

University of Dundee

DOCTOR OF PHILOSOPHY

Visual feature learning with application to medical image classification

Manivannan, Siyamalan

Award date:
2015

Awarding institution:
University of Dundee

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 17. Feb. 2017

**VISUAL FEATURE LEARNING
WITH APPLICATION TO
MEDICAL IMAGE CLASSIFICATION**



Siyamalan Manivannan

A thesis submitted in partial fulfilment for the
degree of Doctor of Philosophy

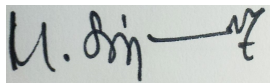
in the
School of Computing,
University of Dundee,
Dundee, UK

July 2015

Declaration of Authorship

Candidate's declaration

I, Siyamalan Manivannan, hereby declare that I am the author of this thesis; that all references cited have been consulted by me; that the work of which this thesis is a record has been done by me, and that it has not been previously accepted for a higher degree.



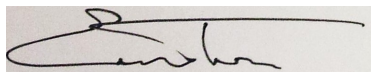
03/08/2015

Signed

Date

Supervisor's declaration

I, Emanuele Trucco, hereby declare that I am the supervisor of the candidate, and that the conditions of the relevant Ordinance and Regulations have been fulfilled.



03/08/2015

Signed

Date

Acknowledgements

First, I would like to thank my supervisor, Prof. Emanuele Trucco, for his continued support, guidance and encouragement throughout my research. I appreciate him for giving me the motivation to achieve more and setting the bar high. I take this opportunity to express my gratitude to him for being an understanding and patient supervisor, and a wonderful person, while not compromising on the quality of the research.

I also wish to thank Dr. Ruixuan (Roy) Wang for collaborations, and innumerable discussions, which went a long way in helping me to ease into the research environment.

I would like to thank Dr. Adrian Hood (Leeds Institute of Molecular Medicine, University of Leeds, UK) for providing annotations for the colonoscopy datasets used in this thesis.

It was a pleasure to be part of the rapidly growing and diverse Computer Vision and Image Processing Group (CVIP) at the University of Dundee. I also would like to express my thanks to the CVIP members for useful technical discussions throughout my PhD. The results based on the cell images reported in Chapter 8 are based on a collaborative work with other members of the CVIP group. I sincerely thank Wenqi Li, Shazia Akbar, Ruixuan Wang, Jianguo Zhang and Stephen J. McKenna for this collaborative work.

The research presented in this thesis is funded by 2011-2016 EU FP7 ERC project "CODIR: colonic disease investigation by robotic hydrocolonoscopy", collaborative between the Universities of Dundee (PI Prof Sir A Cuschieri) and Leeds (PI Prof A Neville).

List of publications and the activities involved

1. List of publications

1.1. Journal publications

- S. Manivannan, R. Wang, and E. Trucco, Feature learning for colonoscopy image classification, *Medical Image Analysis* (under preparation).
- S. Manivannan, W. Li, S. Akbar, R. Wang, J. Zhang, S. J. McKenna, An automated pattern recognition system for classifying Indirect Immunofluorescence images of HEP-2 cells and specimens, *Pattern Recognition* (accepted).

1.2. Conference publications

- S. Manivannan, R. Wang, and E. Trucco, Automatic normal-abnormal video frame classification for colonoscopy, *IEEE International Symposium on Biomedical Imaging (ISBI)*, San Francisco, USA, 2013.
- S. Manivannan, R. Wang, and E. Trucco, Inter-cluster features for medical image classification, *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2014.
- S. Manivannan, and E. Trucco, Learning discriminative local features from image-level labelled data for colonoscopy image classification, *IEEE International Symposium on Biomedical Imaging (ISBI)*, New York, USA, 2015.

1.3. Workshop publications

- S. Manivannan, R. Wang, and E. Trucco, Extended Gaussian-Filtered Local Binary Patterns for Colonoscopy Image Classification, *IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2013.
- S. Manivannan, R. Wang, and E. Trucco, Video-specific SVMs for colonoscopy image classification, *Computer-Assisted and Robotic Endoscopy - CARE (MICCAI workshops)*, 2014.
- S. Manivannan, W. Li, S. Akbar, R. Wang, J. Zhang, S. J. McKenna, HEP-2 Specimen Classification using Multi-Resolution Local Patterns and SVM, *I3A 1st workshop on Pattern Recognition Techniques for Indirect Immunofluorescence Images, ICPR* (invited paper), 2014.

-
- S. Manivannan, W. Li, S. Akbar, R. Wang, J. Zhang, S. J. McKenna, HEp-2 Cell Classification using Multi-Resolution Local Patterns and Ensemble SVMs, I3A 1st workshop on Pattern Recognition Techniques for Indirect Immunofluorescence Images, ICPR (invited paper), 2014.
 - G. Puerto-Souza, S. Manivannan, M. Trujillo, J. Hoyos, E. Trucco, G.L. Mariottini, Enhancing normal-abnormal classification accuracy in colonoscopy videos via temporal consistency, Computer-Assisted and Robotic Endoscopy - CARE (MICCAI workshops), 2015.

2. Contests and awards

- Member of the winning teams for the following competitions,
 - “Detection of Abnormalities in Gastroscopic Images” organised by MICCAI 2015.
 - “Early Barrett’s Cancer Detection” organised by MICCAI 2015.
 - “Performance Evaluation of Indirect Immunofluorescence Image Analysis Systems (I3A)” organised by the 22nd International Conference on Pattern Recognition (ICPR 2014).
- Member of the runner-up team for the “Brain Tumour Digital Pathology Segmentation Challenge” organised by MICCAI 2014.
- Runner-up prize for the outstanding doctoral student, School of Computing, University of Dundee, 2014.

3. Summer school and other activities

- Attended BMVA computer vision summer school, University of Manchester, 2013.
- Attended and presented periodically at the CODIR plenary (Dundee and Leads teams) and monthly (Dundee team only) meetings throughout the PhD (2011-2015).
- Attended and presented regularly at the twice-yearly CVIP technical workshops throughout the PhD (2011-2015).

Abstract

Various hand-crafted features have been explored for medical image classification, which include SIFT and Local Binary Patterns (LBP). However, hand-crafted features may not be optimally discriminative for classifying images from particular domains (e.g. colonoscopy), as not necessarily tuned to the domain's characteristics.

In this work, I give emphasis on learning highly discriminative local features and image representations to achieve the best possible classification performance for medical images, particularly for colonoscopy and histology (cell) images. I propose approaches to learn local features using unsupervised and weakly-supervised methods, and an approach to improve the feature encoding methods such as *bag-of-words*. Unlike the existing work, the proposed weakly-supervised approach uses image-level labels to learn the local features. Requiring image-labels instead of region-level labels makes annotations less expensive, and closer to the data normally available from normal clinical practice, hence more feasible in practice.

In this thesis, first, I propose a generalised version of the LBP descriptor called the *Generalised Local Ternary Patterns* (gLTP), which is inspired by the success of LBP and its variants for colonoscopy image classification. gLTP is robust to both noise and illumination changes, and I demonstrate its competitive performance compared to the best performing LBP-based descriptors on two different datasets (colonoscopy and histology). However LBP-based descriptors (including gLTP) lose information due to the binarisation step involved in their construction. Therefore, I then propose a descriptor called the *Extended Multi-Resolution Local Patterns* (xMRLP), which is real-valued and reduces information loss. I propose unsupervised and weakly-supervised learning approaches to learn the set of parameters in xMRLP. I show that the learned descriptors give competitive or better performance compared to other descriptors such as root-SIFT and Random Projections. Finally, I propose an approach to improve feature encoding methods. The approach captures inter-cluster features, providing context information in the feature as well as in the image spaces, in addition to the intra-cluster features often captured by conventional feature encoding approaches.

The proposed approaches have been evaluated on three datasets, 2-class colonoscopy (2,100 images), 3-class colonoscopy (2,800 images) and histology (public dataset, containing 13,596 images). Some experiments on radiology images (IRMA dataset, public) also were given. I show state-of-the-art or superior classification performance on colonoscopy and histology datasets.

Contents

1	Introduction	1
1.1	Medical image analysis background and motivation	1
1.2	Clinical background and motivation	3
1.2.1	Background on colonoscopy	4
1.2.2	The CODIR project	6
1.3	Contributions	7
1.3.1	Novel feature learning approaches	7
1.3.2	Inter-cluster statistics for feature encoding	8
1.3.3	Experimental evaluations	8
1.4	Thesis organisation	8
2	Related Work	11
2.1	Endoscopy image analysis	11
2.1.1	Purpose	11
2.1.2	Features and representations	14
2.1.3	Classifiers for endoscopy image analysis	16
2.2	Image representation using visual words	24
2.2.1	Feature extraction	25
2.2.2	Feature encoding	26
2.2.3	Pooling	28
2.3	Feature learning approaches	29
2.3.1	Unsupervised feature learning	30
2.3.2	Supervised feature learning	32
2.3.3	Weakly supervised feature learning	36
2.4	Conclusions and discussion	37
3	Datasets and experimental settings	40
3.1	Datasets	40
3.1.1	2 class colonoscopy images dataset	40
3.1.2	3 class colonoscopy images dataset	41
3.1.3	ICPR 2014 cell images dataset	42
3.1.4	IRMA dataset	44
3.2	Experimental settings	46

3.2.1	Image preprocessing	46
3.2.2	Evaluation metric and experimental setup	46
4	The Generalised Local Ternary Patterns	48
4.1	Introduction	48
4.2	LBP and its variants	49
4.2.1	The standard LBP descriptor	49
4.2.2	Uniform and rotation invariant LBP	50
4.2.3	Local Ternary Patterns	51
4.2.4	Scale Invariant Local Ternary Patterns	52
4.2.5	Other variants	52
4.3	Generalised Local Ternary Patterns	53
4.3.1	Definition	53
4.3.2	Effect of parameters	54
4.4	Experiments	57
4.4.1	Experimental setup	57
4.4.2	Parameter selection	58
4.4.3	Comparison of LBP, LTP, SILTP and gLTP	58
4.5	Conclusions and discussion	60
5	Extended Multi-Resolution Local Patterns and unsupervised feature learning	61
5.1	Introduction	61
5.2	Extended Multi-Resolution Local Patterns	63
5.2.1	Extended Local Patterns	63
5.2.2	Extension to multi-resolution version	64
5.2.3	Image-level representation using xMRLP	66
5.3	Parameter learning: unsupervised approach	66
5.3.1	The objective function	67
5.3.2	Optimisation	67
5.4	Experiments	69
5.4.1	Experimental setup	69
5.4.2	Effect of parameter learning	69
5.4.3	Comparison with LBP based features	73
5.5	Conclusions and discussion	75
6	Discriminative Feature Learning using weak labels	77
6.1	Introduction	77
6.2	Notation	79
6.3	Image-to-class distances	79
6.3.1	Learning-based and non-learning based classifiers	79
6.3.2	Image-to-class vs image-to-image distances	80

6.3.3	The Naïve Bayes Nearest Neighbour classifier	81
6.3.4	Extensions of NBNN	83
6.4	Discriminative max-margin parameter learning	84
6.4.1	The objective function	85
6.4.2	Optimisation	86
6.5	Experiments	87
6.5.1	Effect of parameter learning	87
6.5.2	Example classification results	89
6.5.3	Sensitivity of the regularisation parameters	89
6.5.4	The learned xMRLP _s parameters	91
6.6	Conclusions and discussion	92
7	Discriminative feature learning with weak-labels and weighted I2CD	94
7.1	Introduction	94
7.2	The proposed joint learning framework	96
7.2.1	Notation	96
7.2.2	Weighted I2CD	96
7.2.3	Discriminative probabilistic softmax classifier	97
7.2.4	The objective function	97
7.2.5	Optimisation	98
7.3	Experiments	99
7.3.1	Effect of learning	100
7.3.2	Sensitivity of the regularisation parameters	101
7.3.3	The learned parameters	102
7.3.4	Example probabilistic output	102
7.3.5	Comparison of different features using the NBNN classifier	103
7.4	Conclusions and discussion	108
8	MRLP and xMRLP for colonoscopy and cell image classification: experimental evaluation	109
8.1	Introduction	109
8.1.1	Root-SIFT	110
8.1.2	Random Projection	110
8.1.3	Local Colour Histogram	111
8.2	Colonoscopy image classification	111
8.2.1	The proposed system	111
8.2.2	Experiments	113
8.3	Cell image classification	122
8.3.1	xMRLP-based features for cell image classification	122
8.3.2	The proposed system	123
8.3.3	Experiments	127

8.4	Conclusions and discussion	134
9	Inter-cluster features for image classification	136
9.1	Introduction	136
9.2	Inter-cluster features	139
9.2.1	Selection of cluster pairs based on LSA	139
9.2.2	Construction of the term-document matrix	141
9.2.3	Inter-cluster statistics	142
9.2.4	Feature encoding	143
9.3	Experiments	143
9.3.1	Effect of the inter-cluster features	143
9.3.2	Patch-based vs. image-based methods	146
9.3.3	LSA-based pair selection	146
9.3.4	Inter-cluster features for Fisher Vector	146
9.4	Conclusions and discussion	147
10	Conclusion, discussion and future work	148
10.1	Summary of the thesis	148
10.1.1	Novel feature learning approaches	148
10.1.2	Experimental evaluation	150
10.1.3	Inter-cluster statistics for feature encoding	150
10.2	Key contributions	151
10.3	Limitations and analysis	152
10.3.1	Classifying colon images from new videos	152
10.3.2	Multiple annotations	152
10.4	Future work	153
10.4.1	Incorporating temporal information for classification	153
10.4.2	Discriminative inter-cluster features for classification	154
10.4.3	I2CD prototype learning	154
10.4.4	Lesion localisation and multiple instance learning	155
	Bibliography	156

List of Figures

Chapter 1 Introduction

1.1	The major steps involved in the automated classification systems for computer vision and medical image analysis.	2
1.2	Example images from different colonoscopy systems.	5
1.3	Thesis roadmap. Chapters and their relations.	9

Chapter 2 Related Work

2.1	Pipeline of feature encoding approaches.	24
2.2	Local descriptor sampling.	25
2.3	Pooling region candidate rings.	33
2.4	Semantic texton forest features.	34
2.5	Learning image descriptors with boosting.	35

Chapter 3 Datasets and experimental settings

3.1	Examples of normal images.	42
3.2	Examples of abnormal images.	43
3.3	Examples of uninformative images.	43
3.4	Example cell images from the ICPR dataset.	44
3.5	Examples of images from the IRMA dataset.	45

Chapter 4 The Generalised Local Ternary Patterns

4.1	The circular LBP neighbourhoods.	49
4.2	The derivation of the standard LBP codes.	50
4.3	LBP-based image representation.	50
4.4	Some example patterns in the uniform LBP.	51
4.5	A demonstrative example of the effect of noise and illumination changes for different descriptors.	55
4.6	Effect of noise and illumination changes for different descriptors for an example colonoscopy image.	56
4.7	Effect of noise and illumination changes for different descriptors for an example cell image.	56
4.8	Histogram of the selected parameters for different descriptors.	59

Chapter 5 Extended Multi-Resolution Local Patterns and unsupervised feature learning

5.1	An example sampling pattern with 8 sampling points.	63
5.2	A three-resolution sampling pattern.	65
5.3	Compactness of the clusters.	70
5.4	xMRLP vs MRLP for different datasets.	71
5.5	Sensitivity of the regularisation.	72
5.6	Visualisation of the learned parameters.	73

Chapter 6 Discriminative Feature Learning using weak labels

6.1	Sensitivity of the parameters.	81
6.2	Example of correctly classified images.	90
6.3	Examples of images mis-classified.	90
6.4	Example of images which are correctly classified by xMRLP _s but miss-classified by MRLP features.	90
6.5	Sensitivity of the regularisation.	91
6.6	Visualisation of the learned xMRLP _s parameters.	91

Chapter 7 Discriminative feature learning with weak-labels and weighted I2CD

7.1	Example abnormal images.	95
7.2	The I2CD between a normal and an abnormal image to different classes.	95
7.3	Convergence of Algorithm 5.	100
7.4	Sensitivity of the regularisation.	101
7.5	Visualisation of the learned parameters.	102
7.6	Examples of images correctly classified with high confidence.	103
7.7	Examples of images wrongly classified.	104
7.8	Examples of wrongly classified images.	105

Chapter 8 MRLP and xMRLP for colonoscopy and cell image classification: experimental evaluation

8.1	Global vs local colour histograms.	112
8.2	Comparison of xMRLP-based features with different feature encodings.	115
8.3	Comparison of different features with different feature encodings.	116
8.4	Comparison of different features for the 3-class colonoscopy dataset.	121
8.5	Comparison of different features for the cells dataset.	123
8.6	An overview of the proposed system for the cell classification.	124
8.7	Image preprocessing for the ICPR cells dataset.	125
8.8	Cell pyramids.	126

8.9 An overview of the system for data augmentation and training SVM ensemble	126
8.10 Testing an image using the SVM ensemble	127
8.11 Performance of different features and encoding methods for the ICPR cells dataset.	129
8.12 Sample specimen images from ICPR Task-2 dataset.	132
8.13 The MCA at cell level attained by each method on the test set of Task 1. . .	134

Chapter 9 Inter-cluster features for image classification

9.1 High-order co-occurrence.	140
9.2 Term-document matrix from images and patches.	141
9.3 Effect of the inter-cluster features.	144
9.4 BOW and co-occurrence features for the ICPR dataset.	145

List of Tables

Chapter 2 Related Work

- 2.1 Existing endoscopy image analysis systems for different purposes 13
- 2.2 An overview of the existing endoscopy image analysis systems. 23

Chapter 3 Datasets and experimental settings

- 3.1 Detail of the datasets. 45

Chapter 4 The Generalised Local Ternary Patterns

- 4.1 gLTP with different parameter settings. 54
- 4.2 The range of parameters used for different descriptors. 58
- 4.3 Comparison of LBP, LTP, SILTP and gLTP. 59

Chapter 5 Extended Multi-Resolution Local Patterns and unsupervised feature learning

- 5.1 The parameters of the multi-resolution sampling patterns. 65
- 5.2 Cluster compactness. 70
- 5.3 Performance of LBP and its variants. 74

Chapter 6 Discriminative Feature Learning using weak labels

- 6.1 MCA of MRLP, $xMRLP_u$ and $xMRLP_s$ features using NBNN classifier. 88

Chapter 7 Discriminative feature learning with weak-labels and weighted I2CD

- 7.1 Performance of joint learning. 100
- 7.2 Performance of different features using NBNN classifier. 105
- 7.3 Average computational time required to classify an image. 105
- 7.4 Confusion matrices for different features for the 2-class colonoscopy dataset using the NBNN classifier. 106
- 7.5 Confusion matrices for different features for the 3-class colonoscopy dataset using the NBNN classifier. 106
- 7.6 Confusion matrices for different features for the ICPR cells dataset using the NBNN classifier. 107

8.1 Computational time for feature extraction and encoding.	118
Chapter 8 MRLP and xMRLP for colonoscopy and cell image classification: experimental evaluation	
8.2 Combining features for classification	119
8.3 Comparisons with the state-of-the-art for colonoscopy.	120
8.4 Two-fold cross-validation results for the ICPR dataset.	130
8.5 Computational time for feature extraction and encoding for the cell images	130
8.6 Confusion matrix for feature combinations (neither CPM nor data augmentation).	131
8.7 Confusion matrix for feature combinations (without data augmentation). .	131
8.8 Confusion matrix for feature combinations.	131
8.9 Confusion matrix for leave-one-specimen-out experiment.	132
8.10 Confusion matrix of the system trained on Task 1 images and tested on the cell images extracted from Task 2.	133
8.11 Comparison with Gragnaniello et al.	134

Glossary

Adenoma detection rate - The proportion of individuals undergoing a complete colonoscopy screening who have one or more adenomas detected.

Benign - A condition, tumour, or growth that is not cancerous. Such growths can often be removed and, in most cases, they do not come back. Cells in benign tumours do not spread to other parts of the body.

Bin (histogram) - Consecutive, non-overlapping intervals of a variable.

Computer Aided Diagnosis - Computerized techniques that assist doctors in the interpretation of medical images.

Histology - Branch of biology that deals with the microscopic examination of tissue.

Histogram - A bar graph of a frequency distribution in which the horizontal axis lists each unique value (or range of values) in a set of data, and the height of each bar represents the frequency of that value (or range of values).

iid - A sequence or collection of random variables is iid (independent and identically distributed) if each random variable in that collection has the same probability distribution as the others and all are mutually independent.

Malignant - A condition, tumour, or growth that is cancerous and is made up of cells that grow out of control. Cells in these growths can invade nearby tissues and spread to other parts of the body.

Non-invasive procedure - A medical procedure which does not involve the introduction of instruments into the body.

Withdrawal times (colonoscopy) - The amount of time spent viewing the internal wall of the colon as the colonoscope is withdrawn during a colonoscopy.

Acronyms

ADR	Adenoma Detection Rate
BOW	Bag of visual words
CAD	Computer Aided Diagnosis
CNN	Convolutional neural networks
CWC	Colour wavelet covariance
FP	False positive
FN	False negative
FV	Fisher Vector
GLCM	Gray level co-occurrence matrix
I2CD	Image-to-class distance
iid	Independent and identically distributed
LLC	Locality constraint linear coding
LBP	Local binary patterns
LTP	Local ternary patterns
MCA	Mean class accuracy
NN	Nearest neighbour
SC	Sparse coding
SILTP	Scale invariant local ternary patterns
SVM	Support vector machines
TP	True positive
TN	True negative
VLAD	Vector of locally aggregated descriptors

This thesis is mainly focussing on learning highly discriminative local features and image representations to achieve the best possible image level classification performance for medical, particularly, for colonoscopy and histology (cell) images. Some experiments with radiology images were also given. Since this thesis was funded by the 2011-2016 EU FP7 ERC project "CODIR: colonic disease investigation by robotic hydrocolonoscopy"[3], I gave emphasize to the images from the colonoscopy domain.

This chapter explains the background and the motivation for feature learning approaches for medical image classification, the clinical background and the motivation for image analysis systems for colonoscopy, the contributions, and the organisation of the thesis.

1.1 Medical image analysis background and motivation

Three main steps are involved in automated classification systems in computer vision and medical image analysis: feature extraction, feature encoding, and classification (Figure 1.1). In the feature extraction stage, descriptors capturing a variety of local image properties are computed. The descriptors (features) from a set of training images, for which labels are available, are then clustered to generate a dictionary. In the feature encoding stage, this dictionary of features is then used to compute a compact image representation for any given image. In the training stage, the image representations obtained from the training images are used to train a classifier. Finally the classifier learned is used to predict the label of any test image.

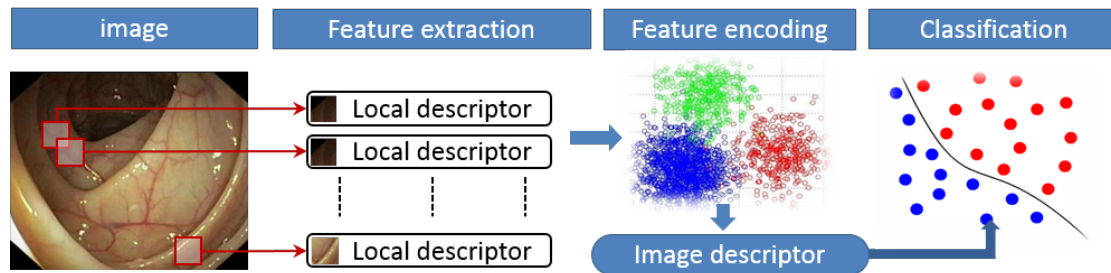


Figure 1.1: The major steps involved in the automated classification systems for computer vision and medical image analysis.

The current approaches proposed for medical image analysis (particularly for colonoscopy and cell image classification) have focussed mainly on identifying appropriate features and classifiers. Various hand-crafted features such as Root-SIFT [13, 113], colour histograms [70], Local Binary Patterns [113], Local Ternary Patterns [113], and classifiers such as SVM [113], Neural Networks [72] have been explored for colonoscopy image analysis. However, hand-crafted features may not be optimally discriminative for classifying images from particular domains (e.g. colonoscopy), as not necessarily tuned to the domain's characteristics.

Recently, feature learning approaches [19, 26, 132, 150, 151, 174, 175] have become popular as they learn domain-specific discriminative features and have been shown to improve the performance of, among others, medical image segmentation [19], image retrieval [132, 150, 151], and interest point matching [26, 174, 175]. These approaches assume that region-based annotations or a training set consisting of matching and non-matching pairs of image patches are available to learn the feature descriptors. For example, Winder et al. [26, 174, 175] proposed to learn the configurations of the local descriptor such as the smoothing factors, number of orientation bins, configuration of the local pooling regions, and others. Philbin et al. [132] proposed to learn the local descriptors by learning a projection matrix such that in the projected descriptor space the matching descriptors are assigned to the same cluster, and non-matching descriptors are assigned to different clusters. These approaches need a training set of matching and non-matching image patches to learn the local descriptors. On the other hand, Simonyan et al. [151] proposed a method to generate the matching and non-matching image patches from a weakly labelled dataset (dataset with image-level labels) for learning the descriptors. However, generating annotations

for any medical training set (region-level or matching and non-matching feature pairs) is a difficult, time-consuming task.

Convolutional neural nets (CNN) [79] have also been used to learn local features (filters). In CNN, a set of filters as well as the image-level classifiers are learned in a unified framework. Usually CNN require a very large amount of training data [122]; when this is not available, CNN may give worse performance than traditional, hand-crafted features and BOW-based feature encoding methods [122].

None of these feature-learning approaches have been explored for colonoscopy image classification yet. In this thesis I investigate novel approaches to learn features based on weak supervision (learning from image-level labels) for discriminative image classification.

1.2 Clinical background and motivation

More than one million new colorectal cancer (CRC) cases are diagnosed yearly worldwide [2]. CRC remains the second leading cause of cancer death in the world and the third most common cancer in the UK [2], although the death rate due to CRC has been dropping for more than 20 years worldwide. One of the reasons is the development of screening programmes identifying and removing polyps and other suspicious lesions before they can develop into cancers [7]. If CRC is diagnosed in its earliest stages, the chance of surviving for five years is 90%, and a complete cure is often possible [1]. Clearly, early identification of colonic abnormalities is crucially important.

Adenoma detection rate (ADR) is a commonly used predictor of the risk of developing colorectal cancer after undergoing a colonoscopy screening [167]. Although colonoscopy remains the gold standard for colorectal cancer screening, miss rates around 6% for CRC, and 12% to 17% for adenomas larger than 1 cm have been reported in different studies [25, 40], posing risk of developing colon cancer due to a failure to detect treatable lesions in time. Several studies have examined the miss rate associated with colonoscopy and its causes, which include inadequate training and experience of the examiners, misinterpretation of the images [157], and shorter-than-average withdrawal times [18]. It is therefore arguable that a reliable computer-aided detection (CAD) system specialised for identifying suspicious colonic abnormalities in colonoscopy

videos could contribute to improve ADR, e.g. by presenting clinicians with a second opinion obtained by objective and repeatable methods.

Some studies support this hypothesis. For example, Baker et al. [15] found a significant improvement in the sensitivity of polyp detection by inexperienced radiologist readers when CAD was used, and Regge et al. [136] reported an improvement of 9% and 2% in CRC detection rates for inexperienced and experienced readers, respectively, when analysing virtual colonoscopy with CAD. Complementarily, Mang et al. [112] found that a 'first-reader' CAD workflow, in which an observer only reviewed the colonic regions identified as suspicious by a CAD system, ignoring the rest of the colon, substantially decreased reading times while enabling accurate detection of colorectal adenomas. Similarly, video recording and post-procedure review improved ADR with flexible endoscopy data in [107]. This background gives the clinical motivation for the colonoscopy related research reported in this thesis.

1.2.1 Background on colonoscopy

Colonoscopy is an endoscopic procedure to inspect the colonic mucosa in a relatively painless way. Colonoscopy is used to investigate the potential cause of symptoms like abdominal pain, rectal bleeding, or changes in bowel habits; and, most relevant for my work, to screen for CRC [4]. Various kinds of colonoscopy/endoscopy systems exist, e.g. white-light endoscopy or flexible endoscopy (FE), wireless capsule endoscopy (WCE), and virtual colonoscopy (VC) [97].

Colonoscopy systems other than VC contain light sources and a camera to capture the images of the colon mucosa. They differ in various ways, e.g. purpose (whether to image the micro or macro structure of the colon), motion control, patient comfort, imaging type (whether zoomed or not, 3D or 2D), etc. The following section briefly explains various kinds of colonoscopy systems and Figure 1.2 shows some example images.

1.2.1.1 Flexible Endoscope

A flexible endoscope (FE) consists of a flexible, hollow tube and a camera with light sources on its tip. A channel in the tube is dedicated to surgical instruments for

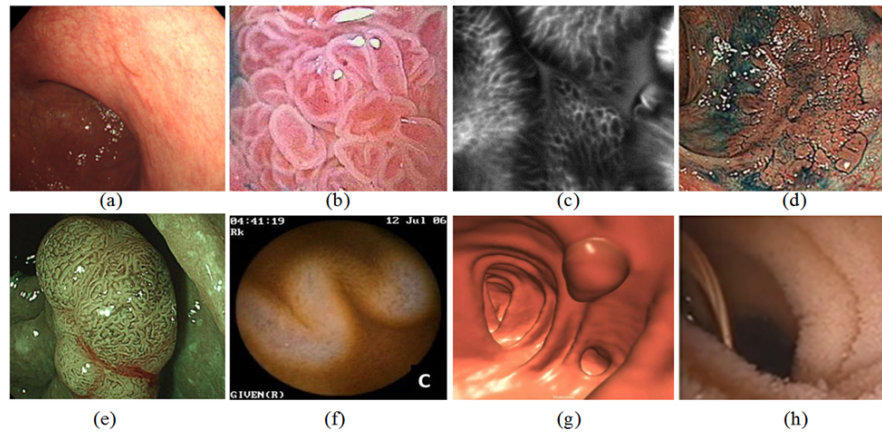


Figure 1.2: Examples of images taken from various endoscopy systems (a) white light (b) zoom (c) confocal laser endomicroscopy (d) chromoendoscopy (e) narrow band imaging (f) wireless capsule endoscopy (g) virtual endoscopy (h) hydro colonoscopy.

excisions, e.g. polyp removal, and biopsies. Disadvantages of FE include the need for sedation, patient discomfort, and a small but real risk of colonic perforation with adverse consequences including, in the most extreme cases, even death.

The following can be used with FE to get some additional advantages. I refer the reader to [17] for a detailed review. (a) *Zoom-endoscope* is a FE with the advantage of zoom-in at interesting regions with a magnification factor of up to 150 times. (b) *Chromoendoscopy* is a procedure sometimes used with FE to improve tissue localisation, characterisation, or diagnosis by applying colour dyes. (c) *Confocal Laser Endomicroscopy (CLE)* is a new diagnostic technique that allows microscopic examination of the digestive mucosa during endoscopy using low-power laser. 'Optical biopsies' are obtained by injecting a fluorescent marker and imaging with a high level of magnification (up to 1000-fold)[5]. (d) *Narrow Band Imaging* is a real time, on-demand image enhancement technique which places narrow-band filters in front of a conventional white light source to obtain tissue illumination in selected narrow wavelength bands.

1.2.1.2 Wireless Capsule Endoscopy

Wireless Capsule Endoscopy (WCE) is not designed particularly for the colon but for the whole gastrointestinal tract. It is a better imaging tool for the small intestine, where the FE has no access. In WCE the patient swallows a small capsule of size about 22×11 mm which contains light source, lens, camera, radio transmitters, and batteries. This capsule

then travels through the digestive system, propelled by peristalsis, and automatically takes images of 5 to 40 f/s during a travelling time of about eight hours [97]. There are several advantages, including very low levels of patient discomfort as the capsule travels in the digestive system, but also drawbacks: an extensive pre-procedure complete bowel preparation lasting 24 hours and use of drugs to promote capsule transit is needed, the motion is passive resulting in uncontrollable images (unlike FE), cannot take biopsy, contamination of the lenses and reluctance of patients to adopt the procedure.

1.2.1.3 Virtual Colonoscopy

Virtual Colonoscopy (VC) is a non-invasive procedure to look for signs of pre-cancerous growths and other diseases in the large intestine. 3D images are taken using computerised tomography (CT) or, less often, magnetic resonance imaging (MRI). Specialised software creates a 3D view of the inside of the large intestine. Advantages include the speed and non-invasiveness of the procedure with no patient discomfort and no sedation required. A limitation is that, since VC is non-invasive, lesions cannot be excised directly as with FE, e.g. taking a biopsy. The major drawback of this procedure is the loss of colour and texture information.

All the above mentioned imaging modalities other than VC and CLE provide colour images. VC is the only procedure providing 3D image models.

1.2.2 The CODIR project

FE is uncomfortable for patients. In FE CO₂ gas is often used to inflate the colon to distend the colonic mucosa. Usually pain medication and sedation (which can cause drowsiness) are given to the patient prior to the procedure to reduce discomfort. Sedation imposes a recovery time burden on patients [83]. To minimise discomfort, the water method [83, 84, 86] uses warm water to inflate the colon in unsedated patients, instead of CO₂ (as in conventional endoscopy). The water-method shows improved adenoma detection rate (ADA) by up to 50%.

Inspired by the water-method, the 2012-2017 EU-funded project CODIR (Colonic Disease Investigation by Robotic hydro colonoscopy)[3], of which this PhD was part,

aims to develop a controllable, tethered swimming/submerging robot inspecting the colon wall in an irrigated environment.

CODIR stems from two considerations: (1) the replacement of the conventional FE with a lower-discomfort system for inspection of the colonic mucosa (internal wall), and (2) the very recent concept of hydro-colonoscopy (or the water method) whereby water or specifically developed chemical solution [85] is used instead of the traditional air insufflation. CODIR aims to enable a breakthrough in patient-compliant complete endoscopic examination and biopsy of the colon for the further study of life threatening disorders [3].

Although the work reported here is a part of the CODIR project, it proved impossible, in the course of the work, to obtain hydro-colonoscopy videos, as the robotic platform is still under development. Therefore, the proposed methods were evaluated on the images taken from white-light colonoscopy.

1.3 Contributions

The contributions of this thesis can be summarised in three points: (1) novel feature learning approaches for discriminative image classification; (2) a novel approach to capture inter-cluster statistical features for feature encoding; and (3) extensive experimental evaluation of the proposed approaches and various state-of-the-art approaches for colonoscopy and histology (cell) image classification. Some experiments with radiology images were also given.

The approaches proposed in this thesis are not particularly designed for colonoscopy, histology or radiology images. Hence, can be applied to other medical imaging domains, e.g. brain tumour segmentation [118].

1.3.1 Novel feature learning approaches

To the best of my knowledge, discriminative feature/descriptor learning approaches have never been explored for colonoscopy images. I propose unsupervised (Chapter 5) and weakly-supervised (Chapters 6 and 7) feature learning approaches for colonoscopy image classification. Unlike existing discriminative feature learning approaches

explained in Section 1.1, where labelled data is required for learning in the form of region-level annotations [19], or matching and non-matching feature pairs [26, 174, 175], I use only weakly labelled data, i.e. training data with image-level labels, to learn the most discriminative local image features for image-level classification.

1.3.2 Inter-cluster statistics for feature encoding

Feature encoding plays an important role for image classification. Intra-cluster features such as bag-of-words (BOW) have been widely used for feature encoding, to capture statistical information within each cluster of local features, but fail to capture the inter-cluster statistics, such as how the visual words co-occur in images or image regions. I propose a new method (Chapter 9) to choose a subset of cluster pairs and propose new inter-cluster statistics to improve the traditional BOW-based feature encoding approaches. Since the cluster pairs are selected based on image regions rather than the whole images, the final representation also captures the local structures present in images.

1.3.3 Experimental evaluations

I provide extensive experimental results to validate the proposed approaches. Comparative experiments with state-of-the-art approaches on various datasets show that the proposed approaches match or surpass the state-of-the-art approaches for colonoscopy as well as cell (histology) image classification. Mean class accuracy (average of per-class accuracies, defined in Section 3.2) was used to measure the classification performance, as it is a widely used measure for classification in medical imaging [102, 172, 173] and computer vision [146, 168].

1.4 Thesis organisation

Figure 1.3 is a map of some paths that the reader may choose to explore the chapters. Chapters 4-7 focus on learning local image descriptors for image classification, and Chapter 9 focuses on improving the traditional feature encoding approaches.

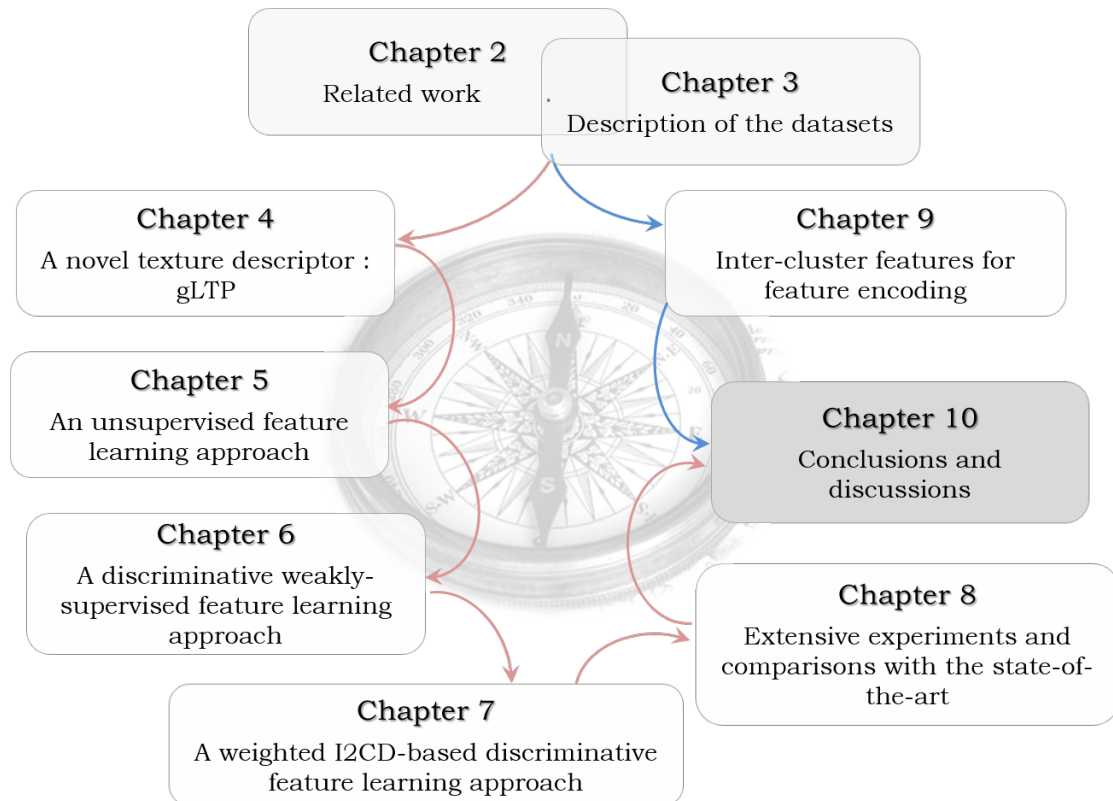


Figure 1.3: Thesis roadmap. Chapters and their relations.

- **Related Work.** Chapter 2 reviews the work related to colonoscopy image analysis, local descriptor learning and image classification. A concise review of BOW related approaches is also given.
- **Datasets and experimental settings.** Chapter 3 explains the datasets as well as the experimental settings used in the experiments throughout in this thesis.
- **The Generalised Local Ternary Patterns.** Inspired by the success of the *Local Binary Patterns* (LBP) descriptor, and its variants for colonoscopy image classification, in Chapter 4 I propose a generalised variant of LBP called the *generalised Local Ternary Patterns* (gLTP). This chapter is a part of the paper which is under preparation for the journal of *Medical Image Analysis* (MIA).
- **Extended Multi-Resolution Local Patterns and unsupervised feature learning.** Although gLTP shows competitive performance for image classification compared to LBP and its variants, gLTP loses information due to the binarisation procedure involved in the feature extraction stage. Therefore, Chapter 5 proposes a novel descriptor called the *Extended Multi-Resolution Local Patterns* (xMRLP), and its simplified variant the *Multi-Resolution Local Patterns* (MRLP). Since xMRLP

contains a set of free parameters an unsupervised approach to learn these parameters is also presented. This chapter is a part of the journal paper which is under preparation for MIA.

- **Discriminative Feature Learning using weak labels.** Chapter 6 is an extension of Chapter 5, where a discriminative weakly-supervised approach is proposed for feature learning using image-to-class distances (I2CD) [23]. This work has been published at International Symposium on Biomedical Imaging 2015 [114].
- **Discriminative feature learning with weak-labels and weighted I2CD.** The I2CD used in Chapter 6 can be affected by the noisy local features as well as the features from the image background. Therefore, in Chapter 7 I propose a feature learning approach based on weighted I2CD, where the I2CD calculated from different classes are weighted differently to learn the local features as well as an image-level classifier. This chapter is a part of the journal paper which is currently under preparation for MIA.
- **MRLP and xMRLP for colonoscopy and cell image classification: experimental evaluation.** In Chapter 8 I provide extensive experiments with the proposed as well as various other state-of-the-art features. Also in this chapter I propose image-classification systems to classify colonoscopy and cell images into predefined classes, and show my systems outperform the state-of-the-art. Some work from this chapter based on cell images has been published at the I3A 1st workshop on Pattern Recognition Techniques for Indirect Immunofluorescence Images, International Conference on Pattern Recognition, 2014 [115, 116]. This chapter is also parts of the journal paper which is accepted by Pattern Recognition [117].
- **Inter-cluster features for image classification** Since the traditional feature encoding approaches (e.g. BOW) capture only the intra-cluster information (statistics of each cluster of features), in Chapter 9 I propose an approach to improve them by capturing the inter-cluster information in addition to the intra-cluster features. The work from this chapter has been published at MICCAI 2014 [120].
- **Conclusion, discussion and future directions.** Chapter 10 concludes this thesis and suggests future directions for exploration.

This section reviews the work related to colonoscopy image analysis, feature learning and feature encoding approaches (e.g. BOW) which have been proposed in the computer vision and medical image analysis literature, and identifies their possible limitations.

