Discriminating dysplasia: optical tomographic texture analysis of colorectal polyps

Wenqi Li\textsuperscript{a}, Maria Coats\textsuperscript{b}, Jianguo Zhang\textsuperscript{a}, Stephen J. McKenna\textsuperscript{a}

\textsuperscript{a}CVIP, School of Computing, University of Dundee, Dundee, UK, DD1 4HN
\textsuperscript{b}School of Medicine, University of Dundee, Dundee, UK, DD1 4HN

Abstract

Optical projection tomography enables 3-D imaging of colorectal polyps at resolutions of $5 - 10 \mu m$. This paper investigates the ability of image analysis based on 3-D texture features to discriminate diagnostic levels of dysplastic change from such images, specifically, low-grade dysplasia, high-grade dysplasia and invasive cancer. We build a patch-based recognition system and evaluate both multi-class classification and ordinal regression formulations on a 90 polyp dataset. 3-D texture representations computed with a handcrafted feature extractor, random projection, and unsupervised image filter learning are compared using a bag-of-words framework. We measure performance in terms of error rates, \textit{F}-measures, and ROC surfaces. Results demonstrate that randomly projected features are effective. Discrimination was improved by carefully manipulating various important aspects of the system, including class balancing, output calibration and approximation of non-linear kernels.

\textit{Keywords:} Volumetric texture, Optical projection tomography, 3-D histopathology, Colorectal polypoid cancer

1. Introduction

Colorectal cancer is the third most common cancer in men (756k new cases per annum, 10.0\%) and the second in women (614k new cases per annum, 9.2\%) worldwide \cite{Ferlay et al., 2013}. Colonic polypoid cancers are the earliest detectable form of colorectal cancer and, if completely excised, can offer a potential cure to patients. Screening programmes have played a major role in keeping cancer mortality rates low. As of 2012, 65,535 polyps had been excised and recorded in the UK NHS bowel cancer screening programme.
Figure 1: (a) An H&E-stained whole mount slide. (b) A virtual section from an OPT image of the same polyp with the section chosen to be close to the cut surface that resulted from removal of the physical section shown in (a).

(BCSP) database (Majumdar et al., 2012). Histological analysis of polyps is essential for accurate diagnosis. However, the current gold standard method, using Haematoxylin and Eosin (H&E) stained sections, has its limitations. This conventional technique involves taking a thin section of tissue from the centre of the polyp and this will not necessarily be representative of the whole specimen. Much ambiguity exists between experienced pathologists when making diagnoses using H&E sections due to features such as epithelial displacement (EPD) in which surface epithelial cells become misplaced into the stalk of the polyp mimicking true invasive cancer. Over-diagnosis of EPD as cancer has a confounding effect subjecting patients to unnecessary treatments and generating misleading epidemiology reports (Loughrey and Shepherd, 2015; Muto et al., 1973).

Optical projection tomography (OPT) is a relatively new 3-D imaging technique first developed to better our understanding of embryo development (Sharpe, 2009). It is a simple and affordable imaging technology that is well-suited for specimens up to 15 mm in size. OPT of colorectal polyps is non-destructive to the original tissue and enables virtual sectioning of the specimen at any orientation. Figure 1 shows a comparison between an H&E
section and a virtual OPT section of the same polyp. A significant advantage of tomography is the flexibility in viewing virtual sections and manipulating the image to gain more information. By contrast, once the cutting angle has been chosen for the H&E section, it cannot be changed. Histology sections are cut from the tissue once it has been embedded in paraffin wax and subsequently stained. These can be viewed at sub-micron resolutions whereas OPT provides a lower spatial resolution of about 5-10µm (Sharpe et al., 2002). Near visible wave-length light is used to obtain OPT images and therefore polyps must be optically cleared in advance of scanning using Benzyl Alcohol Benzyl Benzoate (BABB).

Inter- and intra-observer variation exists when pathologists diagnose colorectal polyps, notably when grading dysplasia whether from H&E or OPT images (Coats et al., 2012). Reliable and repeatable automatic recognition systems are desirable. As a starting point, this paper proposes and evaluates methods for discrimination of dysplasia in OPT images of colorectal polyps. Specifically, the task addressed is that of differentiating between regions of low-grade dysplasia (LGD), high-grade dysplasia (HGD) and invasive cancer (ICA).

The use of automated image analysis for histopathological grading of colorectal polyps has been studied mainly with stained physical sections imaged using 2D microscopy (e.g., Atlamazoglou et al., 2001; Esgiar et al., 1998, 2002; Ficsor et al., 2008; Hamilton et al., 1997, 1995; Kalkan et al., 2012; Ozdemir and Gunduz-Demir, 2013; Shuttleworth et al., 2002). This body of work demonstrates that texture-based analysis is often an important component of lesion detection, segmentation and classification. 3-D texture features have been widely used in medical image analysis more broadly; a comprehensive review of 3-D texture analysis methods is available elsewhere (Depeursinge et al., 2014). Texture representation using local binary patterns (LBPs) and its variants forms an important component of many successful 3-D medical image classification systems e.g., for brain white matter lesion classification in MRI (Oppedal et al., 2012), analysis of lung CT (Sorensen et al., 2010), and retinal optical coherence tomography (Liu et al., 2011). Apart from hand-crafted features as exemplified by LBP, two fundamentally different directions for texture analysis have attracted increased attention in recent years. The first is domain-specific descriptors estimated from observations automatically. For example, deep belief networks have been used to learn features for 3-D brain image segmentation (Brosch and Tam, 2013), and independent subspace analysis (ISA) has been used for feature extraction from
H&E histology images of Glioblastoma Multiforme (Le et al., 2012), MR images (Liao et al., 2013) and video (Le et al., 2011). The second is random projection (RP), a non-adaptive dimensionality reduction tool motivated by compressive sensing theory, applied directly to image patches. RP compared favourably to state-of-the-art texture descriptors in 2D texture classification experiments (Liu and Fieguth, 2012) and it has been used for image classification (Bingham and Mannila, 2001) and for accelerating feature-based registration of 3-D neural ultrastructure (Akselrod-Ballin et al., 2011), for example. In this paper, for the purpose of discriminating dysplasia, these three contrasting approaches to texture representation are compared. Specifically, we compare RP, LBP (exemplifying hand-crafted features), and ISA (exemplifying unsupervised learning).

The main contribution of this paper is that it provides the first study in the literature on automatically discriminating between invasive cancer, high-grade dysplasia and low-grade dysplasia in optical projection tomography images. It builds on and extends earlier conference papers in which preliminary two-class classification experiments discriminating only between LGD and ICA were reported (Li et al., 2013a,b). While the task of discriminating between LGD, HGD and ICA can be cast as a three-class classification problem, this ignores the ordinal structure of these labels. Here, a classification model and an ordinal regression model, both based on margin maximisation, are compared and contrasted for this task. These raise issues of class imbalance and output calibration which are explored empirically. Two state-of-the-art strategies for fast approximation of non-linear kernels are also evaluated. After describing the methods used, results are reported on a data set of 90 polyps. Although the focus is on OPT images of colorectal polyps, the analysis and evaluation methods used should be applicable to other ordinal regression tasks in other image modalities.

