

Texture Analysis of Virtual Slides for Grading Dysplasia in Barrett's Oesophagus.

Afzan Adam¹

scaa@leeds.ac.uk

Andrew J. Bulpitt¹

<http://www.comp.leeds.ac.uk/andyb/>

Darren Treanor²

darrentreanor@nhs.net

¹ School of Computing,

University of Leeds,

LS2 9JT Leeds, UK

² Department of Histopathology,

St. James's University Hospital,

LS9 7TF Leeds

Abstract

Barrett's Oesophagus is a pre-malignant, but treatable condition of the Oesophagus, but only fair-good interobserver agreement is achieved in grading dysplasia by histopathologists. This paper describes a method for automatic classification of dysplasia in Barrett's Oesophagus using the tissue's textural features at two levels of detail. At the first level, tiles of patches are created across the images and the texture features are extracted and clustered for each patch. The texture of these coded patches is then further analyzed, and used to grade dysplasia. Five classes of dysplasia were analysed using 150 annotated images from virtual pathology slides. The overall agreement has increased [average of 71.79% accuracy percentage with fair kappa value (0.37)] without looking at other tissue structure or features.

1 Introduction

Interobserver agreement between pathologists in diagnosing dysplasia has always been a focus area in histopathology. The overall agreement in grading dysplasia in Barrett's Oesophagus was only 57% [10], ranging from fair to good. This is mainly from the lack of a universally accepted definition of Barrett's Oesophagus where the American College of Gastroenterology, German Society of Pathology and British Society of Gastroenterology (to name a few), have different definitions of Barrett's Oesophagus [6]. On top of this various dysplasia criteria, smooth transition between grades and a pathologist's own experience in diagnosing dysplasia cases contributes to the disagreement.

As a result, a significant body of research has investigated how to extract expert knowledge and combine this with image processing capabilities to aid pathologists in detecting, locating and(or) grading tissue abnormalities automatically e.g [2, 9]. A number of approaches attempt to model the tissue's structure using methods such as Delaunay Triangulation, Voronoi Diagram, Neighboring Graph, Colour Graph, Ulam Tree, Minimum Spanning Tree (for example [1, 4]). All these approaches work on a specific type of tissue and require local structure detection before building a model or graph. However, detecting specific local structures such as nuclei, glands, goblet cells etc require a lengthy development time and are also tissue type specific.

This paper describes an alternative approach using only the hierarchy of a tissue's texture to aid the pathologists in analysing Barrett's Oesophagus virtual slides, and to reduce

variations in grading dysplasia. Indirectly, this has the potential to improve both training and confidence levels of Barrett's sufferers in the diagnostic results.

2 Methodology

2.1 Data and image preparation

Oesophagus glass and virtual slides contain several un-arranged tissues from more than one biopsy sample. They vary in size, thickness, shape, the number of tissues per slide, and staining concentrations may also differ. The virtual slides were created from glass pathology slides, scanned with an Aperio Scanner at 40X zoom level. Two expert pathologists reviewed and graded each virtual slide and were asked to annotate and grade regions which show dysplasia accordingly. Six grades of dysplasia were used [10] where all grades can be regrouped in a more general classification as shown in table 1. The transition direction shows the complexity of tissue structure and texture increase from simple Barrett's to Intramucosal Carcinoma (IMC).

Transition direction	Modified classification(2 groups)	Modified classification: 3 groups	Vienna Classification of Dysplasia
↓	Non-dysplasia	1-2	1-Barrett's only
			2-Atypia, probably negative for dysplasia
	Dysplasia	3-4	3-Atypia, probably positive for dysplasia
			4-Low grade dysplasia
			5-High grade dysplasia
			6-Intramucosal carcinoma
	5-6		

Table 1: Range of dysplasia grading which could be used.

30 annotated images were selected for grade G1, G2, G4, G5 and G6; where G3 and G4 were grouped due to the lack of annotated images for G3. Thus, a total of 150 images were retrieved at 40X magnification and reduced to lower magnification to avoid image fractions or jagged effects on the image texture if the image is scaled up. Image size was restricted where size $< 250\text{pix} * 250\text{pix}$ and $> 5000\text{pix} * 5000\text{pix}$. This is to ensure that images are not too small for feature extraction, or too large that they contain the whole Oesophagus tissue.

Colour normalisation with a Hematoxylin and Eosin (H&E) relevance vector model colour classifier [5] was used to ensure all images have a similar colour distribution. Normalised images were reviewed by a pathologist and did not have any significant identifiable artefact in them.

2.2 Texture feature extraction

Three main steps are used to extract textural attributes from the normalised annotated images, as shown in figure 2 below. In the first step, normalized annotated images are analysed at several magnifications starting from 40X and reduced to 20X, 10X and 5X. At each magnification, tiles of patches are created on each image. Different patch sizes ($pz=50\text{pixels}$, 100 pixels, 150 pixels and 200 pixels) were used to obtain the best sub-image for the next step. Sub-images in all annotated images are converted into gray-scaled images and their gray-level co-occurrence matrices (GLCM) are calculated.