2.1 Endoscopy image analysis

As explained in Section 1.2.1, different imaging modalities are used to image the gastrointestinal tract. The images taken from WCE have similar features to conventional, white-light colonoscopy (WLC) images. The images taken from confocal laser endomicroscopy, virtual endoscopy and narrow band imaging are very different from the images taken from WCE and WLC. Therefore, this literature survey only focuses on the systems proposed for WCE and WLC.

I categorize the systems proposed for endoscopy image analysis by *purpose/tasks*, *feature representations* and *classification methods*. The purpose of the systems can be further categorised to consider image-level classification tasks, region segmentation (region-level detection/classification) methods, lesion quantification, image retrieval, and topographic segmentation.

2.1.1 Purpose

Table 2.1 shows an overview of the methods proposed for (a) image classification, (b) region detection/classification, (c) lesion characterisation, (d) image retrieval, and (e)

topographic segmentations for WLC and WCE images.

Image classification aims to assign the whole image to one of several classes, e.g. normal vs abnormal [72, 113, 158], informative vs non-informative[70]. Supervised classifiers are generally used.

Region segmentation (region detection/classification) approaches segment regions of interest in images, e.g. abnormal region segmentation, polyp segmentation. Both supervised and unsupervised approaches have been reported.

Supervised approaches require region-level annotations in order to train classifiers, e.g. labelling regions into normal vs abnormal [91, 92] or normal vs bleeding [88, 89]. Unsupervised approaches rely on pre-defined rules, e.g. thresholding schemes on colour values [63], and do not therefore need annotated images. The rules must be devised in close collaboration with clinical experts.

Table 2.1 shows that a higher amount of work has been directed to region detection and classification than to other tasks. Since bleeding is an important indicator (e.g. bowel cancer), the majority of the work reviewed concerns “normal vs. bleeding” classifiers. Polyp detection is also much investigated in the literature. Polyps are abnormal growths and carry a high risk of developing into cancer [8]. Crohn’s disease, ulcers and cancers have been considered for automatic WCE and endoscopy image analysis [65, 109].

Lesion characterisation / quantification systems can be thought as a special kind of classifiers assuming a specific type of abnormality and classifying images into a set of abnormality varieties, e.g. classifying Crohn’s disease images into mild or severe [76], automatic grading of celiac disease images [32]. Image ranking methods have also been proposed for lesion quantification (e.g. Crohn’s disease [75]) where a ranking score indicates the lesion severity.

Image retrieval from large image repositories has been reported for endomicroscopy [11, 12], but, to my knowledge, not for WLC or WCE images. The image of a specific lesion observed during or after the examination could be used as query to retrieve similar, annotated images, to assist with the final clinical decision. For instance, [176] reports an iterative process which updates the weights of the features based on the user’s interaction with the system for retrieval of endoscopy images.

Purpose	endoscopy type	Reference
Image classification		
Normal vs bleeding	WCE	[16, 39, 70]
Normal vs abnormal	WCE	[72]
	endoscopy	[158]
	colonoscopy	[113, 119]
Normal vs ulcer	WCE	[182, 184]
Polyp vs non-polyp	WCE	[183, 185, 190]
Informative vs non-informative	WCE	[70]
Normal, Crohn's disease and other abnormalities	WCE	[20]
Normal, bleeding, polyps, ulcer and other abnormalities	WCE	[98]
Region segmentation (region detection/classification)		
Bleeding detection	WCE	[63, 64, 77, 80, 88–90, 106, 135]
Polyp detection	WCE	[105, 149]
	endoscopy	[165]
	colonoscopy	[10, 44, 57, 74, 121]
Tumour region	endoscopy	[67]
Polyp and ulcer region	WCE	[65]
Normal vs Abnormal	endoscopy	[91, 92]
Malignant vs benign tumour	endoscopy (gullet)	[109]
Normal vs Cancer	endoscopy	[66, 154]
Normal vs gastritis	endoscopy (stomach)	[153]
Normal, Crohn's disease and others	WCE	[76]
Lesion quantification		
Crohn's disease(mild, severe)	WCE	[75, 76]
Celiac disease	endoscopy	[32]
Tumour (malignant vs benign)	endoscopy (gullet)	[109]
Image retrieval		
Image retrieval	endoscopy	[176]
Topographic segmentation		
Topographic segmentation	WCE	[36, 37]

Table 2.1: Existing endoscopy image analysis systems for different purposes

Topographic segmentation approaches segment the gastrointestinal tract into major topographic areas, for example, stomach, small intestine, and large intestine. According to medical specialists, this can reduce exam annotation times by up to 12% [36, 41]. For instance, MPEG 7 visual descriptors were used for topographic segmentation in [36, 41].

2.1.2 Features and representations

Texture, colour, shape, or combination of them have been used for endoscopy image analysis. Table 2.2 shows an overview of the existing approaches proposed for endoscopy image analysis.

2.1.2.1 Texture features

Texture is an important property for medical images, capturing the structure of the variation of intensity patterns. Various texture features have been proposed in the literature for general image analysis [124, 177, 189]. For example, Zhang et al. [189] categorized these features into (1) statistical, (2) model-based, and (3) structural features. Statistical texture features capture statistical information about the local spatial distribution of pixel values. e.g. gray-level co-occurrence matrices, LBP. In the model based approaches, a texture image is modelled as a probability model or as a linear combination of a set of basis functions. Model based approaches include, Markov model [35], Gabor filters [58], etc. In structural methods, texture is viewed as consisting of many textural elements arranged according to some placement rules. Commonly used element properties are average element intensity, area, perimeter, eccentricity, orientation, etc [14]. However, in this section I only focus on the texture features which were proposed for endoscopy image analysis.

Gray level co-occurrence matrices (GLCM), LBP and texture spectrum (TS) have been used as texture features in the spatial as well as in the frequency domain in endoscopy image analysis. A global image representation is normally obtained by computing statistical representations (e.g. histogram) of these features.

GLCM [50] captures the distribution of co-occurrence of intensity values of pixels in local neighbourhoods. Statistical measures (e.g. contrast, energy) on GLCM in frequency or spatial domain are often a basis to generate more complex texture features [67, 98], although these statistics have also been used directly, e.g. to detect pre-cancerous polyps [121]. A GLCM-based feature is the colour wavelet covariance (CWC) [67] whose entries capture the covariance of statistical measures of co-occurrence of wavelet coefficients between colour channels. CWC has been used to detect polyps [67] and to classify WCE images into 5 different categories (normal,

bleeding, ulcer, polyps and unclassified defects) [98]. CWC-related measures have been used to detect colorectal lesions [121].

LBP [126] and TS [52] are features describing the local texture around each pixel by comparing and thresholding pixel difference in local neighbourhood. In contrast to LBP, three-level thresholding is used in TS. Statistics of LBP/TS have been used, e.g. for polyp detection [190], bleeding detection [89], and normal-abnormal image classification [66, 72, 113].

Gabor filters are Gaussian functions modulated by oriented complex sinusoidal signals [188]. Gabor filters capture low-level oriented edges and are widely used for texture segmentation [188, 189]. They have also been explored for detecting polyps in WCE images [149] and to classify chromoendoscopy and narrow-band imaging into a set of predefined categories (normal, precancerous, and cancerous)[137–140].

SIFT features, on the other hand, captures the histograms of local intensity gradients. SIFT features have been employed for ulcer vs normal image classification of WCE images [182].

2.1.2.2 Colour features

Colour is a salient feature for bleeding detection. Histograms and other statistics in different colour spaces (e.g. RGB or HSI) have been used for bleeding detection in WCE images [80][16][39][158]. Colour histograms reportedly perform better than CWC features for classifying colonoscopy frames as informative (e.g. with folds, lumen, abnormalities) or uninformative (i.e. poor-quality images not useful for decision making), and as containing bleeding or not [70].

Most of the endoscopy image classification approaches compute global colour histograms (from the entire images), failing to capture local image properties. In [70] colour histograms computed from non-overlapping image blocks are concatenated to get a feature representation, which encodes spatial/location information of the blocks and therefore is not desirable for endoscopy images.

2.1.2.3 Shape-based features

Shape-based image features have also been tested in endoscopic image analysis, e.g. ellipses to approximate the contours of polyps [57], edge orientation histograms, part of MPEG-7 visual descriptors for Crohn's disease classification [76], and for topographic segmentation of WCE videos [36, 38].

2.1.2.4 Combination of texture, colour and/or shape-based features

Combining different features may capture richer image representation than any individual feature type. In practice, colour and texture features are often combined together, e.g. colour histograms are combined with TS-based features for normal-abnormal image classification [158], with the statistics of discrete wavelet transformations (DWT) for lesion detection [91], and with LBP histograms to detect bleeding [89]. Edge orientation histograms, part of MPEG-7 visual descriptors, together with colour and texture features such as dominant colour and homogeneous texture, have also been used for Crohn's disease classification [76] and for topographic segmentation of WCE videos [36, 38].

Finally, a few approaches based on feature encoding (e.g. BOW) with features such as SIFT have been reported, e.g. for ulcer-normal WCE image classification [182], and normal-abnormal colonoscopy image classification [113].

In contrast to the above approaches, Yuan et al. use saliency information for polyp classification [183, 185] and ulcer detection [184] in WCE images. In their approach the encoded features computed from the salient and non-salient regions of an image are combined together to get the final image representation. This approach show improved performance compared to the ones where the saliency information is not considered.

2.1.3 Classifiers for endoscopy image analysis

In endoscopy image analysis, SVM and artificial neural networks (ANN)-based approaches have been widely used to classify and/or detect lesions. Table 2.2 summarizes the different classifiers which have been used by existing approaches.

Poh et al. [135] propose a hierarchical ANN ensemble for bleeding detection in WCE images. In their approach, first each image is divided into blocks (small rectangular image regions), then blocks into cells. An ANN is trained on cells to classify them into bleeding/non-bleeding cells. Then the cell-level classification responses are used to train a block-level ANN to classify blocks. The final classification decision of each block has been computed based on the outputs of the cell and the block-level classifiers.

Kodogiannis et al. [72] propose an ANN-based classifier fusion approach, which combines multiple classifiers from the features extracted from different colour channels for normal-abnormal image classification. However, as there is no comparison given with other classifiers such as SVM, or ANN (without fusion) it is hard to say whether this classifier fusion approach improves the classification performance or not.

Since the shape and the size of the possible abnormalities vary, single-size patch analysis may not provide sufficiently discriminative feature vectors. Therefore, Peng et al. [91, 92] propose a multi-size patch-based classifier ensemble to detect abnormal regions (polyps, tumours, inflammation, bleedings, ulceration and diverticula) in colonoscopic images. Different sizes of overlapping patches are extracted, and a classifier is trained for each patch size independently. In the classification stage, features from each patch size are passed to this ensemble classifier to get the classification score, and then these scores are then aggregated to get the final decision. This multi-size patch classifier ensemble approach shows improved performance over the classifier which is trained on single-size patches.

Classifier cascades using SVM classifiers have also been used in [20, 76] to classify images into normal, mild and severe Crohn's disease. In this approach, first the images are classified into normal vs lesions, and then the lesion images are further classified into mild vs severe.

Ref.	Purpose	Feature	Colour Space	Classifier	Colonoscopy type	Dataset	Results
Colour features							
[106]	bleeding detection	mean & variance of intensity values	HSI	SVM	WCE	6,416 images	acc 97.8%
[63]	bleeding detection	histogram of weighted R values	RGB	decision rules	WCE	1,000 images	se 92% sp 85%
[135]	bleeding detection	histogram of intensity values	HSI	NN ensemble	WCE	200 images	acc 93%
[90]	bleeding detection	average colour in each channel	RGB, HSI	SVM	WCE	150 bleeding, 1,000 images without blood	se 96.6% sp 99.5%
[77]	bleeding detection	saturation & luminance values	HSI	decision rules	WCE	5 videos (~5,000 images)	se 88.3%
[80]	bleeding detection	ratios of R to G and B, ratio between G and B, saturation, normalised R values	HSI, RGB	decision rules	WCE	42 images	se 87% sp 90%
Texture features							
[66]	normal vs cancer region classification	statistics of texture spectrum	gray scale	discrimination measure	lung cancer	-	-
[169]	unsupervised clustering of endoscopic image regions	LBP histogram	gray scale	self organising maps	endoscopy	-	-

Colour and texture features									
1. Frequency domain techniques									
	polyp detection	CWC features	RGB	Linear Analysis	Discrimination	endoscopy	images from 60 videos (5-10s duration each), 1,380 images	sp 99.3% se 93.6%	
[98]	classification into 5 classes (normal N, bleeding B, polyp P, tumour T, undefined U)	CWC-based statistical features	RGB	RBF ANN		WCE	images from 6 videos, training 2,000 N, 23 B, 54 P, 123 T, 58 U; testing on 85,000 images.	se 93%, sp 95%	
[92]	abnormal region detection	colour histogram and DWT-based statistical features in each colour channel	CIELab	multi-size patch ensembles. SVM/ANN for each patch size.	patch classifier	colonoscopy	58 images (46 abnormal and 12 normal) - LOU	acc 83.4%	
[91]	abnormal region detection, comparison between binary and one-class SVMs	colour histogram and DWT-based statistical features in each colour channel	CIELab	multi-size patch ensembles. SVM for each patch size.	patch classifier	colonoscopy	58 images (46 abnormal and 12 normal) - LOU	2-class SVM performs better than one-class SVM	
[121]	precancerous polyp detection	GLCM statistics from wavelet coefficients	gray scale	ANN		colonoscopy	8 frames	acc 95%	
[190]	polyp detection	colour moments, LBP histograms obtained on contourlet transformed images	OPC	SVM		WCE	-	acc 97% (training accuracy)	

2. Spatial domain techniques							
[72]	normal vs abnormal image classification	statistics on texture spectrum on each channel (R,G,B,H,S,V)	RGB, HSI	ANN for each channel and Fuzzy Integral for the final decision	WCE	140 images (half for train, half for test)	acc 95.7%
[158]	normal-abnormal image classification	statistics on texture spectrum obtained in each channel, colour histogram	RGB, HSI	ANN	colonoscopy	54 abnormal, 12 normal images (half for train, half for test).	acc 96.9%
[106]	bleeding detection/region classification (bleeding, lesion & normal)	colour histogram, 3D LBP histogram	HSI	2 stage classification, (1) identify suspicious regions using adaptive colour histograms and decision rules, (2) classify them using SVM	WCE	84 videos, training on 83 videos (50,000 frames for each video)	-
[16]	bleeding detection	histogram, dominant colours, co-occurrence of dominant colours	HSI	SVM ensemble (SVM for each feature type)	WCE	5 videos, each contains 55,000 frames	se 80%, sp 92%
[188, 89]	bleeding detection	colour moments, statistics of LBP	HSI	ANN	WCE	1800 bleeding, 1800 non-bleeding blocks. half for training	acc 90%
[183]	polyp vs normal classification	SIFT features with saliency information, together with BOW	-	SVM	WCE	872 images (436 normal, 436 polyps)	acc 90%, se 87.9%, sp 93%
[185]	polyp vs normal classification	SIFT and LBP features with saliency information, together with BOW	-	SVM	WCE	2,500 images (2,000 normal, 500 polyps)	acc 93.2%, se 90.8%, sp 94.5%

[184]	Ulcer vs normal classification	SIFT, HOG and LBP features with saliency information, together with Locality Constrained Linear Coding	-	SVM	WCE	340 images (170 normal, 170 ulcer)	acc 92.6%, se 94.1%, sp 91.2%
3. Comparative studies							
[10]	comparison of GLCM and LBP for polyp detection	GLCM obtained from gray scale patches, colour LBP, OPC-LBP	RGB, OPC, gray scale	SVM	colonoscopy	1,736 images from 4 videos	OPC-LBP performs better than GLCM
[70]	comparison of colour histograms (CH) vs texture (T) for informative frame detection and bleeding detection	CWC, concatenated colour histograms obtained from image regions	HSI	SVM, ANN	WCE	200 informative frame detection, 300 images for bleeding detection	acc (1) info. frame detection CH = 94.1%, T = 73.9%, (2) bleeding detection CH = 93.4%, T = 92.3%
[44]	comparison of different features (GLCM, LBP, DWT) for polyp detection.	GLCM(colour, gray scale), LBP(colour, gray scale, rotation invariant), wavelet-GLCM (GLCM features on 3-level wavelets obtained from each colour channel)	RGB, OPC, gray scale	SVM, KNN	colonoscopy	1,600 images	GLCM and LBP perform equally well. SVM performs better than KNN. best acc 84.3% by colour GLCM with KNN

[154]	comparison of different classifiers for cancer region segmentation	adapted colour histograms, LBP	HSV	decision trees, naive Bayes, KNN, SVM	endoscopy	176 images	SVM performs better than others, acc ~ 90%
Edge/shape features							
[65]	polyp detection	dominant texture segments (using log-Gabor filters), edge features such as curvatures	gray scale	SVM	WCE	50 frames (10 contains polyps and 40 normal frames)	se 97%, sp 94%
[57]	polyp detection	elliptical shapes	RGB	decision rules	endoscopy	8,621 frames contains 27 polyps.	TP = 26, FP = 1, FN = 5
[74]	abnormality detection	curvature changes in edges	gray scale	decision rules	-	-	-
Combination of features (colour, texture and edge features)							
1. MPEG-7 Visual descriptors							
[76]	Crohn's disease 3 class classification (normal N, mild M, and severe S)	edge orientation histograms, dominant colour, homogeneous texture	LUV	SVM cascade(1 st for N vs lesion, 2 nd for lesion images into M vs S)	WCE	513 images (212 N, 213 M, 108 S)	acc 93.8% (N vs S), 90% (M vs S)
[20]	normal, Crohn's disease and others	MPEG 7 visual descriptors and GLCM features	LUV	SVM cascade(1 st for normal vs lesion, 2 nd for lesion images into M vs S)	WCE	images from 10 videos, number of images vary for each videos (500 - 2000). train on 10% of data from each video	

[38]	normal, polyp, or blood	MPEG 7 visual descriptors	YCrCb, HMMD	statistical measures	WCE	~ 60,000 images	
[36, 41]	topographic segmentation	MPEG 7 visual descriptors		Bayesian classifier, SVM	WCE	60,000 images	

Table 2.2: An overview of the existing endoscopy image analysis systems. (acc - accuracy, se - sensitivity, sp - specificity, OPC - opponent colour)

2.2 Image representation using visual words

In image classification, image descriptors/features play an important role as they capture image/region properties, such as colour, shape, edges, texture, etc. In general, there are two approaches to describe an image using descriptors, global and local. The global descriptor captures the overall statistics of an image. On the other hand, the local descriptors capture the local image properties. Various global descriptors have been proposed, for example colour histograms [155], GIST descriptors [128]. These global representations may not well capture the local image properties, and may not be invariant to image and object transformations. On the other hand, local descriptors (e.g. SIFT [87, 104]) capture the local image properties such as local shape, texture, etc. and they are designed to be robust to image transformations. Since many local descriptors can be extracted from each image, an aggregation strategy is necessary to get an image-level representation from the extracted local features. Feature encoding approaches such as BOW together with feature pooling strategies are often used for this purpose.

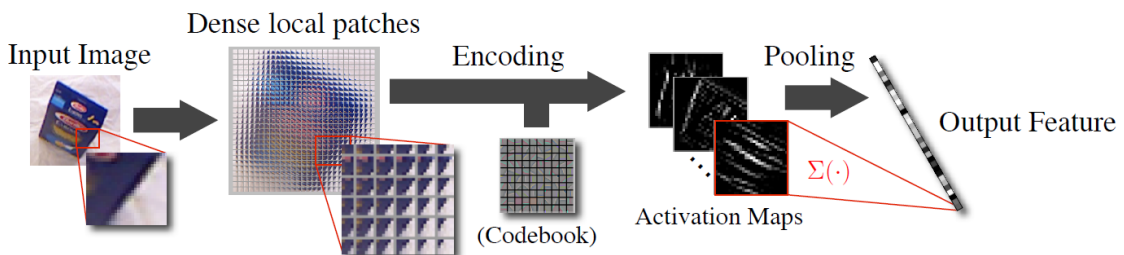


Figure 2.1: Pipeline of feature encoding approaches. Local features extracted from an input image are encoded using a pre-trained dictionary. The encoded-features are then pooled to get an image-level representation, on which the classification is based on (image was extracted from [47]).

The major steps involved in feature encoding-based image classification approaches are (Figure 2.1) local feature extraction, encoding, aggregation/pooling and classifier learning. In the feature extraction stage, low-level features which describe the local image properties are extracted. Then they are vector-quantised using a pre-trained dictionary to obtain a fixed size mid-level representation for each local feature, called the *encoded-features*. These encoded local features from each image are then aggregated by a pooling step to get a vector which describes each image (image-level representation). Finally the image-level representations from the training images are used to build a

classifier to separate different classes. In the testing, the image-level representation of a test image is given as the input to the trained classifier to predict its label.

Various approaches have been proposed to improve each step of this pipeline. They mainly focus on: what kind of, and how the local features are extracted (e.g. sparse vs dense feature extraction [125]); how the dictionary is built (e.g. unsupervised manner [33, 34], supervised manner [108]), how to effectively encode each feature using the dictionary (e.g. use of different cluster-statistics [60]), and how to aggregate the encoded features effectively so that the final image representation will capture discriminative information (e.g. different pooling mechanisms [61, 181]).

2.2.1 Feature extraction

Various local descriptors have been proposed in the literature to effectively capture the local image properties, for example, SIFT [104] descriptors capture local shape/texture, LBP [126] and BRISK [87] descriptors capture texture features.

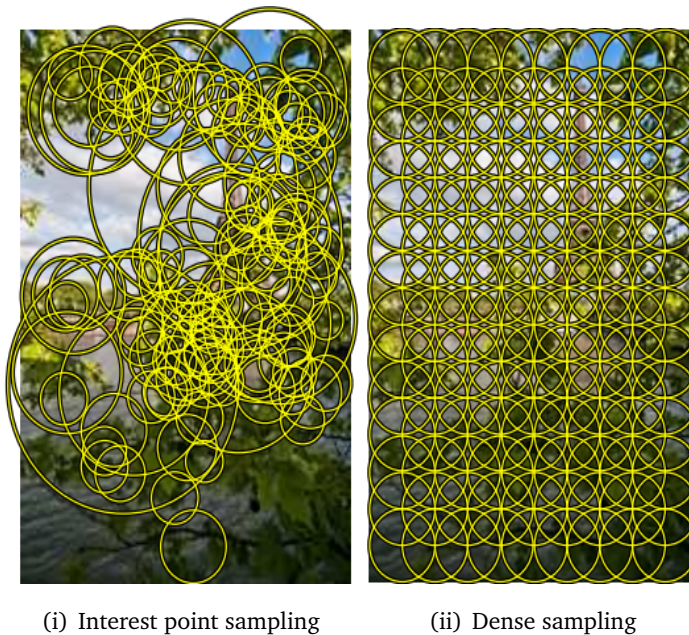


Figure 2.2: Example descriptor sampling strategies; (a) interest points-based sampling, and (b) dense sampling (image was obtained from [47]).

There are two sampling methods generally used for feature extraction (Figure 2.2): (i) *dense sampling*, where the feature extraction is based on a regular grid of points placed over the images, and (ii) *interest points*, where special points in the images are identified by a detector (e.g. Harris detector [51]) and feature descriptors computed

around those points. Dense feature sampling seems to lead to better performance compared to interest points detectors for image classification [125].

2.2.2 Feature encoding

Feature encoding methods transform the low-level (or local) image descriptors into a mid-level representation called the encoded features using a pre-trained dictionary. Various feature encoding methods such as BOW [152], *Sparse Coding* [168, 178], *Fisher Vectors* [129], and *Vector of Locally Aggregated Gradients* [60] have been proposed in the literature.

2.2.2.1 Bag-of-Words

BOW is widely applied as a feature encoding method for medical [113, 120] as well as natural [146] image classification. In BOW local features sampled from training images are clustered to build a dictionary (codebook). This dictionary represents a set of visual words (clusters or dictionary elements) which are then used to compute a BOW frequency histogram as a feature vector representation of any given image. BOW uses hard quantisation where each local image descriptor is assigned to only one visual word.

2.2.2.2 Sparse coding

Sparse coding (SC) has shown improved performance over BOW for image classification [178]. In SC each local image descriptor is reconstructed using weighted combination of a few dictionary elements.

Locality-constrained linear coding (LLC) [168], on the other hand, enforces locality instead of sparsity. LLC utilizes the local linear property of manifolds to project each descriptor into its local coordinate system [168]. Let $X_i \in \mathbb{R}^{d \times N_i}$ be a matrix in which each of the N_i columns is a d -dimensional local descriptor extracted from an image I_i , i.e. $X_i = [\mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{iN_i}]$. Given a codebook with M entries, $B = [\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_M] \in \mathbb{R}^{d \times M}$, LLC uses the following criterion to compute the codes $C = [\mathbf{c}_{i1}, \mathbf{c}_{i2} \dots \mathbf{c}_{iN_i}]$.

$$\begin{aligned} \underset{\mathbf{C}}{\operatorname{argmin}} \sum_{j=1}^{N_i} \|\mathbf{x}_{ij} - B\mathbf{c}_{ij}\|^2 + \lambda \|\mathbf{d}_{ij} \odot \mathbf{c}_{ij}\|^2 \\ \text{s.t. } \mathbf{1}^T \mathbf{c}_{ij} = 1, \quad \forall ij \end{aligned} \quad (2.1)$$

where \odot denotes the element-wise multiplication and,

$$\mathbf{d}_{ij} = \exp\left(\frac{\operatorname{dist}(\mathbf{x}_{ij}, B)}{\sigma}\right) \quad (2.2)$$

where $\operatorname{dist}(\mathbf{x}_{ij}, B) = [\|\mathbf{x}_{ij} - \mathbf{b}_1\|_2^2, \dots, \|\mathbf{x}_{ij} - \mathbf{b}_M\|_2^2]^T$ and σ is a decay parameter. A fast approximation to LLC was described in [168] to speed up the encoding process. Specifically, instead of solving Problem (2.1), the K (with $K < d < M$) nearest neighbours of \mathbf{x}_{ij} in B were considered as the local bases \bar{B}_{ij} and a much smaller linear system (Equation (2.3)) was solved to get the local linear codes.

$$\begin{aligned} \underset{\mathbf{C}}{\operatorname{argmin}} \sum_{j=1}^{N_i} \|\mathbf{x}_{ij} - \bar{B}_{ij}\mathbf{c}_{ij}\|^2 \\ \text{s.t. } \mathbf{1}^T \mathbf{c}_{ij} = 1, \quad \forall ij \end{aligned} \quad (2.3)$$

The image representation of an image I_i is then obtained by aggregating (pooling) the sparse codes associated with the local descriptors.

2.2.2.3 Fisher vectors

Fisher vectors (FV) capture additional information about the distribution of the image descriptors compared to the count (0^{th} -order) statistics in BOW. FV has shown improved performance over BOW and SC for image classification in [130]. In FV, the dictionary is first modelled as a Gaussian mixture model (GMM) $p(\mathbf{x}|\Theta)$:

$$\begin{aligned} p(\mathbf{x}|\Theta) &= \sum_{m=1}^M \pi_m p(\mathbf{x}|\mu_m, \Sigma_m) \\ p(\mathbf{x}|\mu_m, \Sigma_m) &= \frac{\exp^{-\frac{1}{2}(\mathbf{x}-\mu_m)^T \Sigma_m^{-1} (\mathbf{x}-\mu_m)}}{\sqrt{(2\pi)^d \det(\Sigma_m)}} \end{aligned} \quad (2.4)$$

where $\Theta = (\pi_1, \mu_1, \Sigma_1, \dots, \pi_M, \mu_M, \Sigma_M)$ are the parameters of the GMM. $\pi_m \in \mathbb{R}^+$ ($\sum_m \pi_m = 1$), $\mu_m \in \mathbb{R}^d$ and $\Sigma_m \in \mathbb{R}^{d \times d}$ are respectively the weight, the

mean and the covariance of the m^{th} Gaussian. GMM uses a soft descriptor-to-cluster assignment:

$$q_m(\mathbf{x}_{ij}) = \frac{\pi_m p(\mathbf{x}_{ij} | \mu_m, \Sigma_m)}{\sum_{l=1}^M \pi_l p(\mathbf{x}_{ij} | \mu_l, \Sigma_l)} \quad (2.5)$$

In FV each cluster is then represented based on the derivative of the GMM with respect to its parameters $\{\mu_m\}$ and $\{\Sigma_m\}$ (1st and 2nd order statistics), i.e.

$$\begin{aligned} \mathcal{G}_{\mu_m}^i &= \frac{1}{N \sqrt{\pi_m}} \sum_{j=1}^{N_i} q_m(\mathbf{x}_{ij}) \Sigma_m^{-\frac{1}{2}} (\mathbf{x}_{ij} - \mu_m) \\ \mathcal{G}_{\Sigma_m}^i &= \frac{1}{N \sqrt{2\pi_m}} \sum_{j=1}^{N_i} q_m(\mathbf{x}_{ij}) \left[(\mathbf{x}_{ij} - \mu_m)^T \Sigma_m^{-1} (\mathbf{x}_{ij} - \mu_m) - 1 \right] \end{aligned} \quad (2.6)$$

The final image description is the concatenation of $\mathcal{G}_{\mu_m}^i$ and $\mathcal{G}_{\Sigma_m}^i$ for all $m = 1, \dots, M$, leading to a dimensionality of $2Md$. For e.g. SIFT features with a dictionary size of 64 will lead to an image level representation of size $2 \times 64 \times 128$.

2.2.2.4 Vector of locally aggregated descriptors

The vector of locally aggregated descriptors (VLAD) [60], an approximation of FV, uses k-means to learn the dictionary. VLAD uses the 1st-order statistics to represent each cluster Q_m ; the m^{th} cluster representation of an image I_i can be given as:

$$\mathbf{v}_m^i = \sum_{\mathbf{x}_{ij} \in Q_m} \mathbf{x}_{ij} - \mu_m \quad (2.7)$$

The image representation by VLAD is the concatenation of \mathbf{v}_m^i for all $m = 1, \dots, M$, leading to a dimensionality of Md .

2.2.3 Pooling

BOW, FV and VLAD use *sum* pooling to aggregate the local codes to get an image-level representation. In addition to the sum pooling, *max* pooling is also used in SC [24].

Let $\mathbf{z}_i = [z_i^1, \dots, z_i^M]^T$ be the image level representation of an image I_i , where M is the dimensionality of \mathbf{z}_i , e.g. for BOW M represents the size of the dictionary. c_{ij}^k represents the j^{th} encoded feature from image I_i .

Max-pooling can be defined as:

$$\mathbf{z}_i^k = \max |c_{ij}^k|, \quad j = 1, \dots, N_i. \quad (2.8)$$

and sum pooling as:

$$\mathbf{z}_i^k = \sum_{j=1}^{N_i} |c_{ij}^k| \quad j = 1, \dots, N_i. \quad (2.9)$$

where, \mathbf{z}_i^k and \mathbf{c}_{ij}^k are respectively the k^{th} element of \mathbf{z}_i and \mathbf{c}_{ij} .

The sum or max pooling extracts statistical information from all the encoded feature vectors over the entire image, without considering information on the spatial layouts of local features in the image. This may reduce the discriminative power of the representation. To solve this issue, region-based pooling was proposed, which firstly divides an image into fixed spatial pyramid (SPM) regions, and then the pooled image features from each region are concatenated to get the final image representation [178]. Various region-based versions include learning a set of rectangular regions instead of fixed ones [61], applying weights to different regions based on saliency [179], and assigning each local feature to multiple SPM regions with weights [27]. Since the final image representation using these region-based pooling methods is usually obtained via direct concatenation of region-based pooled representations, the final image representation can capture the location information. This location or global structure information is very useful for natural images, for example the sky is always in the upper part of the image. Unlike natural images, colonoscopy images have less or no spatial structures therefore this direct concatenation region-level pooled results may not be appropriate.

2.3 Feature learning approaches

Local image descriptors play an important role in many computer vision systems. Various descriptors, e.g. SIFT [104], LBP [126], BRISK [87] have been proposed for different purposes, including image classification [113, 168], and feature matching [87, 103]. Among them SIFT [104] is the most widely used descriptor capturing a set of local orientation histograms. Since most of these descriptors are hand-crafted, they may not be optimally discriminative for classifying/retrieving images from a particular domain,

e.g. colonoscopy or histology. On the other hand, recent machine learning techniques have explored learning domain-specific descriptors, showing improved performance compared to hand-crafted ones, for example, in medical image segmentation [19], image retrieval [132, 150, 151], and interest point matching [26, 174, 175]. However, these approaches assume that labelled data, e.g. region-based annotations or a training set of labelled image patches, are available to learn the feature descriptors.

The approaches proposed so far for feature learning can be categorised into unsupervised, supervised and weakly-supervised, based on the annotations used for learning. Let \mathbf{x}_{ij} represents the j^{th} local feature extracted from image I_i . The unsupervised approaches do not require labels to learn the feature representations, hence the training set used to learn the features in the unsupervised approaches is in the form of $\{\mathbf{x}_{ij}\}$. Supervised approaches need labels for each individual feature; they require a training dataset in the form of $\{\mathbf{x}_{ij}, y_{ij}\}$, where y_{ij} is the label of \mathbf{x}_{ij} . On the other hand, weakly supervised approaches do not require labels for individual features, but assume that the image level labels are given for the training set, i.e. $\{\{\mathbf{x}_{ij}\}, y_i\}$, where y_i is the label of the image I_i . The following section concisely reviews the existing feature learning approaches under these categories.

Various discriminative feature learning approaches for recognizing faces have been proposed, e.g. [81, 82]. However, face recognition is a different problem as face images are usually aligned with each other. The approaches in [81, 82], which were proposed to learn discriminative features for face recognition, make use of this alignment. Differently from face images, colonoscopy and histology images do not have such a property, hence these approaches cannot be applied, and are excluded from this review.

2.3.1 Unsupervised feature learning

Unsupervised feature learning aims to learn a set of filters which efficiently represent the data using an unlabelled dataset. Various unsupervised feature learning approaches have been proposed, which includes Principal Component Analysis (PCA), Local Linear Embedding (LLE), and unsupervised dictionary learning such as K-means clustering, and SC. These techniques are often applied on raw pixel values as well to features extracted from image patches.

PCA and LLE are dimensionality reduction techniques, projecting high-dimensional inputs into a low-dimensional space while preserving important information. After the projection a supervised classifier is often learned based on the projected feature space to separate different classes. PCA is a linear dimensionality reduction technique which finds a low-dimensional space in which the variance of the data after the projection is maximal. PCA-based approaches have been used for image classification, e.g. in [28, 68, 142]. LLE [144], on the other hand, is a non-linear, neighbourhood-preserving low-dimensional embedding technique, successfully applied to image classification. Applications include digit classification [42] and prediction of Alzheimer's disease from brain MRI data [101].

Unlike PCA and LLE, unsupervised dictionary learning approaches aim to find a set of prototypes which best represent the original data. These prototypes are subsequently used to get an image representation for a given image, based on the local patches (raw pixels or features extracted from patches) from that image.

K-means is one of the widely applied technique to learn these prototypes, by grouping the inputs into a set of clusters. K-means techniques are widely used for classification, e.g. K-means on raw image patches [33] and features extracted from image patches [146].

SC aims to reconstruct the data using an over-complete dictionary (number of dictionary elements $>$ size of the input). SC represents the optimal dictionary that can be used to reconstruct a set of training samples under sparsity constraints on the feature vector. For given data X (e.g. a set of vectorised patches), and the matrix B whose columns are the dictionary elements, feature vectors Z^* is obtained by minimising the following energy function [178],

$$E(X, Z, B) = \|X - BZ\|_2^2 + \lambda \|Z\|_1, \quad (2.10)$$

$$Z^* = \arg \min_Z E(X, Z, B).$$

where λ is a regularisation parameter. SC and its variants (e.g. Locality Constrained Linear Coding [168]) are widely applied for natural [33, 59, 168, 178] as well as medical [120] image classification.

Auto-encoders (AE) [54, 145] also aims to reconstruct the data using an encoder-decoder paradigm. Compare to SC, AE use non-linear activation functions in the encoding stage. The encoder maps an input \mathbf{x} to hidden representation $\mathbf{z} = \sigma(B\mathbf{x})$, where B is a weight matrix, and σ is a non-linear activation function, typically a sigmoid function $\sigma(B\mathbf{x}) = \frac{1}{1+\exp^{-B\mathbf{x}}}$. The decoder maps the hidden representation \mathbf{z} back to a reconstruction $\mathbf{u} = B^T\mathbf{z}$. AE learns the weight matrix B by minimising the reconstruction error of the training set:

$$\arg \min_B \sum_{i=1}^N \|B^T \sigma(B\mathbf{x}_i) - \mathbf{x}_i\|_2^2. \quad (2.11)$$

where N represents the number of training samples. Several variants of AE have also been proposed to improve the original version, e.g. [141].

Inspired by the multi-layer architectures of deep neural networks, various approaches have been proposed which stack simple learning blocks (such as AE), and show improved performance compared to the single-layer approaches [59, 78].

2.3.2 Supervised feature learning

Various approaches in the form of supervised feature learning have been proposed to learn descriptors which are discriminative for tasks such as feature matching [26, 55, 174, 175], image retrieval [132], object [148] and vessel [19] segmentation. These approaches use labelled data for learning in the form of matching and non-matching feature pairs [26, 55, 132, 174, 175], or segment-level labels [19, 148].

These approaches mainly learn a set of descriptor parameters or configurations of the descriptors such that the learned descriptors are discriminative for specific tasks. For example, Winder et al. [174, 175] examine a series of building blocks for descriptor construction which consists blocks such as, pre-smoothing, transformation, spatial pooling, and normalisation. The optimal configuration of each block (e.g. the configuration of the pooling regions in the pooling block) is selected in a way that the resultant descriptor maximizes the area under the ROC curve for descriptor matching. Selecting the optimal pooling region configurations from a set of pre-defined ones is an optimisation problem which cannot be solved analytically. In [26], an approach similar to that of Hua et al. [55] is introduced, which replaces the pooling block by

an embedding block. The embedding block applies a dimensionality reduction on the original local image patch, hence the optimisation can be done analytically.

A similar idea, where feature learning is formulated as selecting and weighting a set of pooling regions (PR) among a large set of candidate ones, has been proposed by Simonyan et al. [150, 151]. The PR configurations are constrained to be symmetric, and are grouped into rings (Figure 2.3). Hence, the PR selection is performed at the ring level. To determine which rings to select, a non-negative weight w_k is assigned to each ring, and these weights are learnt by minimising the distance between the matching feature pairs while maximizing the distance between the non-matching feature pairs. The learned descriptors show improved performance compared to the hand-crafted SIFT features for large-scale image retrieval on two public datasets, Oxford buildings¹ and Paris buildings².

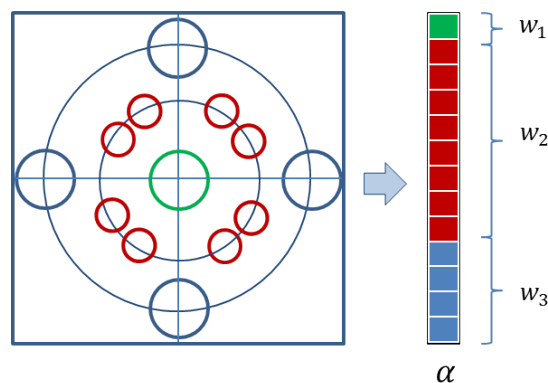


Figure 2.3: Pooling region candidate rings. The green circle shows a ring of single PR, the red and the blue circles show a ring with eight PRs, and a ring of four PRs respectively. Each ring corresponds to the sub-vector in the final descriptor, and the weight w_i is applied to the i^{th} ring, which determines whether that ring has to be selected ($w_i > 0$) or not ($w_i = 0$) (shown on the right).

Shotton et al. [148] propose semantic texton forest features, where the local features are given by the output of a learned decision forest. A decision forest from the given training data has been learned, where the nodes in each tree of that forest apply simple functions on raw pixels which are inside image patches (Figure 2.4). These simple functions include sum, difference, and absolute difference of a pair of pixels. This approach shows state-of-the-art performance for segmentation on the Pascal VOC-2007 dataset³.

¹<http://www.robots.ox.ac.uk/~vgg/data/oxbuildings/>

²<http://www.robots.ox.ac.uk/~vgg/data/parisbuildings/>

³<http://host.robots.ox.ac.uk/pascal/VOC/>

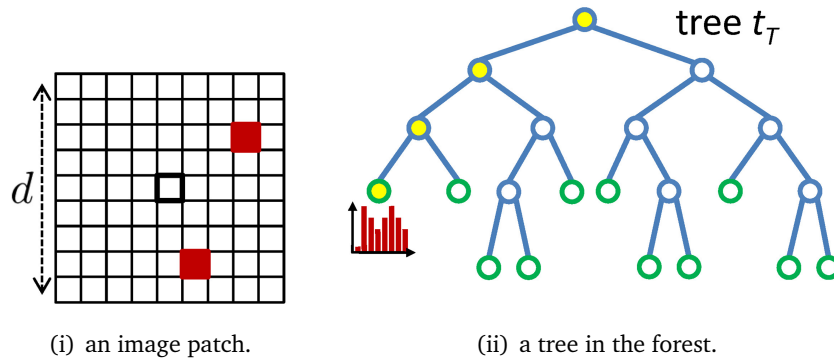


Figure 2.4: Semantic texton forest features. The split nodes in semantic texton forests use simple functions of raw image pixels within a $d \times d$ patch: either the raw value of a single pixel, or the sum, difference, or absolute difference of a pair of pixels (red) (figures extracted from [148]).

A similar approach which transforms regions (instead of image pixels) which are inside the local image patches into a non-linear space has been proposed by Trzcinski et al. [160, 161] for viewpoint and illumination invariant descriptor matching. Here a metric which transforms the original Euclidean space to some other space, where the similar and dissimilar patches can be easily separated is learned together with a set of local non-linear filters. In this approach, image patch appearance is modelled using non-linear filters evaluated within the image patch that are effectively selected with boosting (Figure 2.5).