2. Materials and Methods

2.1. Tissue Processing, Imaging, and Annotation

Ninety colorectal polyps were selected from the NHS Tayside Tissue Bank archive to be representative of the dysplasia subgroups: invasive cancer (ICA), high-grade dysplasia (HGD) and low-grade dysplasia (LGD). Thirty samples were selected for each of these three groups to give a balanced dataset. The H&E stained sections taken from each specimen were re-diagnosed by an experienced gastro-intestinal histopathologist according to
Figure 2: Direct renderings of OPT polyp images with polyp voxels rendered as opaque: (a), (b) a low-grade dysplasia (LGD) polyp, (c), (d) a high-grade dysplasia (HGD) polyp, (e), (f) an invasive cancer (ICA) polyp. Top row: viewing angles adjusted to view the natural surfaces of the polyps. Bottom row: viewing angles adjusted to view artefactual surfaces due to physical cuts.
Figure 3: Images showing slices with regions annotated as (a) LGD, (b) HGD, and (c) ICA

the NHS BCSP and WHO guidelines to reduce intra-observer bias (Hamilton et al., 2000; NHS BCSP, 2007). Images were acquired using OPT in emission mode under ultraviolet light and Cy3 dye at a voxel resolution of $6.7 \mu m^3$. Each image was of one colorectal polyp and had $1024^3$ voxels. Figure 2 shows renderings of three of these colorectal polyp OPT images.

Each 3-D image was manually annotated with 3-D regions by an individual experienced in interpreting OPT images. Characteristic regions were annotated in each polyp, i.e., regions of ICA were annotated in polyps labelled as ICA, regions of HGD were annotated in polyps labelled as HGD, and regions of LGD were annotated in polyps labelled as LGD. Each region’s boundary was delineated such that the annotator had high confidence that all tissue within the region was correctly labelled. The H&E slide corresponding
to the cut surface of each polyp was used as guidance for this annotation. Annotations were performed using the software tool ITK-SNAP (Yushkevich et al., 2006) by delineating 2D regions every 4 or 5 slices and then interpolating between them. Figure 3 shows some examples of annotated slices. Figure 4 summarises the quantities of voxels annotated per slice.

2.2. Patch Encoding

In the following, patch denotes a cube-shaped image region to be classified; window denotes a smaller cube-shaped region of $d^3$ voxels from which local texture features are extracted. Figure 5 shows example patches from LGD, HGD and ICA. LGD tends to have somewhat regular texture as shown in Figure 5(a); LGD normally has tubular morphological structure. ICA shown in Figure 5(c) contains more homogeneous micro-texture patterns corresponding to more dense tissue with villous structure. HGD is intermediate in appearance as shown in Figure 5(b). The bag-of-words framework is adopted to encode patches with a pre-learned visual dictionary (Zhang et al., 2007). The procedure of encoding a patch is illustrated in Figure 6. A window slides through the patch and at each location, a texture feature vector is extracted from the window. The feature vectors are quantised into visual words by matching with the most similar visual word in a pre-learned dictionary. Finally, an L1-normalised histogram of visual word frequencies constitutes the patch representation. Window step size was set to half the window width. Dictionaries of 200 visual words were obtained with k-means++ (Arthur and Vassilvitskii, 2007). Bag-of-words provides a compact representation of fixed dimensionality regardless of the number of local windows used. It offers a uniform approach to comparative
evaluation of local feature extraction methods. The next section elaborates the feature extraction methods used.

2.3. Extracting Local 3-D Texture Features

Random projection (RP), local binary patterns (LBP) and independent subspace analysis (ISA) techniques were adopted to compute 3-D texture representations. These methods were selected from the vast literature on texture features mainly because they are among the most advanced and potentially effective for OPT classification. Computational efficiency was also considered when choosing these methods since the OPT datasets in our experiments are large. Random projection was reported effective in representing general 2-D textures [Liu and Fieguth 2012]. It can be extended to 3-D and computed efficiently. Local binary patterns are invariant to local contrast changes and local rotations. These properties are potentially capable of rep-
representing micro-textures in OPT images. An approximation of 3-D LBP was employed which is not fully invariant to arbitrary 3-D rotations but computationally feasible. Independent subspace analysis can be used to learn a set of phase- and shift-invariant filters that have similarities to features computed by complex cells in the V1 area of primate visual cortex. ISA learning scales well to large training sets (e.g., [Le et al., 2011]). The following sections present technical details of the feature extraction procedures.

2.3.1. Random Projection

Let \( X \) be a \( d^3 \times N \) matrix in which the elements in each column are the voxel values of one of \( N \) windows. The random projection method maps windows onto a \( k \)-dimensional subspace using a suitably generated \( k \times d^3 \) random projection matrix \( R \) (Equation (1)).

\[
\hat{X} = RX.
\] (1)

Each element in the matrix \( R \) is an independent sample from a standard normal distribution, i.e., a Gaussian with zero mean and unit variance. After projection, the columns in \( \hat{X} \) are considered as the window descriptors. According to the Johnson-Lindenstrauss lemma ([Dasgupta and Gupta, 2003]), data points in \( \mathbb{R}^{d^3} \) are embedded into the lower dimensional Euclidean space \( \mathbb{R}^k \) such that pair-wise distances between columns in \( X \) are approximately preserved. The computational complexity of RP is only \( O(d^3kN) \). In experiments reported in this paper, \( k \) was set to 200 unless \( d^3 < 200 \) in which case the window was transformed with a square \( (d^3 \times d^3) \) random projection matrix.

2.3.2. Local Binary Patterns

Computing local binary patterns (LBP) from 2D images involves thresholding each \( 3 \times 3 \)-pixel neighbourhood at the value of its central pixel thus
obtaining an 8-bit binary code. A histogram of these codes over an image window can then be used as a local descriptor. This representation, known as LBP$_{8,1}$, (i.e., 8-bit LBP with radius 1 neighbourhood), is not rotation invariant. [Ojala et al., 2002] found that the vast majority of binary codes in a local neighbourhood are so called uniform patterns. To achieve rotational invariance (around the central pixel) using uniform patterns, all non-uniform LBP patterns are stored in a single bin in the histogram computation. The length of the uniform LBP$_{8,1}$ descriptor is 59 bins, which is smaller than the permutations ($2^8$ bins). In a 3-D volumetric image, designing LBPs that are invariant to arbitrary rotations is not straightforward as the ordering of 3-D neighbourhood is undefined. [Fehr and Burkhardt, 2008] addressed the problem by computing spherical correlations in the frequency domain. The approach is robust to 3-D rotations however it is computationally expensive. We choose to approximate the 3-D LBP with uniform LBP$_{8,1}$ descriptors computed in each of three orthogonal planes, taken to be aligned with the image axes for convenience ([Zhao and Pietikainen, 2007]). This is computationally feasible in our experiments. In addition, this approximated 3-D LBP has been demonstrated to be competitive with a full volumetric LBP for encoding texture features in microscopic images ([Majtner and Svoboda, 2014]). The 3-D uniform LBP$_{8,1}$ operator encodes a window as a histogram with 177 bins.