Four GLCM features; contrast, energy, correlation and homogeneity in four directions are calculated. Many textures have no correlation value, so two more additional sets of GLCM features are calculated. The first set without correlation and the second additional set where the no correlation values are replaced by a small positive correlation value of 0.0001. On top of the GLCM features, local binary pattern (LBP) features were also explored for each

sub image but produced data overhead even after feature selection. The results from using *GLCM* were found to be better.

Next, all sub-images are clustered using k-means clustering based on the similarity of their *GLCM* features using square Euclidean distance with k tested was 3,4,5,6 and 7. Some samples of the clustered sub-images in 20X magnification with $k=5$ and $p_z=100*100$ pixels are shown in figure 1 below. The clusters obtained from this process are then used to produce cluster coded images (*CCI*) for each annotated image.






Clusters	Sample images	Clusters	Sample images
C1		C2	
C3		C4	
C5			

Figure 1: Clustered sub-images across all annotated images.

Examples of *CCIs* are shown in figure 2 with each cluster number replaced with colour for easier visualization. In order to encode the spatial relationships between textures, cluster co-occurrence matrices (*CCM*) for each *CCI* are calculated. *CCM* features are calculated using the same principle as *GLCM*.

Three sets of *CCM* features are calculated. The first set contains the energy, contrast, homogeneity and correlation between each patch for all directions. The second set contains the direction invariant features where the same four features are averaged over all four directions, and the final set contains only the cluster frequencies in each image.

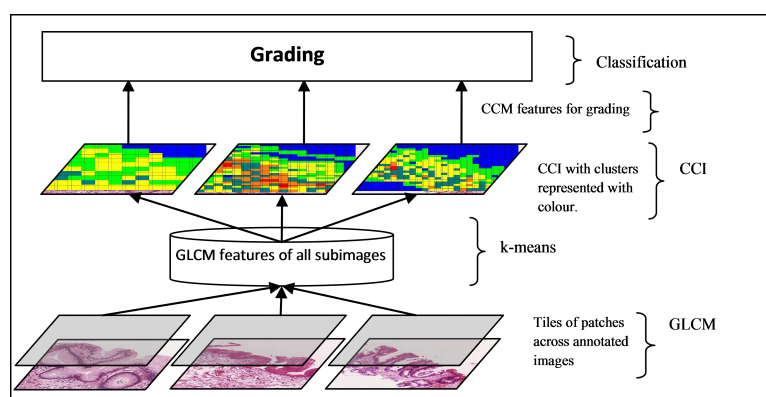


Figure 2: Texture features are extracted in two layers; the *GLCM* features represent texture at patch level while *CCM* features represent texture at a more general level.

The *CCM* features are used to classify the images into its dysplasia grading, using random forest (*RF*) and Decision Tree *DT* classifiers. The results are compared to the grade given by pathologists for evaluation.

3 Results

Annotated images were classified into a)5 specific grades, b)3 groups of grades and c)2 specific grades, as shown in table 1. Interobserver agreement between pathologists and the classification result are used to measure the performance, and is compared with the existing agreement score achieved by pathologists in grading similar cases. The interobserver agreement between the classifier and pathologist grades is measured using kappa statistics value

(*KV*). *KV* has long been used to access the degree of interobserver variation in pathology; where <0.21 , $0.21-0.40$, $0.41-0.60$, $0.61-0.80$ and >0.80 are commonly accepted interpretation of poor, fair, moderate, good and very good agreement respectively.

Different combinations of parameters were tested at many levels of this research to obtain the best texture features to represent the tissue images. The best results with consistent performance were achieved using 20X magnification with $pz=100*100$ pixels. For *GLCM*, the most consistent results were achieved from a very small correlation value with $n=10$ for the neighbouring pixels and for *CCI*, $n=2$. For unsupervised clustering, the best number of clusters was found to be $k=5$. Histograms of cluster frequency give the most consistent and best classification results.

Table 2 shows cases which are frequently discussed in literature. Poor results were achieved when classifying images into 5 specific dysplasia grades. Classifying images into 3 groups (no dysplasia, dysplasia and severe dysplasia) produced better results where fair agreement was achieved with the pathologists. Grading dysplasia into 2 classes produced even better results, particularly between groups of consecutive grades (table 1).