Let \mathbf{u}_i and \mathbf{v}_i be two image patches, and y_i the label indicating whether they are similar (+1) or dissimilar (-1). The method proposed by Trzcinski et al. [160, 161] learns the descriptor $H(\mathbf{u}_i)$ as a non-linear transformation of the original image patch \mathbf{u}_i . The non-linear transformation $H(\mathbf{u}_i)$ is formed by a collection of non-linear response functions $\{h_k\}_{k=1}^K$. Each function $h_k(\mathbf{u}_i)$ operates on a rectangular region r within the patch \mathbf{u}_i as shown in Figure 2.5. This non-linear mapping is learned by minimising the following exponential loss:

$$L = \sum_{i=1}^N \exp(-y_i f(\mathbf{u}_i, \mathbf{v}_i)), \quad (2.12)$$

Where N is the total number of training patch pairs. The similarity function $f(\mathbf{u}_i, \mathbf{v}_i)$ is given by

$$f(\mathbf{u}, \mathbf{v}) = \mathbf{h}(\mathbf{u})^T A \mathbf{h}(\mathbf{v}) \quad (2.13)$$

$$= \sum_{i,j=1}^K \alpha_{ij} h_i(\mathbf{u}) h_j(\mathbf{v}) \quad (2.14)$$

where $A \in \mathbb{R}^{K \times K}$ is a distance metric defining a space where any two patches can be easily separated into similar (matching) or dissimilar (non-matching). α_{ij} are the entries of A . In [160, 161] the non-linear response functions ($\{h_k\}_{k=1}^K$) are obtained via simple weak-learners (decision stumps) which threshold the region-based pooled responses (Figure 2.5). K represents the considered number of local regions inside the patch (or the number of weak learners). The regions, the thresholds of the weak learners which acts on the selected regions, and the metric A have to be selected/learned using the training set with labelled patches.

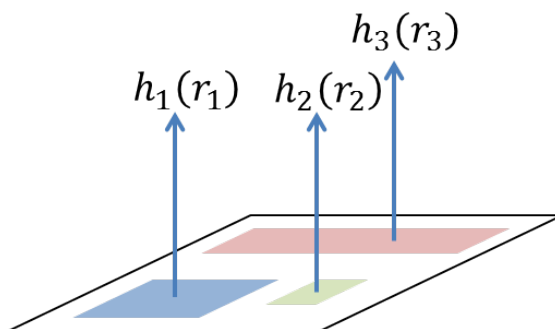


Figure 2.5: Learning image descriptors with boosting. An image patch (\mathbf{u}) is represented by the black rectangle. Blue, red and green rectangles (r_i) show the selected pooling regions inside the image patch. The pooled responses from these regions are mapped to a non-linear space ($h_i(r_i)$). The resultant descriptor $H(\mathbf{u})$ can be given as $[h_1(\mathbf{u}), \dots, h_K(\mathbf{u})]^T$.

Supervised feature learning approaches have also been explored for medical image analysis. An approach similar to [160, 161] has been proposed by Becker et al. [19] to segment vessels in the retina and confocal microscopy images. In [160, 161] the pooling function to aggregate the information from a region r is pre-defined. Therefore, Becker et al. [19] learn these functions automatically in addition to the boosting classifier which classify pixels into vessel or non-vessel using a dataset with pixel-level labels. Since learning both may lead to over-fitting, a regularisation term which encourages the learned adjacent kernels to be similar to each other is added to the objective function.

This joint learning shows improved performance compared to existing approaches for vessel segmentation in the retina and confocal microscopy images. However, it is unclear whether learning these functions instead of fixing them as in [160, 161] improves the classification as no comparisons are given.

The approaches proposed in [160, 161] and [19] use boosting to select the regions (support of the kernels) and the weak-learners (kernels) which acts on these regions. However, boosting is a greedy approach: once a kernel or its support is selected in a particular boosting iteration, its values cannot be changed in subsequent iterations. These may need to learn an increased number of kernels, compared to an approach where the kernels can be updated during the learning process.

2.3.3 Weakly supervised feature learning

Unlike the approaches reviewed in Section 2.3.2, weakly-supervised approaches can learn discriminative local features using image-level labels. Therefore, they avoid the need for expensive annotations in the form of patch-level or region-level labels, which are needed for supervised feature learning. They are therefore, in principle, more applicable to the medical imaging domain, where obtaining such labels is difficult due to the limited time available from specialists.

Few approaches try to automatically identify similar or dissimilar patches as a pre-processing step for feature learning for descriptor matching. For example Simonyan et al. [151] detect correspondences between images which contain a common image part, and use the identified patches to learn the features in a supervised manner. A similar approach has been proposed by Philbin et al. [132] to generate the training data for learning a discriminative projection for natural image retrieval.

Shotton et al. [148] assign the label of an image as the label for a patch which is extracted from that image to learn the texton forest (Section 2.3.2) to represent the local features. The weakly learned texton forest features in this manner give worse performance compared to supervised learning approach where each patch is individually labelled as belonging to a particular object or not.

Convolutional neural nets (CNN) [79] have also been used to learn local features in a supervised or weakly-supervised way. In CNN, a set of convolutional filters and

a classifier are learned in a unified framework. CNN requires a careful design, and it is computationally expensive to train, even on the GPU [19]. Usually CNN requires a very large amount of training data [122]; when this is not available, CNN gives worse performance than traditional, hand-crafted features and BOW-based feature encoding methods [122].

Differently from the above approaches, recently, Zuo et al. [192] proposed an approach which learns a set of filters transforming the local image patches into features. These filters are learned based on three objectives: (1) the learned features should preserve relevant information in the original data; (2) the learned filters should be shareable across different categories; and (3) the learned filters should be discriminative for different categories. This approach shows state-of-the-art performance for scene image classification on different datasets such as Scene-15 ⁴, UIUC Sports ⁵, and MIT Indoor ⁶.

2.4 Conclusions and discussion

In this chapter I concisely reviewed the work related to endoscopy image analysis, feature encoding approaches, and feature learning approaches. It should be noted that feature learning approaches (supervised/weakly-supervised) reported so far have not yet been explored for endoscopy image analysis. The approaches proposed for endoscopy image analysis mainly use a combination of hand-crafted features, which capture texture, colour and/or shape information. Below I discuss the main limitations associated with the existing approaches for endoscopy image classification.

- **Limits of concatenated colour representations.** Colour is a salient feature for colonoscopy. Some works, e.g. [16, 158], use histograms of intensity values to capture colour properties, but such global representations fail to capture local image properties well. In [70], the image is divided into blocks, and the histograms obtained from the blocks are concatenated to get an image representation, which are then concatenated into a feature vector. Such concatenated histogram representations have drawbacks when used with

⁴http://www-cvr.ai.uiuc.edu/ponce_grp/data/

⁵http://vision.stanford.edu/lijiali/event_dataset/

⁶<http://web.mit.edu/torralba/www/indoor.html/>

colonoscopy images. First, they encode spatial/location information of the local features. Spatial information is useful in natural images; for instance, the sky appears always in the upper part of an image. In colonoscopy images such information is absent, limiting the use of concatenated representations. Second, such representations increase the length of the feature vector, and therefore increase the computational and memory cost of a classifier. Instead, colour features could be extracted locally, from image patches, to capture the local image properties, and can be encoded using any feature encoding approach to compute the image representation. This representation will avoid the limitations raised by the global colour histograms, as well as the concatenated colour representations. Therefore, in Chapter 8 I propose to use local colour histograms together with feature encoding approaches for colonoscopy image classification, and show considerable improvements in MCA over the global and concatenated colour histogram representations.

- **Recent feature encoding approaches not yet investigated for colonoscopy.** BOW and SC approaches have been explored for colonoscopy image classification (e.g. [183, 185]). However, in computer vision various feature encoding approaches such as VLAD and FV (Section 2.2) have been proposed, and show improved performance over BOW and SC ([131]) as they capture additional information of the images compared to BOW and SC. These approaches have not yet been explored for colonoscopy image analysis. Therefore, in Chapter 8 I explore these approaches for colonoscopy image classification.
- **Limitations with the feature encoding approaches.** The traditional feature encoding approaches capture *intra-cluster* features, for example, BOW capture information about how many local features fall into a particular cluster. BOW or other feature encoding methods such as SC, FV and VLAD do not capture how the local features co-occur in the local image regions. This information may be important for medical images, for example, a particular type of cell may densely appear with another type of cell in cancer regions compared to healthy regions in histology images. This local co-occurrence can be captured using *inter-cluster* features. To overcome this, in Chapter 9, I propose an approach to capture inter-cluster features, in addition to the intra-cluster features which are often

captured by the traditional feature encoding approaches. My approach shows improved performance over the traditional feature encoding approaches.

- **Supervised/weakly-supervised feature learning approaches not yet investigated for colonoscopy.** The features explored in the current colonoscopy image classification literature are hand-crafted, hence are not tuned to the specific characteristics of the problem domain, possibly limiting their discriminative power. On the other hand, feature learning approaches (Section 2.3) are becoming popular in computer vision as they automatically learn features which capture discriminative domain-specific properties. These approaches have not yet been explored for colonoscopy image analysis. Hence, I propose to learn the local features based on the given training data and show improved performance over the hand-crafted features in Chapters 5, 6 and 7.

DATASETS AND EXPERIMENTAL SETTINGS

3

In this thesis I used four datasets, 2-class colonoscopy, 3-class colonoscopy, ICPR cell images, and IRMA radiology images. This section explains how these datasets were generated, the experimental settings, and in which chapters these datasets were used.

3.1 Datasets

3.1.1 2 class colonoscopy images dataset

I collected white light, air colonoscopy images and 82 white light, air colonoscopy video clips (each one is less than a minute long) from various sources from the Internet (mainly from [6]). These videos are in various resolutions and illumination conditions. Normal videos represent the videos taken from healthy colons and the abnormal videos represent the videos taken from unhealthy colons. Abnormal videos contain frames which show different lesions including polyps, cancers, Crohn's disease, ulcerative colitis, etc.

Images were extracted from videos in 30 frames/sec rate. The unusable (e.g. very blurred) images were identified and removed manually. Since the resulting dataset contains a large set of redundant images, I applied a clustering approach to get a subset of representative frames from each video. I extracted a set of colour and texture features to represent each image. Statistics such as mean, standard deviation, skewness, kurtosis and entropy in each colour channel in the RGB space were considered as the

colour features. Local Binary Pattern (LBP) histograms were considered as the texture feature. LBP histograms computed from 3 different scales and from each colour channel of RGB colour space were concatenated to get the image representation (detail about LBP can be found in Section 4.2 of Chapter 4). At the first, second and third scales LBP features were extracted around each pixel from a neighbourhood of radius 1, 2 and 3 respectively. At each scale eight sampling points were considered. Uniform LBP patterns were considered as they capture frequently occurring local image patterns such as edges, spots, etc. The dimensionality of the resulting image representation was 546 (3×5 for colour, and 3×59 for texture). After feature extraction, I used k-means to get a set of representative clusters for the images obtained from each video. It is observed that the movement of the colonoscope is fast in normal videos compared to the abnormal ones as the corresponding colon segments do not need a careful inspection of the colonic walls. Therefore the number of clusters were experimentally set to $\frac{V_i}{7}$ for normal and $\frac{V_i}{10}$ for abnormal videos, where V_i is the total number of frames extracted from video i . After clustering, one image per cluster is randomly selected and added to the final dataset. Each image in the resulting dataset was annotated into normal (healthy) or abnormal (containing any lesion, including polyp, ulcer, bleeding, cancer) by a clinician (thanks to Dr. Adrian Hood, surgical research fellow, Leeds Institute of Molecular Medicine, University of Leeds, UK), who was blind to the video labels. In total I was able to obtain 1050 normal and 1050 abnormal images. Some example images from the dataset are shown in Figures 3.1 and 3.2. Since the videos are very short (< 1 min), I reported experiments based on random splits (see experimental settings in Section 3.2). These experiments show the effectiveness of the proposed features compared to the baseline features. However, future work will focus on training and testing images from different videos (see Sections 10.3 and 10.4 for the limitations of this dataset and for the future work).

3.1.2 3 class colonoscopy images dataset

This dataset is an extension of the 2-class colonoscopy dataset. In this dataset very unclear (or uninformative) images from the normal and the abnormal classes were removed manually. In this dataset, the uninformative frames which were manually identified when generating the 2-class dataset were added into a new class called the "uninformative". In these uninformative images the mucosa is largely invisible due

to blur, overexposure, smoke, etc. After removing the unclear images (where the colonic mucosa is hardly visible) from the normal and abnormal classes, the new 3-class colonoscopy dataset contains 1000 abnormal, 900 normal, and 900 uninformative images. Some images from the uninformative class are shown in Figure 3.3.

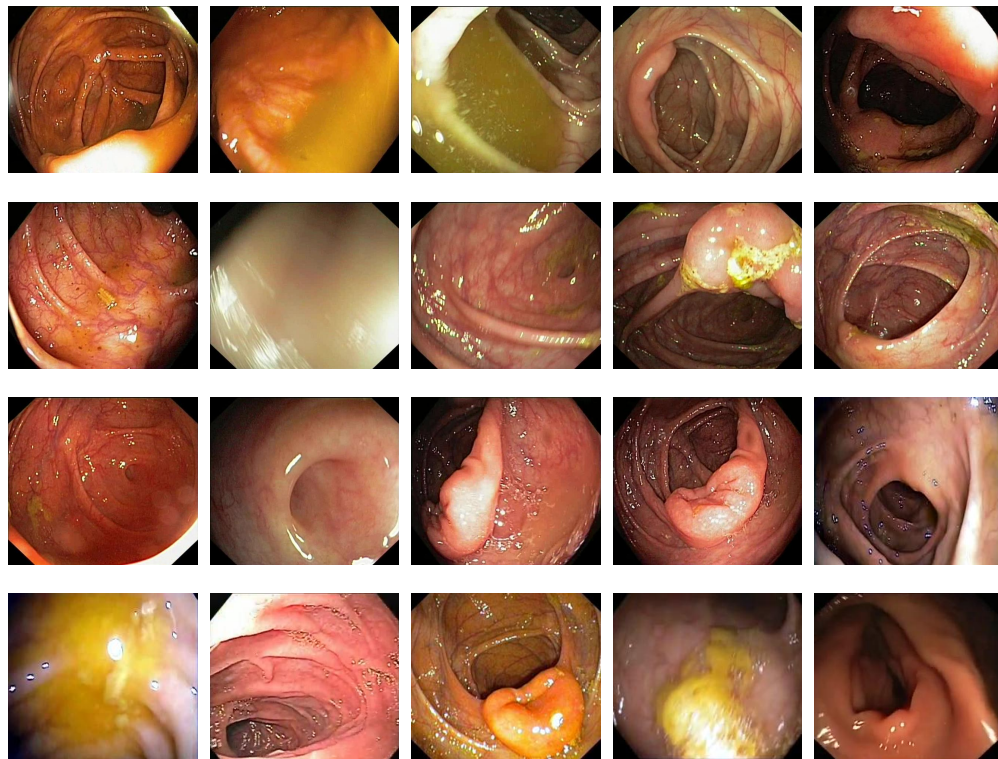


Figure 3.1: Examples of normal images from the 2-class colonoscopy dataset.

3.1.3 ICPR 2014 cell images dataset

This dataset is a part of the contest "*Performance Evaluation of Indirect Immunofluorescence Image Analysis Systems*"¹ organised by ICPR 2014. This competition consists of two tasks, Task 1 - classifying individual cell images, and Task 2 - classifying specimens which contains groups of cells. In this thesis I use only the Task 1 dataset.

The images in this dataset (Task 1) were collected between 2011 and 2013 at the Sullivan Nicolaides pathology laboratory, Australia. For this task, a set of training images was provided to the contest participants. Submitted systems were then evaluated on a separate hidden test set which was privately maintained by the contest organisers and not released to the participants.

¹<http://i3a2014.unisa.it/>

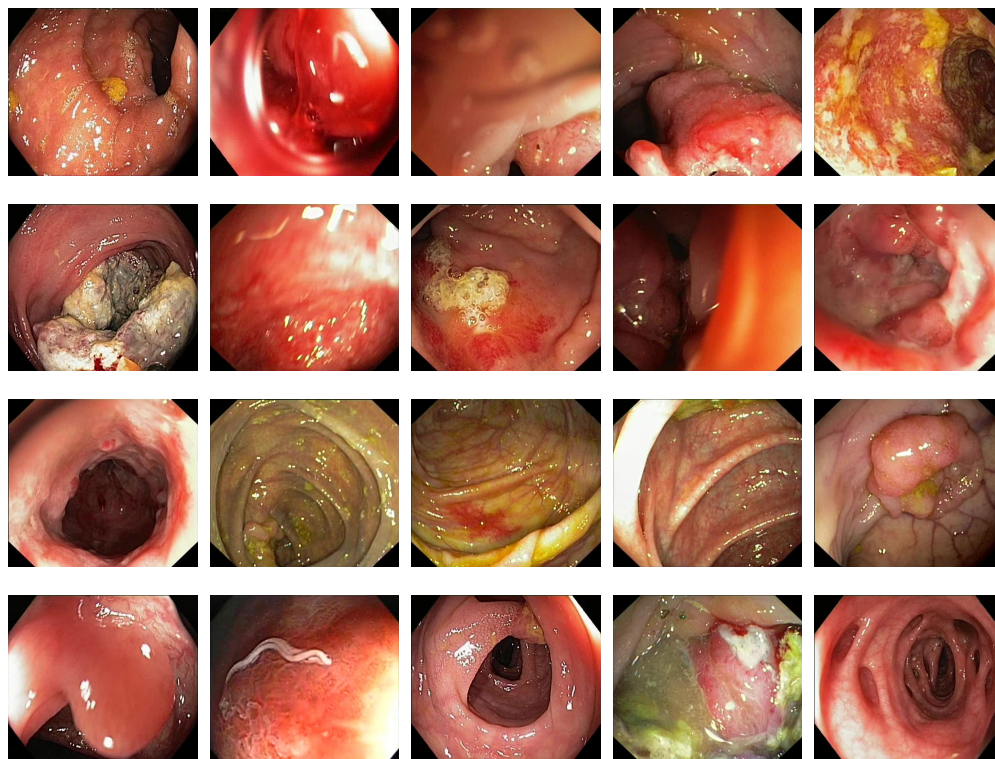


Figure 3.2: Examples of abnormal images from the 2-class colonoscopy dataset.

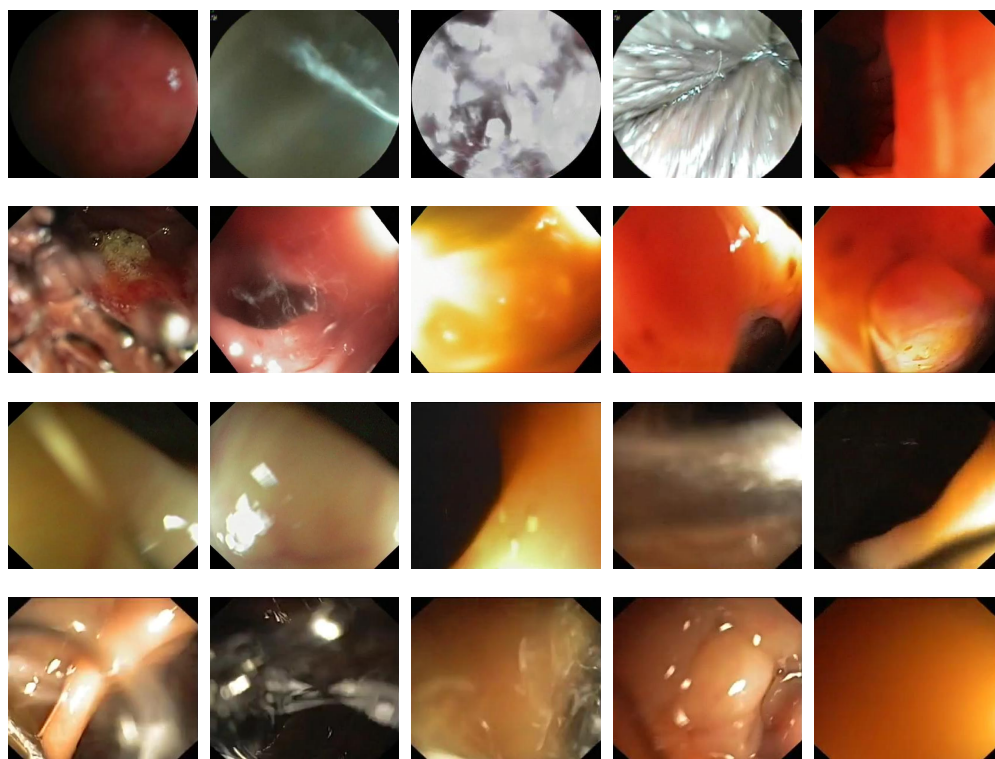


Figure 3.3: Examples of uninformative images from the 3-class colonoscopy dataset.

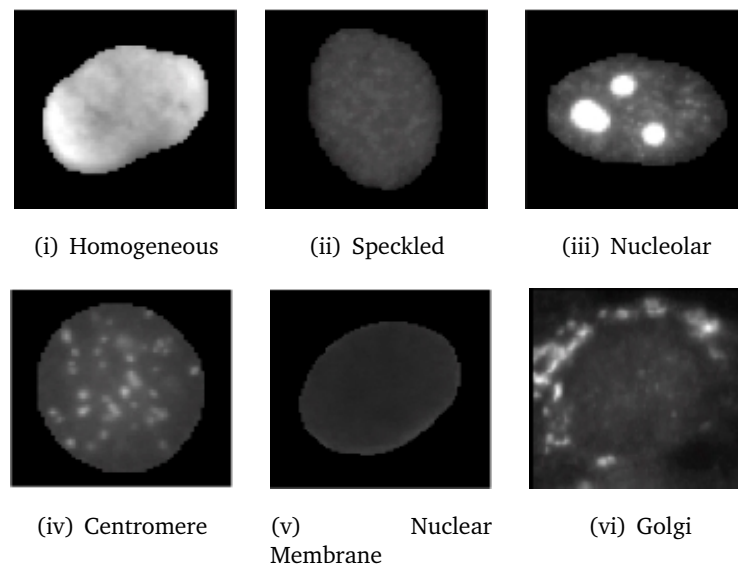


Figure 3.4: Sample images from the ICPR 2014 Task 1 dataset (individual cell classification)

The Task 1 dataset consists of 68,429 images of individual cells extracted from 419 patient positive sera (approximately 100 – 200 cell images per patient serum) along with their binary segmentation masks. 13,596 images were available during training. The remaining 54,833 images were used for the hidden test set to evaluate performance of systems submitted to the contest. The specimens were automatically photographed using a monochrome high dynamic range cooled microscopy camera. Cell images are approximately 70×70 pixels in size. The dataset has six pattern classes: homogeneous, speckled, nucleolar, centromere, nuclear membrane, golgi. An example image from each of the six classes is given in Figure 3.4.

3.1.4 IRMA dataset

The Image Retrieval in Medical Applications dataset² (IRMA) contains 15,363 anonymous radiographs from 57 classes of various human body parts. Since the number of images is very unbalanced across the classes, only 20 classes which contain at least 200 images were randomly selected. The images in each of the selected class were randomly sampled such that each class will contain 200 images. Some examples of images from different classes of this dataset is given in Figure 3.5.

²courtesy of TM Deserno, Dept. of Medical Informatics, RWTH Aachen, Germany. http://ganymed.imib.rwth-aachen.de/irma/index_en.php

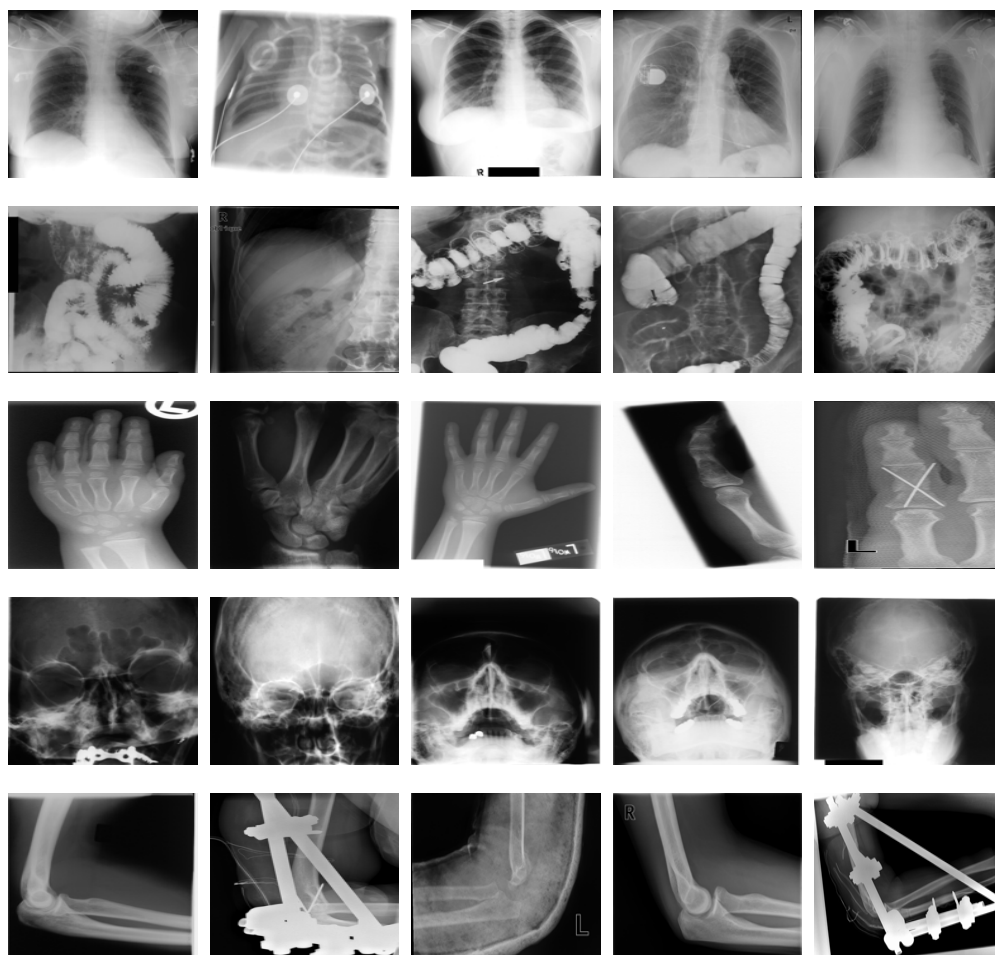


Figure 3.5: Examples of images from the IRMA dataset.

Dataset	Class name	No of images	Chapters
2-class colonoscopy	Abnormal	1,050	Chapter 4, 5, 6
	Normal	1,050	
3-class colonoscopy	Abnormal	1,000	Chapter 4, 5, 6
	Normal	900	
	Uninformative	900	
ICPR cells	Homogeneous	2,494	Chapter 4, 5, 6, 9
	Speckled	2,831	
	Nucleolar	2,598	
	Centromere	2,741	
	Nuclear Membrane	2,208	
	Golgi	724	
IRMA	20 classes	200/class	Chapter 9

Table 3.1: Detail of the classes in different datasets, and the chapters in which these datasets were used.

3.2 Experimental settings

To facilitate the comparisons (between different features), I follow the same experimental settings throughout this thesis. This section explains the settings for different datasets in detail.

3.2.1 Image preprocessing

For the 2-class colonoscopy, 3-class colonoscopy and the IRMA radiographs datasets the images were rescaled by keeping their row to column aspect ratio unchanged, such that the maximum dimension (row or column) of the images was 300 pixels.

The masks of the cells provided with the cell images of the ICPR dataset were not used in Chapters 4, 5, 6, 7 and 9. In these chapters, prior to the feature extraction, the cell images were intensity normalised such that 2% of pixels in each cell became saturated at low and high intensities. Cell masks were used in Chapter 8 when comparing the proposed method with the state-of-the-art methods.

3.2.2 Evaluation metric and experimental setup

In this thesis Mean Class Accuracy (MCA) was used as the evaluation metric. It is defined as,

$$\text{MCA} = \frac{1}{K} \sum_{k=1}^K \text{CCR}_k \quad (3.1)$$

where CCR_k is the correct classification rate for class k and K is the number of classes.

I randomly sample 300 and 200 images respectively from each class of the 2-class and the 3-class colonoscopy datasets for training, and I use the rest of the images for testing. This process is repeated 10 times. The mean and the standard deviation of the MCAs obtained from these 10 experimental runs are reported.

For the ICPR cells dataset I report the average mean and the standard deviation of the MCA obtained from two-fold cross-validation experiments, which were repeated 5 times.

For the IRMA dataset 30 images per class are selected for training and the rest are used for testing; the averaged MCA over 10 iterations are reported.

I experimented different approaches for normalising the image-level feature representations prior to classification. This approaches include (1) un-normalised representation, (2) L1-normalisation, (3) L2-normalisation, and (3) L2 and power normalisations [130]. I experimentally found that, overall, the L2 and power normalisation performs better than other approaches regardless of the feature and encoding types. Therefore, in all the reported experiments the final image-level feature representations are normalised by the L2 and power normalisation. Let $\mathbf{z}_i \in \mathbb{R}^d$ represents the image-level representation of an image I_i , where d is the size of the representation, the *L2-and-power* normalisations can be given as.

$$\mathbf{z}_i \leftarrow \frac{\text{sign}(\mathbf{z}_i) |\mathbf{z}_i|^{\frac{1}{2}}}{\|\mathbf{z}_i\|_2} \quad (3.2)$$

where $|\mathbf{z}_i|^{\frac{1}{2}}$ applies the square root to each component of \mathbf{z}_i .

THE GENERALISED LOCAL TERNARY PATTERNS

*Local Binary Patterns (LBP) are widely used for texture classification. Several variants of LBP have been proposed, e.g. Local Ternary Patterns (LTP) to make LBP resilient to noise, Scale Invariant LTP (SILTP) to make LBP resilient to illumination changes. But neither LTP nor SILTP are resilient to **both** noise **and** illumination changes. This chapter proposes a generalised variant of LBP called the Generalised Local Ternary Patterns (gLTP) which captures edge and blob-like features and makes the LBP resilient to both noise and illumination changes. Experiments on two datasets (Normal/Abnormal colonoscopy and ICPR cell images) show that neither LTP nor SILTP gives better performance on either dataset. On the other hand, the proposed gLTP descriptor gives competitive performance compared to the best performing descriptors in the datasets, confirming that gLTP is resilient to noise and illumination.*

4.1 Introduction

LOCAL Binary Patterns, proposed by Ojala et al. [126], have proved a very powerful texture descriptor which has been widely applied for e.g. texture classification [126], face recognition [56], medical image classification [113, 190]. LBP describes the local texture around each pixel by comparing and thresholding pixel differences in a local image neighbourhood. A global image representation of an image is normally obtained by computing statistical representations (e.g. histogram) of the LBP-based pixel representations. Several variations of LBP have been proposed, e.g. Local Ternary

Patterns (LTP) [156] makes LBP resilient to noise, Scale Invariant LTP (SILTP) [94] makes LBP resilient to illumination changes, uniform LBP[127] captures informative patterns such as edges, bright and dark spots, and reduces the dimensionality of the histogram image representation, BlockLBP[96] captures information from larger local neighbourhood (e.g. larger than 3x3 which is often used by LBP).

In the following sections first LBP and its major variants will be concisely reviewed, then the proposed gLTP descriptor described in detail, and finally several experiments comparing gLTP with baseline representations reported.

4.2 LBP and its variants

4.2.1 The standard LBP descriptor

Consider N sampling points distributed uniformly on a circle of radius R around a 2D point \mathbf{p}_c in a gray image I (Figure 4.1). LBP can be defined as:

$$LBP_{N,R}(\mathbf{p}_c) = \sum_{n=1}^N q_n \times 2^{n-1} \quad \text{where } q_n = \begin{cases} 1 & I_n \geq I_c \\ 0 & I_n < I_c. \end{cases} \quad (4.1)$$

I_c and I_n represent the intensity values at the centre point (\mathbf{p}_c) and at the n -th sampled image point, respectively. I_n is bilinearly interpolated when the sampling point does not coincide with a pixel. Since this operator gives 2^N different labels, an image can be represented as a histogram with 2^N bins.

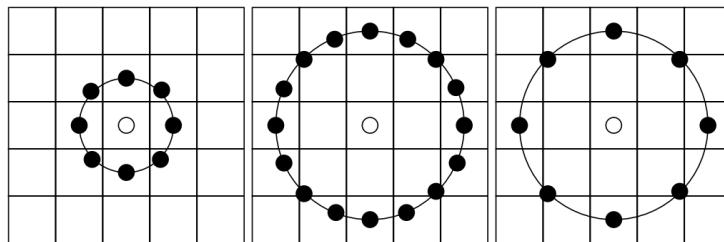


Figure 4.1: The circular (8, 1), (16, 2) and (8, 2) neighbourhoods. The pixel values are bilinearly interpolated whenever the sampling point is not in the centre of a pixel.

The generation of the standard LBP codes from a 8 neighbourhood is illustrated in Figure 4.2. Figure 4.3 shows an example image, its corresponding LBP representation

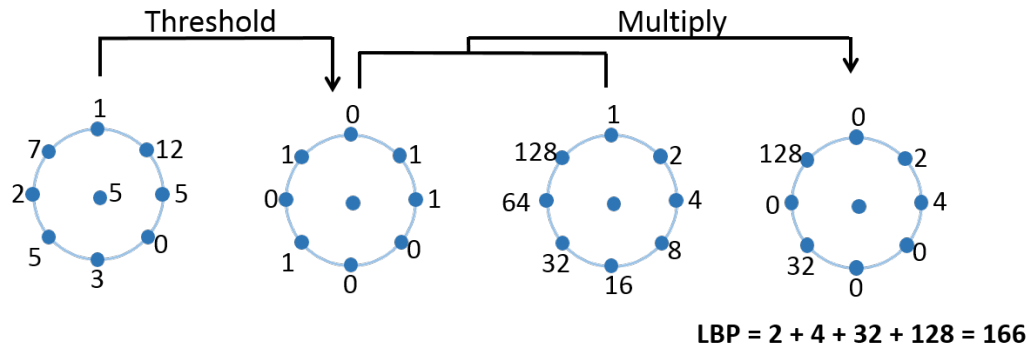


Figure 4.2: Example illustrating the derivation of the standard LBP codes. The pixels in this block are thresholded by its centre pixel value, multiplied by powers of two and then summed to obtain a value for the centre pixel.

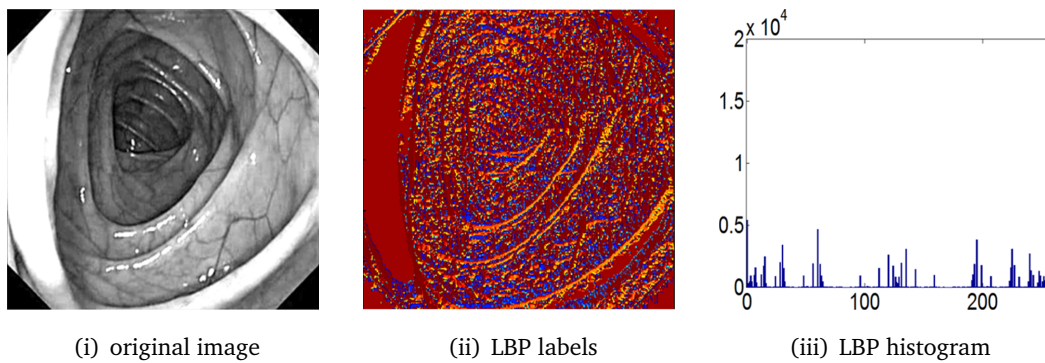


Figure 4.3: Example of input image, its corresponding LBP labels ($R=1$, $N=8$) and LBP histogram. (The higher values (red) in the LBP image correspond to LBP labels with value 255 and the lower values (blue) correspond to LBP labels with value 0.)

and the LBP histogram. The LBP image (Figure 4.3(ii)) was obtained by transforming each 3×3 image patch (Figure 4.3(i)) by its LBP representation (Equation 4.1).

4.2.2 Uniform and rotation invariant LBP

As some of the binary patterns occur more commonly in texture images than others, an extension of LBP, called the *uniform LBP* [127], has been proposed. The uniform patterns describe frequently occurring basic features such as bright spots, dark spots and edges. A LBP is called uniform if the binary pattern (Equation 4.1) contains at most two bitwise transitions from 0 to 1 or vice versa when the bit pattern is traversed circularly. For example, the patterns 00000000 (0 transitions), 01110000 (2 transitions) and 11001111 (2 transitions) are uniform whereas the patterns 11001001 (4 transitions) and 01010010 (6 transitions) are not. Figure 4.4 shows some examples of uniform and non-uniform patterns. The resultant LBP representation has a separate label for

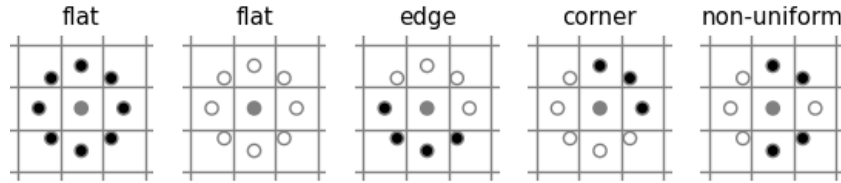


Figure 4.4: Some example uniform (first four images) and non-uniform (last image) patterns.

each uniform pattern and all the non-uniform patterns are assigned to a single label. For example when $N = 8$ the number of patterns produced by the uniform LBP is 59 compared to the total number of 256 patterns produced by the standard LBP ([133]).

Uniform patterns have several advantages: (1) they capture more commonly occurring local structures, (2) considering the uniform patterns makes the number of possible LBP labels significantly lower, hence require fewer samples to estimate their distribution reliably and (3) the dimensionality of the final image representation is reduced, hence reducing the classification complexity. It has been observed that considering only the uniform patterns instead of all the possible patterns produces better recognition results for many applications [133].

Rotation-invariant LBP makes the LBP descriptors invariant to **local image region** rotations by rotating the LBP binary codes in a circular bit-wise manner so that the resultant LBP label will have the minimum value [127].

4.2.3 Local Ternary Patterns

To make LBP robust to noise, a three-level thresholding has been applied in *Local Ternary Patterns (LTP)* [156] by the introduction of a user-specified threshold τ (Equation (4.2)). The LTP histogram representation of an image is obtained by splitting each LTP into two LBP and then concatenating the two LBP-based representations.

$$q_n(\tau) = \begin{cases} 1 & I_n - I_c \geq \tau \\ 0 & |I_n - I_c| < \tau \\ -1 & I_n - I_c \leq -\tau. \end{cases} \quad (4.2)$$

An ideal LBP should be robust to illumination changes and (at least) Gaussian noise. While the introduction of τ makes LTP robust to noise, LTP may be sensitive to changes in illumination (Figure 4.5(iii)).

4.2.4 Scale Invariant Local Ternary Patterns

To counteract illumination variations, a variant of LTP called the *Scale¹ Invariant Local Ternary Pattern* (SILTP) has been proposed in [94], i.e.

$$SILTP_{N,R}(\mathbf{p}_c, a) = \oplus_{n=1}^N q_n(a) \quad (4.3)$$

$$\text{where } q_n(a) = \begin{cases} 01 & I_n > (1+a)I_c \\ 10 & I_n < (1-a)I_c \\ 00 & \text{otherwise,} \end{cases}$$

where a is a scale factor and \oplus denotes concatenation of the 2-bit binary strings q_n . Note that SILTP is not designed particularly for image classification, but the ‘2-bit’ codes can be converted to ‘ternary’ patterns to generate a histogram representation for an image. Since SILTP is designed to cope with the changes in illumination it may be sensitive to noise (Figure 4.5).

4.2.5 Other variants

Inspired by the success of LBP in various computer vision applications, different variants of LBP have been proposed to increase robustness and discriminative power. Since LBP uses zero as the threshold to compare a pixel with its neighbourhood, several alternative thresholding techniques have been proposed, e.g. median and mean of the local neighbourhood is used as the threshold in [49] and [62] respectively. Usually LBP operates on a small image neighbourhood (3×3). To capture larger image neighbourhoods Gaussian filtering is applied to collect intensity information from an area larger than the original single pixel in [123], the averaged pixel values in small image blocks were used in [95]. I described only the major relevant variants. A complete review on LBP variants can be found in [133].

¹The term “scale” here means gray scale pixel value, not spatial scale.

4.3 Generalised Local Ternary Patterns

LTP and SILTP make the LBP representations resilient to noise and illumination changes respectively. But neither LTP nor SILTP are resilient to **both** noise **and** illumination changes. Therefore I propose a generalised variant of LBP called the Generalised Local Ternary Patterns (gLTP) which makes the LBP resilient to **both** noise **and** illumination changes.

4.3.1 Definition

When a scene is illuminated by a single distant light source, the observed luminance image $I(x, y)$ at point (x, y) can be approximated as the product of the reflectance image $R(x, y)$ and the illuminance image $S(x, y)$ [69], i.e.

$$I(x, y) = S(x, y)R(x, y) + G(x, y). \quad (4.4)$$

Consider two pixels $I(x_1, y_1)$ and $I(x_2, y_2)$ in image I , and the difference $D = I(x_1, y_1) - I(x_2, y_2)$. Under a different illumination, using Equation (4.4) D becomes:

$$\begin{aligned} D &= [a_1 I(x_1, y_1) + \tau_1] - [a_2 I(x_2, y_2) + \tau_2] \\ &\propto I(x_1, y_1) - a I(x_2, y_2) - \tau. \end{aligned} \quad (4.5)$$

where a_1 and a_2 represent the non-uniform illumination applied to the pixels $I(x_1, y_1)$ and $I(x_2, y_2)$, τ_1 and τ_2 are the sensor noise due to the image capturing device at those pixels, and $a = \frac{a_2}{a_1}$, $\tau = \frac{\tau_2 - \tau_1}{a_1}$.