2.3.3. Independent Subspace Analysis

Independent subspace analysis (ISA) is an unsupervised feature learning method based on natural image statistics ([Hyvärinen and Hoyer, 2000]). In the ISA method, image filters $W = \{w_l\}_{l=1}^L$ can be learned from a set of image windows $X = \{x_n\}_{n=1}^N$. The invariances are achieved by grouping the image filters into subspaces: the filter responses within the same subspace are pooled together, while among different groups the filter responses are treated independently. The model is estimated by maximising the log-likelihood function ([Hyvärinen et al., 2009]):

$$\log L(X|W) = -\sum_{n=1}^N \sum_{l=1}^L \sqrt{\sum_{i \in g(l)} (w_i^T x_n)^2},$$

subject to: $WW^T = I$,  \hspace{1cm} \text{(2)}

where each image filter $w_l$ is a $d^3$-dimensional vector which is applied to the image window; $g(l)$ is the set of indices of $l$-th group in the total $L$ groups; $I$
is an identity matrix. The constraint (3) is introduced to reduce the number of free parameters and leads to more stable solutions (Hyvärinen, 2013).

Figure 7: (a-f) Six groups of 4 filters learned from 2700 windows (filter size: $13 \times 13 \times 13$). Filters within the same subspace (group) shared some similarities while filters in different subspaces had different orientations and frequencies.

This optimisation was performed using stochastic gradient descent after applying whitening transformations to remove correlations between voxels and then PCA to reduce the dimensionality to 300 (whenever the number of voxels in the window was greater than 300). A set of 300 filters forming 75 independent subspaces (4 filters per subspace) were simultaneously learned by optimising Formula (2). For each optimisation process 200 passes of stochastic gradient descent with adaptive learning rate were applied through the entire set of training windows. After training, windows are represented as vectors containing 300 filter responses. Some filters learned from the training window set are visualised in Figure 7. These 300 responses are further pooled into a 150-dimensional vector as the final window representation. The pooling process takes the square root of squared sum of two responses in the same subspace (illustrated in Figure 8). Filters within a given subspace are similar.

2.4. Multi-Class Patch Classification

When formulated as a three-class classification problem, the task of discriminating between LGD, HGD and ICA was addressed using a set of three
support vector machine (SVM) binary classifiers (Allwein et al., 2001). Each classifier was trained to discriminate one class from the others. This one-vs-rest approach leads to unbalanced data sets for each of the classifiers. Furthermore, it is important to calibrate the outputs of the classifiers before comparing them in order to infer the class label. A further consideration is the use of non-linear kernels in the classifiers which, given large data sets, can be approximated for practical reasons. This section describes the classification method used and considers available solutions to each of these issues.

2.4.1. Binary Subproblem

Given a training set \( \{x_i, y_i\}_{i=1}^l \) where \( x_i \) is the \( i^{th} \) bag-of-words feature vector and \( y_i \in \{-1, 1\} \) is its corresponding class label, a classification function \( f \) that maps the feature vector to the label set \( \{-1, 1\} \) can be found using the widely used SVM formulation by minimising

\[
L(w, \xi) = \frac{1}{2} ||w||^2 + C \sum_{i=1}^l \xi_i, \tag{4}
\]

subject to : \( y_i(w^T x_i + b) \geq 1 - \xi_i, \ \xi_i \geq 0, \tag{5} \)

where \( C \) is a parameter controlling the trade-off between model complexity and training errors; \( \xi_i, \ i = 1, \ldots, l \) are slack variables; \( w \) is the weight vector, \( b \) is a bias weight, and \( f(x) = w^T x + b \) is the SVM hyperplane to be optimised. This optimisation was performed using the primal form solver provided in the LIBLINEAR package (Fan et al., 2008). \( C \) was searched over the set \( \{2^\lambda | \lambda \in \mathbb{Z} \land [-15, 15]\} \).
The binary classification sub-problems do not have balanced data sets because in each case one class is being discriminated against all other classes. Class imbalance also arises because HGD is a less commonly assigned label than LGD or ICA (see Figure 1). This problem can be addressed by replacing the free parameter $C$ with $C_p$ and $C_n$ for positive and negative classes respectively. Formula (4) then becomes:

$$\text{Minimise} \quad \frac{1}{2}||w||^2 + C_p \sum_{i=1}^{l^+} \xi_i^+ + C_n \sum_{i=1}^{l^-} \xi_i^-,$$

subject to:

$$w^\top x_i^+ + b \geq 1 - \xi_i^+,$$  \hspace{1cm} (7)

$$-w^\top x_i^- - b \geq 1 - \xi_i^-,$$  \hspace{1cm} (8)

$$\xi_i^+ \geq 0, \quad \xi_i^- \geq 0,$$  \hspace{1cm} (9)

where $x_i^+, x_i^-$ are positive and negative training examples in the one-vs-rest setting; $l^+$ and $l^-$ are the numbers of such examples; $\xi_i^+, i = 1, \ldots, l^+$ and $\xi_i^-, i = 1, \ldots, l^-$ are slack variables; $w$ is the weight vector, $b$ is a bias weight, and $f(x) = w^\top x + b$ is the SVM hyperplane to be optimised; $C_p$ and $C_n$ are set as $C_p = \frac{l^+ + l^-}{2}C_0$ and $C_n = \frac{l^+ + l^-}{2}C_0$. $C_0$ was searched over the set $\{2^\lambda \mid \lambda \in \mathbb{Z} \land [-15, 15]\}$.

Three binary classifiers are trained for the three-class classification problem. Traditionally in the one-vs-rest scheme the final output score for a test example is the highest among the scores given by the three classifiers. However, since the three classifiers are trained independently, the scores are not necessarily comparable. This situation is helped by calibrating the scores prior to making this comparison by using Platt’s scaling method to obtain values that can be treated as class probability estimates (Lin et al., 2007). Calibration maps the binary classifier output ($w^\top x$) onto values that can be treated as probabilities with a parameterised sigmoid function:

$$P(y = 1|x; w, A, B) = \frac{1}{1 + \exp(Aw^\top x + B)},$$  \hspace{1cm} (10)

where $A$ and $B$ are learned from a validation set. For $k$-class classification, $k$ such sigmoid functions are estimated, one per binary classifier. Unbalanced datasets mean that the fitting of these sigmoid functions is more heavily...
effected by over-represented classes. This problem was recently addressed by Wallace and Dahabreh (2014) who introduced bagging of under-sampling data estimators to refine the calibration procedure. Similarly, we apply the bagging method to one-vs-rest classifier calibration. More specifically, $T$ balanced patch sets are formed by randomly discarding patches from over-represented classes and $T$ calibration models are learned from these sets. The final probability estimate from a binary classifier is an average of the $T$ models, i.e.,

$$ P(y = 1 | \mathbf{x}; \mathbf{w}) = \frac{1}{T} \sum_{t=1}^{T} P^{(t)}(y = 1 | \mathbf{x}; \mathbf{w}, A^{(t)}, B^{(t)}), \quad (11) $$

where $P^{(t)}(y = 1 | \mathbf{x}; \mathbf{w}, A^{(t)}, B^{(t)})$ is obtained by applying Equation (10) with $A^{(t)}$ and $B^{(t)}$ estimated from the $t$-th balanced patch set.