No of Classes	Cases	Classifier				Literature review
		<i>DT</i>		<i>RF</i>		
		<i>AP</i>	<i>KV</i>	<i>AP</i>	<i>KV</i>	
5	G1 vs G2 vs G4 vs G5 vs G6	31.13	0.14	28.2	0.1	0.25-.27
3	G1+G2 vs G4 vs G5+G6	47.41	0.21	44.44	0.17	0.24
2	G5 vs G6	71.47	0.42	74.13	0.47	0.42
2	G6 vs G1+G2+G4+G5	79.53	0.41	81.47	0.37	*
2	G1+G2 vs G4+G5+G6	72.40	0.38	67.00	0.31	0.33
2	G4 vs G5	62.05	0.24	63.27	0.26	*
2	G4 vs G1+G2	72.4	0.38	63.27	0.24	<50%
2	G4 vs G1+G2+G5+G6	69.73	0.09	72.67	0.03	72%
2	G5 vs G1+G2+G4+G6	78.60	0.00	75.27	0.14	*

Table 2: Comparison of classification between *DT* and *RF* achieved at 20X. (* not found in literature)

In addition, frequency of clusters in *CCI* gives consistently high accuracy percentage(*AP*) and *KV* for certain grades. These was achieved across all value of k in 20X. The texture with highest frequency contribute to classify images into the correct grade, particularly in 3 cases, as shown in table 3.

Cases	k value	Average <i>AP</i>	Average <i>KV</i>	Texture of:
HGD vs IMC	3,5,6,7	69.27	0.39	C5
IMC vs Non	3,5,6,7	81.06	0.37	C2
No dysp vs dysp	3,5,6,7	64.20	0.24	C3

Table 3: Significant texture to classify certain grade.

4 Discussion and conclusions

Classifying regions into 5 different grades obtained only slightly better results than random (31.13%). However there is moderate agreement in grading dysplasia into three groups, similar to the kappa value scored by pathologists [3, 7]. The best classification results with high agreement were achieved when grading between 2 grades of dysplasia. The results also depend on the grades of dysplasia used. This is in line with interobserver variations reported in [8], among others. More features are needed to distinguish middle grades from the rest of its neighbouring grades.

Poor results were obtained when the system is used to distinguish the texture between LGD and non-LGD, as well as between HGD and non-HGD. This is because both LGD and

HGD are in the middle of the transition process from no dysplasia to becoming IMC. The texture of these two grades is believed to have combinations of the classes either side of them in the transition process, thus differentiating them is difficult. High accuracy percentages were shown because of the unbalanced data between the two grades.

The results achieved in this set of images have demonstrated that tissue changes across a few grades of dysplasia in Barrett's Oesophagus can be measured using only the texture features with kappa values achieved in diagnosing certain grades equal or higher than the value achieved between pathologists themselves.

Acknowledgements

Thanks to The Leeds Institute of Molecular Medicine (LIMM) and Prof. Mike Dixon and Dr. Nigel Scott from St. James's University Hospital.

References

- [1] S. Doyle, M. Hwang, K. Shah, and A. Madabhushi. Automated grading of prostate cancer using architectural and textural image features. In *Proc. of IEEE International Symposium on Biomedical Imaging*, pages 1284–1287, 2007.
- [2] P.W. Hamilton, P.H. Bartels, D. Thompson, N. H.. Anderson, R. Montiront, and J.M Sloan. Automated location of dysplastic fields in colorectal histology using image texture analysis. *Journal of Pathology*, 182, 1997.
- [3] M. Kerkhof, H.V Dekken, and E.W. Steyerberg et al. Reproducibility of the diagnosis of dysplasia in barrett's oesophagus: a reaffirmation. *Histopathology*, 50(7):368–378, 2007.
- [4] G. Landini and I.E. Othman. Architectural analysis of oral cancer, dysplastic and normal epithelia. *Cytometry, Part A*(61A):45–55, 2004.
- [5] D. Magee, D. Treanor, D. Crellin, M. Shires, K. Mohee, and P. Quirke. Colour normalisation in digital pathology images. In *Proc. of Optical Tissue Image Analysis in Microscopy, Histopathology and Endoscopy*, pages 100–111, 2009.
- [6] D.M. Maru. Barrett's esophagus: Diagnostic challenges and recent developments. *Annals of Diagnostic Pathology*, 13:212–221, 2009.
- [7] E. Montgomery, M. Bronner, and J. Goldblum et al. Reproducibility of the diagnosis of dysplasia in barrett's oesophagus: a reaffirmation. *Human Pathology*, 32(4):368–378, 2001.
- [8] A.H. Ormsby, R.E Petras, W.H Henricks, T.W Rice, L.A Rybicki, J.E Richter, and J.R Goldblum. Observation variation in the diagnosis of superficial oesophagus adenocarcinoma. *Int. Journal of Gastroenterology and Hepatology*, pages 671–676, 2002. doi: 10.1136/gut.51.5.671.
- [9] R. Susomboon, D. Raicu, J. Furst, and T.B. Johnson. A co-occurrences texture semi-invariance to direction, distance and patient size. In *Proc. of SPIE Medical Imaging Conf.*, pages 1–6, 2008.
- [10] D. Treanor, C.H. Lim, A. Bulpitt, and D. Magee. Tracking with virtual slides: A tool to study diagnostic error in histopathology. *Histopathology*, 2009.