From Equation (4.5) we can observe that LBP is not resilient to noise (Figure 4.5(ii)), as it assumes $\tau = 0$. LTP is not robust to illumination changes (Figure 4.5(iii)) as it considers $a = 1$. SILTP is robust to illumination changes but it assumes the noise dependent on pixel values (SILTP can be rewritten as, e.g. $q_n(a) = 01$ when $I_n - aI_c > \tau$, where $\tau = I_c$). To make the LBP robust to noise *and* illumination changes, my formulation considers $a \in \mathbb{R}$ and $\tau \geq 0$ (Gaussian noise). The proposed formulation,

generalised Local Ternary Patterns (gLTP), becomes:

$$q_n(a, \tau) = \begin{cases} 1 & I_n - aI_c \geq \tau \\ 0 & |I_n - aI_c| < \tau \\ -1 & I_n - aI_c \leq -\tau. \end{cases} \quad (4.6)$$

The standard LBP ($a = 1, \tau = 0$), LTP ($a = 1, \tau > 0$) and the SILTP ($a \in \mathbb{R}, \tau = I_c$) can be seen as special cases of gLTP. The proposed formulation outputs ternary patterns; I convert each ternary pattern into two binary patterns as in the standard LTP.

parameters	LBP	LTP	SILTP	gLTP
a	1	1	$\in \mathbb{R}$	$\in \mathbb{R}$
τ	0	$\in \mathbb{R}$	I_c	$\in \mathbb{R}$

Table 4.1: The gLTP with different parameter settings (Equation (4.6)). LBP is not resilient to noise, as it assumes $\tau = 0$. LTP is not robust to illumination changes as it considers $a = 1$. SILTP is robust to illumination changes but it assumes the noise dependent on the value of the centre pixel. gLTP is a generalisation of LBP, LTP and SILTP and robust to *both* noise *and* illumination changes.

Figure 4.5 shows an image patch, its three transformed versions (changed illumination, noise and both) and its LBP, LTP, SILTP and gLTP codes. This figure shows that LTP is resilient to noise but not to illumination changes and SILTP is resilient to illumination changes but not to noise. Changing illumination and noise (last row, Figure 4.5), gLTP yields the most stable output compared to the other methods.

4.3.2 Effect of parameters

Here I show qualitatively that the gLTP can capture edge and blob-like features in an image by appropriate parameter setting. The first row of Figure 4.6 shows a colonoscopy image from a standard in-air procedure and its LBP, LTP and gLTP codes. The remaining rows show the original image under different illumination and noisy conditions, and the codes computed by LBP, LTP, SILTP and gLTP. To mimic the spot illumination which is often used in colonoscopy procedure, first an un-normalised Gaussian filter was created with a window size which is equal to $3w$ and a standard deviation equals to $\frac{2w}{3}$, where w is the width of the image. Then I randomly selected a point in the image and placed this Gaussian filter. The pixel values of the image are then multiplied by this filter, and the resultant values are then clipped at 255 to make sure they are in $[0, 255]$.

S = 1 G = 0	63 64 68	0 1 1	0 0 0	00 00 00	-1 -1 0
	80 64 70	1 1	1 1	01 00	1 0
	57 100 60	0 1 0	-1 1 0	10 01 00	-1 1 -1
S = 1 G ≠ 0	65 62 68	1 0 1	0 0 0	00 00 00	-1 -1 0
	80 64 72	1 1	1 1	01 01	1 0
	59 100 60	0 1 0	-1 1 0	00 01 00	-1 1 -1
S = 2 G = 0	126 128 136	0 1 1	0 0 1	00 00 00	-1 -1 0
	160 128 140	1 1	1 1	01 00	1 0
	114 200 120	0 1 0	-1 1 -1	10 01 00	-1 1 -1
S = 2 G ≠ 0	128 126 136	1 0 1	0 0 1	00 00 00	-1 -1 0
	160 128 142	1 1	1 1	01 01	1 0
	116 200 120	0 1 0	-1 1 -1	00 01 00	-1 1 -1
	(i) Image patches	(ii) LBP	(iii) LTP	(iv) SILTP	(v) gLTP

Figure 4.5: A demonstrative example for the effect of noise and illumination changes on (b) LBP, (c) LTP ($\tau = 5$), (d) SILTP ($w = 0.1$), and (e) the proposed gLTP ($w = 0.9, \tau = 5$). (a) First row: original image patch with 3×3 pixels (i.e. $s = 1$ and $G = 0$ in Equation 4.4); Second row: noise added to the original image patch ($s = 1$ and $G \neq 0$) shown in red; Third row: the original image patch under a different illumination ($s = 2$ and $G = 0$); Fourth row: the original image patch under a different illumination with noise added ($s = 2$ and $G \neq 0$) (noisy pixels are shown in red);. LTP is robust to noise but not to illumination changes. LBP and SILTP are robust to illumination but not to noise. LBP, LTP and SILTP are sensitive to both illumination and noise (last row). The proposed gLTP is robust to both noise and illumination.

It is clear that LBP, LTP and SILTP (second, third and fourth columns) gives different output codes under different conditions (noise, illumination or both). On the other hand the codes generated by gLTP capture edge-like features and are less affected by illumination and noise transformations. Figure 4.7 shows an example histological image and the resulting LBP, LTP, SILTP and gLTP codes. Since the original cell image itself is very noisy, LBP and SILTP gives very different outputs under different conditions. gLTP captures blob-like features and is less affected by the illumination and/or noise transformations.

Figures 4.6 and 4.7 show the uniform patterns generated by LBP-based representations computed from $N = 8$ and $R = 2$ (Section 4.2.1). Only one binary code representation is shown for LTP, SILTP and gLTP. The blue and the red colours represent the uniform LBP labels with values 0 and 59 respectively.

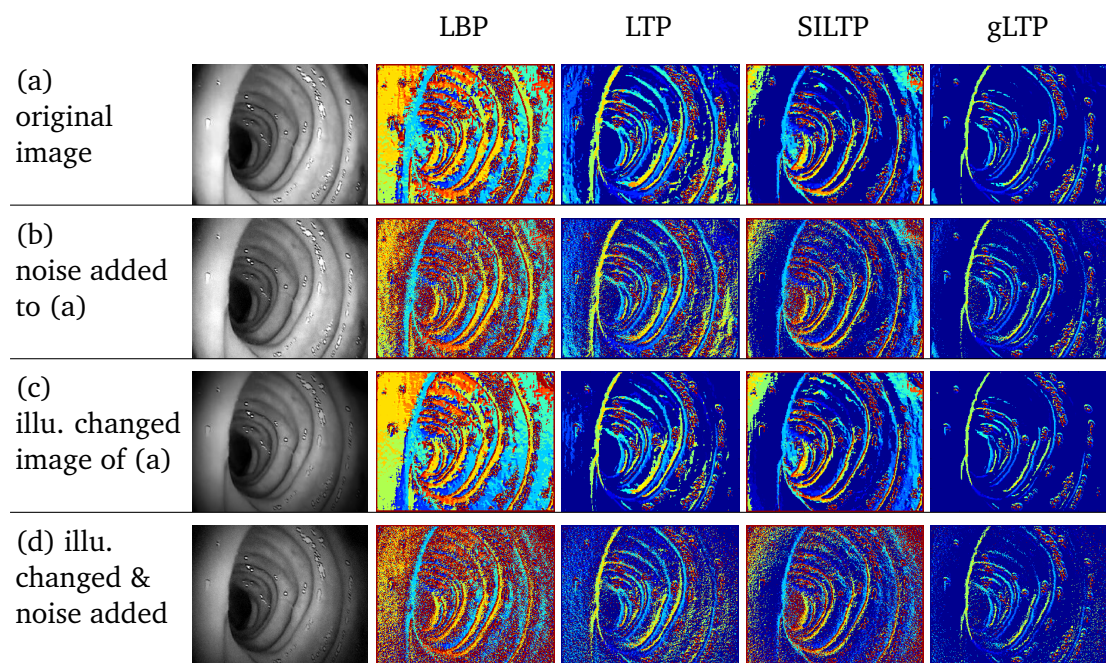


Figure 4.6: The effect of noise and illumination changes on LBP, LTP ($\tau = 10$), SILTP ($w = 0.1$), and the proposed gLTP ($w = 0.9, \tau = 10$) for an example colonoscopy image. The colonoscopy image under different transformations and the corresponding LBP based labels. See text (Section 4.3.2) for more details.

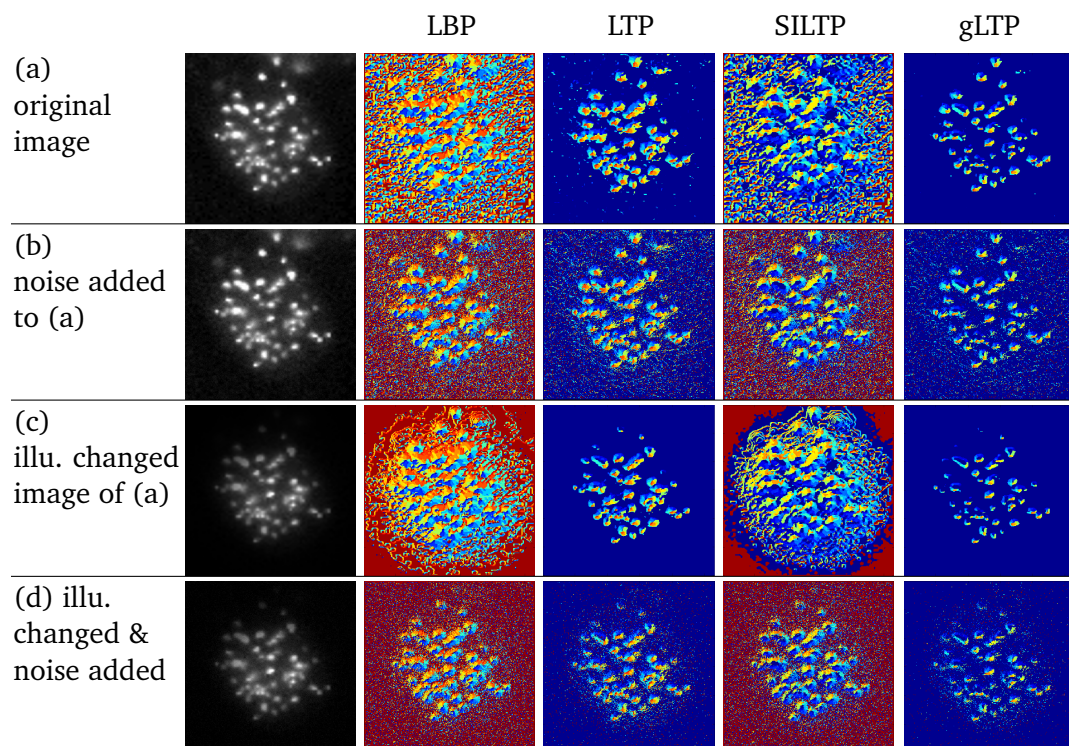


Figure 4.7: The effect of noise and illumination changes on LBP, LTP ($\tau = 10$), SILTP ($w = 0.1$), and the proposed gLTP ($w = 0.9, \tau = 10$) for an example cell image. The cell image under different transformations and the corresponding LBP based labels. See text (Section 4.3.2) for more details.

4.4 Experiments

This section compares the proposed gLTP with the baselines such as LBP, LTP and SILTP, on two qualitatively different datasets from two independent domains, Normal/Abnormal colonoscopy and the ICPR cell images, and shows that gLTP gives competitive performance compared to the best-performing descriptors in the datasets. The colonoscopy image dataset contains images which were taken under different illumination conditions, and the images in the ICPR cell dataset are severely affected by noise.

4.4.1 Experimental setup

Two sets of experiments with different sizes of the local image neighbourhood for LBP-based descriptors were carried out. The first one uses $N = 8$ sampling points on a circle of radius $R = 1$ around each pixel. To capture larger local neighbourhood the second one uses 3 sets of sampling points $\{(N = 8, R = 1), (N = 12, R = 2), (N = 16, R = 3)\}$ around each pixel. From each image the LBP-based codes were extracted densely with a step size of S pixels in the horizontal and the vertical directions. Since the colonoscopy images are relatively large compared to the cell images (Section 3.2) the step size S was set to $S = 4$ and $S = 2$ for the colonoscopy and the cell images respectively. In all the reported experiments in this chapter, the uniform patterns were used for all the LBP-based (LBP, LTP, SILTP and gLTP) descriptors to get the final image-level histogram representations as they capture commonly occurring patterns and reduce the dimensionality of the image representations.

I follow the experimental setup explained in Section 3.2 and report the results. For colonoscopy dataset I use a SVM classifier (LibSVM [29]) with the exponential Chi-square kernel defined as:

$$K(\mathbf{H}_1, \mathbf{H}_2) = \exp\left(-\frac{\gamma}{2} \sum_{i=1}^d \frac{(H_{1i} - H_{2i})^2}{H_{1i} + H_{2i}}\right). \quad (4.7)$$

where \mathbf{H}_1 and \mathbf{H}_2 are d -dimensional histograms representing two images, and H_{1i} is the i -th component of \mathbf{H}_1 . The high number of training images in the ICPR cell dataset

(~ 6000) makes it computationally expensive to use the exponential chi-square kernel. Therefore I use a SVM classifier (LibLinear [45]) with the explicit chi2 kernel mapping [164]. Since it is an explicit mapping a linear SVM can be used for training, which makes the training procedure much faster to learn and evaluate than the non-linear SVM particularly for larger datasets.

The SVM and the kernel parameters were learned using a 3-fold cross validation applied on the training set of each experimental run.

4.4.2 Parameter selection

At each iteration of the experiment I apply a 3-fold cross validation on the training set to select the parameters for LTP, SILTP and gLTP. The parameters which give the best average MCA over these 3-folds of the training set were selected as the best parameters. Table 4.2 shows the ranges of the parameters used for parameter selection.

LBP variant	a	τ
LTP	-	[5, 10, 20, 30, 40]
SILTP	[0.05, 0.1, 0.2, 0.3, 0.5, 0.7, 0.9]	-
gLTP	[0.95, 0.9, 0.8, 0.7, 0.5, 0.3, 0.1]	[5, 10, 20, 30, 40]

Table 4.2: The range of parameters used for LTP, SILTP, and gLTP.

Figure 4.8 shows the histogram of the parameters which were selected in the repeated experiments.

4.4.3 Comparison of LBP, LTP, SILTP and gLTP

Table 4.3 reports the experimental results. Using the larger neighbourhood ($\{(N = 8, R = 1), (N = 12, R = 2), (N = 16, R = 3)\}$) improves the MCA of all the descriptors regardless of the dataset. It is clear that LBP gives modest performance compared to the best performing descriptor as it is sensitive to noise and illumination changes. As expected SILTP gives better performance than LBP and LTP for colonoscopy images as they are affected by illumination, and LTP gives better performance compared to LBP and SILTP on cell dataset as the images in that dataset are severely affected by noise. Neither LTP nor SILTP gives better performance on either dataset. The proposed

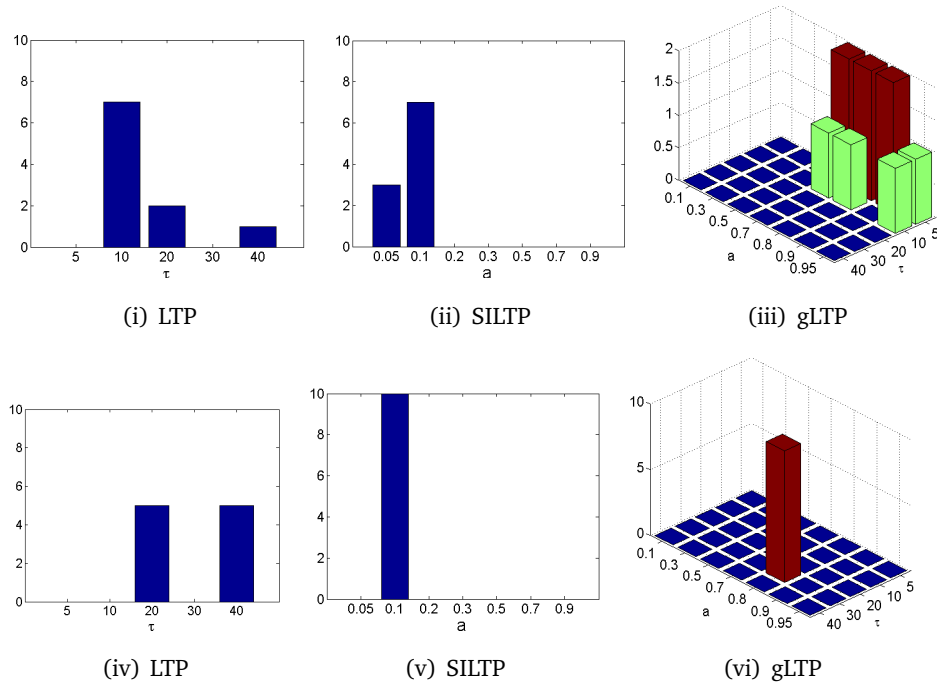


Figure 4.8: Histogram of the selected parameters for LTP, SILTP and gLTP for colonoscopy images (top row) and the ICPR cell images (second row). The vertical axis represents the number of iterations (out of 10) the values in the horizontal axis (or axes) were selected as the best parameters by the cross-validation process using the sampling points $\{(8, 1), (12, 2), (16, 3)\}$.

gLTP descriptor gives similar performance compared to the best performing SILTP on colonoscopy, and performs significantly better than others on the cell images. Note that the aim of the experiments based on gLTP is not to beat other LBP-based descriptors, but to show that under different conditions gLTP gives stable performance compared to others.

	Colon images		ICPR cell images	
	(8, 1)	$\{(8, 1), (12, 2), (16, 3)\}$	(8, 1)	$\{(8, 1), (12, 2), (16, 3)\}$
LBP	80.97 ± 1.25	84.64 ± 1.18	41.25 ± 0.51	63.29 ± 0.63
LTP	79.79 ± 2.41	85.50 ± 1.55	57.47 ± 0.82	72.28 ± 0.52
SILTP	82.25 ± 1.11	87.56 ± 1.05	51.97 ± 0.51	68.09 ± 0.29
gLTP	83.75 ± 0.77	88.14 ± 1.15	67.51 ± 0.68	78.48 ± 0.42

Table 4.3: Comparison of LBP, LTP, SILTP and gLTP, showing that gLTP gives competitive or better results than the baselines.

4.5 Conclusions and discussion

In this chapter I presented a generalised version of LBP, LTP and SILTP descriptor called the *generalised LTP* (gLTP). LBP are sensitive to noise as well as illumination changes; LTP are robust to noise but not to illumination changes; SILTP are robust to illumination changes but not to noise. Instead the proposed gLTP are robust to *both* noise *and* illumination changes. I experimentally showed on two datasets, colonoscopy and cell images, where the colonoscopy images were taken under different illumination conditions and the cell images were severely affected by noise, that neither LTP nor SILTP gives better performance on either dataset. On the other hand, the proposed gLTP gives competitive performance compared to the best performing descriptors on the datasets, confirming that the gLTP is robust to noise and illumination changes.

Since there are two parameters in the gLTP, tuning those based on cross validation is a time consuming process. This could limit the use of gLTP on larger datasets. Although gLTP is robust to illumination and noise, it is also affected by information-loss - a common property of the LBP and its variants usually observed due to binarisation. To overcome these limitations the next chapter (Chapter 5) proposes a novel descriptor called the *Extended Multi-Resolution Local Patterns* (xMRLP) and proposes an unsupervised learning approach to learn its parameters.

EXTENDED MULTI-RESOLUTION LOCAL PATTERNS AND UNSUPERVISED FEATURE LEARNING

Local Binary Patterns (LBP) and its variants lose information due to binarisation involved in the descriptor construction. This chapter proposes a novel descriptor called the Extended Multi-Resolution Local Patterns (xMRLP) inspired by the generalised Local Ternary Patterns (gLTP) descriptor proposed in Chapter 4. Unlike LBP and its variants xMRLP avoids information loss and captures larger local image neighbourhood (e.g. 16×16). Since xMRLP contains a set of parameters, this chapter proposes an unsupervised approach to learn them. A simplified variant of the xMRLP feature, the Multi-Resolution Local Patterns, was also proposed, where the parameters were assigned fixed values, hence avoiding the learning stage. Experiments on colonoscopy as well as the ICPR cell image datasets show that MRLP gives improved performance compared to LBP and its variants, and that the learned descriptor xMRLP gives considerable improvements over MRLP on the colonoscopy datasets.

5.1 Introduction

LBP and its variants are widely applied for image classification in various domains, e.g. face recognition [56] and medical image classification [113, 190]. However,

they have a few limitations.

- Information loss : LBP-based descriptors lose information due to the binarisation procedure involved in the descriptor building stage.
- High dimensionality of the image-level representation when the number of sampling points is large : the standard LBP with d sampling points leads to an image representation of size 2^d . Increasing d will lead to a larger dimensionality of the image-level histogram representation, hence increasing the computational complexity of the classification. E.g. when $d = 16$ there will be $2^d = 65,536$ possible LBP labels, leads to a histogram of dimension 65,536.
- LTP, SILTP and gLTP double the size of the image-level representation: these variations make the standard LBP robust to noise and/or illumination changes by introducing free-parameters and produce a set of ternary patterns. Usually the image representation is obtained by splitting each ternary pattern into two binary patterns and obtaining histograms from these binary patterns. Splitting each ternary code into two binary codes and histogramming them doubles the size of the image-level representation compared to the one obtained by LBP.
- Difficulty with defining the uniform patterns: uniform LBP uses heuristics to define the commonly occurring LBP patterns in the images. This heuristic process makes it difficult to define the uniform patterns when d is large.

To overcome these limitations and to capture larger local image neighbourhoods, I propose a novel descriptor called the *Extended Multi-Resolution Local Patterns* (xMRLP) which is a multi-resolution and non-binarised version of gLTP proposed in Chapter 4. xMRLP avoids information loss by avoiding the binarisation step which is always included in LBP based descriptors. Since the xMRLP descriptor contains a set of parameters, I propose an unsupervised approach to learn them. Also I propose a simplified variant called the *Multi-Resolution Local Patterns* (MRLP), where the parameters are fixed to their default values (explained in Section 5.2.2), hence avoiding the learning stage. The final image-level representation using MRLP or xMRLP can be obtained using feature encoding approaches such as bag-of-words.

In the following first the xMRLP descriptor is introduced, and then the unsupervised approach to learn the parameters of xMRLP is explained, finally experiments and the results are reported.

5.2 Extended Multi-Resolution Local Patterns

First I introduce the single-resolution version of the descriptor called the *Extended Local Patterns* (xLP) and then I extend it to multi-resolution version called the *Extended Multi-Resolution Local Patterns* (xMRLP).

5.2.1 Extended Local Patterns

Let's assume I_i is the i^{th} image in the training dataset $\mathcal{D} = \{(I_i, y_i)\}$, $i = 1, \dots, N$, where N is the number of images in \mathcal{D} , $y_i \in \{1, \dots, C\}$ is the label of image I_i , and C represents the number of classes. Let's consider a circular sampling pattern shown in Figure 5.1, where I_{ij} represents the intensity value at the j^{th} location of I_i (to reduce the notation cluttering I use a single index j to represent the pixel at a location $I(x, y)$), and I_{ij}^s is the intensity value at the s^{th} sampling point around I_{ij} , where $s = 1, \dots, d$, and d is the number of sampling points in the local neighbourhood ($d = 8$ in Figure 5.1). The intensity values at the sampling points are bilinearly interpolated whenever a sampling point does not coincide with a pixel of I_i .

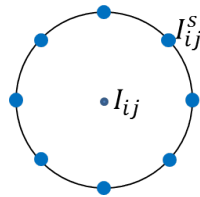


Figure 5.1: An example sampling pattern with 8 sampling points.

As explained in Section 4.3.1 of Chapter 4 the difference D between two pixels, I_{ij} and its s^{th} neighbour I_{ij}^s , can be written using the following equation:

$$\begin{aligned} D(I_{ij}, I_{ij}^s) &= [a_1 I_{ij} + \tau_1] - [a_2 I_{ij}^s + \tau_2] \\ &\propto I_{ij} + a I_{ij}^s + \tau, \end{aligned} \tag{5.1}$$

where a_1 and a_2 represent the non-uniform illumination applied to the pixels I_{ij} , I_{ij}^s and τ_1, τ_2 model the sensor noise due to the image capturing device at those pixels, and $a = -\frac{a_2}{a_1}$, $\tau = \frac{\tau_1 - \tau_2}{a_1}$. Unlike gLTP which assumes a common value for each of the parameters a and τ for all the images, xLP relaxes this constraint by assuming a set of values $\mathbf{a} = [a_1, \dots, a_d]^T$ and $\boldsymbol{\tau} = [\tau_1, \dots, \tau_d]^T$ each for each of the sampling points.

I define $\mathbf{x}_{ij} \in \mathbb{R}^d$ which describes the local neighbourhood around the j^{th} pixel of image I_i as a concatenation of the differences $D(I_{ij}, I_{ij}^s)$, $s = 1, \dots, d$, as follows:

$$\mathbf{x}_{ij}(\mathbf{a}, \boldsymbol{\tau}) = \begin{bmatrix} D(I_{ij}, I_{ij}^1) \\ \vdots \\ D(I_{ij}, I_{ij}^d) \end{bmatrix} = \begin{bmatrix} I_{ij} + a_1 I_{ij}^1 + \tau_1 \\ \vdots \\ I_{ij} + a_d I_{ij}^d + \tau_d \end{bmatrix}. \quad (5.2)$$

In the above equation, $\mathbf{a} = [a_1, \dots, a_d]$ weight the importance of the neighbourhood pixels, while $\boldsymbol{\tau}$ are the biases in different directions. Unlike LBP and its variants, the descriptor defined in Equation (5.2) is a real valued descriptor, hence the bias parameters ($[\tau_1, \dots, \tau_d]$) may not capture much information. I experimentally found that adding these biases does not significantly improve the classification performance, but it increases the computational complexity and the number of parameters to be learned. Therefore I discarded these bias parameters and define xLP descriptor as follows:

$$\mathbf{x}_{ij}(\mathbf{a}) = \begin{bmatrix} I_{ij} + a_1 I_{ij}^1 \\ \vdots \\ I_{ij} + a_d I_{ij}^d \end{bmatrix}. \quad (5.3)$$

5.2.2 Extension to multi-resolution version

Usually LBP based descriptors operate on a small neighbourhood (e.g. 3×3). To effectively capture a larger context with a reduced number of sampling points, I adopt a sampling pattern inspired by the spatial structure of receptive fields of the human retina, widely adopted in recent work on visual descriptors, e.g. FREAK [9], BRISK [87], and DAISY [159].

Figures 5.2(i) and 5.2(ii) show two examples of sampling patterns, where the local neighbourhood is quantised radially into three resolutions (radii), and at each resolution

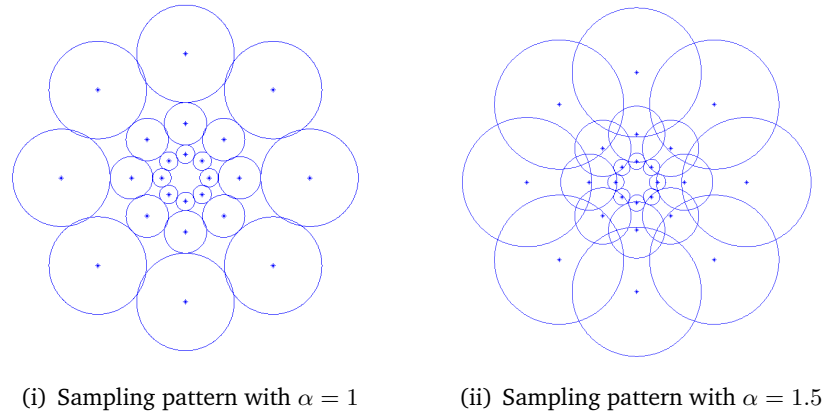


Figure 5.2: A three-resolution sampling pattern ($Q = [8, 8, 8]$).

8 sampling points are considered. At each sampling point, a Gaussian filter is applied to capture a local neighbourhood which is larger than one pixel.

I propose the parameters in Table 5.1 to define the sampling pattern. The equations in Table 5.1 are obtained such that the circles (support regions of Gaussian filters at the sampling points) at resolution r and $r + 1$ touch each other when $\alpha = 1$ (Figure 5.2(i)), or they overlap when $\alpha > 1$ (Figure 5.2(ii)), where α determines the overlap between the support regions around two adjacent sampling points. I set $R_1 = 1$ and $\alpha = 1.5$ for all the experiments reported in this thesis. Note that throughout this thesis, the support regions are fixed to $\mu \pm 2\sigma$ in both x and y directions, where μ is the location of the centre pixel. However, future work will focus on analysing the sensitivity of these values for classification.

Parameter	Description
Q_i	The number of sampling points at resolution i . ($Q = [8, 8, 8]$ in Figure 5.2(i) and 5.2(ii))
$R_i = \frac{Q_i(Q_{i-1} + \pi)}{Q_{i-1}(Q_i - \pi)} R_{i-1}$	The distance between the centre of the sampling pattern and any sampling point at resolution i . (R_1 is set to 1)
$r_i = \alpha \frac{R_i \pi}{Q_i}$	The radius of the circle around a sampling point at resolution i .
$W_i = 2\lceil r_i \rceil - 1$	The window size of the Gaussian filter around a sampling point at resolution i
$\sigma_i = \max(1, W_i/4)$	The standard deviation of the Gaussian filter around a sampling point at resolution i . When $\sigma_i \leq 1$ no smoothing is performed.
α	determines the overlap of the Gaussian filtering regions.

Table 5.1: The parameters of the multi-resolution sampling patterns.

When xLP (Equation (5.3)) is combined with this multi-resolution sampling pattern, I call the resultant descriptor the *Extended Multi-Resolution Local Patterns* (xMRLP).

When the parameters in Equation (5.3) are fixed to their default values, which assigns equal importance to the neighbourhood pixels, i.e. $\mathbf{a} = [-1, \dots, -1]$ I call the resulting descriptor the Multi-Resolution Local Patterns (MRLP).

5.2.3 Image-level representation using xMRLP

I use a feature encoding approach (such as BOW, SC) to aggregate the xMRLP features to get an image-level representation. As explained in Section 2.2 the major steps to get an image representation using a BOW based model include feature extraction and feature encoding.

I use a dense feature extraction strategy to compute the xMRLP features as they show improved performance for classification compared to the feature extraction based on interest points detectors [125]. From each image I extract overlapping small patches (e.g. 16×16). The sampling patterns shown in Figure 5.2(ii) are rescaled so that the sampling points in the external ring lie inside the patch. The xMRLP descriptors extracted from each colour channel of a particular patch (whenever the colour information is available) are concatenated to get the final description of that patch. For example, using the sampling patterns of Figure 5.2(ii) with $Q = [8, 8, 8]$ from a RGB colour patch and a gray scale patch will yield xMRLP descriptors of size 72 (i.e. 3×24) and 24 respectively.

After extracting the xMRLP features from an image, I use the feature encoding methods explained in Section 2.2.2 to obtain an image-level representation from them. The normalisation procedure given in Equation (3.2) was applied to get the final image representation.

5.3 Parameter learning: unsupervised approach

This section proposes an unsupervised approach to learn the parameters of xMRLP (Equation (5.3)), by modifying the optimisation function of the k-means clustering

algorithm.

Clustering refers to the task of partitioning unlabelled data into meaningful groups (clusters) [53]. K-means is a classic clustering algorithm, extensively studied and applied due to its simplicity and robustness. Cluster compactness is one of the quantitative criteria to measure the quality of the result [53]. K-means maximizes compactness in terms of the summed squared distance between the features and the cluster centres to which they are assigned.

5.3.1 The objective function

Let $\mathcal{X} = \{\mathbf{x}_{ij}\}$ be a set of M descriptors sampled from the training images. I learn the xMRLP parameters \mathbf{a} using the following optimisation promoting resulting clusters more compact than the initial ones (i.e. the ones obtained when parameters are fixed).

$$\arg \min_{\{\boldsymbol{\mu}_k\}, \mathbf{a}} \frac{1}{M} \sum_{k=1}^K \sum_{\mathbf{x}_{ij} \in C_k} \|\mathbf{x}_{ij} - \boldsymbol{\mu}_k\|_2^2. \quad (5.4)$$

where $\boldsymbol{\mu}_k \in \mathbb{R}^d$ is the centre of the k^{th} cluster C_k , and K is the number of clusters. Note that $\mathbf{x}_{ij} \equiv \mathbf{x}_{ij}(\mathbf{a})$. The objective in Equation (5.4) along with a regularisation term can be rewritten as:

$$L(\mathbf{a}, \{\boldsymbol{\mu}_k\}) = \frac{1}{M} \sum_{k=1}^K \sum_{\mathbf{x}_{ij} \in C_k} \|\mathbf{x}_{ij} - \boldsymbol{\mu}_k\|_2^2 + \beta \|\mathbf{a} + \mathbf{1}\|_2^2. \quad (5.5)$$

where $\mathbf{1} = [1, 1, \dots, 1]^T \in \mathbb{R}^d$. The first term in Equation (5.5) is exactly the same as the k-means objective function which maximizes the compactness of the clusters when the parameters \mathbf{a} are fixed. The last term is inspired by the LBP and its variants, which makes the values of \mathbf{a} close to -1 .

5.3.2 Optimisation

The partial derivatives of Equation (5.5) w.r.t. to the parameters \mathbf{a} and $\{\boldsymbol{\mu}_k\}$ can be written using Equation (5.7), where $\mathbf{J}_{ij} = [I_{ij}^1, \dots, I_{ij}^d]^T$ and \odot represent the element-wise multiplication (Hadamard product) between two vectors.

$$\nabla_{\mathbf{a}} L = \frac{2}{M} \sum_{k=1}^K \sum_{\substack{\mathbf{x}_{ij} \in C_k \wedge \\ \mathbf{x}_{ij} \in \mathcal{X}}} (\mathbf{x}_{ij} - \boldsymbol{\mu}_k) \odot \mathbf{J}_{ij} + 2\beta(\mathbf{a} + \mathbf{1}). \quad (5.6)$$

$$\nabla_{\boldsymbol{\mu}_k} L = -\frac{2}{M} \sum_{k=1}^K \sum_{\substack{\mathbf{x}_{ij} \in C_k \wedge \\ \mathbf{x}_{ij} \in \mathcal{X}}} (\mathbf{x}_{ij} - \boldsymbol{\mu}_k). \quad (5.7)$$

Algorithm 1 Parameter learning

Input: unlabelled training set $\{I_i\}$, $i = 1, \dots, N$
size of the dictionary K

Output: \mathbf{a} , $\{\boldsymbol{\mu}_k\}$, $k = 1, \dots, K$

- 1: initialize : $\mathbf{a} = [-1, \dots, -1]^T$
 - 2: **while** not converged **do**
 - 3: $\mathbf{a} \leftarrow$ learn \mathbf{a} using Algorithm 2
 - 4: $\boldsymbol{\mu}_k \leftarrow$ learn $\boldsymbol{\mu}_k$, $k = 1, \dots, K$ using Algorithm 3
 - 5: **end while**
-

Algorithm 2 Learn \mathbf{a}

Input: images $\{I_i\}$, \mathbf{a} , size of the dictionary K

Output: \mathbf{a}

- 1: **while** not converged **do**
 - 2: compute \mathbf{x}_{ij} using Equation (5.3)
 - 3: calculate $\nabla_{\mathbf{a}} L$ using Equation (5.6)
 - 4: $\mathbf{a} \leftarrow \mathbf{a} - \eta_{\mathbf{a}} \nabla_{\mathbf{a}} L$
 - 5: **end while**
-

Algorithm 3 Update the dictionary

Input: $\{\mathbf{x}_{ij}\}$, $\{\boldsymbol{\mu}_k\}$, $k = 1, \dots, K$.

Output: $\{\boldsymbol{\mu}_k\}$, $k = 1, \dots, K$.

- 1: **while** not converged **do**
 - 2: **for** $k = 1, \dots, K$ **do**
 - 3: calculate $\nabla_{\boldsymbol{\mu}_k} L$ using Equation (5.7)
 - 4: $\boldsymbol{\mu}_k \leftarrow \boldsymbol{\mu}_k - \eta_{\boldsymbol{\mu}} \nabla_{\boldsymbol{\mu}_k} L$
 - 5: **end for**
 - 6: **end while**
-

I use an iterative alternating optimizing technique described in Algorithm 1 to optimize Equation (5.5), whereby I iteratively optimize one parameter (e.g. \mathbf{a}) at a time while keeping the other (e.g. $\{\boldsymbol{\mu}_k\}$) constant. $\eta_{\mathbf{a}}$ in Algorithm 2 and $\eta_{\boldsymbol{\mu}}$ in Algorithm 3 are the learning rates for \mathbf{a} and $\{\boldsymbol{\mu}_k\}$ respectively. At each iteration I use a line search method to determine these learning rates. The learning was stopped when there is no

further reduction in the cost function (i.e. $|L_i - L_{i-1}| < 10^{-4}$, where L_i and L_{i-1} are the objective values at the i^{th} and $i - 1^{\text{th}}$ iterations respectively).

5.4 Experiments

In this section, first I investigate the effect of learning the parameters of xMRLP features using the proposed unsupervised approach and then compare xMRLP features with LBP based features such as LBP, LTP, SILTP and gLTP. Three datasets (2-class Normal/Abnormal colonoscopy, 3-class Normal/Abnormal/Uninformative colonoscopy and ICPR cell image dataset) were used to compare the proposed features with others.

5.4.1 Experimental setup

To guarantee a fair comparison, all the local descriptors were computed from patches of size 16×16 pixels with an overlap of 4 pixels in horizontal and vertical directions for colonoscopy and 12×12 pixels with an overlap of 2 pixels for ICPR cell datasets. In all the following experiments, the final image representation is normalised as explained in Section 3.2.2.

5.4.2 Effect of parameter learning

Let xMRLP_u denotes the xMRLP descriptor with the parameters learned using the unsupervised approach proposed. In this section I compare the xMRLP_u features with its direct baseline MRLP, where the parameters \mathbf{a} were fixed to its default values ($\mathbf{a} = [-1, \dots, -1]^T$).

I use 100,000 local descriptors sampled randomly from training images to learn the dictionary and the feature parameters for xMRLP features. The β value in Equation (5.5) was set to $\beta = 10$ for the colonoscopy datasets and $\beta = 1000$ for the ICPR cell dataset (Section 5.4.2.3 reports the sensitivity of the β values). The size of the dictionary was set to 200.

5.4.2.1 Cluster compactness

Figure 5.3 shows the ordered standard deviation of clusters (Equation (5.9)) using k-means and the proposed approach (Algorithm 1) to construct the clusters.

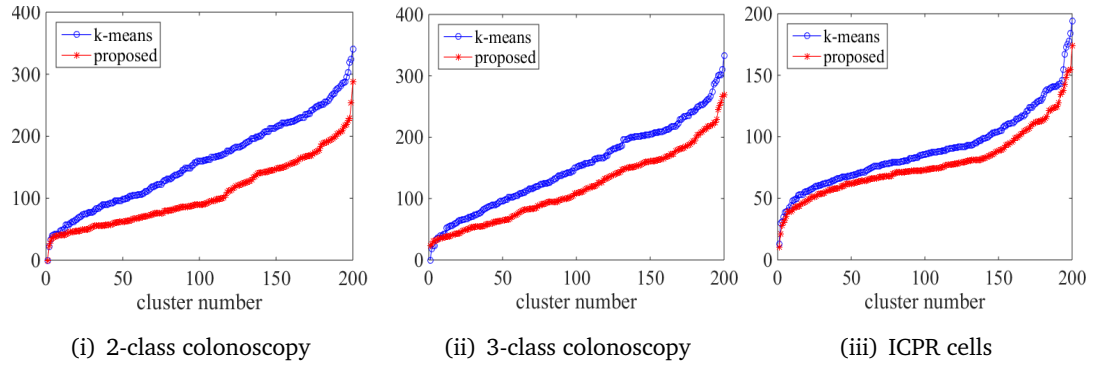


Figure 5.3: The ordered standard deviation of each cluster by the proposed method and k-means for the (a) 2-class colonoscopy, (b) 3-class colonoscopy and the (c) ICPR cells image datasets.

Dataset	Cluster compactness	
	K-means	proposed
2-class colonoscopy	1.0042	0.4631
3-class colonoscopy	1.0897	0.6564
ICPR cells	0.6263	0.5437

Table 5.2: Cluster compactness using the K-means and the proposed approach.

The proposed approach gives lower intra-cluster standard deviation compared to k-means, suggesting that the resulting clusters are more compact compared to k-means. Further, I use the cluster compactness measure proposed in [53] to measure the quality of the clustering by the proposed algorithm. Cluster compactness evaluates how well the output clusters are redistributed by the clustering system, compared to the input set, in terms of the data homogeneity reflected by the distance metric used by the clustering system, and can be defined as [53]:

$$Cm = \frac{1}{K} \sum_{k=1}^K \frac{\sigma(c_k)}{\sigma(\mathcal{X})}. \quad (5.8)$$

where $\sigma(c_k)$ is the standard deviation of the cluster c_k , and $v(\mathcal{X})$ is the standard deviation of all the features in the data set \mathcal{X} . $\sigma(c_k)$ can be defined as:

$$\sigma(c_k) = \sqrt{\frac{1}{N_k} \sum_{\mathbf{x}_{ij} \in C_k} \|\mathbf{x}_{ij} - \mu_k\|_2^2} \quad (5.9)$$

where N_k is the number of features which are assigned to cluster C_k . By fixing $K = 200$, the proposed approach gives more compact clusters compared to k-means (Table 5.2).

5.4.2.2 *xMRLP* vs MRLP descriptors

After learning the parameters \mathbf{a} of \mathbf{xMRLP}_u , I use the BOW encoding method described in Section 2.2.2 to compute the feature representation of images and use a SVM classifier for classification.

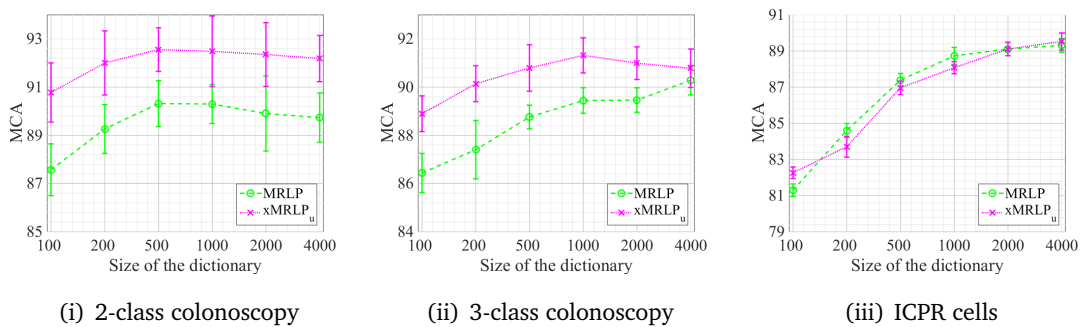


Figure 5.4: *xMRLP* vs MRLP for different datasets (size of the dictionary vs MCA). The vertical bars show the standard deviations calculated over the repeated experiments (experimental setup is described in Section 3.2).