2.4.4 Non-Linear Kernel Approximation

Using appropriate non-linear kernels that map feature vectors into a high dimensional space can improve the classification performance of bag-of-words encoding. For example, the $\chi^2$ kernel has been used for classification of endoscopic images (Tamaki et al., 2013). However, this is computationally prohibitive for large-scale problems due to the expensive operation of constructing a Gram matrix over all training data. Recently the method of approximating kernels with explicit feature maps enabled the use of non-linear kernels with relatively low computational cost on large scale datasets (Maji et al., 2013; Vedaldi and Zisserman, 2012). The main idea is that for a homogeneous kernel $k(\mathbf{x}, \mathbf{y}) : \mathbb{R}^M \times \mathbb{R}^M \mapsto \mathbb{R}$, it is possible to use an approximation function $\Phi(\cdot) : \mathbb{R}^M \mapsto \mathbb{R}^{M'}$ so that $k(\mathbf{x}, \mathbf{y}) \approx \Phi(\mathbf{x})^T \Phi(\mathbf{y})$. The $\Phi(\cdot)$ can be data-independent. A linear SVM can then be used directly with $\Phi(x)$ for classification. This leverages the high performance of non-linear kernels while maintaining scalability to large scale problems.

Given two OPT patches encoded with bag-of-words, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^M$, the $\chi^2$ kernel is computed as $k_{\text{chi}}(\mathbf{x}, \mathbf{y}) = \sum_{m=1}^{M} (2x_m y_m)/(x_m + y_m)$, and the histogram intersection kernel is computed as $k_{\text{hist}}(\mathbf{x}, \mathbf{y}) = \sum_{m=1}^{M} \min(x_m, y_m)$. These two kernels were approximated with methods proposed by Vedaldi and Zisserman (2012) and Maji et al. (2013) respectively.

2.4.5 Inherently Multi-class Classifiers

The one-vs-rest SVM classifications were compared with three inherently multi-class classifiers: random forest (RF) (Breiman, 2001), $k$-nearest neigh-
bours (KNN), and multi-layer perceptrons (MLP). A non-linear classification
function \( \hat{f} \) that maps the feature vector to the label set \{LGD, HGD, ICA\}
was trained for each method to discriminate three classes. For RF classifier we report performances of 2,000 randomised classification trees. \( k \) was
searched over the set \{5, 10, 20\} in the KNN classifier. We set one hidden
layer with 20 neurons for the MLP\(^1\). For the one-vs-rest SVM in the compari-
sion, the procedures described in Section 2.4.2, Section 2.4.3 and Section 2.4.4
were used (denoted as K SVM). The non-linear kernel in K SVM was computed
as the 1400-dimensional \( \chi^2 \) approximation (the Chi\_2 method in Figure 15)
described in Section 2.4.4.

2.5. Ordinal Patch Regression

The labels of dysplastic change are qualitative measurements that reflect
increasing severity in diagnosis from LGD to HGD to ICA. This suggests
that a patch label \( r \in \{\text{LGD, HGD, ICA}\} \) should be an ordinal variable rather
than a nominal one; the order is LGD \( \prec \) HGD \( \prec \) ICA.

2.5.1. Large Margin Formulation

Given a training set \( \{x_i, r_i\}_{i=1}^l \) where \( x_i \) is the \( i \)th bag-of-words feature
vector and \( r_i \in \{\text{LGD, HGD, ICA}\} \) is its corresponding ordinal label, a
function \( g \) that maps the feature vector to the label set \{LGD, HGD, ICA\}
can be found using a structural risk minimisation formulation of ordinal
regression. [Herbrich et al. (1999)] show how to estimate \( g \) by learning a
ranking function \( f(x) \in \mathbb{R} \) so that the pair-wise orders are preserved, i.e.,
\( f(x_i) \prec f(x_j) \iff r_i \prec r_j \). The score \( f(x_i) \) is thresholded to determine
the ordinal category of \( x_i \), i.e., \( g(x) = k \iff f(x) \in [\theta_k, \theta_{k+1}] \) where \( \theta_k \) and
\( \theta_{k+1} \) are the learned thresholds of the \( k \)-th category. In the case of a linear
mapping \( f(x) = w^\top x \), the weight vector \( w \) can be learned by minimising
empirical risk on the pair-wise order set \( \mathcal{P} = \{(i, j) | r_i \prec r_j\} \):

\[
L(w, \xi) = \frac{1}{2}||w||^2 + C \sum_{(i,j) \in \mathcal{P}} \xi_{ij},
\]

subject to:
\[
t_{ij}(w^\top x_j - w^\top x_i) \geq 1 - \xi_{ij}, \xi_{ij} \geq 0, \forall (i, j) \in \mathcal{P},
\]

\(^1\)We used the newpr function from the Matlab Neural Network Toolbox. (URL: [http://uk.mathworks.com/products/neural-network/](http://uk.mathworks.com/products/neural-network/))
where \( t_{ij} \) is the order label, i.e., \( t_{ij} = 1 \) iff \( r_i \prec r_j; t_{ij} = -1 \) otherwise; \( \xi_{ij}, \forall (i, j) \in \mathcal{P} \) are slack variables.

There are three types of pair-wise ordering in \( \mathcal{P} \), i.e., LGD \( \prec \) HGD, LGD \( \prec \) ICA and HGD \( \prec \) ICA (The number of pairs of each type is denoted as \( N_1, N_2, \) and \( N_3 \) respectively). Similar to the balancing of multi-class classification in Section 2.4.2, the balance problem is also considered on the pair-wise order set. Specifically, we replace \( C \) with three cost parameters to be proportional to the number of pair-wise preferences of each type respectively, i.e., \( C_{\text{LGD} \prec \text{HGD}} = C_0(N_1 + N_2 + N_3)/N_1, C_{\text{LGD} \prec \text{ICA}} = C_0(N_1 + N_2 + N_3)/N_2, \) and \( C_{\text{HGD} \prec \text{ICA}} = C_0(N_1 + N_2 + N_3)/N_3, \) where \( C_0 \) is a free parameter; the value is searched within the set \( \{2^\lambda | \lambda \in \mathbb{Z} \land [-15, 15]\} \).

2.5.2. Solving the Optimisation Problem

The objective function in Formula (12) is a standard quadratic programming problem that can be solved by many existing convex optimisation packages. However, the size of the pair-wise order set grows quadratically with the number of samples, e.g., 100 LGD and 100 ICA patches gives 10,000 pair-wise orders. Most solvers are not feasible due to the scale of our problem. Here we choose the fast rank SVM solver proposed by Chapelle and Keerthi (2010) which tackles the primal form of rank SVM with Newton’s method. Instead of computing the inverse Hessian matrix in the Newton step, the fast rank SVM approximates the inverse with a conjugate gradient method. This approximation is both fast and memory-efficient.

