Murtaza, S. Khan, and N. Rajpoot. Villagefinder: Segmentation of nucleated villages in satellite imagery. *British Mission Vision Conference*, 2009.

[10] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8):1226–1238, 2005.

[11] N.M. Rajpoot. Texture classification using discriminant wavelet packet subbands. In *Circuits and Systems, 2002. MWSCAS-2002. The 2002 45th Midwest Symposium on*, volume 3, pages III–300. IEEE, 2002.

[12] R. Wilson and M. Spann. *Image segmentation and uncertainty*. John Wiley & Sons, Inc., 1988.