Figure 5.4 reports the performance of \mathbf{xMRLP}_u and MRLP descriptors with different dictionary sizes. It's clear that learning the feature parameters improves the classification performance for the colonoscopy datasets. For the 2-class colonoscopy dataset the \mathbf{xMRLP}_u feature gives about 2% improvement compared to the MRLP features (92.18 ± 0.9 by \mathbf{xMRLP}_u compared to 89.73 ± 1.0 by MRLP), when the dictionary size is 4000. But this unsupervised feature learning gives similar performance for the ICPR dataset, suggesting that MRLP already captures discriminative information needed for classification, and learning the parameters or apply different weights to different neighbourhood pixels does not improve the performance.

In this experiment a SVM classifier with exponential Chi-square kernel was used for the colonoscopy dataset. For the ICPR cell image dataset I use a SVM classifier with linear kernel as the number of training images in the ICPR cell dataset is high which increases the computational complexity of the kernel matrix computations.

5.4.2.3 Sensitivity of the regularisation

The cost function in Equation (5.5) contains a regularisation parameter β which makes the values of α close to -1 (as in LBP). This section investigates the sensitivity of the results to this parameter.

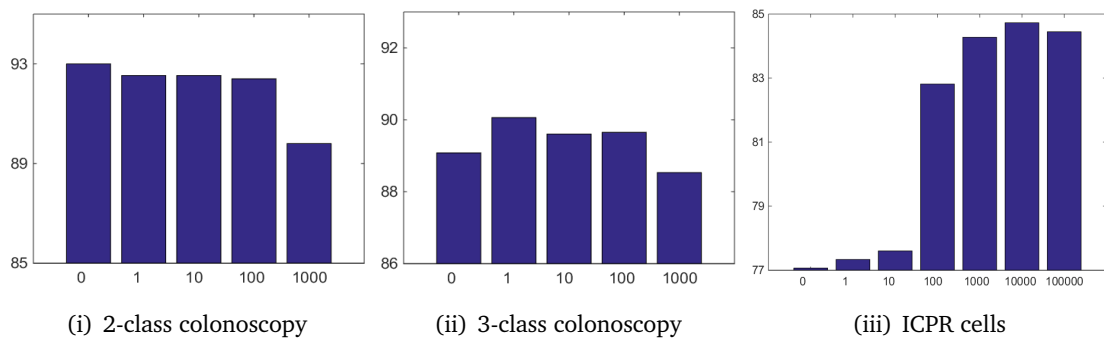


Figure 5.5: Sensitivity of the regularisation: different β values (horizontal axis) vs MCA (vertical axis).

Figure 5.5 reports the performance (MCA) for different value of β when the dictionary size = 200. Small β values give higher performance than large β ($=1000$) values for the 2-class colonoscopy dataset, on the other hand, large β value ($=1000$) gives better performance than small values ($\beta \in \{1, 10, 100\}$) for the ICPR cell dataset. This is expected as β controls the parameter α which weights the importance of the neighbourhood pixels; α close to -1 ($\alpha_s = -1, \forall s$) are preferable for the images which are less affected by illumination changes and contains no edge-like structures. ICPR cell images does not have significant edge-like structures and less affected by illumination. On the other hand colonoscopy dataset contains images which are severely affected by illumination and contain many edge-like structures.

The parameter β is not sensitive in a large range ($\{0 - 100\}$) for the 2-class colonoscopy dataset. But it is sensitive for the ICPR dataset; larger β value ($\beta = 1000$) gave better performance than smaller values.

5.4.2.4 The learned parameters

Figure 5.6 visualizes the value of the learned parameters (α) for the two datasets. The horizontal axis of Figure 5.6 shows different sampling points of the descriptor (Section 5.2.2) and the vertical axis reports their learned values. For example, in Figure 5.6(i) the first 24 sampling points correspond to the first resolution, and the next 24 (25-48) correspond to the second resolution, and so on; and first 8 sampling points (1-8) corresponds to the points at the first resolution and computed from the red colour channel.

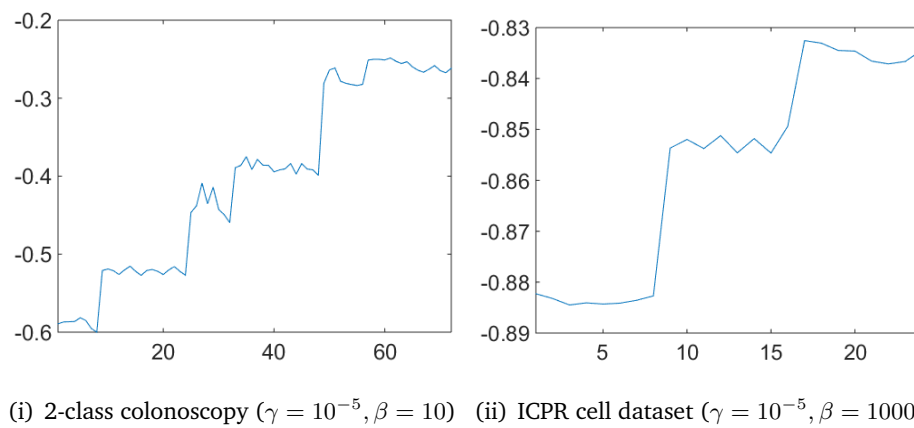


Figure 5.6: Visualisation of the learned parameters: The values (vertical axis) of the learned parameters α at different sampling points (horizontal axis).

From Figure 5.6 it is clear that the learning applies different weights ($[a_1, \dots, a_d]$) to different sampling points, and it also weights the importance of the colour channels as well as the local neighbourhoods (resolutions) differently.

5.4.3 Comparison with LBP based features

This section compares the $xMRLP_u$ features with LBP based features such as standard LBP, LTP, SILTP and gLTP with the sampling patterns shown in Figure 5.2(ii) and shows that MRLP and $xMRLP_u$ features give improved performance over the LBP based features.

Experimental setup with multi-resolution sampling patterns

For each image I compute LBP, LTP, SILTP and gLTP descriptors using a three-resolution version ($Q = [8, 8, 8]$) of the sampling patterns described in Section 5.2.2 (Figure

Feature	2-class colonoscopy	ICPR cells	
	exp chi2 kernel	chi2 kernel	linear kernel
MR-LBP	87.18 ± 1.07	69.73 ± 0.47	69.17 ± 0.43
MR-LTP	88.87 ± 1.04	83.43 ± 0.25	83.29 ± 0.50
MR-SILTP	90.8 ± 1.31	77.55 ± 0.38	77.12 ± 0.33
gMR-LTP	90.26 ± 1.01	83.61 ± 0.30	83.00 ± 0.73
MRLP (BOW, dict size 500)	90.32 ± 0.95	–	87.39 ± 0.35
xMRLP_u (BOW, dict size 500)	92.56 ± 0.90	–	86.96 ± 0.37

Table 5.3: Classification performance (MCA \pm std) of MR-LBP, MR-LTP, MR-SILTP, gMR-LTP and xMRLP_u features on 2-class colonoscopy and ICPR cell image datasets.

5.2(ii)). Since the multi-resolution sampling points were applied for LBP, LTP, SILTP and gLTP they are referred to henceforth as MR-LBP, MR-LTP, MR-SILTP and gMR-LTP. In all cases I consider the uniform pattern histogram representation [127], as it captures meaningful patterns such as edges, corners, etc. and reduces the dimensionality of the final histogram representation.

For the colonoscopy dataset I compute these descriptors independently from each colour channel of RGB colour space, then concatenate their histogram representation to get the final image description. The size of the histogram features for MR-LBP is n_c colours \times 3 resolutions \times 59 histogram bins, and the size for MR-LTP, MR-SILTP and gMR-LTP is double ($2 \times n_c \times 3 \times 59$). n_c is the number of colour channels, $n_c = 3$ for colonoscopy and $n_c = 1$ for the cell dataset. In each experimental run the parameters of MR-LTP, MR-SILTP and gMR-LTP were learned based on a 3-fold cross validation on the training set.

In this experiment, an exponential Chi-square kernel was used for the colonoscopy dataset, and linear as well as Chi-square kernel (explicit feature mapping [164]) was used for the ICPR dataset.

Observations: Table 5.3 shows the classification performance of the LBP based descriptors on 2-class colonoscopy and the ICPR cells datasets. Mainly there are three observations.

1. Overall the multi-resolution sampling patterns (Figure 5.2(ii)) improve the performance of LBP based descriptors compared to the sampling patterns used

in the experiments of Chapter 4 (Table 4.3), as the multi-resolution sampling patterns (Figure 5.2(ii)) capture larger local context and reduce the noise in the neighbourhood by applying Gaussian filters as explained in Section 5.2.2.

2. Similar observations as in Chapter 4 apply: neither MR-LTP nor MR-SILTP gives the highest performance on either dataset. But gMR-LTP descriptors give competitive performance compared to the best performing descriptors (among MR-LBP, MR-LTP and MR-SILTP) in the datasets.
3. MRLP and $xMRLP_u$ improves the performance of LBP based descriptors as they avoid information loss due to binarisation in the descriptor construction. The ICPR dataset shows about 4% improvement by MRLP features compared to gMR-LTP descriptors with reduced size of the image representation (500 by MRLP vs 1064 by gMR-LTP), proving that MRLP features capture more information than the gMR-LTP features.

5.5 Conclusions and discussion

Inspired by the success of LBP and its variants, a novel descriptor called the $xMRLP$ and its simplified variant the MRLP were proposed in this chapter. $xMRLP$ was designed to overcome the limitations of LBP and its variants (e.g. information loss). Since the $xMRLP$ descriptor contains a set of parameters, an unsupervised learning approach was proposed to learn those parameters. MRLP uses a set of default parameters, hence is parameter free and has no need for the learning step.

I experimentally showed that MRLP descriptor gives competitive performance compared to LBP and its variants on the colonoscopy dataset, and gives a significant improvement of $\sim 4\%$ over LBP and its variants on the ICPR cells dataset.

The $xMRLP$ features give considerable improvements compared to MRLP features (92.18 ± 0.9 vs 89.73 ± 1.0) and LBP based descriptors on the 2-class colonoscopy dataset. But for the ICPR cells dataset $xMRLP$ descriptor gives similar performance to the MRLP descriptor suggesting that the additional parameters add no further information for this specific type of images. This would require further investigations into the suitability of the descriptor for specific image characteristics.

To improve the performance of xMRLP descriptors, the next chapter proposes a discriminative learning approach, where the parameters of xMRLP are learned using a set of training images with image-level labels.

DISCRIMINATIVE FEATURE

6

LEARNING USING WEAK LABELS

In the previous chapter I proposed a novel descriptor called the Extended Multi-Resolution Local Patterns (xMRLP) and an unsupervised learning approach to learn its parameters. In this chapter I propose a discriminative approach based on the Naïve Bayes Nearest Neighbour (NBNN) classifier to learn the parameters of xMRLP features.

In contrast to existing discriminative feature learning approaches, in which a set of labelled data is available for learning in the form of region-level annotations or matching and non-matching feature pairs, I use weakly labelled data, i.e. training data with image-level labels, to learn the local image features which are discriminative for image-level classification. Requiring image-level instead of region-level labels or matching and non-matching image patch pairs makes annotations less expensive to generate, hence more feasible in practice.

6.1 Introduction

As explained in Section 1.1, the current approaches proposed for automated colonoscopy image analysis have been mainly focussing on identifying appropriate features; various hand-crafted features such as Root-SIFT (rSIFT) [13, 113], colour histograms [70], LBP [113], and LTP [113] have been explored. For example LBP and GLCM features was used for normal/abnormal classification [98, 113], CWC features was explored for polyp detection [67], and for classification [98], colour histograms and other statistics was used for bleeding detection [70].

All these approaches deploy sets of hand-crafted features. However, hand-crafted features may not be optimally discriminative for classifying images from particular domains (e.g. colonoscopy), as not necessarily tuned to the domain's characteristics. Since I learn the features from the colonoscopy images I expect them to be more discriminative than the hand-crafted ones for colonoscopy image classification. I show experimental evidence in support of this claim in Section 6.5.

Recently feature learning approaches [19, 26, 132, 151, 174, 175] have become popular as they learn domain-specific discriminative features and have been shown to improve the performance of medical image segmentation [19], image retrieval [132, 151], and interest point matching [26, 174, 175]. These approaches assume that region-based annotations or a training set consisting of matching and non-matching pairs of image patches are available to learn the feature descriptors (reviewed in Chapter 2).

Convolutional neural nets (CNN) [79] have also been used to learn local features. In CNN, a set of filters as well as the image-level classifiers are learned in a unified framework. Usually CNN require a very large amount of training data [122]; when this is not available, CNN gives worse performance than traditional, hand-crafted features and BOW-based feature encoding methods [122].

None of these feature learning approaches have been explored for colonoscopy image classification. Since generating annotations for any medical training set (region-level, or matching and non-matching feature pairs) is a difficult, time-consuming task, and clinical time is notoriously at a premium I propose a novel feature learning approach which uses only weakly-labelled data, namely image-level labels. Requiring image-labels instead of region-level labels or matching and non-matching image patch pairs makes annotations less expensive, and closer to the data normally available from normal clinical practice, hence more feasible in practice. Unlike CNN, the proposed approach does not require large amounts of training data, hence it is more suitable for images from the medical domain.

In the following, first I introduce the notation, then review concisely I2CD distances and NBNN classifier. Finally I report experiments based on the proposed descriptor and comparisons with different features for colonoscopy and cell images.

6.2 Notation

Let I_i be the i^{th} image in the training set characterised by a set of local features $\{\mathbf{x}_{ij}\}$, $j = 1, \dots, N_i$, where N_i is the number of local features in I_i and $\mathbf{x}_{ij} \in \mathbb{R}^d$. Let's consider the case of weak labels, whereby an image I_i is associated with a single image-level label, y_i , indicating its class membership. The goal of this chapter is to learn the local features (the parameters of xMRLP) based on the given training data, which is formed by the set of tuples $\mathcal{D} = \{(I_i, y_i)\}$, $i = 1, \dots, M$, where M is the number of images in \mathcal{D} , and $y_i \in \{1, \dots, C\}$ corresponds to the label of the i^{th} training image associated with the C classes.

6.3 Image-to-class distances

Image-to-class distance (I2CD) was first introduced by Boiman et al. [23] in the NBNN classifier, and subsequently used, among others, for distance metric learning [171] and discriminative subspace learning [191]. This section explains the motivation and the derivation of NBNN, its extensions and applications.

6.3.1 Learning-based and non-learning based classifiers

Image classification methods can be roughly divided into two broad families of approaches: (1) *learning-based classifier*, where the classifier contain a set of parameters which have to be learned based on the training data, e.g. SVM, (2) *classifiers without a learning stage*, for which the classification decision depends directly on the data, and requires no training phase, e.g. nearest neighbour (NN) classifier with fixed parameters. The classifiers without a learning stage have several advantages compared to the classifiers with a learning stage: (1) they can naturally handle a huge number of classes; (2) they avoid overfitting of parameters, which is a central issue with the learning-based classifiers; (3) and they require no training or learning phase [23].

On the other hand, however, classifiers without a learning stage show reduced performance for BOW-based classification compared to ones with a learning stage [23]. In BOW-based approaches first the local features extracted from images are

clustered to generate a dictionary, then this dictionary is used to compute compact image representations for each image [152]. Although vector quantisation gives a significant dimensionality reduction, it also degrades the discriminative power of the resulting image representation as some information is lost. After obtaining an image representation, image-to-image distances are often employed in connection with kernel-based methods such as SVM for image classification. As argued by Boiman et al. [23], the errors in the feature quantisation stage and the image-to-image comparisons make the classifiers which have no learning stage less useful than the classifiers which have a learning stage. The learning-based classifiers, since, they have a learning stage, can compensate for the information loss, leading to good classification results.

To improve the performance of classifiers without a learning stage, Boiman et al. [23] suggest using image-to-class distances rather than image-to-image distances.

6.3.2 Image-to-class vs image-to-image distances

Usually image-to-image distances (I2ID) are fundamental for the kernel based methods such as SVM. When the images are represented by BOW histograms, I2ID becomes the distance between the descriptor distributions of two images. Such distances can be measured via histogram intersection, Chi-square distance, or KL divergence [23]. An I2ID based classifier compares the query image with the labelled images in the training set; when the query image is similar to one of the training images, the I2ID becomes small, hence it provides good classification results. The performance of I2ID-based classification depends significantly on the number of training images and on intra-class variations. Therefore to get a better generalisation Boiman et al. propose *image-to-class distances* (I2CD) which compares the descriptor distribution of an image I and the descriptor distributions of different classes (using all the images in each class).

Figure 6.1 shows an example of image-to-image and image-to-class comparisons: even though the features of two images are not very similar according to the image-to-image distance, their distances to the class distribution are similar, enabling correct classification.

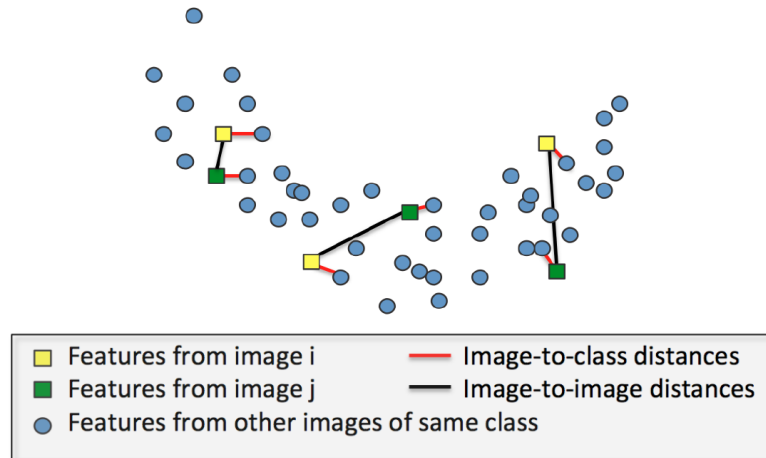


Figure 6.1: The image-to-class concept: even though the features of the two images are not very similar (close), their distances to the class distribution are similar, and that is what counts for NBNN (adapted from [162]).

6.3.3 The Naïve Bayes Nearest Neighbour classifier

The NBNN classifier uses the I2CD to classify a test image I_i into one of the predefined class $\{1, \dots, C\}$, where C is the number of classes [23]. The label \hat{y}_i of a test image I_i is found according to maximum-a-posteriori (MAP) estimation:

$$\hat{y}_i = \arg \max_c p(c|I_i). \quad (6.1)$$

Assuming a uniform prior $p(c)$ over classes and applying Baye's rule, the MAP classifier reduces to maximum-likelihood (ML):

$$\hat{y}_i = \arg \max_c \log(p(I_i|c)). \quad (6.2)$$

Let $\{\mathbf{x}_{ij}\}$, $j = 1, \dots, N_i$ denotes the local features from image I_i , where N_i is the number of local features extracted from I_i . Assuming local features are independent, and taking the log-likelihood,

$$\hat{y}_i = \arg \max_c \log \left(\prod_{j=1}^{N_i} p(\mathbf{x}_{ij}|c) \right) \quad (6.3)$$

$$= \arg \max_c \sum_{j=1}^{N_i} \log p(\mathbf{x}_{ij}|c). \quad (6.4)$$

Assuming a Parzen window estimator with kernel \mathcal{K} which approximates $p(\mathbf{x}_{ij}|c)$:

$$\hat{p}(\mathbf{x}_{ij}|c) = \frac{1}{L_c} \sum_{l=1}^{L_c} \mathcal{K}(\mathbf{x}_{ij}, \mathbf{x}_{ij}^{cl}). \quad (6.5)$$

where \mathcal{K} is typically a Gaussian $\mathcal{K}(\mathbf{a}, \mathbf{b}) = \exp^{-\frac{\|\mathbf{a}-\mathbf{b}\|^2}{h^2}}$, L_c represents the number of local descriptors in the training set for class c , and \mathbf{x}_{ij}^{cl} is the l^{th} nearest neighbour (NN) of \mathbf{x}_{ij} in class c . When L_c approaches infinity and h (width of \mathcal{K}) reduces accordingly, $\hat{p}(\mathbf{x}_{ij}|c)$ converges to the true density $p(\mathbf{x}_{ij}|c)$ [23].

Since most of the local features are far away from each other in the feature space, most of the terms in the summation of Equation (6.5) become negligible as \mathcal{K} exponentially reduces with the distance. Therefore Equation (6.5) can be accurately approximated by the few, say R , largest elements in the sum [23]. These R largest elements correspond to the R nearest neighbours of \mathbf{x}_{ij} in class c . Hence, Equation (6.5) can be rewritten as:

$$\hat{p}(\mathbf{x}_{ij}|c) = \frac{1}{L_c} \sum_{l=1}^R \mathcal{K}(\mathbf{x}_{ij}, \mathbf{x}_{ij}^{cl}). \quad (6.6)$$

As shown by Boiman et al. [23], even when using a small R (as small as $R = 1$) a very accurate approximation of the complete Parzen window estimate can be obtained. When $R = 1$, using the Equation (6.4) and dropping constants which do not influence the optimisation gives,

$$\hat{y}_i = \arg \max_c \sum_{j=1}^{N_i} \log \frac{1}{L_c} \exp^{-\frac{\|\mathbf{x}_{ij} - \mathbf{x}_{ij}^c\|^2}{h^2}} \quad (6.7)$$

$$= \arg \min_c \sum_{j=1}^{N_i} \|\mathbf{x}_{ij} - \mathbf{x}_{ij}^c\|^2 \quad (6.8)$$

$$= \arg \min_c D_{ic}. \quad (6.9)$$

where, \mathbf{x}_{ij}^c is the 1st NN descriptor of \mathbf{x}_{ij} in class c and D_{ic} is the I2CD between an image I_i to class c .

$$D_{ic} = \sum_{j=1}^{N_i} \|\mathbf{x}_{ij} - \mathbf{x}_{ij}^c\|^2. \quad (6.10)$$

The I2CD distance D_{ic} in Equation (6.10) between image I_i and a class c is the sum of the Euclidean distances between each local feature in that image and its NN in c . NBNN identifies the class \hat{y}_i of an image I_i by the class which minimises D_{ic} .

6.3.4 Extensions of NBNN

Various extensions have been proposed to improve the performance of NBNN/I2CD, for example, the NBNN kernel [162], discriminative projection learning [191], and distance metric learning [171].

6.3.4.1 The relaxed version of I2CD

Since some noisy local features may affect the NN calculations (Equation (6.10)), a relaxed version of the I2CD distance considering a set of NN instead of one was proposed in [191], showing improved performance over the original version for complex datasets. The relaxed version of I2CD is given as:

$$D_{ic} = \sum_{j=1}^{N_i} \sum_{r=1}^R \|\mathbf{x}_{ij} - \mathbf{x}_{ij}^{cr}\|_2^2. \quad (6.11)$$

where \mathbf{x}_{ij}^{cr} is the r^{th} NN of \mathbf{x}_{ij} in the c^{th} class, R is the number of considered neighbours and N_i is the number of local features in the image I_i .

6.3.4.2 The NBNN kernel

BOW-based approaches compute the overall distribution of the local features in an image and uses I2ID comparisons in the classification stage. Unlike BOW, the I2CD are considered in NBNN. Hence BOW and NBNN could be complementary to each other, and to combine the advantages of BOW and NBNN a kernelised version of NBNN, called the *NBNN kernel*, was proposed in [162]. This NBNN kernel can be easily incorporated with other kernels and can be trained discriminatively with SVM classifiers. State-of-the-art classification performance for different natural datasets have been reported in [162] by combining BOW based and NBNN based kernels.

6.3.4.3 Discriminative projection learning

Computing I2CD is time consuming especially when the number of local features and their dimensionality are high. The performance of I2CD based classifier can be easily affected by the noisy features in the training set. To reduce the computational burden and to improve the performance of I2CD distances, a discriminative subspace learning method was proposed in [191]. A projection matrix was learned such that in the projected low-dimensional space the I2CD between an image and its own class will be minimum compared to the I2CD between that image and any other classes. This approach showed improved performance over other dimensionality reduction approaches such as PCA and LDA for different action recognition datasets (e.g. KTH [147], and UCF YouTube dataset [99] for action recognition in videos) with reduced time complexity.

6.3.4.4 Distance metric learning

The I2CD in Equation (6.10) assumes that the local features are in a Euclidean space. To improve the performance of I2CD, a distance metric learning approach was proposed in [171]. Class-specific Mahalanobis distance metrics were learned using a max-margin framework, where the I2CD distance between an image I_i and its own class c was minimised while maximizing its distance to all other classes. When classifying a testing image the class which gives the smallest Mahalanobis distance is selected as the predicted class. However, due to the class-specific metrics the number of parameters to be learned becomes high ($= d \times d \times C$, where d is the dimensionality of the local features and C is the total number of classes) which may make the learning complex and require a large amount of data.

6.4 Discriminative max-margin parameter learning

This section proposes a discriminative weakly-supervised max-margin approach to learn the parameters of xMRLP features using the I2CD. Note that, unlike existing feature learning approaches where a set of labelled data in the form of region level labels [19]

or matching and non-matching feature pairs [26, 174, 175] is used to learn the features, I use only image-level labels, which makes the annotation process less expensive.

Unlike the discriminative projection learning or distance metric learning approaches reviewed in Section 6.3.4, my work focusses on learning discriminative local features (particularly the parameters of the proposed xMRLP features) using weak labels. Since my feature contains only a few parameters, the number of parameters to be learned is much smaller than the one required by discriminative projection learning [191] or distance metric learning approaches. Let d be the dimensionality of the local features; then the number of parameters to be learned in discriminative projection learning is $d \times D$ (where $D < d$ is the dimensionality of the new discriminative subspace), and $d \times d$ in metric learning approaches; instead I need to learn only d parameters (Equation (5.3)).

6.4.1 The objective function

Motivated by the soft-margin loss function of SVM and the distance metric learning framework [171], I propose the optimisation framework in Equation (6.12), such that the I2CD from image I_i to its class c is smaller than the distance to any other class \bar{c} with a large margin:

$$\begin{aligned} \arg \min_{\mathbf{a}} \quad & \sum_{c=1}^C \frac{1}{N_c} \left[\sum_{i \in c} \lambda D_{ic} + \xi_{ic\bar{c}} \right] \\ \text{s.t.} \quad & D_{i\bar{c}} - D_{ic} \geq 1 - \xi_{ic\bar{c}} \\ & \xi_{ic\bar{c}} \geq 0. \end{aligned} \tag{6.12}$$

Since the number of local features in different images is different I use a normalised variant of the I2CD given in Equation (6.10) for D_{ic} :

$$D_{ic} = \frac{1}{N_i} \sum_{j=1}^{N_i} \|\mathbf{x}_{ij} - \mathbf{x}_{ij}^c\|_2^2. \tag{6.13}$$

In Equation (6.12) the non-negative slack variable $\xi_{ic\bar{c}}$ is introduced to accommodate the degree of misclassification as in SVM. N_c represents the number of images in class c . The constraints and objective function in Equation (6.12) along with

the regularisation terms can be rewritten as:

$$L(\mathbf{a}) = \sum_{c=1}^C \frac{1}{N_c} \sum_{i \in c} [\lambda D_{ic} + \max(0, 1 - (D_{i\bar{c}} - D_{ic}))] + \beta \|\mathbf{a} + \mathbf{1}\|_2^2. \quad (6.14)$$

where the first and the second terms in Equation (6.14) make the intra-class distances (the distances between images and their own classes) small, while maximizing their inter-class distances (the difference between D_{ic} and $D_{i\bar{c}}$). The last term makes the parameters \mathbf{a} close to -1 as in LBP. The parameters λ and β control the effects of different terms in the cost function given in Equation (6.14). Section 6.5.3 reports the sensitivity of these parameters to the classification results. These parameters can be selected based on e.g. cross validation experiments.

6.4.2 Optimisation

The gradients w.r.t. to the parameter \mathbf{a} can be written as:

$$\nabla_{\mathbf{a}} L = \sum_{c=1}^C \left[\frac{\lambda}{N_c} \sum_{i \in c} \nabla_{\mathbf{a}} D_{ic} + \frac{1}{N_c} \sum_{\substack{i \in c \wedge \\ i \in \mathcal{S}_c}} (\nabla_{\mathbf{a}} D_{ic} - \nabla_{\mathbf{a}} D_{i\bar{c}}) \right] + 2\beta(\mathbf{a} + \mathbf{1}). \quad (6.15)$$

where \mathcal{S}_c is the set of images from class c which are around the margin, i.e. $\mathcal{S}_c = \{i | D_{i\bar{c}} - D_{ic} < 1\}$. $\nabla_{\mathbf{a}} D_{ic}$ can be written as:

$$\nabla_{\mathbf{a}} D_{ic} = \frac{2}{N_i} \sum_{j=1}^{N_i} (\mathbf{x}_{ij} - \mathbf{x}_{ij}^c) \odot (\mathbf{J}_{ij} - \nabla_{\mathbf{a}} \mathbf{x}_{ij}^c). \quad (6.16)$$

where $\mathbf{J}_{ij} = [I_{ij}^1, \dots, I_{ij}^d]^T$, and \odot represents the element-wise multiplication (Hadamard product) between two vectors.

I use a gradient descent approach to optimize Equation (6.12). The overall algorithm to learn \mathbf{a} is given in Algorithm 4, where η_a is the learning rate. The learning rates for different datasets are given in Section 6.5.1. The learning is stopped when there is no further reduction in the cost function (i.e. $\|L_i - L_{i-1}\| < 10^{-2}$, where L_i and L_{i-1} are the objective values at the i^{th} and $(i-1)^{\text{th}}$ iterations respectively).

Algorithm 4 Update \mathbf{a} : supervised learning**Input:** image-level labelled data $\{I_i, y_i\}, i = 1, \dots, N$ **Output:** \mathbf{a}

- 1: initialize $\mathbf{a} = [-1, \dots, -1]^T$
- 2: **while** not converged **do**
- 3: compute \mathbf{x}_{ij} using $(\mathbf{a}, \{I_i\})$ Eqn. (5.3)
- 4: compute the I2CD D_{ic} and $D_{i\bar{c}}$ using Eqn. (6.11)
- 5: calculate $\nabla_{\mathbf{a}}L$ using Eqn. (6.15)
- 6: $\mathbf{a} \leftarrow \mathbf{a} - \eta_a \nabla_{\mathbf{a}}L$
- 7: **end while**

6.5 Experiments

Let MRLP denote the descriptor where the parameters are fixed to their default values ($\mathbf{a} = [-1, \dots, -1]$ in Equation (5.3)) and xMRLP_s denote the learned descriptor using the discriminative approach proposed in this chapter.

In this section, first I investigate the effect of learning the parameters of xMRLP_s using the proposed learning approach and the effect of different regularisation terms in the cost function given by Equation (6.12). Then I compare the MRLP, MRLP_u , and xMRLP_s features using the NBNN classifier and show that xMRLP_s outperforms MRLP.

I follow the same experimental setup explained in Section 5.4.1 in Chapter 5. I use the kd-tree implementation from *Vlfeat* [163] for the NN search.

6.5.1 Effect of parameter learning

This section investigates the effect of parameter learning by comparing the learned xMRLP_s features with its direct baseline MRLP, and shows that the learned features (xMRLP_s) give considerably improved performance compared to MRLP using the NBNN classifier.

To learn the parameters I randomly sample (from the training set of each experimental run) 70, 50 and 70 images respectively from each of the classes of three datasets, 2-class colonoscopy, 3-class colonoscopy and ICPR cells. The value of β was set to 1, 1 and 1000 respectively for the 2-class colonoscopy, 3-class colonoscopy and the ICPR cells datasets. The parameter λ was set to $\lambda = 1 \times 10^{-5}$ for all the datasets.

I2CD calculations are computationally expensive due to the NN search. To reduce the computational burden, I extract features using large step sizes (both at the training and testing phase); the step sizes for the colonoscopy and the cells datasets were set to 16 and 4 respectively. In the classification stage, at each experimental run I randomly sample a maximum number of $\frac{150,000}{N_c}$ features (where N_c is the number of classes) from the training set to perform the classification. The learning rate in Algorithm 4 was experimentally fixed to 2×10^{-3} and 1×10^{-4} for the colonoscopy datasets and the ICPR cell dataset respectively, and the maximum number of iterations for convergence was set to 100 for both datasets.

Dataset	Feature		
	MRLP	xMRLP _u	xMRLP _s
2-class colonoscopy	82.68 ± 0.93	86.55 ± 1.26	86.63 ± 0.99
3-class colonoscopy	80.80 ± 0.83	86.01 ± 0.82	85.73 ± 1.16
ICPR cells	67.90 ± 0.73	68.17 ± 0.57	69.01 ± 0.94

Table 6.1: Classification performance (MCA ± std) using MRLP, xMRLP_u and xMRLP_s features with NBNN classifier.

Table 6.1 reports the performance (MCA and standard deviation over different experimental runs) of the MRLP as well as the learned features (xMRLP_u and xMRLP_s) for all three datasets. For the colonoscopy datasets the learned features considerably improve ($\sim 4\%$) the performance compared to MRLP. Modest improvement was observed for the ICPR dataset, indicating that weighting the neighbourhood pixels captures less additional information. There are two main reasons for this: (1) Edge-like structures can be emphasised by weighting the neighbourhood pixels. Unlike the colonoscopy datasets, the cell images have fewer/no edge-like structures, hence no improvement in the classification results; (2) The cell image dataset contains more classes than the colonoscopy dataset. Therefore, the number of local features from the background regions are higher than the colonoscopy dataset. The higher amount of background features may dominate the I2CD calculations, leading to no improvement in the classification. It is interesting to see that the unsupervised approach performs similarly to the supervised one, even though it is not designed to discriminate different classes. This suggests that when making the clusters compact by learning the feature parameters, the feature distributions of different classes are moving from each other.

6.5.2 Example classification results

Figure 6.2 shows some example images from the 2-class colonoscopy dataset which were correctly classified by both MRLP and xMRLP_s features with the NBNN classifier. Although some images (see examples in Figures 6.2(d-f) and Figures 6.2(j-l)) are difficult to classify as they are genuinely borderline even for experts, the MRLP and xMRLP_s features were able to correctly classify them.

Figure 6.3 shows some example images which were mis-classified by MRLP and xMRLP_s features. This may be due to some local features in the normal images (Figure 6.3) which show similar properties to the local features in the abnormal images (Figure 6.2). For example, the abnormal image in Figure 6.2(g) and the normal images in Figure 6.3(b)(c) contains similar local features. The abnormal regions in the Figures 6.3(g), 6.3(i) and 6.3(k) are small compared to the normal regions; which make the I2CD bias to normal class, hence the images were wrongly classified.

Figure 6.4 shows some example images which were wrongly classified by MRLP features but correctly classified by xMRLP_s features. This is due to the properties captured by xMRLP_s features during learning: such features are able to capture better discriminative properties of the images compared to MRLP.

6.5.3 Sensitivity of the regularisation parameters

The cost function defined in Equation (6.14) contains two parameters, λ and β ; λ controls the contribution of the intra-class distances and β prevents the parameters α from becoming arbitrary high and makes their values close to -1 (as in LBP). This experiment was run to investigate the sensitivity of these parameters.

The 2-class colonoscopy and the ICPR cells datasets were used in this experiment. A small subset of images (30 and 40 images from each class of colonoscopy and ICPR cells respectively) from the training set of a particular experimental run was used as the training set to learn the parameters α . The test images from that run were used to evaluate the performance using the NBNN classifier.

Figure 6.5 reports the MCA for different parameter settings. For both datasets, assigning λ to any of the values $\{0, 1 \times 10^{-5}, 1 \times 10^{-3}\}$ does not affect the MCA

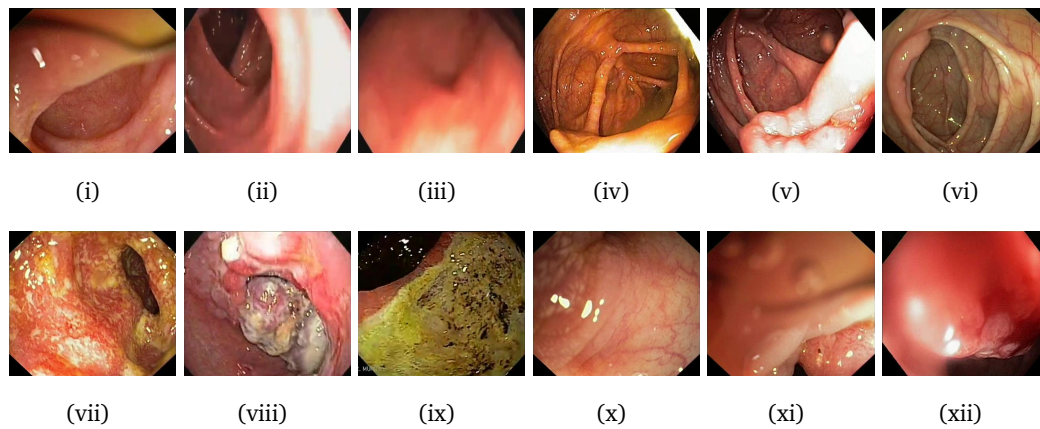


Figure 6.2: Example of correctly classified Normal (top) and Abnormal (bottom) images by both MRLP and $xMRLP_s$ features.

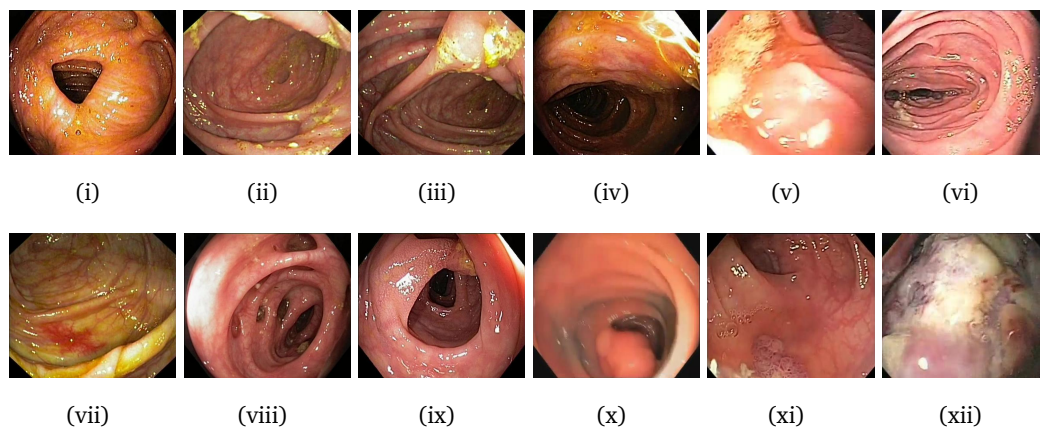


Figure 6.3: Examples of images mis-classified by both MRLP and $xMRLP_s$. i.e. Top row: examples of normal images misclassified as abnormal. Bottom row: examples of abnormal images which misclassified as normal.

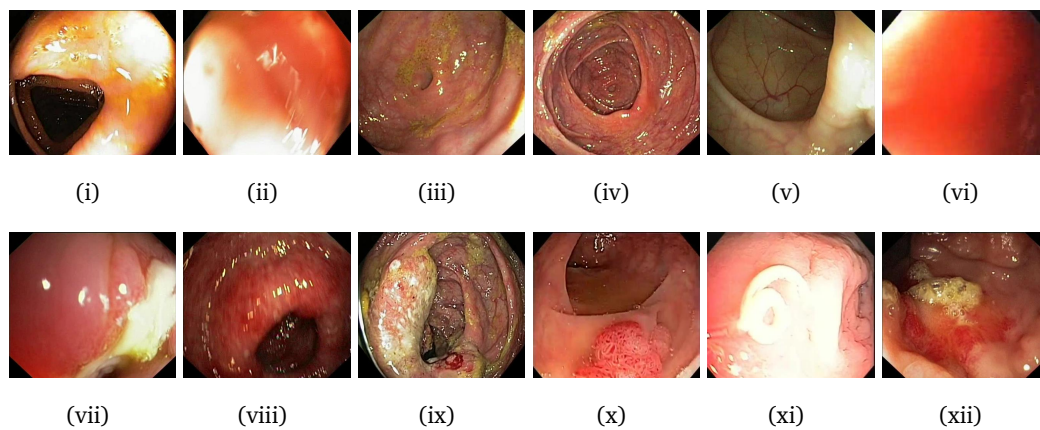


Figure 6.4: Example of normal (top) and abnormal (bottom) images which are correctly classified by $xMRLP_s$ but wrongly classified by MRLP features.

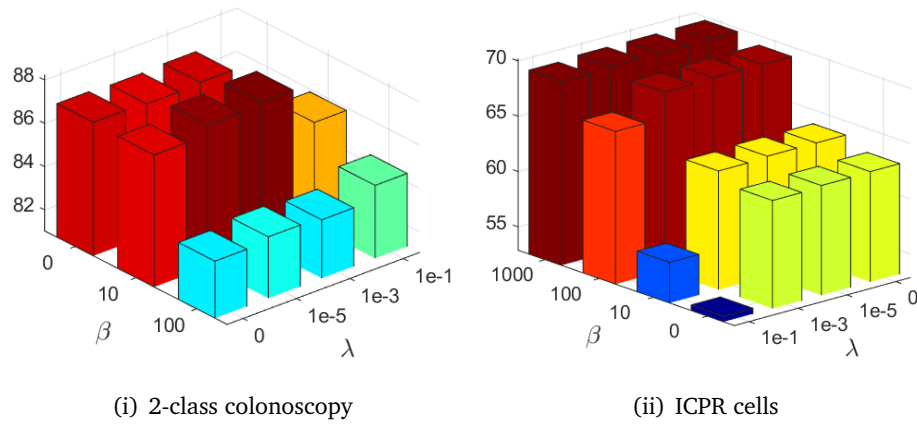


Figure 6.5: Sensitivity of the regularisation parameters: λ and β (Equation (6.14)) vs MCA.

significantly, but setting it to larger values reduces the MCA; suggesting that the discriminative term in Equation (6.14) is more important than the term which minimises the intra-class distances. On the other hand, changing the parameter β affects MCA; small values for the colonoscopy dataset and larger values for the ICPR cells dataset give good classification. This is because unlike the ICPR cells dataset, the colonoscopy dataset contains images which are severely affected by illumination changes (argued in Section 5.4.2.2). Similar observations were also observed in the experiments reported in Section 5.4.2.2 of Chapter 5.