With the estimated \( \mathbf{w} \), the optimal threshold \( \theta_k \) is set to be in the middle of the closest correctly separated training pair in the \( k \)-th and \( (k + 1) \)-th category, i.e.,

\[
\theta_k = \frac{\mathbf{w}^\top \mathbf{x}_{i_k} + \mathbf{w}^\top \mathbf{x}_{j_k}}{2},
\]

where:

\[
(i_k, j_k) = \arg\min_{(i,j) \in \mathcal{P}_k} (\mathbf{w}^\top \mathbf{x}_j - \mathbf{w}^\top \mathbf{x}_i),
\]

\[
\mathcal{P}_k = \{(i, j) | r_i = k \land r_j = k + 1 \land (\mathbf{w}^\top \mathbf{x}_j - \mathbf{w}^\top \mathbf{x}_i) \geq 1\}.
\]

2.6. Patch Sampling and Cross-validation

Empirical evaluations used patches sampled with a systematic uniform random sampling (SORS) strategy. The major advantage of SORS over repeated uniform random sampling is its statistical efficiency. The sampling method is illustrated in Figure [9] The displacement between points
on the sampling grid was set to 31 voxels along each of the three image axes. Patches were sampled with their centres at the grid points. A set of 20,000 patches was sampled from the 90 polyps at each of ten patch sizes, \( d \in \{11, 21, 31, \ldots, 111\} \), giving 200,000 patches in total. This enabled exploration of the effect of patch size as a system parameter.

Experiments based on 10-fold cross-validation with stratified splitting were carried out to estimate generalisation capability of the systems as follows. (1) All patches from the same image are in the same fold; this ensures that no part of the polyp participating in the test is in the training set (cross-polyp generalisation); (2) The testing fold is generated so that it always contains the same proportions of polyp images in the three classes as the original dataset to ensure that there is no dominating class in the test dataset. Within the training folds, we further applied 3-fold cross-validation to search the appropriate system parameters (e.g., SVM calibration parameters, cost parameters). This experimental design is illustrated in Figure 10.

The results in the following section are based on this cross-validation scheme unless otherwise specified.

2.7. Performance Metrics

There are six types of mis-classification, i.e., LGD as ICA, LGD as HGD, ICA as HGD, ICA as LGD, HGD as LGD, HGD as ICA. Error rates were calculated for each error type respectively. Performance was also compared with measures of overall mis-classification rate, absolute error rate and \( F \)-measure.
Overall mis-classification rate is the number of mis-classified cases divided by the total number of test cases, $N$, without considering ordinal information. The absolute error rate is $(\sum_{n=1}^{N} |e_n|)/N$, where $e_n$ is a scalar error value. In the case of correct prediction, $e_n = 0$. In the case of an out-by-one error (LGD confused with HGD or HGD confused with ICA), $e_n = 1$. In the case of an out-by-two error (LGD confused with ICA) $e_n = 2$. For experiments on handling class imbalance, averaged $F$-measure was used. The averaged $F$-measure is an average over $F$-measures with respect to each class. The 95% confidence intervals of each type of measurement were obtained by bootstrapping \cite{DiCiccio1996} with $n = 10,000$; specifically we generated $n$ bootstrap replicates of the classifier outputs, and calculated average $F$-measures of each bootstrap replicate. The confidence intervals were computed with $n$ average $F$-measures using the \texttt{boot.ci} function from the \texttt{boot} package (\texttt{R} implementation). When evaluating window and patch size parameters, and kernel approximations, performance was measured with Averaged Area under the ROC curve (AAUC) which is not affected by specific choices of thresholds on the raw SVM outputs. Error bars indicating 95% confidence intervals were estimated with the method proposed by \cite{DeLong1988}. When comparing features with ordinal regression formulations, ROC surfaces were constructed with true positive rates. Volumes under the
Figure 11: Cobweb diagrams showing number of mis-classifications for multi-class classification and ordinal regression formulations.

surfaces were further calculated using the algorithm proposed by Luo and Xiong (2012).

3. Results

3.1. Overall Comparison of Formulations and Feature Types

Figure 11 shows cobweb diagrams for multi-class classification and ordinal regression for each of the three feature types. Table 1 reports confusion matrices. Table 2 reports the misclassification rates, absolute error rates, and average F-measures. These results were obtained using a patch size of $81 \times 81 \times 81$ and a window size of $13 \times 13 \times 13$ using the procedures described in Sections 2.4.2 and 2.4.3 (Experiments exploring the effect of varying such parameters are reported in the following section). The values in Table 2 suggest that RP outperforms ISA and LBP; the performance rank of RP is consistent for all measures considered for both problem formulations. Table 3 compares classification performance of the four types of multi-class classifier described in Section 2.4.5. In most cases, RF, KNN and MLP performed worse than one-vs-rest KSVM. MLP gave the most competitive performance among the three inherently multi-class classifiers.
Table 1: Multi-class classification (left) and ordinal regression (right) confusion matrices.

(a) Random projection features.

<table>
<thead>
<tr>
<th>Actual labels</th>
<th>ICA</th>
<th>HGD</th>
<th>LGD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pred. ICA</td>
<td>5397</td>
<td>1356</td>
<td>702</td>
</tr>
<tr>
<td>Pred. HGD</td>
<td>1604</td>
<td>2406</td>
<td>1041</td>
</tr>
<tr>
<td>Pred. LGD</td>
<td>569</td>
<td>794</td>
<td>6257</td>
</tr>
</tbody>
</table>

(b) Local binary patterns features.

<table>
<thead>
<tr>
<th>Actual labels</th>
<th>ICA</th>
<th>HGD</th>
<th>LGD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pred. ICA</td>
<td>5465</td>
<td>1547</td>
<td>644</td>
</tr>
<tr>
<td>Pred. HGD</td>
<td>1271</td>
<td>1833</td>
<td>954</td>
</tr>
<tr>
<td>Pred. LGD</td>
<td>834</td>
<td>1176</td>
<td>6402</td>
</tr>
</tbody>
</table>

(c) Independent subspace analysis features.

<table>
<thead>
<tr>
<th>Actual labels</th>
<th>ICA</th>
<th>HGD</th>
<th>LGD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pred. ICA</td>
<td>4752</td>
<td>1242</td>
<td>669</td>
</tr>
<tr>
<td>Pred. HGD</td>
<td>1643</td>
<td>2189</td>
<td>1537</td>
</tr>
<tr>
<td>Pred. LGD</td>
<td>1175</td>
<td>1125</td>
<td>5794</td>
</tr>
</tbody>
</table>

3.2. One-Vs-Rest Classification

Figure 12 plots AAUCs averaged over 10 folds for each one-vs-rest classifier for various window and patch sizes. These plots suggest that patch size has greater impact than window size. At patch size $81 \times 81 \times 81$ all features reached high performance. At large window size ($13 \times 13 \times 13$) we generated a smaller number of windows than with smaller window sizes; this was computationally more efficient without significantly decreasing the AAUC. These parameter values were used in other experiments unless otherwise specified. RP gave generally good results compared to the other two types of feature.

3.2.1. Classifier Calibration

Figure 13 shows calibration of an HGD-vs-rest SVM classifier. Thresholding the raw output scores at 0 gave an $F$-measure of 0.2126 with 95% confidence interval (0.1992, 0.2259). After calibration using Equation (10) (see
Figure 12: Averaged Area under the ROC curve (AAUC) using different types of feature with varied window size and patch size. Error bars show 95% confidence intervals. (Best viewed in colour.)
Figure 13: Calibrating the HGD-vs-rest classifier output of all testing folds. Figure best viewed in colour. (a) Histogram and (b) Gaussian kernel density estimates of raw SVM scores. (c–e) Histogram, Gaussian kernel density estimates and reliability diagram of SVM outputs after calibration using Equation (10). (f–h) Histogram, Gaussian kernel density estimates and reliability diagram of SVM outputs after calibration using Equation (11). The diagonal solid lines in the reliability diagrams indicate perfectly calibrated classifiers.
Table 2: Multi-class classification (Multi.) and ordinal regression (Ordi.) results for different features (with 95% confidence intervals).

<table>
<thead>
<tr>
<th>Feature type</th>
<th>Mis-classification rate</th>
<th>Absolute error rate</th>
<th>Average F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>RP - Mult.</td>
<td><strong>0.3038</strong> (0.2974, 0.3102)</td>
<td><strong>0.3659</strong> (0.3578, 0.3742)</td>
<td><strong>0.6675</strong> (0.6610, 0.6740)</td>
</tr>
<tr>
<td>ISA - Mult.</td>
<td>0.3522 (0.3457, 0.3589)</td>
<td>0.4312 (0.4223, 0.4399)</td>
<td>0.6172 (0.6105, 0.6238)</td>
</tr>
<tr>
<td>LBP - Mult.</td>
<td>0.3457 (0.3391, 0.3523)</td>
<td>0.4227 (0.4141, 0.4313)</td>
<td>0.6216 (0.6149, 0.6283)</td>
</tr>
<tr>
<td>RP - Ordi.</td>
<td><strong>0.4253</strong> (0.4185, 0.4321)</td>
<td><strong>0.4481</strong> (0.4406, 0.4556)</td>
<td><strong>0.5735</strong> (0.5667, 0.5802)</td>
</tr>
<tr>
<td>ISA - Ordi.</td>
<td>0.4397 (0.4328, 0.4466)</td>
<td>0.4965 (0.4882, 0.5047)</td>
<td>0.5542 (0.5475, 0.5608)</td>
</tr>
<tr>
<td>LBP - Ordi.</td>
<td>0.4280 (0.4211, 0.4348)</td>
<td>0.4729 (0.4648, 0.4809)</td>
<td>0.5595 (0.5530, 0.5662)</td>
</tr>
</tbody>
</table>

Table 3: Multi-class classification results for different classifiers and features (with 95% confidence intervals).

<table>
<thead>
<tr>
<th>Feature type</th>
<th>Mis-classification rate</th>
<th>Absolute error rate</th>
<th>Average F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>RP - RF.</td>
<td><strong>0.3110</strong> (0.3046, 0.3174)</td>
<td><strong>0.4024</strong> (0.3935, 0.4113)</td>
<td><strong>0.6238</strong> (0.6171, 0.6308)</td>
</tr>
<tr>
<td>ISA - RF.</td>
<td>0.3373 (0.3308, 0.3439)</td>
<td>0.4177 (0.4089, 0.4265)</td>
<td>0.5994 (0.5927, 0.6060)</td>
</tr>
<tr>
<td>LBP - RF.</td>
<td>0.3234 (0.3168, 0.3300)</td>
<td>0.4199 (0.4109, 0.4291)</td>
<td>0.5913 (0.5846, 0.5981)</td>
</tr>
<tr>
<td>RP - KNN.</td>
<td><strong>0.3572</strong> (0.3507, 0.3638)</td>
<td>0.4671 (0.4577, 0.4764)</td>
<td><strong>0.6026</strong> (0.5958, 0.6094)</td>
</tr>
<tr>
<td>ISA - KNN.</td>
<td>0.3621 (0.3554, 0.3686)</td>
<td><strong>0.4669</strong> (0.4574, 0.4763)</td>
<td>0.5832 (0.5764, 0.5900)</td>
</tr>
<tr>
<td>LBP - KNN.</td>
<td>0.3650 (0.3583, 0.3717)</td>
<td>0.4768 (0.4673, 0.4863)</td>
<td>0.5772 (0.5704, 0.5840)</td>
</tr>
<tr>
<td>RP - MLP.</td>
<td><strong>0.3031</strong> (0.2968, 0.3095)</td>
<td><strong>0.3716</strong> (0.3633, 0.3799)</td>
<td><strong>0.6558</strong> (0.6491, 0.6625)</td>
</tr>
<tr>
<td>ISA - MLP.</td>
<td>0.3373 (0.3308, 0.3437)</td>
<td>0.4135 (0.4048, 0.4222)</td>
<td>0.6117 (0.6050, 0.6184)</td>
</tr>
<tr>
<td>LBP - MLP.</td>
<td>0.3462 (0.3395, 0.3527)</td>
<td>0.4619 (0.4524, 0.4714)</td>
<td>0.5976 (0.5906, 0.6046)</td>
</tr>
<tr>
<td>RP - KSVM.</td>
<td><strong>0.2987</strong> (0.2924, 0.3051)</td>
<td><strong>0.3641</strong> (0.3558, 0.3742)</td>
<td><strong>0.6746</strong> (0.6680, 0.6812)</td>
</tr>
<tr>
<td>ISA - KSVM.</td>
<td>0.3227 (0.3162, 0.3292)</td>
<td>0.3851 (0.3767, 0.3933)</td>
<td>0.6484 (0.6418, 0.6550)</td>
</tr>
<tr>
<td>LBP - KSVM.</td>
<td>0.2996 (0.2932, 0.3059)</td>
<td>0.3655 (0.3573, 0.3739)</td>
<td>0.6636 (0.6570, 0.6703)</td>
</tr>
</tbody>
</table>

Figure 13(c-e)), thresholding probabilities at 0.5 gave an improved F-measure of 0.3286 with 95% confidence interval (0.3090, 0.3368). However, the probability of HGD is systematically underestimated by the sigmoid function in Equation (10) because the model is biased towards the “rest” class due to the imbalanced training set. This can be seen from Figures 13(a-e). In Figures 13(f-h) we visualise the probability outputs of the bagging method. The under-sampling of the dominating class and the ensemble strategy mitigated the bias problem. In the reliability diagrams (Figures 13(e, h)), the observed probabilities are grouped into 25 bins and the observed frequency of positives are plotted against mean probability in each bin. For a perfect reliability, the observed frequency and the predicted probability should be equal (shown as dotted line in diagonal direction). By using the bagging method, the probability outputs from 0.8 to 1.0 are more reliable compared to the original Platt’s method. We are able to further improve the F-measure.
Figure 14: Box plots summarising the distribution of $F$-measures over cross-validation folds with ISA features (patch size: $81 \times 81 \times 81$, window size: $13 \times 13 \times 13$). Balanced denotes use of the balanced cost function. Scaled denotes use of output calibration. Scaled and Balanced denotes use of both a balanced cost function and output calibration.

from 0.3286 to 0.4637 with 95% confidence interval (0.4526, 0.4748). In the other experiments reported here we applied Equation (11) for the calibration procedure.