6.5.4 The learned xMRLP_s parameters

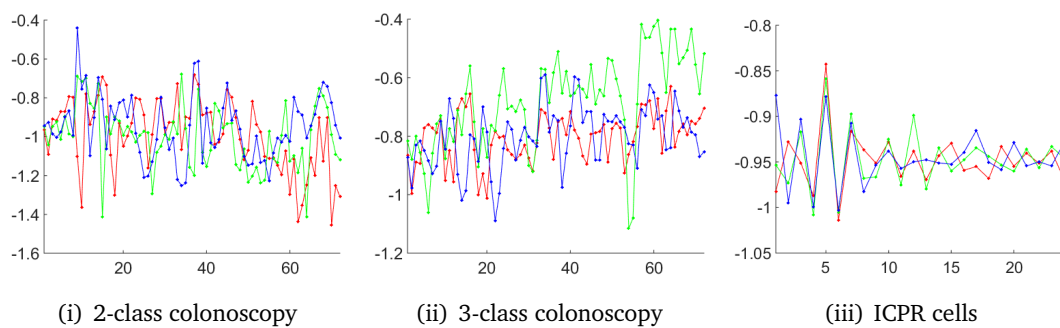


Figure 6.6: Visualisation of the learned xMRLP_s parameters: The learned parameter values (vertical axis) at different sampling points (horizontal axis) (Figure 5.2(ii)). Different colours correspond to different experimental runs.

Figure 6.6 visualizes the learned parameters of the xMRLP_s feature for three randomly selected experimental runs for different datasets. For all the datasets the learning adds different weights to different sampling points (Sampling points of the

$xMRLP_s$ are shown in Figure 5.2(ii)). The discriminative approach gives different parameter values than the ones given by the unsupervised approach described in Chapter 5.

6.6 Conclusions and discussion

In this chapter I presented a weakly-supervised approach for learning the parameters of the $xMRLP$ features ($xMRLP_s$) and showed improved classification performance in terms of MCA over the $MRLP$ features using the NBNN classifier. Unlike existing feature learning approaches, where a set of labelled data in the form of region-level annotations or matching and non-matching feature pairs are used for learning the features, I use only weak labels, which makes annotations less expensive, hence more feasible in practice.

Although the learned features show considerable improvement on the colonoscopy datasets, these give modest improvement on the ICPR cells dataset. I identified a few limitations associated with the proposed approach.

1. For the ICPR cells dataset I experimentally found that small values of α (close to 0) lead to small D_{ic} (Equation 6.13). This results in a smaller objective value for the objective function defined by Equation 6.14 compared to the objective value when $\alpha = 1$. The parameter learning tries to reduce the objective value, hence it reduces the values of α instead of maximizing the discriminative term defined in Equation (6.14) (the second term in Equation (6.14)).
2. The regularisation parameter β in Equation (6.14) is sensitive for the ICPR cells dataset as shown in Figure 6.5. This is mainly due to the reason explained in 1.
3. The margin which separates D_{ic} and $D_{i\bar{c}}$ is fixed to 1 in Equation 6.14. However this is a free parameter that has to be tuned for different datasets.
4. The ICPR cells dataset contains more classes than the colonoscopy datasets (6 classes in ICPR cells vs 2 and 3 classes in 2-class and 3-class colonoscopy datasets respectively). Hence the background features (the features which are common to different classes) in the cell dataset are high compared to the colonoscopy datasets. The noisy local features as well as the background features can easily dominate

the I2CD calculations (explained in Section 7.1 of Chapter 7), leading to wrong classifications.

To overcome these limitations, the next chapter (Chapter 7) proposes an approach which learns the parameters by maximizing the posterior probability of the images; it applies weights to different classes, and learns these weights and the features together, and shows improved performance compared to the method reported in this chapter for the ICPR cells dataset.

DISCRIMINATIVE FEATURE

LEARNING WITH WEAK-LABELS

7

AND WEIGHTED I2CD

In Chapter 6 I presented a weakly-supervised approach to learn discriminative local features based on I2CD. As argued by Zhen et al. [191] noisy local features as well as the local features from the background may degrade the performance of I2CD calculations. To overcome this I propose an approach which applies weights to different classes. I propose a joint learning approach to learn these weights as well as the local features. Our approach uses weak-labels for learning the local features, supports multi-class classification, and has probabilistic outputs. Using the proposed approach I show improved performance on the ICPR cells dataset with the NBNN classifier compared to the results reported in Chapter 6.

7.1 Introduction

Chapter 6 presented a weakly-supervised approach to learn discriminative local features (the parameters of the xMRLP feature) based on I2CD. I2CD is the key element of the NBNN classifier proposed by Boiman et al. [23]. NBNN is a non-parametric approach: it classifies an image by comparing the distance between that image and different classes, and assigning the class which gives the smallest I2CD as the label of that image. But as argued by Zhen et al. [191] noisy local features as well as the local features from the background may degrade the performance of I2CD calculations.

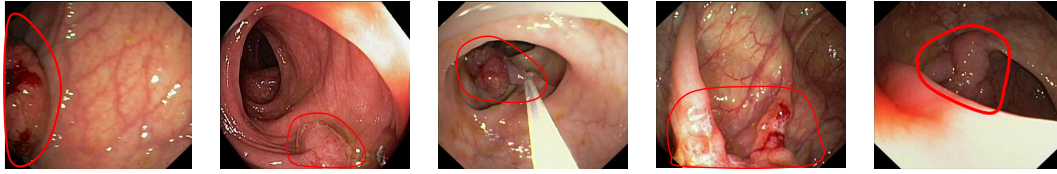


Figure 7.1: Example abnormal images from the 2-class colonoscopy dataset (abnormal regions are indicated by red).

Figure 7.1 shows some example abnormal images from the 2-class colonoscopy dataset. The abnormal colonoscopy images contain not only abnormal regions, but also normal regions. These may cover the majority of the image, so that the normal regions in the abnormal images may dominate the I2CD calculations and lead to wrong classification.

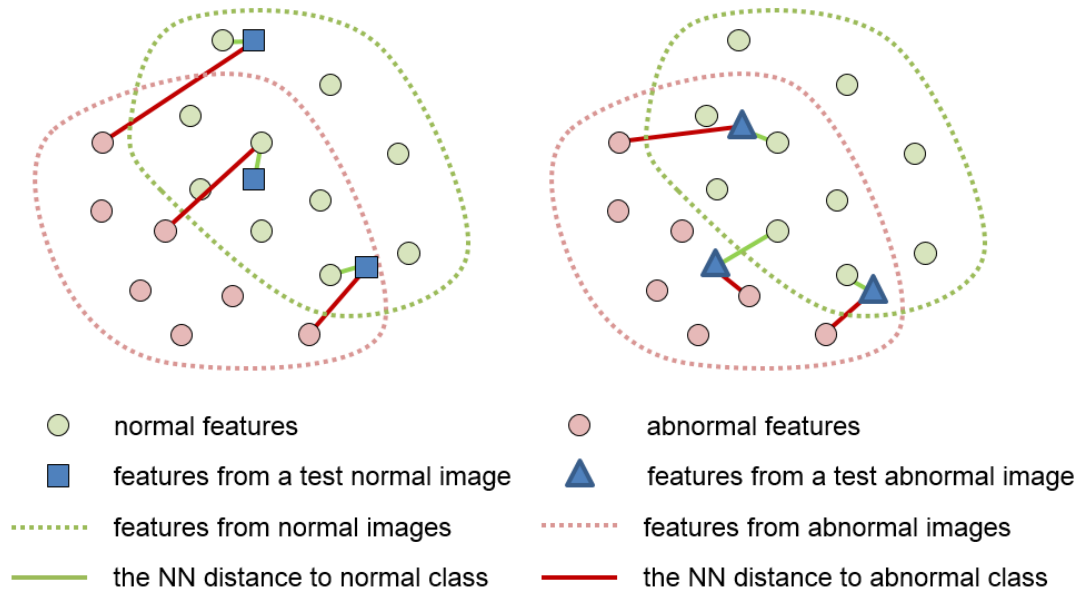


Figure 7.2: The I2CD between an imaginary normal (left) and an abnormal (right) image and to different classes. In the abnormal image the number of abnormal features are less compared to the normal features; it biases the NBNN classifier towards the normal class and classified the abnormal as normal.

Figure 7.2 illustrates this in a simple feature space. It shows a set of local features from normal and abnormal images. In the abnormal test image (Figure 7.2 right) the number of normal features is higher than the number of abnormal features, and this biases the NBNN classifier towards the normal class, predicting the abnormal image as normal.

To overcome this, I apply different weights to different classes. I propose a learning approach to learn these weights as well as the local features (the parameters

of the xMRLP feature). Our approach uses weak labels for learning local features, and supports multi-class classification, and probabilistic outputs. Importantly, I show improved performance compared to the results reported in Chapter 6.

7.2 The proposed joint learning framework

First I introduce the notation and then propose the learning framework to learn the class weights and the local image features.

7.2.1 Notation

Let I_i be the i^{th} image in the training set characterised by a set of local features $\mathbf{X}_i = \{\mathbf{x}_{ij}\}$, $j = 1, \dots, N_i$, where N_i is the number of local features in I_i . Let's consider the general case of weak labels, whereby an image is associated with an image-level soft label indicating, for e.g. class probabilities. The goal of this chapter is to learn the local features, as well as a probabilistic multi-class classifier based on the given training data, which is formed by the set of tuples $\mathcal{D} = \{(I_i, \tilde{\mathbf{p}}_i)\}$, $i = 1, \dots, M$, where M is the number of images in \mathcal{D} , and $\tilde{\mathbf{p}}_i \in [0, 1]^C$ corresponds to a C -dimensional vector of soft labels ($\in \mathcal{R}$) of the i^{th} training image associated with the C classes. I assume that $\sum_{c=1}^C \tilde{p}(y_i = c) = 1$, where $\tilde{p}(y_i = c)$ is the latent class assignment of the image I_i to class c .

7.2.2 Weighted I2CD

The normalised variant of the I2CD between image I_i and class c can be given as (Equation (6.10)):

$$D_{ic}(\mathbf{a}) = \frac{1}{N_i} \sum_{j=1}^{N_i} \|\mathbf{x}_{ij} - \mathbf{x}_{ij}^c\|_2^2. \quad (7.1)$$

where, \mathbf{x}_{ij}^c is the nearest neighbour of \mathbf{x}_{ij} in class c . The NBNN classifier identifies the class \hat{y}_i of an image I_i by the class which minimises D_{ic} :

$$\hat{y}_i = \arg \min_c D_{ic}. \quad (7.2)$$

The NBNN classifier with weighted I2CD can be defined as:

$$\hat{y}_i = \arg \min_c w_c D_{ic}. \quad (7.3)$$

where, w_c is the weight applied to class c .

7.2.3 Discriminative probabilistic softmax classifier

Equation (7.4) below defines a discriminative probabilistic classifier. This classifier outputs the posterior probability of an image I_i belonging to a class c based on weighted I2CD.

$$p(y_i = c | \mathbf{X}_i) = \frac{\exp^{-\gamma w_c D_{ic}}}{\sum_{l=1}^C \exp^{-\gamma w_l D_{il}}}. \quad (7.4)$$

where γ is a decay parameter. In Equation (7.4) the maximum probability for an image belonging to a class c will be obtained for the class which gives small weighted I2CD compared to other classes. Equation (7.4) can be rewritten as:

$$p(y_i = c | \mathbf{X}) = \frac{\exp^{-\gamma_c D_{ic}}}{\sum_{l=1}^C \exp^{-\gamma_l D_{il}}}. \quad (7.5)$$

where $\gamma_c = \gamma w_c$, which has to be learned from the training data.

7.2.4 The objective function

Assuming iid data, I use a maximum a posteriori (MAP) approach to learn the feature parameter \mathbf{a} , and the classifier parameters $\gamma_c, c = 1, \dots, C$ such that the posterior probabilities (Equation (7.5)) of the images in the training data is maximised.

$$\arg \max_{\gamma, \mathbf{a}} \sum_{i=1}^N \sum_{l=1}^C \tilde{p}(y_i = l) \log p(y_i = l | \mathbf{X}). \quad (7.6)$$

where $\gamma = [\gamma_1, \dots, \gamma_C]^T$.

The objective function in Equation (7.6) along with regularisation terms can be given using the following functional:

$$L(\gamma, \mathbf{a}) = \frac{1}{M} \sum_{i=1}^M \sum_{l=1}^C \bar{p}(y_i = l) [\lambda D_{il} - \log p(y_i = l | \mathbf{X})] + \beta \|\mathbf{a} + \mathbf{1}\|_2^2. \quad (7.7)$$

Here, the first term (D_{il}) can be seen as an intra-class distance measure, which minimises the I2CD between an image I_i to a class l weighted by its membership assignments ($\bar{p}(y_i = l)$). The second term minimises the negative log energy defined by Equation (7.6), and the last term, a regularisation term, prevents the feature parameters from becoming arbitrarily large, and keeps them closer to $a_s = -1, \forall s$ as in LBP and its variants.

7.2.5 Optimisation

Letting $p_{ic} = p(y_i = c | \mathbf{X})$, the partial derivative of Equation (7.7) w.r.t the parameters \mathbf{a} and γ can be given as:

$$\nabla_{\mathbf{a}} L = \frac{1}{M} \sum_{i=1}^M \sum_{l=1}^C \bar{p}(y_i = l) \left[(\lambda + \gamma_l) \nabla_{\mathbf{a}} D_{il} - \sum_{c=1}^C p_{ic} \gamma_c \nabla_{\mathbf{a}} D_{ic} \right] + \beta (\mathbf{a} + \mathbf{1}). \quad (7.8)$$

$$\frac{\partial L}{\partial \gamma_c} = -\frac{1}{M} \sum_{i=1}^M \bar{p}(y_i = c) \frac{1}{p_{ic}} \frac{\partial p_{ic}}{\partial \gamma_c} - \frac{1}{M} \sum_{i=1}^M \sum_{\substack{l=1 \\ l \neq c}}^C \bar{p}(y_i = l) \frac{1}{p_{il}} \frac{\partial p_{il}}{\partial \gamma_c}. \quad (7.9)$$

where,

$$\nabla_{\mathbf{a}} D_{ic} = \frac{2}{N_i} \sum_{j=1}^{N_i} \sum_{p=1}^P (\mathbf{x}_{ij} - \mathbf{x}_{ij}^{cp}) \odot (\mathbf{J}_{ij} - \nabla_{\mathbf{a}} \mathbf{x}_{ij}^{cp}). \quad (7.10)$$

$$\frac{\partial p_{il}}{\partial \gamma_c} = \begin{cases} p_{ic} (p_{ic} - 1) D_{ic}, & \text{if } l = c \\ p_{ic} p_{il} D_{ic}, & \text{if } l \neq c. \end{cases} \quad (7.11)$$

$$\frac{\partial p_{ic}}{\partial \mathbf{a}} = -p_{ic} \left[\gamma_c \nabla_{\mathbf{a}} D_{ic} - \sum_{l=1}^C \gamma_l p_{il} \nabla_{\mathbf{a}} D_{il} \right]. \quad (7.12)$$

Algorithm 5 Parameter learning**Input:** training data with image-level labels ($\{I_i, \bar{\mathbf{p}}(y_i)\}$, $i = 1, \dots, M$)**Output:** \mathbf{a}, γ

1: initialize :

$$\mathbf{a} = [-1, \dots, -1]^T$$

$$\gamma_l = 1 \times 10^{-2}, l = 1, \dots, C$$

2: **while** not converged **do**3: $\gamma_l \leftarrow$ learn γ_l , $l = 1, \dots, C$ 4: $\mathbf{a} \leftarrow$ learn \mathbf{a} using Algorithm (6)5: **end while****Algorithm 6** Update \mathbf{a} **Input:** $\{I_i, \bar{\mathbf{p}}(y_i)\}$, $i = 1, \dots, M$, \mathbf{a}, γ **Output:** \mathbf{a} 1: **while** not converged **do**2: compute \mathbf{x}_{ij} using Equation (7.1)3: compute $p(y_i = l | \mathbf{X})$, $l = 1, \dots, C$ using Equation (7.5)4: calculate $\nabla_{\mathbf{a}} L$ using Equation (7.8)5: $\mathbf{a} \leftarrow \mathbf{a} - \eta_a \nabla_{\mathbf{a}} L$ 6: **end while**

The overall algorithm to learn the feature parameter \mathbf{a} and the classifier parameters γ is given in Algorithm 7. Algorithm 6 provides the steps for learning \mathbf{a} , where η_a is the learning rate (learning rates for different datasets are given in Section 7.3.1). The classifier parameters γ can be learned in a similar manner illustrated by Algorithm 6. The learning rates for different datasets are reported in Section 7.3.1. The learning is stopped when there is no further reduction in the cost function (i.e. $\|L_i - L_{i-1}\| < 10^{-1}$, where L_i and L_{i-1} are the objective values at the i^{th} and $(i-1)^{\text{th}}$ iterations respectively). Figure 7.3 shows the convergence of Algorithm 5 for an example experimental run. Algorithm 5 takes about 4 hours to converge (Matlab 2014b and Windows 7 running on a machine with a Core i7 processor and 8GB RAM).

7.3 Experiments

Let MRLP denote the descriptor where the parameters are fixed to their default values ($\mathbf{a} = [-1, \dots, -1]$ in Equation (5.3)). Let xMRLP_s and xMRLP_{s2} denote the learned descriptors using the approaches proposed in chapter 6 and this chapter respectively.

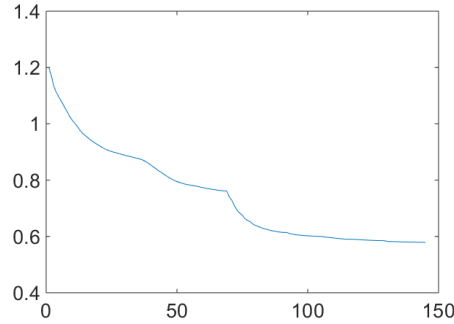


Figure 7.3: The convergence of the Algorithm 5 for an example experimental run (objective value vs iteration number).

I use three datasets, 2-class colonoscopy, 3-class colonoscopy, and the ICPR cell images to evaluate the proposed approach. These datasets were introduced in Chapter 3. In the following, I investigate the effect of learning parameters for both features and classifier, and the sensitivity of the regularisation parameters given in Equation (7.7). Then I compare different features using the NBNN classifier, and show that the proposed features give better performance than the widely-used features in the computer vision such as SIFT and RP, with reduced computational time for classification.

7.3.1 Effect of learning

This section intends to show that learning both the features (\mathbf{a} of Equation (5.3)) and the classifier (γ of Equation (7.5)) using the proposed cost function (Equation (7.7)) gives better performance compared to: (1) no learning (MRLP using NBNN classifier), and (2) learning only the γ while keeping the features (MRLP) unchanged.

Dataset	no learning (MRLP + NBNN)	learned classifier (γ)	learned features & classifier (\mathbf{a} , γ)
2-class colonoscopy	82.93 \pm 0.92	82.74 \pm 1.05	86.35 \pm 0.87
3-class colonoscopy	81.42 \pm 0.88	81.78 \pm 2.25	85.45 \pm 1.66
ICPR cells	68.11 \pm 0.35	69.89 \pm 0.49	70.86 \pm 0.37

Table 7.1: Performance (MCA \pm std over experimental runs, refer Section 3.2 for experimental settings) of joint learning: learning both features and the classifier (learning both \mathbf{a} and γ using Equation (7.7)) improves the performance compared to no learning (MRLP + NBNN classifier) and learning only the classifier (MRLP + learning γ using Equation (7.7)).

Table 7.1 reports the classification accuracy for different datasets. Learning both the features (\mathbf{a} of Equation (5.3)) and the classifier ($\{\gamma_c\}, c = 1, \dots, C$ of Equation

(7.5)) considerably improves the MCA for all the datasets. It should be noted that for the colonoscopy datasets learning the classifier does not improve the classification. But for the ICPR dataset, as it contains more classes than colonoscopy, learning the classifier gives modest improvements and learning both the features as well as the classifier improves the MCA by $\sim 2\%$.

In this experiment I followed the same setup given in Section 6.5.1. The learning rate for features (η_a in Algorithm 6) and the classifier was set to 0.5 and 0.01 respectively. The parameters β and λ in Equation (7.7) was set to $\beta = 1 \times 10^{-3}$ and $\lambda = 1 \times 10^{-5}$ respectively.

7.3.2 Sensitivity of the regularisation parameters

The cost function defined in Equation (7.7) contains two free parameters, λ and β , where λ controls the contribution of the intra-class distances between images to different classes weighted by their membership values, and β prevent the feature parameters \mathbf{a} from becoming arbitrarily large, and keeps them closer to $a_s = -1, \forall s$. This section evaluates the sensitivity of these parameters.

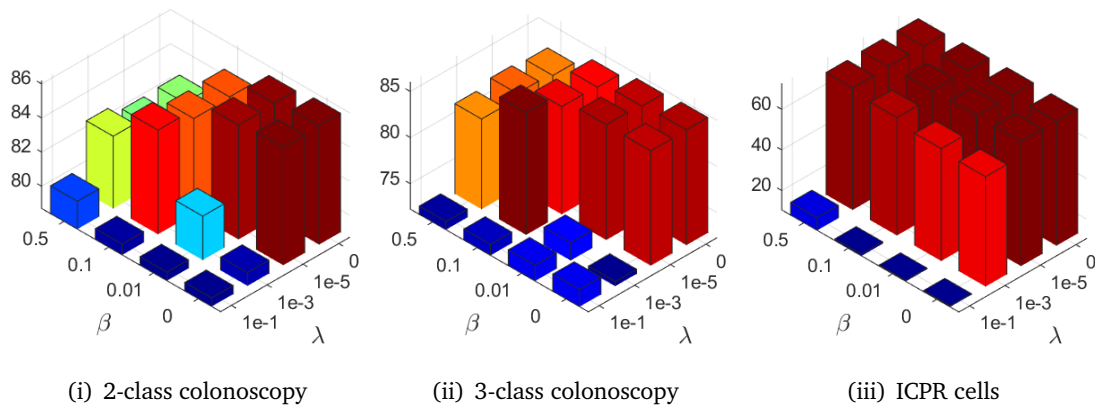


Figure 7.4: Sensitivity of the regularisation parameters: λ and β (Equation (6.14)) vs MCA.

A subset of images (30 and 40 images from the colonoscopy and the ICPR cells dataset respectively) from training set of a particular experimental run was used to learn the feature and classifier parameters. The testing set of that experimental run was used to evaluate the classification performance.

Figure 7.4 reports the MCA for different parameter settings. For all the datasets, setting λ to larger values ($\lambda = 1 \times 10^{-1}$) produces worse performance in comparison with smaller values ($= \{1 \times 10^{-3}, 1 \times 10^{-5}, 0\}$), suggesting that the intra-class distance term in Equation (7.7) is not important. However in Equation (7.7) minimising intra-class distances (D_{ic}) in addition to maximizing the posterior probabilities may reduce overfitting.

Varying β in the range $\{0, 1 \times 10^{-2}\}$ in Equation (7.7) does not affect the classification performance significantly. However setting β to larger values $\{0.1, 0.5\}$ reduces the performance for the colonoscopy datasets. This is because when β getting larger values, the values of α cannot be learned, and will remain closer to initialisation values.

7.3.3 The learned parameters

Figure 7.5 visualizes the learned parameters of the xMRLP_{s_2} features for 3 randomly selected experimental runs for different datasets. As observed in Chapter 6, learning applies different weights to different sampling points. The parameters which were learned at different experimental runs give different values because they were learned using different training dataset.

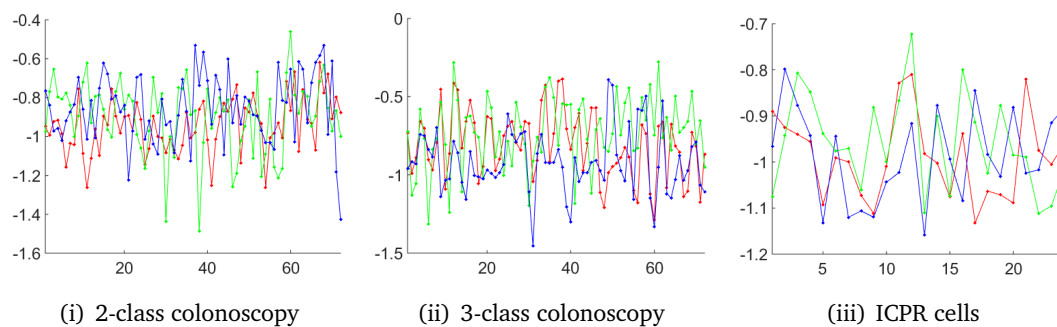


Figure 7.5: Visualisation of the learned parameters: The learned parameter values (vertical axis) at different sampling points (horizontal axis) (Figure 5.2(ii)). Different colours correspond to different experimental runs.

7.3.4 Example probabilistic output

Since the proposed framework can also provide probabilistic outputs for the test images, Figure 7.6 and 7.7 show examples of images from the 3-class colonoscopy dataset which

were correctly and wrongly classified with high confidence.

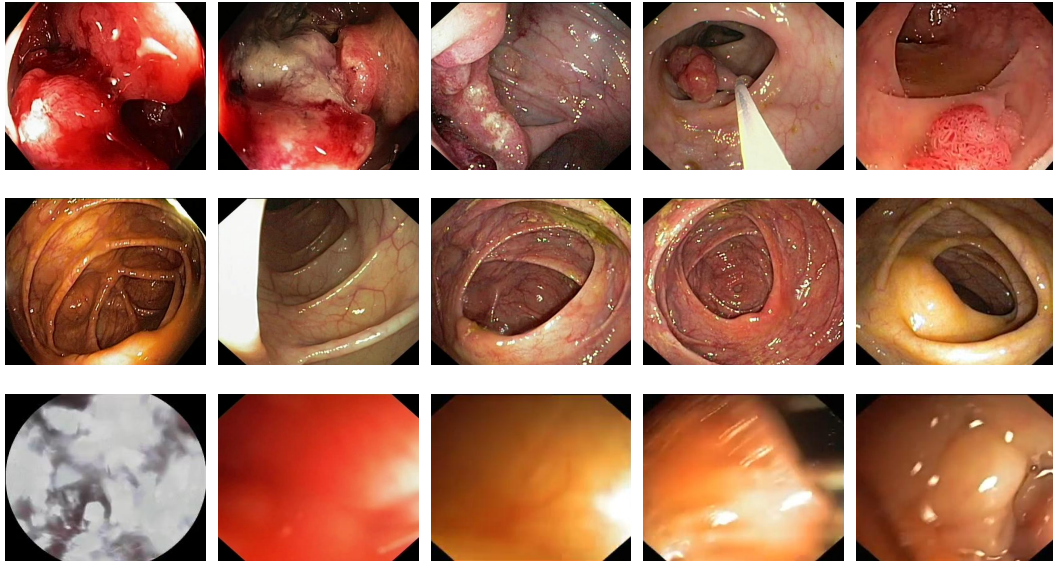


Figure 7.6: Examples of images from the 3-class colonoscopy dataset which were correctly classified with high confidence ($P(\hat{y}_i = y_i) > 0.9$, where, \hat{y}_i is the predicted probability and y_i is the label of I_i). Abnormal, normal and uninformative images are shown in the top, middle and the last rows respectively.

Figure 7.8 shows some example wrongly classified images with low confidence and their confidence values based on the probabilistic soft-max classifier given in Equation (7.5). As can be seen from Figure 7.8 the probabilistic outputs and the wrong classification results are reasonable, as it is hard to assign the difficult images to only one class since they have very similar visual appearance to different classes.

7.3.5 Comparison of different features using the NBNN classifier

Table 7.2 compares the widely used features in computer vision, such as rSIFT and RP with the proposed MRLP, xMRLP_u , xMRLP_s , and xMRLP_{s_2} features for the colonoscopy and the ICPR cells datasets, using the NBNN classifier. xMRLP_u , xMRLP_s and xMRLP_{s_2} performs considerably better than other features for the colonoscopy datasets.

For the colonoscopy datasets we could also observe that the MCA obtained by the weakly-supervised approaches (xMRLP_s and xMRLP_{s_2}) are similar to the MCA obtained by the unsupervised approach (xMRLP_u). The main reason could be the amount of training data used for learning the features. Note that due to computational reasons only 70 and 50 images respectively from the 2-class and 3-class



Figure 7.7: Examples of images wrongly classified with high confidence ($P(\hat{y}_i = y_i) < 0.1$, where, \hat{y}_i is the predicted probability and y_i is the label of I_i) and their predicted probabilities approximated to first decimal place. The top row shows the abnormal images which are wrongly classified into other classes. The middle and the last rows show the wrongly classified normal and uninformative images respectively. The values in the brackets are correspond to $P(\hat{y}_i = \text{abnormal})$, $P(\hat{y}_i = \text{normal})$ and $P(\hat{y}_i = \text{uninformative})$ respectively.

colonoscopy datasets were used to learn the parameters (Section 7.3.1 and Section 6.5.1). Although the weakly-supervised approaches give similar performance compared to the unsupervised approach, in Section 8.2.2.2 I show that the learned features using the weakly-supervised and the unsupervised approaches are complementary to each other, hence improved classification performance can be obtained when combining them using a feature encoding method (experiments can be found in Chapter 8).

The dimensions of the features as well as the computational times required to classify an image for the colonoscopy dataset are reported in Table 7.3. The computational time includes the time needed for feature extraction and NBNN-based classification. Since the dimension of the xMRLP-based features are smaller compared to RP and rSIFT features (Table 7.3), the computational complexity required for feature extraction and I2CD calculations is reduced. It is interesting to note that although the dimension of the MRLP-based features are same, the xMRLP_u features give very low computational time compared to others. This is because the kd-tree

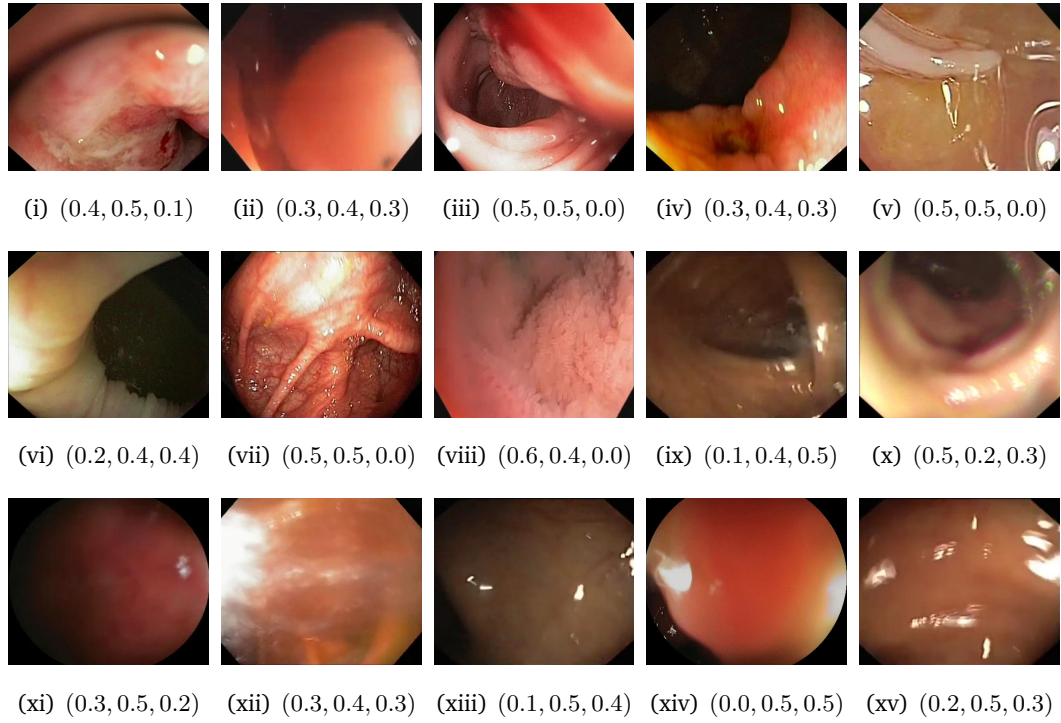


Figure 7.8: Examples of wrongly classified images which are around the classification boundary, and their predicted probabilities approximated to first decimal place. The top row shows the abnormal images which are wrongly classified into other classes. The middle and the last row show the wrongly classified normal and uninformative images respectively. The values in the brackets are correspond to $P(\hat{y}_i = \text{abnormal})$, $P(\hat{y}_i = \text{normal})$ and $P(\hat{y}_i = \text{uninformative})$ respectively.

Feature type	Dataset		
	2-class colonoscopy	3-class colonoscopy	ICPR cells
rSIFT	79.78 \pm 1.29	83.49 \pm 1.74	63.69 \pm 0.50
RP	84.32 \pm 1.62	84.91 \pm 1.20	59.63 \pm 0.46
MRLP	82.68 \pm 0.93	80.80 \pm 0.83	67.90 \pm 0.73
xMRLP _u	86.55 \pm 1.26	86.01 \pm 0.82	68.17 \pm 0.57
xMRLP _s	86.63 \pm 0.99	85.73 \pm 1.16	69.01 \pm 0.94
xMRLP _{s2}	86.35 \pm 0.87	85.45 \pm 1.66	70.86 \pm 0.36

Table 7.2: Classification performance (MCA \pm std over experimental runs, refer Section 3.2 for experimental settings) using different features with NBNN classifier.

	rSIFT	RP	MRLP	xMRLP _u	xMRLP _s	xMRLP _{s2}
time (in sec)	10.42	3.99	1.89	0.52	1.20	1.31
feature dimension	384	200	72	72	72	72

Table 7.3: Average computational time required to classify an image using different features with NBNN classifier for the 2-class colonoscopy dataset. Computational time includes time for feature extraction and classification.

A		N		A		N		A		N	
A	96.3 ± 1.1	3.7 ± 1.1	76.4 ± 3.4	23.6 ± 3.4	90.2 ± 1.8	9.8 ± 1.8					
N	36.7 ± 2.4	63.3 ± 2.4	7.7 ± 1.6	92.3 ± 1.6	24.4 ± 2.5	75.6 ± 2.5					
(i) rSIFT			(ii) RP			(iii) MRLP					
A		N		A		N		A		N	
A	88.4 ± 2.2	11.6 ± 2.2	91.1 ± 1.4	8.9 ± 1.4	86.0 ± 4.0	14.0 ± 4.0					
N	15.3 ± 1.5	84.7 ± 1.5	17.8 ± 2.3	82.2 ± 2.3	13.3 ± 3.0	86.7 ± 3.0					
(iv) xMRLP _u			(v) xMRLP _s			(vi) xMRLP _{s2}					

Table 7.4: Confusion matrices for different features for the 2-class colonoscopy dataset using the NBNN classifier. (A - Abnormal, N - Normal)

A			N			U			A			N			U		
A	96.2 ± 1.7	3.8 ± 1.7	0.1 ± 0.1	72.3 ± 4.7	25.8 ± 4.4	1.9 ± 0.5											
N	33.4 ± 5.1	64.6 ± 4.9	2.0 ± 0.4	7.5 ± 2.0	87.8 ± 2.1	4.6 ± 0.7											
U	4.3 ± 1.7	5.9 ± 1.0	89.8 ± 2.1	1.8 ± 1.0	3.7 ± 0.9	94.6 ± 1.6											
(i) rSIFT						(ii) RP											
A			N			U			A			N			U		
A	86.4 ± 2.3	13.6 ± 2.4	0.1 ± 0.1	86.8 ± 2.9	12.0 ± 3.0	1.2 ± 0.4											
N	23.4 ± 3.0	75.8 ± 3.0	0.8 ± 0.4	16.4 ± 2.0	80.3 ± 2.1	3.3 ± 0.7											
U	6.4 ± 1.2	11.5 ± 2.0	82.1 ± 2.8	3.2 ± 0.9	5.8 ± 1.0	90.9 ± 1.4											
(iii) MRLP						(iv) xMRLP _u											
A			N			U			A			N			U		
A	87.5 ± 3.1	12.1 ± 3.0	0.3 ± 0.3	79.6 ± 6.2	20.1 ± 6.2	0.3 ± 0.3											
N	17.7 ± 3.6	80.6 ± 3.6	1.7 ± 0.6	11.3 ± 4.5	86.0 ± 5.1	2.7 ± 1.0											
U	3.8 ± 1.0	7.4 ± 1.5	88.8 ± 2.1	2.3 ± 1.0	7.0 ± 2.9	90.7 ± 3.5											
(v) xMRLP _s						(vi) xMRLP _{s2}											

Table 7.5: Confusion matrices for different features for the 3-class colonoscopy dataset using the NBNN classifier. (A - Abnormal, N - Normal, U - Uninformative)

algorithm [163] is used for efficient NN retrieval using the clustered features (recall that xMRLP_u features are learned by maximizing cluster compactness). As demonstrated by Maneewongvatana et al. [110, 111] the low-dimensional clustering improves the time needed for kd-tree-based NN search. For the colonoscopy dataset, the learned features not only give improved classification performance but also reduce the computational time required for classification.

Tables 7.4, 7.5 and 7.6 report the confusion matrices for different features for the 2-class colonoscopy, 3-class colonoscopy and the ICPR cells dataset respectively. For the

	Homo.	Spec.	Nucl.	Cent.	NuMe.	Golgi
Homo.	71.0 ± 1.5	17.6 ± 1.5	9.0 ± 1.4	0.0 ± 0.0	2.4 ± 0.6	0.0 ± 0.0
Spec.	17.4 ± 1.5	62.0 ± 1.5	8.9 ± 0.9	0.0 ± 0.0	11.7 ± 1.0	0.0 ± 0.0
Nucl.	4.1 ± 0.5	7.7 ± 1.1	80.4 ± 1.3	1.4 ± 0.6	6.3 ± 0.7	0.0 ± 0.0
Cent.	0.4 ± 0.1	9.2 ± 0.5	10.2 ± 0.8	72.2 ± 0.5	8.0 ± 0.9	0.0 ± 0.0
NuMe.	7.6 ± 1.1	0.3 ± 0.1	0.1 ± 0.1	0.0 ± 0.0	92.1 ± 1.2	0.0 ± 0.0
Golgi	2.7 ± 0.8	10.1 ± 1.5	53.8 ± 2.5	9.3 ± 1.6	19.7 ± 1.8	4.4 ± 0.9
(i) rSIFT						
	Homo.	Spec.	Nucl.	Cent.	NuMe.	Golgi
Homo.	87.7 ± 2.7	9.9 ± 2.2	0.3 ± 0.1	0.0 ± 0.0	2.2 ± 0.8	0.0 ± 0.0
Spec.	34.8 ± 1.8	49.5 ± 1.6	8.4 ± 1.2	0.1 ± 0.1	7.2 ± 1.4	0.1 ± 0.0
Nucl.	1.9 ± 0.5	2.9 ± 1.2	77.3 ± 2.9	0.6 ± 0.2	17.2 ± 2.8	0.1 ± 0.1
Cent.	1.5 ± 0.4	12.2 ± 0.4	25.6 ± 0.7	58.9 ± 0.9	1.8 ± 0.3	0.0 ± 0.0
NuMe.	23.7 ± 1.7	3.9 ± 1.4	1.9 ± 0.3	0.0 ± 0.0	70.2 ± 2.8	0.3 ± 0.1
Golgi	1.5 ± 0.6	10.0 ± 1.7	47.3 ± 2.4	7.9 ± 1.5	19.1 ± 2.6	14.2 ± 1.5
(ii) RP						
	Homo.	Spec.	Nucl.	Cent.	NuMe.	Golgi
Homo.	91.3 ± 1.5	5.1 ± 1.4	3.1 ± 0.8	0.0 ± 0.0	0.4 ± 0.1	0.0 ± 0.1
Spec.	33.8 ± 1.9	51.5 ± 1.9	12.3 ± 0.8	1.4 ± 0.3	0.9 ± 0.2	0.0 ± 0.0
Nucl.	3.0 ± 0.6	3.1 ± 0.6	85.9 ± 0.8	6.9 ± 1.0	0.9 ± 0.3	0.2 ± 0.1
Cent.	1.1 ± 0.2	8.4 ± 0.7	7.1 ± 0.7	83.1 ± 0.6	0.4 ± 0.1	0.0 ± 0.0
NuMe.	19.4 ± 1.7	1.3 ± 0.4	1.8 ± 0.3	0.0 ± 0.0	77.5 ± 1.4	0.1 ± 0.1
Golgi	7.1 ± 1.0	7.8 ± 2.2	34.4 ± 2.6	23.4 ± 1.9	7.7 ± 2.1	19.5 ± 2.2
(iii) MRLP						
	Homo.	Spec.	Nucl.	Cent.	NuMe.	Golgi
Homo.	93.8 ± 1.0	4.6 ± 1.1	1.2 ± 0.3	0.0 ± 0.0	0.3 ± 0.2	0.1 ± 0.1
Spec.	36.7 ± 2.0	48.8 ± 2.4	12.0 ± 1.0	1.2 ± 0.3	1.2 ± 0.3	0.1 ± 0.1
Nucl.	4.0 ± 0.8	3.1 ± 0.8	85.0 ± 0.9	6.4 ± 0.7	1.2 ± 0.4	0.3 ± 0.1
Cent.	1.3 ± 0.3	8.3 ± 0.5	7.6 ± 0.7	82.3 ± 0.9	0.5 ± 0.3	0.0 ± 0.0
NuMe.	20.4 ± 1.5	1.1 ± 0.3	1.5 ± 0.3	0.0 ± 0.0	76.8 ± 1.5	0.2 ± 0.1
Golgi	7.2 ± 0.8	8.0 ± 1.4	32.7 ± 2.7	20.3 ± 2.5	9.4 ± 1.4	22.3 ± 2.2
(iv) xMRLP _u						
	Homo.	Spec.	Nucl.	Cent.	NuMe.	Golgi
Homo.	93.9 ± 0.9	4.2 ± 0.6	1.3 ± 0.4	0.0 ± 0.0	0.4 ± 0.2	0.1 ± 0.1
Spec.	35.7 ± 1.5	49.5 ± 1.7	12.2 ± 1.0	1.4 ± 0.3	1.1 ± 0.3	0.1 ± 0.1
Nucl.	3.5 ± 0.9	2.9 ± 0.6	85.9 ± 1.0	6.3 ± 1.0	1.2 ± 0.4	0.2 ± 0.1
Cent.	1.3 ± 0.3	8.2 ± 0.4	7.4 ± 1.1	82.7 ± 0.9	0.4 ± 0.2	0.0 ± 0.0
NuMe.	18.3 ± 1.7	1.0 ± 0.3	1.7 ± 0.4	0.0 ± 0.0	78.8 ± 1.8	0.1 ± 0.1
Golgi	5.4 ± 1.0	7.3 ± 1.5	34.4 ± 2.6	23.4 ± 4.0	8.8 ± 1.4	20.7 ± 2.8
(v) xMRLP _s						
	Homo.	Spec.	Nucl.	Cent.	NuMe.	Golgi
Homo.	80.1 ± 4.1	15.9 ± 4.3	2.5 ± 1.1	0.0 ± 0.0	1.4 ± 0.9	0.1 ± 0.1
Spec.	18.5 ± 4.0	63.8 ± 4.5	13.8 ± 0.9	3.1 ± 0.9	0.8 ± 0.2	0.1 ± 0.1
Nucl.	1.1 ± 0.4	2.6 ± 0.7	86.8 ± 1.8	8.3 ± 1.1	0.7 ± 0.2	0.4 ± 0.2
Cent.	0.2 ± 0.2	8.7 ± 0.8	6.9 ± 0.9	84.1 ± 1.0	0.1 ± 0.1	0.0 ± 0.0
NuMe.	12.9 ± 1.5	1.0 ± 0.6	2.8 ± 0.5	0.2 ± 0.1	82.7 ± 1.7	0.4 ± 0.2
Golgi	1.5 ± 0.7	4.2 ± 1.6	33.1 ± 3.0	27.7 ± 3.6	5.9 ± 2.0	27.7 ± 2.2
(vi) xMRLP _{s2}						

Table 7.6: Confusion matrices for different features for the ICPR cells dataset using the NBNN classifier.