3.2.2. Effect of Class Balancing and Calibration

We conducted a group of experiments to investigate the effect of balancing the cost function (Section 2.4.2) and output calibration (Section 2.4.3) on classification performance. As a baseline we use the standard SVM (Formula (4)). To obtain decisions from one-vs-rest classifiers, raw SVM scores were thresholded at 0 and calibrated outputs at 0.5. Multi-class classification decisions were obtained by applying the max-win strategy described in Section 2.4.1. Figure 14 shows comparisons with the baseline SVM in terms of $F$-measures of each fold. Classification results after class balancing and calibration were significantly different from the baseline ($p < 0.0001$; McNemar’s test (Dietterich, 1998)). Use of the balanced SVM cost function generally improved performance, especially for HGD-vs-rest classification where the baseline suffers from the very unbalanced dataset (see Figure 4). SVM output calibration helped address the data imbalance problem and improved performance compared to the baseline. Combining the balanced cost func-
tion and output calibration does not lead to further improvement in terms of $F$-measure as compared to each applied individually.

3.2.3. Effect of Kernel Approximation

The effect of approximating the histogram intersection and $\chi^2$ kernels as discussed in Section 2.4.4 was evaluated. Two approximation strategies were compared: 1) B-spline approximations\(^2\) (Maji et al., 2013) of the $\chi^2$ kernel (denoted as $Chi_1$) and the histogram intersection kernel (denoted as $Min_1$); and 2) the homogeneous feature map\(^3\) (Vedaldi and Zisserman, 2012) for the $\chi^2$ kernel (denoted as $Chi_2$) and the histogram intersection kernel (denoted as $Min_2$). Figure 15 gives AUC results using those approximations compared with the baseline result of the plain linear SVM (error bars indicate 95% confidence intervals). It can be seen that, generally, the approximations of non-linear kernels can improve dysplasia classification over the baseline for all three feature types if the feature dimensionality is carefully tuned; the exception was that the performance improvement of B-spline approximations of the histogram intersection kernel was not stable and conclusive. The homogeneous feature map consistently outperformed the B-spline approximation although the improvement was marginal. The $\chi^2$ kernel worked better than the histogram intersection kernel in most cases.

3.3. Ordinal Regression

Figure 16 shows distributions of rank SVM scores obtained along with the two thresholds for three-class ordinal regression. By varying the two thresholds ROC surfaces can be created. Each point on an ROC surface represents a tuple of three true positive rates (TPR). Figure 17(a) shows ROC surfaces for the three feature types.

For visual comparison, each surface was mapped onto the plane $TPR_{LGD} + TPR_{ICA} + TPR_{HGD} = 1$ and the signed distances between points on the surface and the plane were colour-coded (Figure 17(b)). The plane represents the ROC surface for random classification. The signed distances give an indication of how much better the classifier is than random guessing. The

\(^2\)We used the Matlab implementation by Maji et al. (2013). (URL: \url{http://ttic.uchicago.edu/~smaji/projects/libspline-release1.0.tar.gz})

\(^3\)We used the Matlab implementation provided in VLFeat package (Vedaldi and Fulkerson, 2008).
<table>
<thead>
<tr>
<th>Kernel type</th>
<th>Mean class AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear (200)</td>
<td>0.750</td>
</tr>
<tr>
<td>Chi_1 (1400)</td>
<td>0.775</td>
</tr>
<tr>
<td>Chi_1 (2200)</td>
<td>0.800</td>
</tr>
<tr>
<td>Chi_2 (1400)</td>
<td>0.825</td>
</tr>
<tr>
<td>Chi_2 (2200)</td>
<td>0.850</td>
</tr>
<tr>
<td>Min_1 (1400)</td>
<td>0.875</td>
</tr>
<tr>
<td>Min_1 (2400)</td>
<td>0.900</td>
</tr>
<tr>
<td>Min_1 (6400)</td>
<td>0.925</td>
</tr>
<tr>
<td>Min_2 (1400)</td>
<td>0.950</td>
</tr>
<tr>
<td>Min_2 (2200)</td>
<td>0.975</td>
</tr>
</tbody>
</table>

Figure 15: Averaged AUCs for different kernel types (with feature dimensionality given in parentheses). Features were extracted with patch size $81 \times 81 \times 81$ and window size $13 \times 13 \times 13$. “Chi_1” and “Min_1” denote B-spline based kernel approximation of $\chi^2$ and histogram intersection kernels respectively. “Chi_2” and “Min_2” denote homogeneous feature map of $\chi^2$ and histogram intersection kernels respectively. Error bars indicate 95% confidence intervals.

signed distance maps for different feature types were further compared (Figure 17(c)). Compared to ISA features, RP and LBP give better performances in high TPR range of LGD (high LGD specificity range), whereas ISA is better in high TPR range of ICA (high ICA specificity range). In the comparison of RP and LBP, RP performs better when true positive rate of HGD is in low range (low HGD specificity).

The volume under an ROC surface (VUS) is the expected proportion of correctly ranked triplets uniformly drawn from all possible samples of triplets (He and Frey, 2008). It is an extension of area under the ROC curve.

Figure 16: Distributions of training set ranking scores. Dashed and dotted vertical bars are thresholds for the three ordinal categories, obtained using Equation (13).
Figure 17: (a) ROC surfaces for three-class ordinal regression (patch size: 81 × 81 × 81, window size: 13 × 13 × 13). Colour on the \( \text{TPR}_{\text{LGD}} - \text{TPR}_{\text{ICA}} \) plane encodes \( \text{TPR}_{\text{HGD}} \) value. (b) Signed distances from points on the ROC surfaces to the plane \( \text{TPR}_{\text{LGD}} + \text{TPR}_{\text{ICA}} + \text{TPR}_{\text{HGD}} = 1 \). (c) Differences between the maps in (b). (Best viewed in colour.)
to the three-class case. The VUS and its 95% confidence interval for the RP, ISA and LBP methods were 0.5901 (0.5821, 0.5981), 0.5215 (0.5136, 0.5294) and 0.5325 (0.5246, 0.5405) respectively. In this experiment RP features showed significantly better performance than ISA and LBP in terms of VUS.

4. Concluding Remarks

Large margin multi-class classification and ordinal regression techniques were adopted with a focus on practical considerations. Their performance with three types of texture descriptor was demonstrated. The results showed that random projection features and bag-of-words framework was the most effective. Although the RP method is not invariant to rotations, it outperformed carefully designed LBP operators as well as automatically learned ISA filters for all parameter settings. Similar observations were also reported in several general texture classification systems (Liu and Fieguth, 2012; Varma and Zisserman, 2003) where classification of filter-based features did not outperform raw pixels as patch representation. The RP method generates a low-dimensional representation without making strong assumption about the nature of the texture being analysed. These attributes may partially explain its relatively good performance in OPT analysis.

The class balancing problem is important and the performance can be further improved if it is carefully managed. Using the approximation of non-linear kernels, especially $\chi^2$ kernel, also improves the classification performance slightly.

In terms of computational complexity, encoding a window with RP and ISA methods is similar: the output window feature is a simple multiplication of the vectorised window with a matrix (either a random projection matrix or a set of filters learned with ISA). The LBP operator is more expensive as it involves a neighbourhood-pixel thresholding and a uniform pattern matching procedure. ISA requires a training phase in order to estimate filters from

4The complexity of RP method can be further reduced by constructing sparse random projections with a simpler sampling distribution than a standard normal distribution (e.g., Achlioptas, 2003; Li et al., 2006). We use the standard normal distribution for RP matrix because both methods achieve similar performance and a standard normal distribution is easier to implement.

training patches. In our experiments learning a set of 300 filters\(^6\) from a 81 × 81 × 81 patch set required approximately 4 hours; Encoding and classifying 200 test patches with size 81 × 81 × 81 required approximately 30 seconds for the RP and ISA methods, and 4 minutes for the LBP method on a 3.4GHz Intel i7 CPU with 16Gb memory.

LGD was the class most readily distinguished. Unsurprisingly, being the intermediate class, HGD was the most often confused class. In terms of problem formulations, ordinal regression performance was slightly worse than multi-class classification in terms of mis-classification rate, absolute error rate and averaged $F$-measure (Table 2). However ordinal regression makes less confusion between the ICA and LGD classes (Figure 11). The ordinal regression formulation is a better choice over one-vs-rest classification when the focus concentrates on minimising the risk of mis-classification between LGD and ICA. Ordinal regression is also simpler than multi-class classification in the sense that only one model is trained while in multi-class classification three models are trained. Dysplastic change is naturally a continuous phenomenon on which ordinal grading imposes artificial categories. The ordinal regression method works by first mapping samples onto the real line and it would be interesting to investigate using this continuous map in diagnostic histopathology.

Compared to the emerging study of using OPT for colorectal polyps analysis, extensive literature exists on automatic analysis of more traditional histopathology images for a range of analysis tasks, clinical settings, and disease types. However, we are not aware of a suitable study with which to make direct comparisons (i.e., patch discrimination between LGD, HGD and ICA in colorectal polyps). The proposed system achieved promising performance in discriminating OPT image patches. Further improvements and incorporation into software tools would be needed to enable translation or adoption for clinical research. Larger data sets of annotated images would be needed to train and validate such tools because the visual appearance of dysplastic change is complex and the variations across polyps are large.

It is worth noting that inter-observer variation exists in polyp diagnosis with OPT (Coats et al., 2013). In order to minimise the effect of such variation and uncertainty in our study, the ground truth was annotated within

---

\(^6\)For ISA model estimation we use the Matlab implementation by Hyvärinen et al. (2009) (URL: http://www.naturalimagestatistics.net).
high confidence regions by an experienced pathologist rather than trying to
delineate accurate region boundaries. The annotated regions were meanwhile
cross-checked and calibrated with the corresponding H&E slices to ensure the
certainty of the obtained ground truth. Nevertheless, there may still ex-
ist some uncertainty in the ground truth. Quantifying this using multiple
pathologists would be interesting for a future study.

OPT enables access to polyps’ surface morphology and internal structure
at the same time. In the future one interesting direction is to combine mor-
phological analysis of colorectal polyp surfaces (e.g., [Zhang et al., 2014]) with
3-D texture analysis. Recently, deep learning techniques have yielded promis-
ng results in learning image representations from large datasets. Applying
such techniques to OPT datasets would be an interesting direction. Finally,
building accurate classification and regression systems usually requires large
amounts of manually annotated data; it is expensive to obtain such manual
annotations. Another research direction is to develop training algorithms
that efficiently utilise limited amounts of annotation (e.g., [Li et al., 2013c]).

References

Achlioptas, D., 2003. Database-friendly random projections: Johnson-
Lindenstrauss with binary coins. Journal of Computer and System Sciences
66, 671–687.

image registration with the Johnson–Lindenstrauss lemma: Application to
imaging 3-D neural ultrastructure with electron microscopy. IEEE Trans-
actions on Medical Imaging 30, 1427–1438.

Allwein, E.L., Schapire, R.E., Singer, Y., 2001. Reducing multiclass to bi-
nary: A unifying approach for margin classifiers. The Journal of Machine
Learning Research 1, 113–141.

Arthur, D., Vassilvitskii, S., 2007. k-means++: The advantages of care-
1027–1035.

Atlamazoglou, V., Yova, D., Kavantzas, N., Loukas, S., 2001. Texture anal-
ysis of fluorescence microscopic images of colonic tissue sections. Medical
and Biological Engineering and Computing 39, 145–151.


He, X., Frey, E.C., 2008. The meaning and use of the volume under a  
three-class ROC surface (VUS). IEEE Transactions on Medical Imaging  
27, 577–588.

Herbrich, R., Graepel, T., Obermayer, K., 1999. Large margin rank bound-
daries for ordinal regression, in: Advances in Neural Information Processing  
Systems, MIT. pp. 115–132.

Philosophical Transactions of the Royal Society A: Mathematical, Physical  
and Engineering Sciences 371, 20110534.

features by decomposition of natural images into independent feature sub-

and Vision, Springer.

of local patches in colon histopathology, in: International Conference on  
Pattern Recognition, pp. 61–64.

Le, Q.V., Han, J., Gray, J.W., Spellman, P.T., Borowsky, A., Parvin, B.,  
2012. Learning invariant features of tumor signatures, in: IEEE Interna-
tional Symposium on Biomedical Imaging, pp. 302–305.

Le, Q.V., Zou, W.Y., Yeung, S.Y., Ng, A.Y., 2011. Learning hierarchical  
invariant spatio-temporal features for action recognition with independent  
subspace analysis, in: IEEE Conference on Computer Vision and Pattern  
Recognition, pp. 3361–3368.

Li, P., Hastie, T.J., Church, K.W., 2006. Very sparse random projections, in:  
ACM International Conference on Knowledge Discovery and Data Mining,  
pp. 287–296.

Li, W., Coats, M., Zhang, J., McKenna, S.J., 2013a. Comparative analysis of  
feature extraction methods for colorectal polyp images in optical projection  
tomography, in: Medical Image Understanding and Analysis, pp. 67–72.