2-class colonoscopy dataset rSIFT provides a higher TP and a lower TN rate, and RP provides a higher TN and a lower TP rate. On the other hand MRLP, MRLP_u and MRLP_s features give similar TP and TN rates. rSIFT has 36.7% FP, and RP 23.6% FN. MRLP_s, although not beating the best results taken in individual categories (TP, TN, FP, FN), seem to be, numerically, a good compromise at a cheaper computational cost. For the ICPR cells dataset MRLP, MRLP_u and MRLP_s features give higher accuracy for the Golgi classes compared to RP and rSIFT.

7.4 Conclusions and discussion

The I2CD can be easily affected by the noisy local features and the features which are common to different classes (background features). To overcome this problem I propose an approach which applies weights to different classes. I propose a joint learning approach to learn these weights as well as the local features together using data with image-level labels. I experimentally showed that the joint learning framework improves the NBNN-based MCA compared to the approach proposed in Chapter 6 for the ICPR dataset, and gives similar MCA for the colonoscopy datasets, as class weighting is not important for the datasets with small number of classes.

The NBNN classifier is computationally expensive due to the NN search, which limits the amount of local features which should be used for classification. Note that to reduce the computational time required for I2CD calculations I use larger step sizes (16 for colonoscopy and 4 for the ICPR cells datasets) in the training and the testing stages. Extracting features in this way may not capture the discriminative local image properties well. On the other hand, feature encoding approaches are widely used by the computer vision community (e.g. [168, 178]) to get an image-level feature representation. Usually, feature encoding approaches are computationally more efficient than the I2CD calculations. Therefore, after learning the features using the proposed approach, any feature encoding approach could be applied to get a rich image-level representation. The next chapter investigates this together with a learned image-level classifier (e.g. SVM) and shows improved MCA compared to the ones reported in this chapter. The next chapter gives extended experiments with different features as well as different feature encoding approaches. A comparison with the state-of-the-art approaches is also given in Chapter 8.

MRLP AND xMRLP FOR COLONOSCOPY AND CELL IMAGE CLASSIFICATION: EXPERIMENTAL EVALUATION

In Chapters 5 and 6 we proposed a feature, the MRLP, and its extended version called the xMRLP. Chapter 5 and Chapter 7 respectively proposed an unsupervised and a weakly-supervised approach to learn the parameters of xMRLP features. In Chapter 7 we experimentally showed that xMRLP performs considerably better than the hand-crafted features such as MRLP, rSIFT and RP using the NBNN classifier.

In this chapter, we propose two systems based on the proposed features to classify colonoscopy images and cell images into predefined categories. Extended experiments with different features and different encoding methods are given. Comparative experiments with various state-of-the-art systems proposed for colonoscopy and cell image classification show that the proposed approaches outperform the state-of-the art.

8.1 Introduction

In this chapter I describe the automatic systems developed to classify colonoscopy and cell images using the proposed descriptors. I compare the proposed features (MRLP and xMRLP) with widely used features in computer vision for image classification, such

as root-SIFT [13], random projection (RP) [22] and local colour/intensity histograms (LCH) [170], with different feature encoding approaches such as Bag-of-Words (BOW), Locality constrained Linear Coding (LLC), Vector of Locally Aggregated Descriptors (VLAD) and Fisher Vectors (FV). I show that the proposed descriptors (MRLP, xMRLP_u and xMRLP_{s_2}) give better or competitive MCA compared to these baselines. I also compare my approach with various state-of-the-art approaches proposed for colonoscopy and cell image classification and show that the proposed approaches performs better than the state-of-the-art. The datasets used in this chapter are described in Chapter 3.

In the following, first we briefly explain the rSIFT, RP and LCH features, and then we propose an automatic system to classify colonoscopy and cell images respectively and report various comparative experiments.

8.1.1 Root-SIFT

Root-SIFT (rSIFT) is an enhanced variant of SIFT, reported to perform better than SIFT for some image retrieval tasks [13]. The standard SIFT descriptor is a histogram representation of local image derivatives and was originally designed to support comparisons with Euclidean distance. Using the Euclidean distance to compare histograms often yields inferior performance compared to other distance measures such as Chi-squared or Hellinger kernels [21] for texture classification and image categorisation [13]. Therefore, standard SIFT was modified in [13] to create rSIFT so that comparing rSIFT descriptors using Euclidean distance is equivalent to using the Hellinger kernel to compare SIFT vectors.

8.1.2 Random Projection

Random Projection (RP) is a simple yet powerful method for dimensionality reduction [22]. In the case of image analysis, it projects vectors of intensities taken from an image patch from the original patch-vector space \mathbb{R}^D to a compressed space \mathbb{R}^d ($d < D$), using randomly chosen projection directions in the feature space. Such a scheme has been successfully applied, for instance, to texture image classification [100].

Let \mathbf{x} be a D -dimensional patch vector and $\hat{\mathbf{x}}$ be its d -dimensional representation in the compressed space. The RP method simply maps these vectors linearly using a $D \times d$ random projection matrix R , such that:

$$\hat{\mathbf{x}} = R\mathbf{x}. \quad (8.1)$$

Each element in R is sampled from a Gaussian distribution with zero mean and unit variance. The key point of RP is that, when projecting the patch-vectors from the original space to the compressed space, their relative distances are approximately preserved. This allows one to compare distances directly in the compressed space, at a much reduced computational cost.

8.1.3 Local Colour Histogram

Existing approaches for colonoscopy image classification (Section 2.1) use global colour histograms (GCH), i.e. the histogram computed from the whole image as the image representation, which does not capture the local colour information efficiently. Instead, we compute the colour histogram from overlapping local image patches to capture local image properties. Figure 8.1 shows two synthetic images as well as their global and local histogram representations. It is clear that even though the local colour structures are different, the GCH representation gives similar image features. On the other hand, Local Colour Histogram (LCH) gives different image features.

8.2 Colonoscopy image classification

First we explain the proposed system for colonoscopy image classification and then report various comparative experiments.

8.2.1 The proposed system

In Chapter 7 we experimentally showed that the learned xMRLP (xMRLP_u , xMRLP_s and xMRLP_{s2}) features perform considerably better than the hand-crafted features such as MRLP, rSIFT and RP using the NBNN classifier. However, the NBNN classifier is

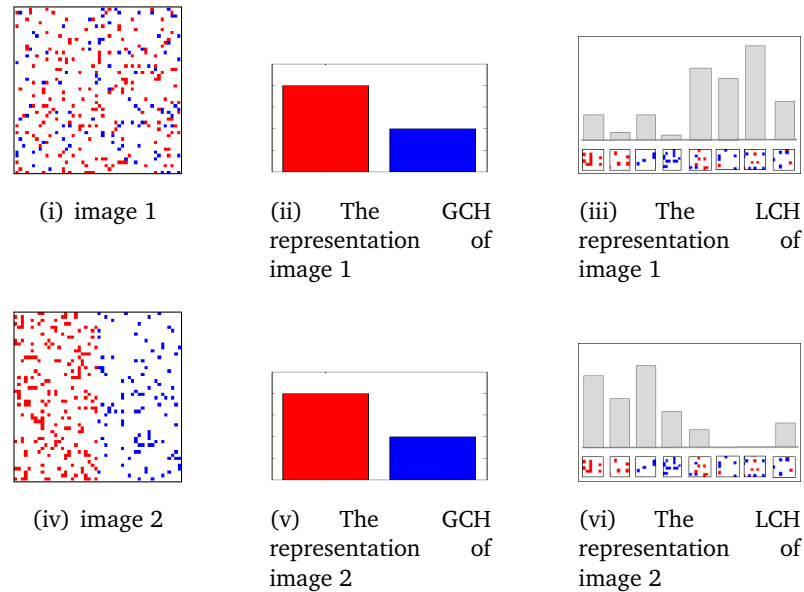


Figure 8.1: Two synthetic images with different local structures and their corresponding GCH and LCH.

computationally expensive due to NN search, which limits the number of features extracted from each image for classification. On the other hand, feature encoding approaches together with the SVM classifier are widely applied in computer vision [168, 178] and medical image analysis literature [120]. Therefore, in this chapter we use a feature encoding approach to get image-level representations based on the local features, and we use a SVM classifier for classifying different image categories.

Since the xMRLP_u and xMRLP_{s_2} features are optimised based on different criteria, they could be complementary to each other, as well as to the MRLP features. Therefore, concatenating the image-level representations obtained by MRLP, xMRLP_u and xMRLP_{s_2} features may give a richer image-level representation than the one obtained individually from any of these. Therefore, in the training stage of the system, first the xMRLP_u and xMRLP_{s_2} features are learned as explained in Chapter 5 and Chapter 7 respectively. Then for each feature type, the image-level representation of a given image is computed based on any feature encoding approach explained in Section 2.2. The normalisation explained in Section 3.2.2 is then applied to normalise each image-level representation. The resultant normalised image-representations computed from the MRLP, xMRLP_u and xMRLP_{s_2} features are then concatenated to obtain the final image-level representation, on which the classification is performed. I call the image-level representation obtained in this way $\text{xMRLP}_{\text{all}}$.

8.2.2 Experiments

This section reports the experiments based on various features, encoding methods and state-of-the-art approaches for the 2-class colonoscopy dataset. Some comparative experiments based on the 3-class colonoscopy dataset are also given in Section 8.2.2.8.

8.2.2.1 Experimental setup

The rSIFT features extracted (patch size 16×16 , overlap 12 pixels) from 3 colour channels were concatenated to get a descriptor of size 3×128 . RP descriptors were extracted in a similar manner. The concatenated vectorised patch representations ($3 \times 16 \times 16$) were projected by a RP matrix to obtain a feature representation of size 200 for each patch. The histograms computed from different colour channels of each patch were concatenated to get a patch-based colour representation for LCH. I applied PCA to reduce the dimensionality of the LCH computed from each patch to 400 as the initial dimension is high (the size of the LCH computed from a 8-bit RGB colour patch is 3×256). The dimension-reduced LCHs were then used for feature encoding.

The k-means algorithm was used for dictionary learning for BOW, LLC and VLAD encoding methods, using a random sample of 200,000 local features from each dataset and feature type. I used the public library, `vfeat`[163], for dictionary learning and feature encoding for BOW, VLAD and FV. For LLC we used the public code provided by the authors of [168]. I used a linear SVM (Lib-Linear [45]) classifier for FV and VLAD as they produce larger image representations (e.g. larger than 5,000). For BOW and LLC approaches we used a SVM with exp-Chi-square kernel [187], and for VLAD and FV we used a SVM with a linear kernel. The parameters of SVM and the kernel were learned based on a 5-fold cross validation on the training data.

8.2.2.2 Comparison of xMRLP-based features

In this section we compare the learned xMRLP_u and xMRLP_{s2} features with the MRLP features based on different feature encoding approaches and a SVM classifier. I also compare the image-level representation obtained by $\text{xMRLP}_{\text{all}}$ with any individual

features (MRLP, xMRLP_u and xMRLP_{s2}) and show that $\text{xMRLP}_{\text{all}}$ outperforms other feature types.

Figure 8.2 reports the performance of the xMRLP-based features. Regardless of the feature encoding approach, the discriminatively learned features xMRLP_{s2} perform better than the hand-crafted one (MRLP). On the other hand, xMRLP_u feature gives improved performance compared to MRLP with BOW and SC, but gives lower performance with VLAD and FV. This is mainly due the additional statistics captured by the VLAD and FV encodings and the features were not learned discriminatively. On the other hand, the combination $\text{xMRLP}_{\text{all}}$ gives considerable improvement compared to any individual feature type. Particularly, when the dictionary size is small, $\text{xMRLP}_{\text{all}}$ outperforms any individual feature type.

BOW and LLC with sum pooling give better performance than other feature encoding approaches. LLC with max pooling gives worse performance than BOW and LLC with sum pooling for smaller dictionary sizes (< 1000). VLAD and FV perform similarly to each other when the dictionary size is large (≥ 32), and FV performs better than VLAD for smaller dictionary sizes (< 32) as it captures additional information compared to VLAD (Section 2.2). However, BOW and LLC with sum pooling performs considerably better than VLAD and FV, and with much smaller image-level feature representation. For example, when xMRLP_s is considered, BOW with dictionary size of 200 gives a MCA of $\sim 91\%$. On the other hand, for the same feature type FV gives similar MCA but with much larger size of the image-level representation (note that, for FV when the dictionary size is 64, the size of the image-level representation is $2 \times 64 \times 72$, where 72 is the dimensionality of the xMRLP_s features).

8.2.2.3 Comparison of different features and encoding methods

Figure 8.3 reports the MCA of rSIFT, RP, LCH, and $\text{xMRLP}_{\text{all}}$ features for different encoding methods.

When BOW and LLC encodings are considered, rSIFT performs similar to RP, and both give modest MCA compared to LCH and $\text{xMRLP}_{\text{all}}$. LCH features give competitive performance compared to the proposed descriptors xMRLP_u and xMRLP_{s2} (Figure 8.2 and Figure 8.3), but, importantly, with much larger patch representations (the

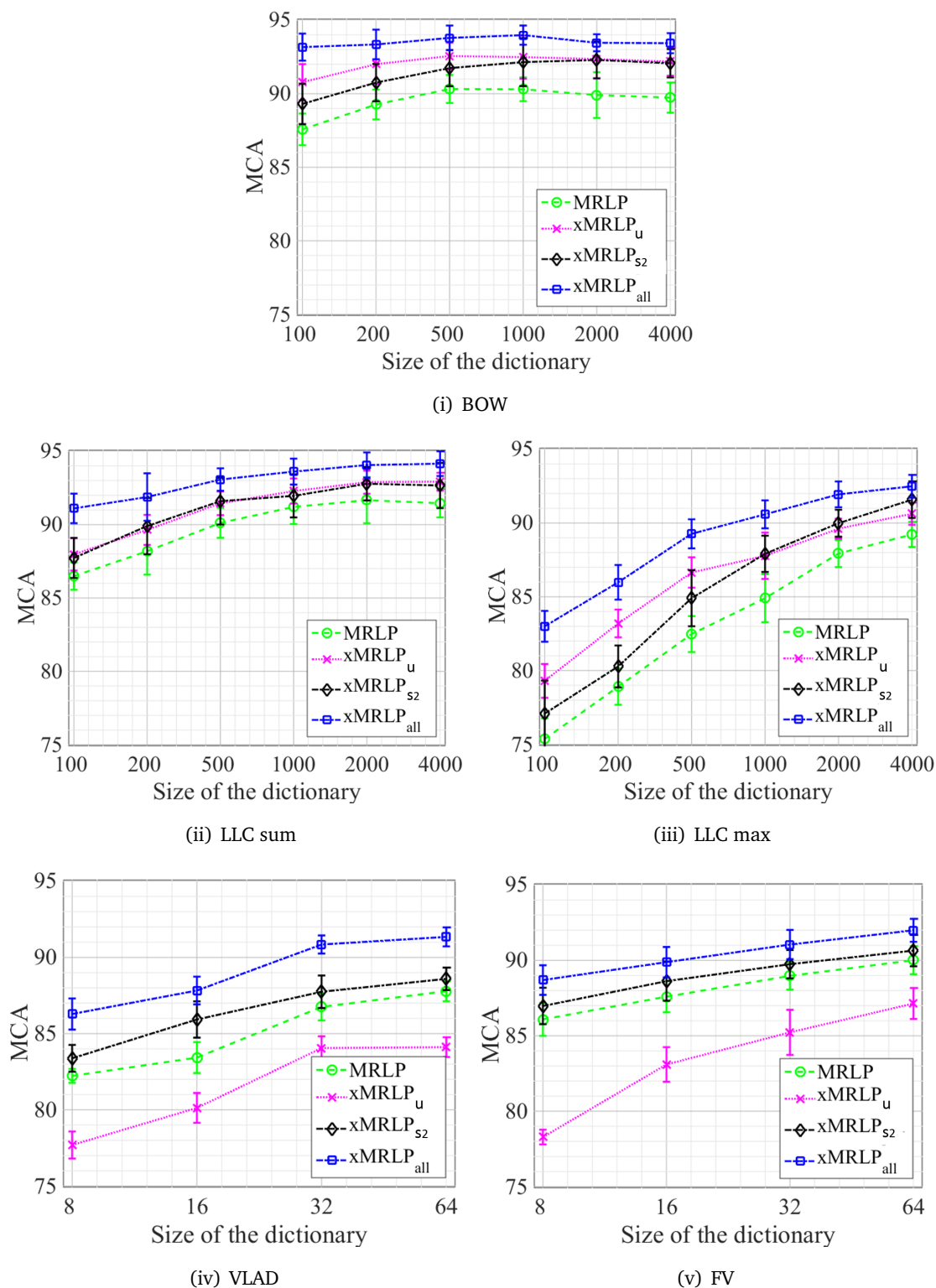


Figure 8.2: Performance (MCA \pm std) of xMRLP-based features with different feature encodings and SVM classifier (size of the dictionary vs MCA).

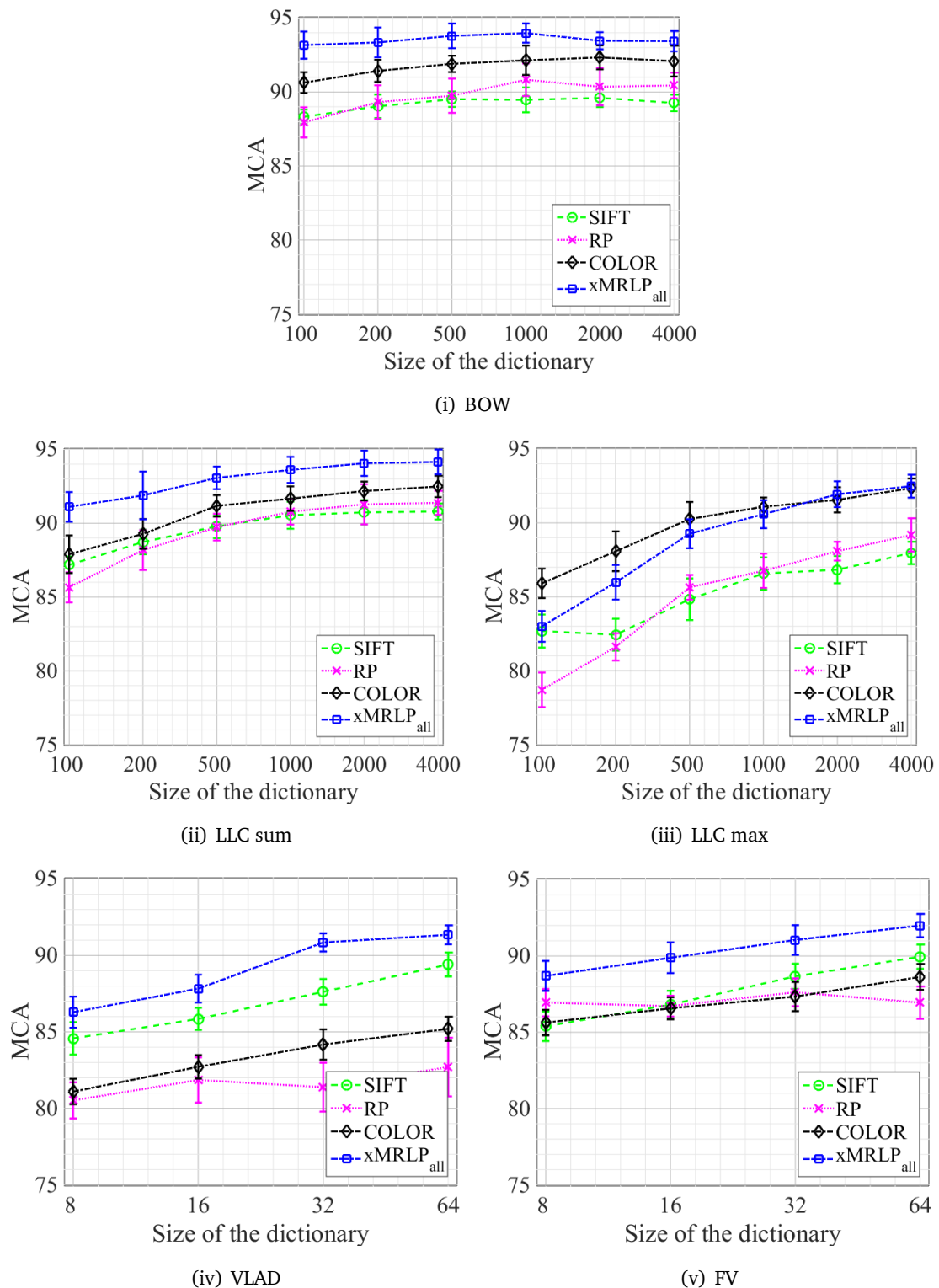


Figure 8.3: Performance (MCA \pm std) of various features with different feature encodings (size of the dictionary vs MCA, COLOR represents the LCH features).

dimensionality of each patch represented by LCH is 400 and of the patch representation by $\text{xMRLP}_u/\text{xMRLP}_{s_2}$ is only 72). On the other hand, regardless of the dictionary size is used with BOW, LLC (with sum pooling), VLAD and FV, $\text{xMRLP}_{\text{all}}$ outperforms LCH, suggesting that the learned features capture discriminative information for classification.

Overall, regardless of the feature type, BOW performs better than LLC when the dictionary size is small; when the dictionary size is large (e.g. 4000) LLC gives similar MCA than BOW. Sum pooling always performs better than max pooling for LLC. FV performs better than VLAD for smaller dictionary sizes (≤ 32) as they capture additional (second-order statistics) information for each cluster.

8.2.2.4 Local vs global colour histograms

This experiment is intended to show that local colour histograms (LCH) for colonoscopy image classification are better than the global one (the histogram obtained from the whole image).

Figure 8.3 reports the classification performance of LCH. Colour histograms computed from whole images (GCH) are widely used for colonoscopy image classification as mentioned in Section 2.1. Therefore we consider this GCH representation as the baseline. I computed colour histograms from each RGB colour channel and concatenated them to represent images. This yields a feature vector of dimension 3×256 . The normalised global histograms computed from training images are used for classifier (SVM with exp-Chi-square kernel) training.

GCH gives a MCA of 86.5 ± 0.7 , which is considerably worse than the MCA 92.3 ± 0.80 (Figure 8.3) obtained by LCH (BOW with dictionary size 2000), suggesting that LCH is more discriminative than GCH as expected, as it captures local information.

8.2.2.5 Computational time required for feature extraction and encoding

Table 8.1 reports the time (in seconds, averaged over 100 images) required for extracting the features and encoding them using a BOW with a dictionary size of 1000. These timings were obtained using Matlab 2014b and Windows 7 running on a machine with a Core i7 processor and 8GB RAM.

Feature type	MRLP	xMRLP _u	xMRLP _{s2}	LCH	rSIFT	RP
Time in sec. for feature extraction	0.10	0.10	0.10	1.79	0.57	0.56
Time in sec. for feature encoding	0.16	0.16	0.16	0.69	0.67	0.36

Table 8.1: Average computational time (averaged over 100 images) required for different features for feature extraction and encoding (BOW with dictionary size of 1000).

For the individual feature types, LCH and the learned features (xMRLP_u, xMRLP_s and xMRLP_{s2}) perform better than other features. It should be noted that the dimensionality of the learned features are much less compared to the dimensionality of LCH (dimensionality of the learned features = 72 vs LCH = 400), which makes computationally efficient for feature extraction and encoding (Table 8.1). The learned features (xMRLP_u and xMRLP_{s2}) not only give better performance compared to rSIFT and RP, but also allow processing ~ 5 frames per second. On the other hand, LCH gives competitive performance compared to the learned features but with much increased computational complexity.

8.2.2.6 Combining features for classification

Since different descriptors may capture complementary information, combining them may improve the MCA. In this experiment we combine different descriptors with each other to check whether they carry complementary information or not. BOW together with the exp-Chi-squared kernel was used in this experiment. The normalised BOW representations computed from different features are combined to get the final image-level feature representation, on which the classification is based on.

Combining different features easily boosts the MCA compared to any individual feature type. Combining xMRLP-based features (xMRLP_{all}) boosts the performance of individual feature types even with much smaller size of the dictionary; MCA of $\sim 93\%$ was obtained with a dictionary size of 100. Although combining other descriptors with LCH improves the MCA, the computational time required by LCH is much higher compared to the time required by xMRLP-based features. The proposed features not only give improved performance but also allow processing more frames compared to LCH (Table 8.1). When combining all the features, the MCA improves by 1% compared to the MCA obtained by xMRLP_{all} (BOW with dictionary size of 100). This is a

Features	Dictionary size		
	100	1000	4000
Single feature type			
rSIFT	88.33 ± 0.48	89.47 ± 0.84	89.26 ± 0.56
RP	87.95 ± 1.02	90.82 ± 1.10	90.44 ± 0.86
LCH	90.63 ± 0.70	92.14 ± 0.99	92.09 ± 1.34
MRLP	87.57 ± 1.08	90.30 ± 0.81	89.73 ± 1.02
xMRLP _u	90.78 ± 1.22	92.48 ± 1.46	92.19 ± 0.96
xMRLP _s	90.50 ± 1.49	91.38 ± 1.10	91.74 ± 0.79
xMRLP _{s2}	89.31 ± 1.38	92.15 ± 1.63	92.07 ± 0.96
Combined features			
xMRLP _u + xMRLP _{s2}	92.74 ± 0.79	93.51 ± 0.86	93.14 ± 0.67
xMRLP _{all}	93.16 ± 0.92	93.98 ± 0.65	93.43 ± 0.68
RP + rootSIFT	92.40 ± 0.83	92.68 ± 0.89	92.09 ± 0.52
LCH + xMRLP _{s2}	92.79 ± 1.08	93.71 ± 0.73	93.46 ± 0.85
LCH + xMRLP _{all}	93.86 ± 0.75	94.11 ± 0.65	93.68 ± 0.69
LCH + RP + rootSIFT	93.96 ± 0.64	93.81 ± 0.74	93.38 ± 0.48
LCH + rootSIFT + RP + xMRLP _{all}	94.52 ± 0.75	94.15 ± 0.72	93.88 ± 0.51

Table 8.2: Classification performance (MCA ± std) of feature combinations (BOW with exp-chi2 kernel).

modest improvement if one considers the computational efforts required. Whether the improvement is worth achieving is ultimately decided in the context of the clinical application, eg, screening for malignant tumours, but this goes beyond the scope of our investigation. When all the features are considered increasing the dictionary size from 1000 to 4000 slightly decreases the MCA, this may be due to the fine partition of the feature space leading to the formation of some noisy clusters, and the hard descriptor-to-cluster assignments of BOW framework.

8.2.2.7 Comparison with state-of-the-art approaches

I considered the following as the baseline features for colonoscopy image classification: global colour histograms (GCH), CWC [67], CWC with higher order statistics (CWC2) [66], GLCM [44], GLCM on wavelet bands (WGLCM) [121] and concatenated colour histograms (CCH) [70]. CWC, CWC2 and CCH are explained in Section 2.1.

Feature	GCH	CCH	CWC	CWC2	GLCM	WGLCM	MR-LBP
M	768	225	216	240	144	144	531
MCA	86.5 ± 0.7	85.1 ± 0.9	76.5 ± 1.0	78.3 ± 1.3	76.7 ± 1.4	77.7 ± 0.9	87.2 ± 1.1

MR-LTP	MR-SILTP	gMR-LTP	rSIFT	RP	xMRLP _{all}	LCH + rSIFT + RP + xMRLP _{all}
1062	1062	1062	1000	1000	300	600
88.8 ± 1.0	90.8 ± 1.3	90.3 ± 1.0	89.5 ± 0.8	90.8 ± 1.1	93.2 ± 0.9	94.52 ± 0.75

Table 8.3: Experimental results (MCA ± std) of the proposed xMRLP-based features and various other features for colonoscopy image classification. M represents the size of the image representation.

GLCM features (energy, entropy, correlation and homogeneity) were extracted from each colour channel (in 4 different directions $\{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$ and 3 different distances $\{1, 2, 3\}$) and concatenated into one vector to get the image representation for GLCM [44]. GLCM (4 directions and 3 distances) features computed on wavelet images were used as the feature for WGLCM [121]. I used our own implementations for the above descriptors, taking care to select parameters to achieve the fairest possible comparison.

I observed that the exponential Chi-squared kernel gave worse performance for most of the baseline approaches. Therefore we used two different kernels, RBF and exponential Chi-square (Equation (4.7)), and report the best performance. The RBF kernel is defined as:

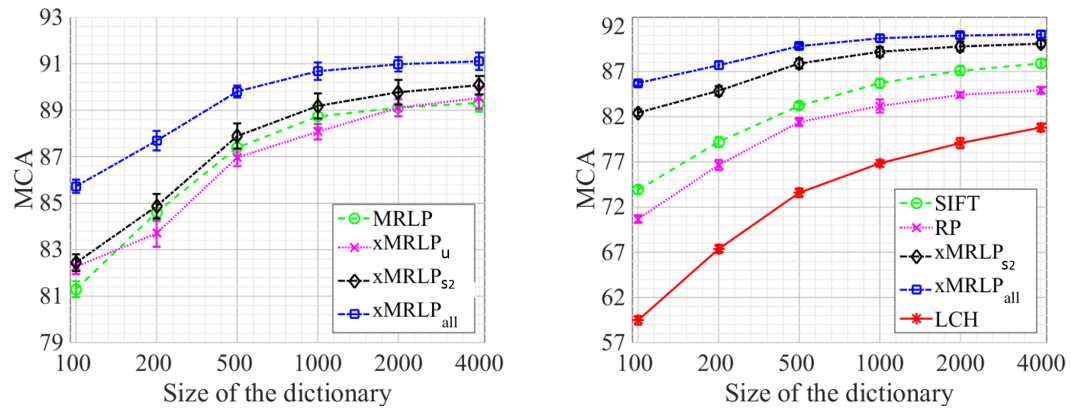
$$K(\mathbf{z}_1, \mathbf{z}_2) = \exp\left(-\frac{\|\mathbf{z}_1 - \mathbf{z}_2\|_2^2}{2\sigma^2}\right) \quad (8.2)$$

where \mathbf{z}_1 and \mathbf{z}_2 are d -dimensional representation of two images. σ is a parameter, which were learned based on a 5-fold cross validation on the training set. For all features other than GCH and CCH, we found that the RBF kernel outperforms the exponential Chi-square kernel.

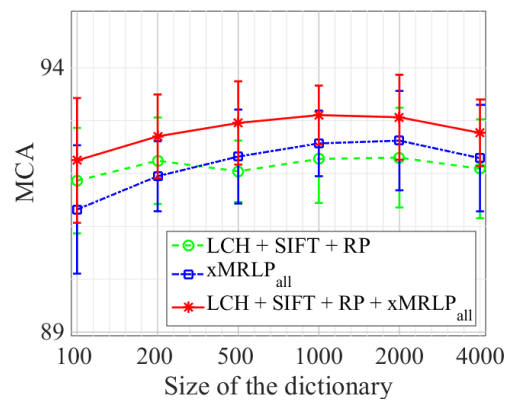
Table(8.3) reports the results of the approaches considered. The proposed learned descriptors (xMRLP_{all}) outperform all the others with a considerable margin and gave state-of-the art results on the 2-class colonoscopy dataset.

8.2.2.8 Experiments on the 3-class colonoscopy dataset

This section compares different features and their combinations on the 3-class colonoscopy dataset. BOW was used for feature encoding.



(i) Comparison of MRLP, xMRLP_u, xMRLP_{s2} and xMRLP_{all} (ii) Comparison of different features with xMRLP-based features



(iii) Performance of feature combinations.

Figure 8.4: Performance (MCA \pm std) of various features for the 3-class colonoscopy dataset (size of the dictionary vs MCA).

Figure 8.4(a) reports the MCA for xMRLP-based features. The learned features xMRLP_u, xMRLP_{s2} performs better than the MRLP features. The combined xMRLP-based features xMRLP_{all} performs better than any individual feature type.

Figure 8.4(b) compare xMRLP_{s2} and xMRLP_{all} features with other features such as LCH, RP and rSIFT. It is clear that xMRLP_{s2} and xMRLP_u performs better than any other individual features, and the combination xMRLP_{all} outperform any other features.

Figure 8.4(c) reports the MCA for different feature combinations. $\text{xMRLP}_{\text{all}}$ performs similar to the MCA obtained by combining LCH + RP + rSIFT features, however, obtaining an image representation using the $\text{xMRLP}_{\text{all}}$ features takes considerably less time compared to any other individual feature type (Table 8.1).

8.3 Cell image classification

In this section first we compare the performance of xMRLP-based features with other features such as rSIFT, LCH, and RP, and then explain the proposed system for cell image classification. Finally we report experiments investigating the effect of different system components and choice of feature representation and encoding.

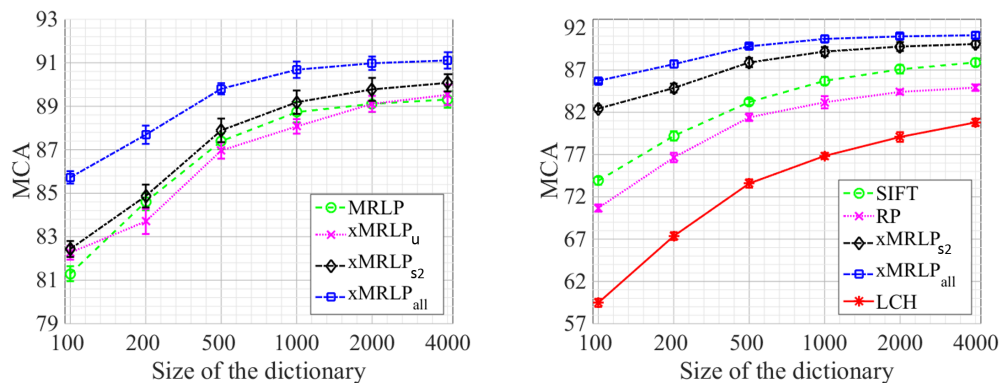
8.3.1 xMRLP-based features for cell image classification

This section compares different features and their combinations on the ICPR cells dataset.

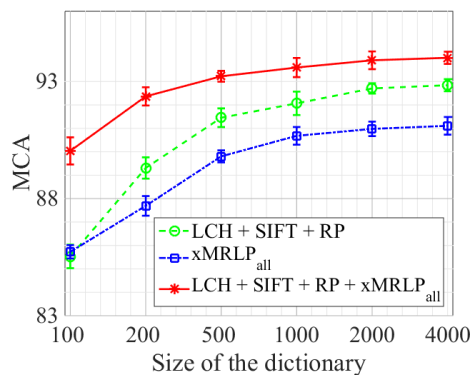
In this experiment, features were extracted densely from 12×12 patches with an overlap of 10 pixels in the horizontal and vertical directions, and the features were extracted from the entire cell images, i.e. the cell masks were not used. BOW was used for feature encoding.

Figure 8.5(a) reports the MCA for xMRLP-based features. The discriminatively learned features xMRLP_{s_2} gives modest improvements over the MRLP features, and the unsupervised feature learning (xMRLP_u) gives similar/reduced performance compared to the MRLP features. However combining them ($\text{xMRLP}_{\text{all}}$) outperforms any individual feature type. Figure 8.5(b) compare xMRLP_{s_2} and $\text{xMRLP}_{\text{all}}$ features with other features such as LCH, RP and rSIFT. It is clear that xMRLP_{s_2} and $\text{xMRLP}_{\text{all}}$ outperform any other features.

Figure 8.5(c) reports the MCA for different feature combinations. $\text{xMRLP}_{\text{all}}$ performs similar to the MCA obtained by combining LCH + RP + rSIFT features, but, note that, obtaining an image representation using the $\text{xMRLP}_{\text{all}}$ features takes considerably less time compared to any other individual feature type.



(i) Comparison of MRLP, xMRLP_u, xMRLP_{s2} and xMRLP_{all} (ii) Comparison of different features with xMRLP-based features



(iii) Performance of feature combinations.

Figure 8.5: Performance (MCA \pm std) of various features for the ICPR cells dataset (size of the dictionary vs MCA).

In our proposed system for cell image classification we use MRLP features instead of xMRLP_{all}. Future work will explore the combination of xMRLP_{all} and other features for cell image classification.

8.3.2 The proposed system

The system to classify ICPR cell images into 6 predefined classes (homogeneous, speckled, nucleolar, centromere, nuclear membrane, golgi) has been developed as a team with other members of the CVIP group. I sincerely thank Wenqi Li, Shazia Akbar, Ruixuan Wang, Jianguo Zhang and Stephen J. McKenna for this collaborative work.

Figure 8.6 gives an overview of the system used for generating a feature representation from an image of a cell for input to a classifier. Each cell image is intensity-normalised. Sets of local features are then extracted and a feature encoding

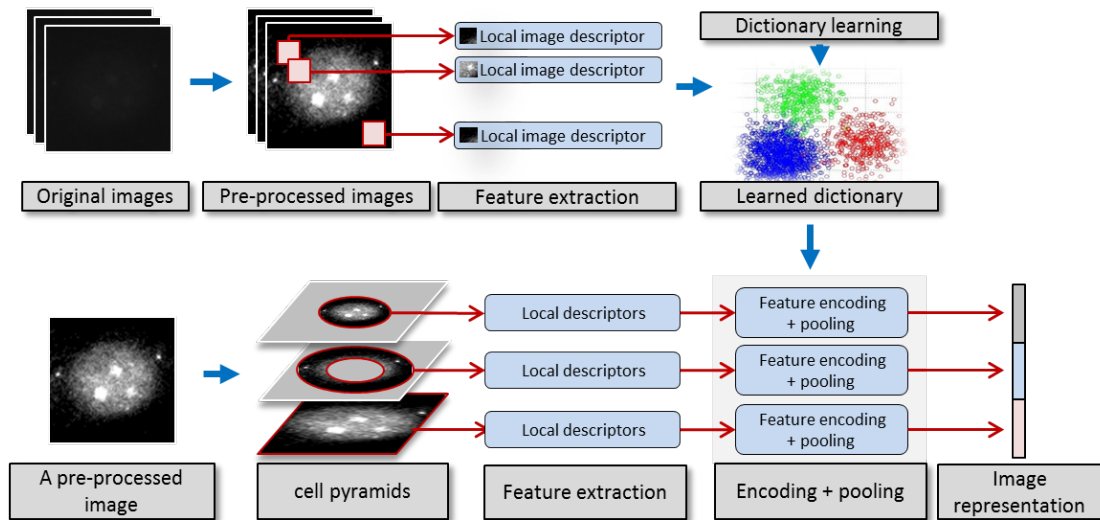


Figure 8.6: An overview of the system for generating the image-level feature representation using only one feature type: Learning dictionary from training images (first row) and feature encoding to obtain the image-level feature representation (second row). The final image representation is a concatenation of the image level representations obtained by different types of features.

method (e.g. LLC) is employed to aggregate the local features into a cell image representation. A two-level cell pyramid is used to capture spatial structure of cell images. An ensemble of SVM classifiers is then used to classify images of cells. The following sections describe these system components in detail.

8.3.2.1 Local feature extraction

Prior to feature extraction, each cell image is intensity normalised; specifically, the segmentation mask is dilated (using a 5×5 structuring element) and the intensity values within each cell's dilated mask region are then linearly rescaled so that 2% of pixels in each cell became saturated at low and high intensities (Figure 8.7).

Local features are extracted densely from each pre-processed cell image. Four types of local features are considered, MRLP, rSIFT, RP and LCH. Since the size of the images in this dataset is small ($\sim 70 \times 70$) and all the images are in grayscale, we use a set of patch sizes (Section 8.3.3) to capture the local image properties.

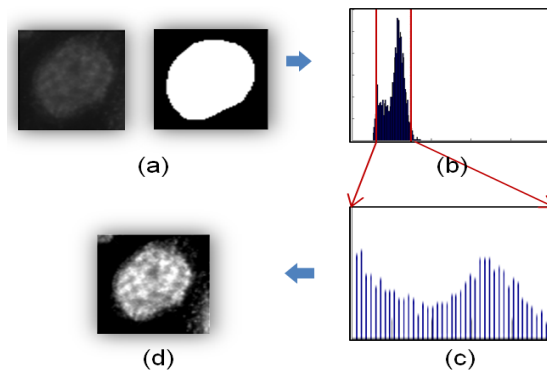


Figure 8.7: Image preprocessing: (a) an example cell image and its mask, (b) histogram of intensity values inside cell region in (a), (c) normalised histogram, (d) preprocessed image.

8.3.2.2 Feature encoding

For each feature type we learn separate dictionaries of size M (a parameter which is varied in our experiments to investigate its effect, Section 8.3.3) using randomly sampled 300,000 local features from training images. Experiments with different feature encoding methods are reported in Section 8.3.3.2.

8.3.2.3 Cell pyramids

Unlike the colonoscopy image dataset, the cell dataset contains some classes which have spatial structure, e.g. Golgi class. This means that the relative position of image elements may be useful for classification; in video colonoscopy images, instead, the unpredictable motion and orientation of the scope, and the unpredictable location of the lesions in the image, make it impossible to use spatial structure for detection and frame classification.

To capture spatial structure within a cell, a 2-level cell pyramid is used in a similar fashion to the dual-region in [172, 173].

After learning the dictionaries, the dictionary-encoded local features are pooled to get an image representation. At the first level of the cell pyramid, the encoded features from the whole cell are pooled to get a feature vector of size P (e.g. for BOW $P = M$). At the second level, feature vectors are pooled from the inner region and from the border region of each cell respectively (see Figure 8.6 and 8.8). The inner cell region and the border region are identified based on eroding and dilating the provided masks

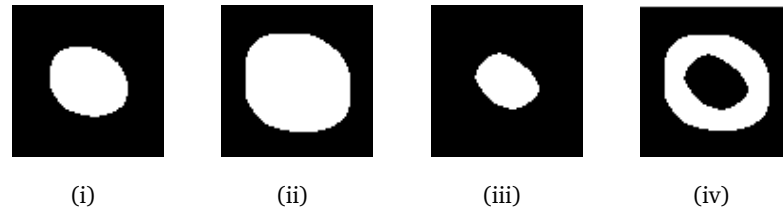


Figure 8.8: Identification of inner and border regions for cell pyramid: (a) an example mask, (b) the inner-region identified by dilating (a) using a structuring element of size 8×8 pixels, (c) erosion of (a), (d) the border-region identified by subtracting (c) from (b).

of the cell images (Figure 8.8). These three feature vectors are concatenated to give a $3P$ -dimensional vector. Finally, encoded features from each of the four feature types are concatenated to give a $12P$ -dimensional vector on which classification was based.

8.3.2.4 Ensemble classifier

Augmenting the classifier's training set with rotated versions of the images may improve the MCA, but it also increases memory requirements to train a multi-class SVM classifier. Instead we used an ensemble of one-vs-rest, multi-class, linear SVMs; the ensemble consisted of four SVMs, one trained on the original training set images, and others trained on images after they are rotated through 90° , 180° , and 270° respectively. The overall system which includes data augmentation as well as the ensemble training is shown in Figure 8.9.

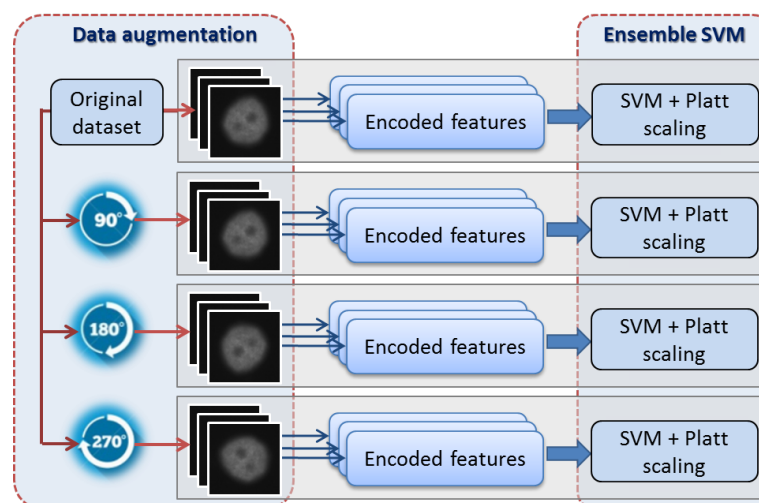


Figure 8.9: An overview of the system for data augmentation and training SVM ensemble. Each image can be encoded as shown in Figure 8.6

At test time, each test image was rotated by 0° , 90° , 180° , and 270° , and each rotated image was then given to the ensemble. This resulted in a set of 16 classification scores for each class ($4 \text{ rotations} \times 4 \text{ SVMs in the ensemble}$). Scores were treated as probabilities using Platt rescaling [134]. The final classification decision was made by averaging these probabilistic scores and selecting the highest scoring class. Figure 8.10 illustrates the process of classifying a cell image in detail.

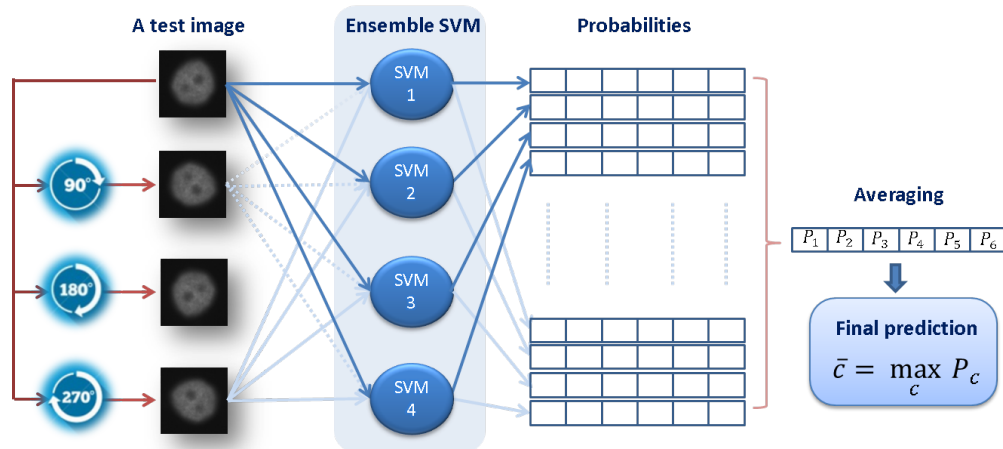


Figure 8.10: Testing an image using the SVM ensemble for single cell classification.

8.3.3 Experiments

8.3.3.1 Experimental setup

From each cell image, each of the feature types was densely extracted from patches of size 12×12 , 16×16 , and 20×20 pixels with a step-size of 2 pixels. For the RP feature the dimension D of each linearised patch was reduced to $d = 300$ whenever $D > 300$. For LCH histograms of 256 bins were used.

8.3.3.2 Comparison of different features and encoding methods

To compare the MCA of different features and encoding methods two-fold cross-validation experiments were carried out, and each was repeated 10 times. Figure 8.11 reports the MCAs for different dictionary sizes. rSIFT gave a slightly better MCA than the other features. LCH gave the worst results. For all encoding methods, larger dictionaries gave higher MCA. LLC with sum pooling always gave better MCA than other

encoding and pooling methods. For all features except LCH, FV performed better than VLAD indicating that the additional (2nd order) information it captured was useful. When the dictionary size was 64, FV obtained similar MCA to LLC with sum pooling with a dictionary size of 4000, but with an increased feature dimensionality. For example, using rSIFT the dimensionality of an FV image representation was 16,384 compared to 4000 using LLC with sum pooling.

8.3.3.3 Combined features for classification

I investigated the performance of combinations of different features. I used BOW and LLC encodings for this purpose as they gave better MCA than VLAD and FV in Experiment 8.3.3.2. The dictionary size was fixed to 1,500. Table 8.4 reports the results (see columns 2-4). Similar MCA was observed using BOW and LLC when combining all four types of feature. An improvement of more than 3% was obtained when combining other features with rSIFT (Figure 8.11 and Table 8.4 columns 2-4). Table 8.6 reports the confusion matrix obtained when combining all the features and encoding with LLC and max-pooling. The Golgi class was the least accurately classified; about 8% of Golgi images were misclassified as nucleolar.

8.3.3.4 Effect of Cell Pyramids

To improve classification accuracy, particularly of the Golgi class, we incorporated spatial structure into the feature encoding process via cell pyramids (CPM). Table 8.4 reports the MCA of different feature combinations *with* and *without* CPM using BOW and LLC approaches (see columns 5-7). When combining all the features and using CPM, the overall MCA was improved by about 1%. In particular, CPM improves the classification accuracy of the Golgi images by about 3% (see Tables 8.6 and 8.7). However this increases the computational time required for feature encoding (Table 8.5).

8.3.3.5 Effect of data augmentation

I investigated the effect of augmenting the training set by including rotated images as explained in Section 8.3.2.4. An ensemble SVM was used for classification. Augmenting the dataset improved the classification accuracy (see Table 8.4 columns 8-10 vs. columns

Features	original dataset without CPM			original dataset with CPM			augmented dataset with CPM		
	BOW	LLC-sum	LLC-max	BOW	LLC-sum	LLC-max	BOW	LLC-sum	LLC-max
rSIFT + MRLP	90.4 ± 0.4	91.0 ± 0.4	90.2 ± 0.4	91.1 ± 0.4	92.0 ± 0.5	91.9 ± 0.4	93.6 ± 0.4	94.1 ± 0.3	94.0 ± 0.5
rSIFT + RP	89.6 ± 0.3	90.6 ± 0.3	89.7 ± 0.4	90.6 ± 0.4	91.9 ± 0.4	91.6 ± 0.3	93.1 ± 0.3	93.9 ± 0.3	93.7 ± 0.3
rSIFT + LCH	91.0 ± 0.4	91.2 ± 0.3	89.9 ± 0.4	92.6 ± 0.4	93.2 ± 0.3	92.7 ± 0.4	94.2 ± 0.3	94.3 ± 0.3	94.1 ± 0.3
all	92.6 ± 0.3	93.1 ± 0.5	92.6 ± 0.4	93.6 ± 0.4	94.1 ± 0.4	94.1 ± 0.4	95.2 ± 0.3	95.2 ± 0.2	95.2 ± 0.2

Table 8.4: Two-fold cross-validation results (MCA ± std) for different feature combinations with and without CPM and data augmentation (dictionary size of 1500).

Features	original dataset without CPM			original dataset with CPM			augmented dataset with CPM		
	feature extraction	feature encoding	total	feature extraction	feature encoding	total	feature extraction	feature encoding	total
MRLP	0.02	0.47	0.49	0.02	0.88	0.91	0.10	3.54	3.64
LCH	0.79	0.54	1.33	0.79	1.06	1.85	3.16	4.15	7.31
SIFT	0.06	0.55	0.61	0.06	1.02	1.08	0.23	4.19	4.42
RP	0.55	0.49	1.04	0.54	0.93	1.46	2.16	3.72	5.88

Table 8.5: Computational time (in sec. averaged over 500 cell images) required for different descriptors for feature extraction and encoding (SC, max pooling, dictionary size of 1500)

8.3.3.6 Computational time for feature extraction and encoding

Table 8.5 reports comparisons of the computational time required for feature extraction and encoding in order to compute the cell-level representations. This was by far the most time consuming part of the proposed system. These timings were obtained using Matlab 2014b and Windows 7 running on a machine with a Core i7 processor and 8GB RAM. LCH took more time than other features while resulting in lower MCA (see Figure 8.11). On the other hand, MRLP took the least time and resulted in competitive MCA. When all feature types were used along with data augmentation and CPM, the system took approximately 21 seconds to compute the cell-level representation for one image.

8.3.3.7 Leave-one-specimen-out experiments

The above experiments discarded the identities of the specimens from which cells had been extracted. To test the generalisation performance of our system across different specimens, we conducted an experiment in a *leave-one-specimen-out* setting. Specifically, we used the specimen IDs to split the data into training and validation sets. Since 83 different specimens were available, we used images from 82 specimens for

	Homo.	Spec.	Nucl.	Cent.	NuMe.	Golgi
Homo.	93.5	05.1	00.3	00.1	00.8	00.1
Spec.	04.6	90.6	01.7	02.2	00.6	00.3
Nucl.	01.0	02.1	94.8	01.0	00.6	00.5
Cent.	00.2	03.4	01.6	94.5	00.1	00.1
NuMe.	02.2	01.1	00.8	00.1	95.1	00.6
Golgi.	01.3	01.3	07.7	01.3	01.5	87.0

Table 8.6: Confusion matrix obtained using all features combined, LLC with max pooling, and dictionary size of 1500. (neither CPM nor data augmentation were used here).

	Homo.	Spec.	Nucl.	Cent.	NuMe.	Golgi
Homo.	94.7	04.2	00.3	00.1	00.6	00.1
Spec.	04.0	91.9	01.4	01.7	00.8	00.2
Nucl.	00.9	01.7	95.5	00.9	00.6	00.5
Cent.	00.1	02.8	01.5	95.5	00.0	00.1
NuMe.	01.9	00.8	00.6	00.1	96.0	00.6
Golgi.	00.9	00.8	05.2	00.6	01.7	90.9

Table 8.7: Confusion matrix obtained using all features combined, LLC with max pooling, dictionary size of 1500, and CPM. No data augmentation was used here).

	Homo.	Spec.	Nucl.	Cent.	NuMe.	Golgi
Homo.	95.5	03.3	00.3	00.1	00.5	00.3
Spec.	04.1	92.1	01.3	01.4	00.8	00.4
Nucl.	00.8	01.2	96.2	00.5	00.5	00.8
Cent.	00.1	02.3	01.5	96.0	00.0	00.1
NuMe.	01.7	00.4	00.6	00.1	96.5	00.8
Golgi.	00.4	00.3	03.1	00.1	01.0	95.0

Table 8.8: Confusion matrix obtained using all features combined, LLC with max pooling, dictionary size of 1500, CPM, and data augmentation

training in each split, and the images from the remaining specimen for testing. In this experiment we used the combination of all feature types, the augmented dataset, CPM, LLC, max-pooling, and dictionary size of 1,500. Table 8.9 reports the confusion matrix. An MCA of 81.1% was obtained. The Golgi class had poor results (66.7%). This class exhibits high intra-class variability and was poorly represented in the available data set; only 4 Golgi specimens were in the training set.

8.3.3.8 Performance on images extracted from Task 2 dataset

As explained in Section 3.1 the ICPR image cell dataset was obtained from the ICPR 2014 HEp-2 cell and specimen image classification challenge¹. In that challenge two tasks

¹<http://i3a2014.unisa.it/>

	Homo.	Spec.	Nucl.	Cent.	NuMe.	Golgi
Homo.	81.8	14.8	00.8	00.2	02.0	00.4
Spec.	09.0	75.5	03.7	10.6	00.8	00.4
Nucl.	01.1	03.4	89.4	02.5	01.3	02.3
Cent.	00.3	10.7	03.4	85.4	00.0	00.2
NuMe.	05.8	01.9	01.5	00.0	87.9	02.8
Golgi.	04.8	02.1	17.4	01.5	07.5	66.7

Table 8.9: Confusion matrix for leave-one-specimen-out experiment. (All features, CPM, data augmentation, LLC, max pooling, dictionary size of 1500).

(Task 1 and Task 2) were proposed to the participants: Task 1 (cell classification) was to classify pre-segmented immunofluorescence images of individual HEP-2 cells into six classes (homogeneous, speckled, nucleolar, centromere, golgi, and nuclear membrane) and Task 2 (specimen classification) was to classify HEP-2 specimen images into seven classes (homogeneous, speckled, nucleolar, centromere, golgi, nuclear membrane and mitotic spindle). Two example specimen images are shown in Figure 8.12(a) and 8.12(c).

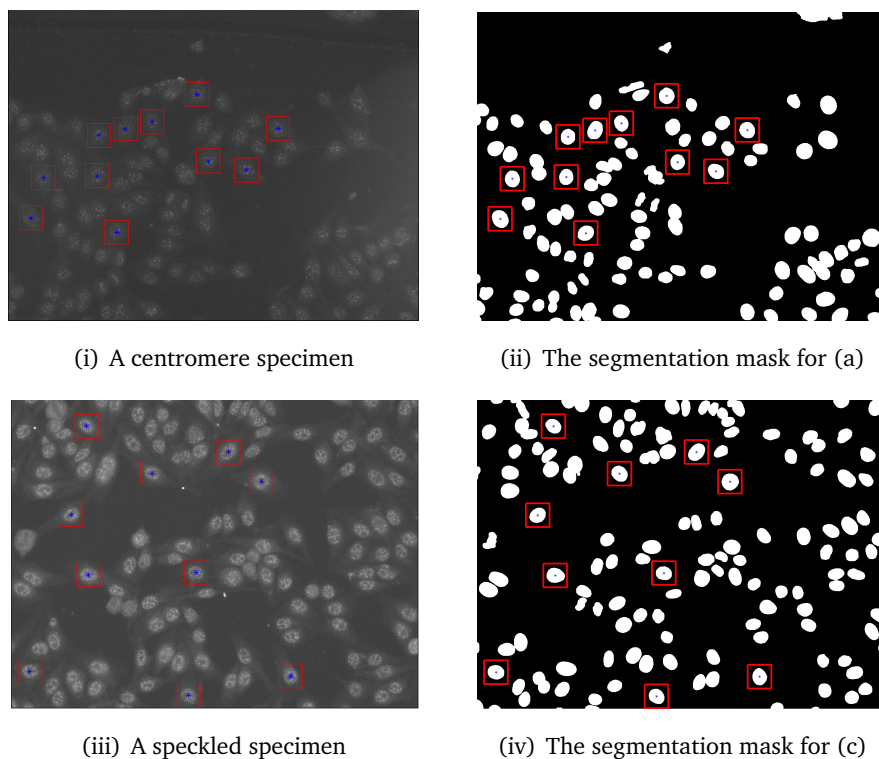


Figure 8.12: Sample specimen images from ICPR Task-2 dataset. The red bounding boxes indicate the cell images which are automatically extracted from these specimen images.

In this experiment we made use of cell images segmented from the Task 2 dataset to investigate the performance of our system on those images (we did not use the *mitotic*

	Homo.	Spec.	Nucl.	Cent.	NuMe.	Golgi
	65.4	28.5	01.5	00.0	03.8	00.7
	04.5	90.8	00.6	01.7	02.2	00.2
	01.2	01.9	95.7	00.0	00.3	00.9
	00.1	11.1	06.5	82.0	00.2	00.1
	03.7	01.9	00.3	00.0	92.0	02.2
	00.0	01.4	03.1	00.2	05.1	90.2

Table 8.10: Confusion matrix of the system trained on Task 1 images and tested on the cell images extracted from Task 2 (LLC with max pooling, dictionary size of 1500).

spindle images in the experiment reported in this Section). An automatic procedure was used to select cells from the Task 2 dataset given the segmentation masks provided with that dataset. Firstly, all disjoint regions were identified in the segmentation mask images using connected component analysis. Secondly, eccentricity values were calculated for each connected component. Finally, low-eccentricity components that could be bounded by an 80×80 square with which no other component overlapped were selected. Approximately 5000 isolated cells were selected in this way. This is illustrated in Figure 8.12 where red bounding boxes denote cell images that were extracted.

I trained an ensemble classifier using all the images from Task 1 training dataset and then tested it on the cell images extracted from the Task 2 dataset. I used the combination of all feature types with the augmented dataset, CPM, LLC and max-pooling (dictionary size of 1,500). The results are reported in Table 8.10; an MCA of 86% was obtained.

8.3.3.9 Comparison with state-of-the-art approaches

The test data of the ICPR challenge was withheld by the organisers. I participated in the ICPR cell image classification challenge by submitting two systems for Task 1; the first system used only data made available in the Task 1 training set; the second system trained on a data set consisting of the Task 1 training set together with the additional 5000 cell images extracted from the Task 2 training set (see Section 8.3.3.8). Both systems used all the features together with LLC (max-pooling, dictionary size 1,500), the rotated versions of the images, and CPM.

Figure 8.13 reports the MCAs obtained by all of the methods submitted to the contest on the Task 1 test set. Our first submission which made use of only the Task 1 training data obtained an MCA of 84.2%, higher than all the other teams' entries. Our

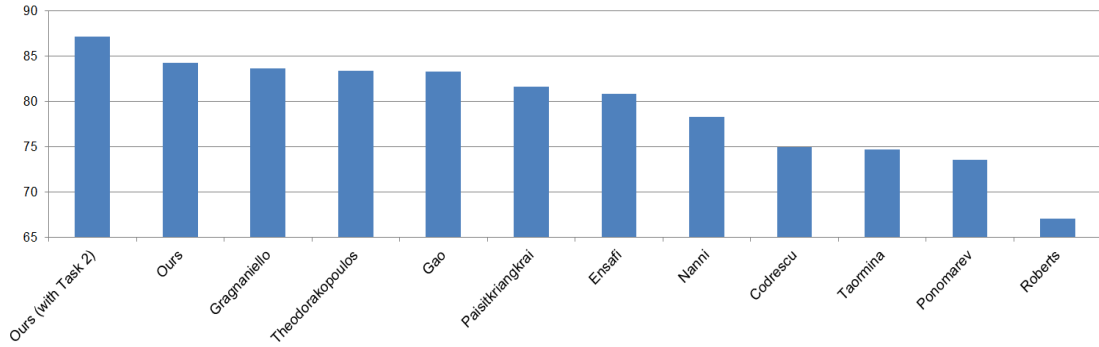


Figure 8.13: The MCA at cell level attained by each method on the test set of Task 1.

	Cent.	Golgi	Homo.	Nucl.	NuMe.	Spec.
Cent.	97.5	00.1	00.5	00.8	00.2	00.9
Golgi.	00.1	82.0	05.4	03.6	08.2	00.7
Homo.	00.2	00.8	82.6	05.7	04.5	06.1
Nucl.	00.8	00.5	01.4	94.8	01.3	01.3
NuMe.	00.1	00.6	04.9	00.7	92.2	01.4
Spec.	10.3	00.5	12.0	02.0	01.5	73.6

(i) Proposed method (trained with cell images from Task 1 and Task 2 training sets)

	Cent.	Golgi	Homo.	Nucl.	NuMe.	Spec.
Cent.	95.5	00.4	00.2	01.2	00.1	02.7
Golgi.	00.0	71.8	04.7	07.3	14.6	01.6
Homo.	00.1	00.8	78.6	04.9	08.1	07.6
Nucl.	00.8	01.6	02.0	92.5	01.7	01.5
NuMe.	00.1	00.8	03.1	00.9	93.3	01.8
Spec.	13.4	00.7	11.1	02.7	02.2	70.0

(ii) Gragnaniello et al. [48]

Table 8.11: Confusion matrices for the proposed method and that of Gragnaniello et al. [48] on the Task 1 test set.

second submission which used additional data (cells extracted from the Task 2 dataset) achieved an MCA of 87.1%. The next best entry, that of Gragnaniello et al. [48], obtained an MCA of 83.6%. Table 8.11 reports confusion matrices from our method and the method of Gragnaniello et al. The reader is referred to the I3A report [102] for detailed results of other entries.

8.4 Conclusions and discussion

In this chapter we proposed a system to classify colonoscopy images into two classes (normal, abnormal) as well as a system to classify HEP-2 cell images into six classes (homogeneous, speckled, nucleolar, centromere, golgi, and nuclear membrane). Our

system for colonoscopy image classification make use of the proposed features (MRLP, xMRLP_u or xMRLP_{s2}), and the system for cell image classification use the MRLP features. I empirically studied different local feature extraction and encoding methods for colonoscopy as well as cell image classification.

I found that:

- The learned features xMRLP_u and xMRLP_{s2} perform better than the baseline features MRLP, rSIFT and RP for colonoscopy images.
- For both datasets, a combination of different features improves performance (i.e. MCA) compared to using only individual features.
- LLC with sum pooling performs better than LLC with max pooling, particularly for smaller size of the dictionaries.
- FV and VLAD could achieve similar accuracies to BOW and LLC but only with feature representations of much higher dimensionality.
- For the cell image classification adding spatial information from the cell images via the use of cell pyramids improves MCA (an improvement of $\sim 3\%$ was observed for Golgi images) and augmenting the training set by the use of rotated training images further improves the MCA.

Overall, comparative experiments with state-of-the-art approaches for colonoscopy and cell images show that our approach outperforms the state-of-the-art.

INTER-CLUSTER FEATURES FOR IMAGE CLASSIFICATION

Feature encoding plays an important role for image classification. Such representations are based on the statistical information within each cluster of local features and therefore fail to capture the inter-cluster statistic, such as how the visual words co-occur in images. This information brings further discriminative power to a feature-based representation. This chapter proposes a new method to choose a subset of cluster pairs based on Latent Semantic Analysis (LSA), and proposes a new inter-cluster statistic which captures richer information than the traditional co-occurrence information. Since the cluster pairs are selected based on image patches rather than the whole images, the final representation also captures local structures. Experiments on medical datasets (ICPR cells and IRMA radiographs, Chapter 3) show that explicitly encoding inter-cluster statistics in addition to intra-cluster statistics significantly improves the classification performance.

9.1 Introduction

The BOW approach is widely applied as a feature encoding method for medical [113] as well as non-medical [180, 181] image classification. In BOW, local features such as SIFT [104] extracted from training images are used to build a dictionary. This dictionary represents a set of visual words (or clusters of features) which are then used to compute a BOW frequency histogram as a feature vector for any given image. BOW captures the simplest *intra-cluster* statistics of each cluster by just counting the

number of local features falling into that cluster (0th-order statistics). VLAD [60] and Fisher Vectors (FV) [129] represent the intra-cluster information by a richer statistical representation compared to BOW. In VLAD, a distance measure between the cluster centre and the features in that cluster is used as the intra-cluster information (1st-order statistics). In addition to the 0th and 1st order statistics, FV also considers 2nd order statistics (i.e. variance for each feature component) [129] *within* each cluster. All the above encoding methods (BOW, VLAD and FV) consider that local features extracted from images independently from each other; none of them captures (1) the *inter-cluster* statistical information (e.g. how two visual words co-occur in each image) and (2) the local structure information of images. Such information can add useful discriminative power to a representation.

Various methods have been proposed in the computer vision literature to add information to these representations. For example, spatial pyramids (SPM) capture spatial information in the images by partitioning the images into fine grids and computing a feature vector from each grid [146]. In this representation, an image is represented as a concatenated feature vector which is obtained from each partition. SPM shows promising performance on natural images. SPM has been adapted by other researchers, e.g. [61] by learning the relevant regions for classification instead of using the fixed partitions as in SPM.

Recently, co-occurrence between visual words have been considered for classification in [180, 181]. Here, co-occurrence of visual words at different partitions of the spatial pyramids are used as the feature vector to capture the co-occurrence as well as the spatial information. In this representation co-occurrence between all the visual words are used as features, leading to a very high-dimensional feature vector (e.g. co-occurrence features from a dictionary of size 100 leads to a dimensionality of $\frac{(100-1)*100}{2}$). A mutual information criterion is used to select the pairs in [31] and the number of co-occurrence of each pairs is used for classification. This method only considers the dependency between two visual words and does not consider any dependency based on higher-order co-occurrence (discussed in Section 9.2). In co-occurrence based methods, the inter-cluster information is represented as the number of co-occurrences between two clusters. In comparison, I show that adding richer inter-cluster statistics performs better than only considering the co-occurrence frequency information as the inter-cluster statistic feature. Beside inter-cluster information,

there is an additional advantage in co-occurrence image representations: since the co-occurrence is obtained from local image regions, the final representation captures some local structure information present in the images, which is not captured by the standard feature encoding approaches.

The co-occurrence of visual words within local image regions has been considered by other researchers as well (e.g. [71] [46]). However, their work focuses on representing objects as a set of parts by building a mid-level feature representation, while ours focuses on extracting inter-cluster statistics.

To capture inter-cluster information, co-occurrences between all pairs of visual words are considered as features for classification [180, 181]. However, this leads to a very high-dimensional feature vector. Including inter-cluster features from pairs of clusters which do not have relevant information for classification may decrease classification performance. Recently a mutual information based criterion has been used to select cluster pairs whose co-occurrence information was then used for classification [31]. However, all these methods [31, 180, 181] only consider the dependency between two visual words (first-order co-occurrence) and failed to consider any higher-order dependencies (discussed in section 9.2). The inter-cluster information in these methods is represented merely as the number of co-occurrence between two clusters. In contrast, I make use of higher-order co-occurrence information to select the informative cluster pairs and encode the inter-cluster features using a richer representation.

As a summary, I propose the following: (1) a new method to select a subset of cluster pairs based on Latent Semantic Analysis (LSA), by considering higher-order co-occurrence of visual words; (2) a patch-based method to construct the term-document matrix in the LSA framework, which can capture structural information of objects in images; (3) A new inter-cluster feature to capture rich statistical information between selected pairs of clusters, which performs better than co-occurrence frequency. I experimentally show that adding inter-cluster statistics (even from a small subset of cluster pairs) improves the performance of medical image classification (ICPR cells and the IRMA radiology images).

9.2 Inter-cluster features

This section focuses on adding inter-cluster statistical information to intra-cluster statistics (e.g. BOW) to represent images. A new method is proposed to choose a subset of cluster pairs by considering the higher-order co-occurrence of visual words within local image regions and introduces an inter-cluster feature which captures rich statistical information between any chosen cluster pairs.

9.2.1 Selection of cluster pairs based on LSA

Latent Semantic Analysis (LSA) is a well-known technique applied to a wide range of tasks such as search and retrieval [43] and classification [186].

Let \mathbf{A} be a *term-document matrix* with t rows (terms) and d columns (documents), where the element $a(i, j)$ represents the frequency of the occurrence of term i in document j . In image analysis, terms correspond to visual words and documents often (but not always, see Section 9.2.2) to images. In this chapter terms and words are used interchangeably. An example term-document matrix is shown in Figure 9.2. In LSA, a low-rank (e.g. rank k) approximation \mathbf{A}_k of matrix \mathbf{A} is obtained by keeping the k largest non-zero singular values in the SVD of \mathbf{A} ($\mathbf{A} = \mathbf{T}\mathbf{S}\mathbf{D}^T$), i.e.

$$\mathbf{A}_k = \mathbf{T}_k \mathbf{S}_k \mathbf{D}_k^T. \quad (9.1)$$

where the t -by- k matrix \mathbf{T}_k , the k -by- k diagonal matrix \mathbf{S}_k , and the d -by- k matrix \mathbf{D}_k are, respectively, the truncated versions of the original matrices \mathbf{T} , \mathbf{S} , and \mathbf{D} . Then the i -th row in $\mathbf{T}_k \mathbf{S}_k$ can be used to represent the meaning of the i -th term (or word) in the so-called k -dimensional latent semantic space, where noise can be largely suppressed by discarding the smaller singular values in \mathbf{S} . Based on such semantic representation of terms, the similarities (correlations) between terms can be captured by the term-term (co-occurrence) matrix [73]:

$$\mathbf{C}_k = \mathbf{T}_k \mathbf{S}_k (\mathbf{T}_k \mathbf{S}_k)^T. \quad (9.2)$$

where each element $C_k(i, j)$ represents the similarity between the i -th and the j -th terms, with higher positive value representing stronger similarity (or positive

correlation) between terms and the lower negative value representing stronger anti-similarity (or negative correlation) between terms.

More importantly, it has been shown that term-term matrix C_k from the truncated matrix $T_k S_k$ can additionally capture *higher-order co-occurrence* information (Figure 9.1) between terms compared to the original co-occurrence matrix (i.e. a matrix where each element (i, j) represents how many times the words i and j co-occur in a document) which is obtained directly from documents [73]. As shown in Figure 9.1, terms t_1 and t_2 , t_2 and t_3 , and t_3 and t_4 respectively co-occur in three different documents. With the original co-occurrence matrix, only the first order co-occurrence was captured and therefore the similarity between terms t_1 and t_3 (also t_2 and t_4 , and t_1 and t_4) will be zero. But there is a relationship between t_1 and t_3 via t_2 . Such higher-order co-occurrence can be captured by the term-term matrix C_k where the corresponding entries won't be zero.

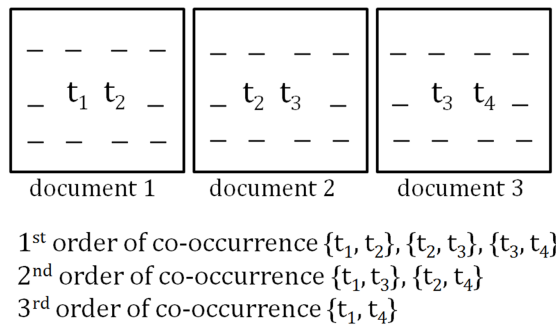


Figure 9.1: High-order co-occurrence.

I propose to select a subset of cluster (or term) pairs which have corresponding larger values in the term-term matrix C_k . As explained above, the use of the truncated term-term matrix C_k instead of the original co-occurrence matrix can help choose the cluster pairs which are semantically similar. In addition, by using a small subset of cluster pairs for inter-cluster feature extraction, richer (in general with higher-dimensional) inter-cluster statistics can be extracted from the selected pairs. Instead, if all the cluster pairs are used for inter-cluster feature extraction as in [181], richer inter-cluster statistics will make feature dimensionality too high to be practically applicable for classifier training.

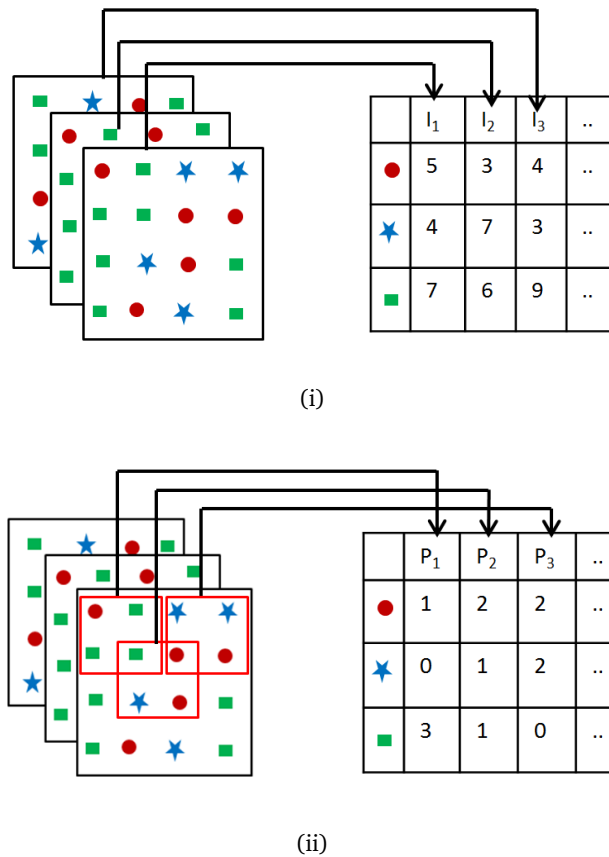


Figure 9.2: Term-document matrix obtained from images (a) and patches (b).

9.2.2 Construction of the term-document matrix

Note that the truncated term-term matrix C_k is obtained from the term-document matrix A . To construct A , in general, each image corresponds to one document and the occurrence of each visual word is counted within the whole image (Figure 9.2a). However, such term-document matrix construction does not consider any spatial relationship (e.g. far from or close to each other) between the corresponding image regions to any two visual words. As a result, the term-term matrix C_k won't contain any information about the spatial relationships between any two visual words. In order to make C_k capture spatial relationship between visual words, I propose to use each image patch (with a certain size) as one document (Figure 9.2b). In this way, the term-term matrix only considers the co-occurrence information between visual words whose corresponding image regions are within the same image patches (therefore close to each other in the image).

By selecting word pairs (i, j) for which the corresponding absolute values of $C_k(i, j)$ are larger in the patch-based term-term matrix C_k , I expect that the selected highly co-occurring word pairs within image patches (i.e. local image regions) will capture certain structural information of objects in an image (e.g. teeth and nose in radiographic images of heads are often close to each other and therefore more likely appear within an image patch). The statistical information between such cluster (word) pairs may implicitly convey such structural information which cannot be captured within each cluster. What's more, the patch-based term-term matrix C_k can also capture the larger-scale structural information (if it exists) by the higher-order co-occurrence information within C_k (e.g. eye balls with teeth via nose).

9.2.3 Inter-cluster statistics

After selecting a subset of word (or cluster) pairs, I need to extract the inter-cluster information based on these pairs. Let W denote the dictionary which contains K visual words $\{w_i\}$, and Π denote the selected subset of word pairs. Given any image, a number of L local descriptors (e.g. SIFT) $X = \{x_l, l = 1, \dots, L\}$ will be extracted from each image patch. Let cluster C_i denote the subset of X such that the nearest visual word for each x_l in C_i is w_i . I consider the two measures explained in the following subsection to capture these inter-cluster statistics.

9.2.3.1 Co-occurrence of visual words

Co-occurrence is a simple measure of how many times a pair of visual words co-occur locally in each image. Consider an image patch within which visual word w_i occurs a times and visual word w_j occurs b times, and the word pair (i, j) is in the selected subset Π . The co-occurrence statistics $f(i, j)$ of these two visual words inside the image patch will be $f(i, j) = \min(a, b)$.

9.2.3.2 Statistical difference between two clusters

For each cluster C_i , the VLAD [60] descriptor v_i is first computed using $v_i = \sum_{x \in C_i} (x - w_i)$. Then for every word pair (i, j) in Π , the inter-cluster statistics is computed as $f(i, j) = \|\frac{v_i}{\sigma_i} - \frac{v_j}{\sigma_j}\|^2$, where σ_i and σ_j are the standard deviations of

the clusters i and j which are computed in the dictionary learning phase. $\|\cdot\|^2$ is a component-wise squared distance measure, and therefore $\mathbf{f}(i, j)$ is a vector and will contain richer statistical information than the scalar co-occurrence value.

9.2.4 Feature encoding

Given an image, I encode the image using both intra-cluster and inter-cluster statistics. First I compute the intra-cluster statistics using the existing approaches such as BOW or VLAD. Then I compute the inter-cluster statistics for image patches in the image as described above. Finally I apply sum pooling over all image patches for the inter-cluster statistics, obtaining a feature vector which represents the inter-cluster statistical information for the whole image. The feature vectors obtained based on the intra and inter-cluster statistics are normalised individually (I use the power and $L2$ normalisations as in [131]) and concatenated together as the final image descriptor.

9.3 Experiments

ICPR cells and the IRMA datasets were used to evaluate the proposed method. The datasets and the experimental settings were explained in Chapter 3.

I used one-vs-rest multi-class SVM with linear and intersection kernels [45] for classification. SVM parameters were learned using 5-fold cross-validation on the training set. The value of k is chosen such that the \mathbf{A}_k keeps 95% of its column-wise variance.

BOW and VLAD features were respectively used as two intra-cluster features based on the local descriptor SIFT, where for each image, dense SIFT descriptors were extracted from each small regions of size 16×16 pixels over a grid with spacing of 4 pixels along both directions, and every 7×7 neighbouring regions compose one image patch (i.e. 49 SIFT features in each patch).

9.3.1 Effect of the inter-cluster features

When using BOW as intra-cluster feature and co-occurrence frequency of visual words as inter-cluster features, Figures 9.3(a) and (b) show that adding inter-cluster features

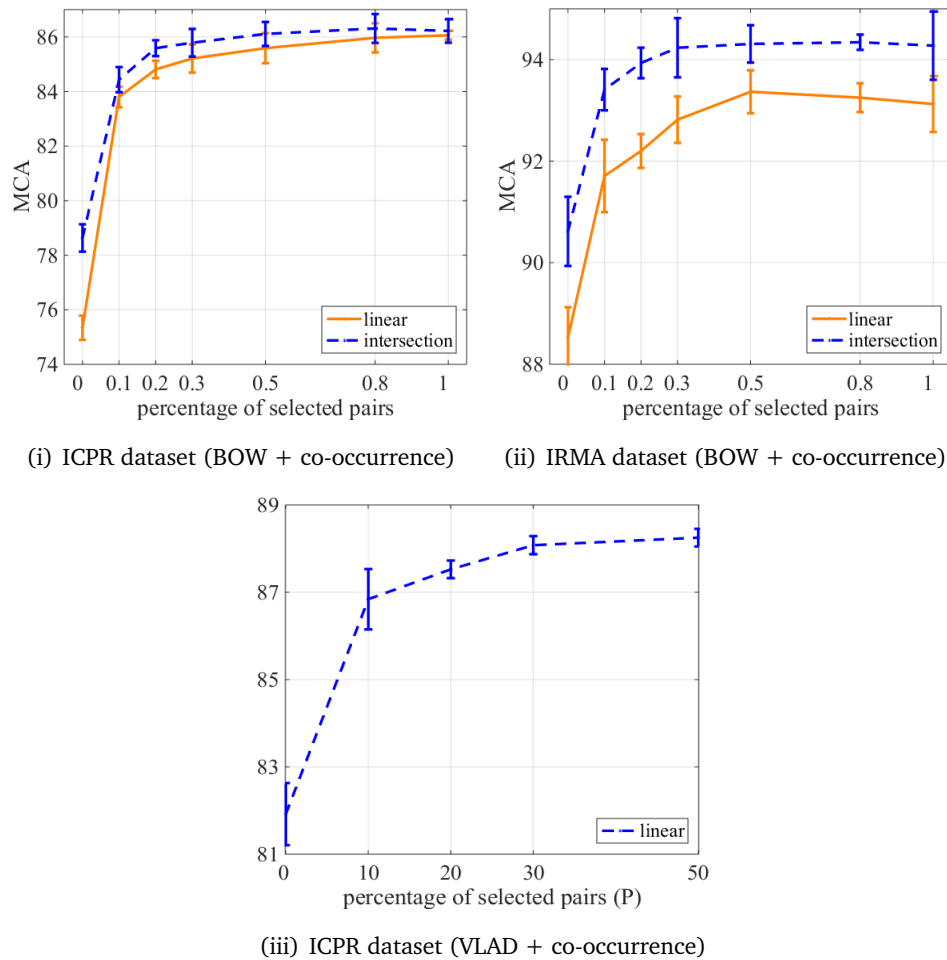


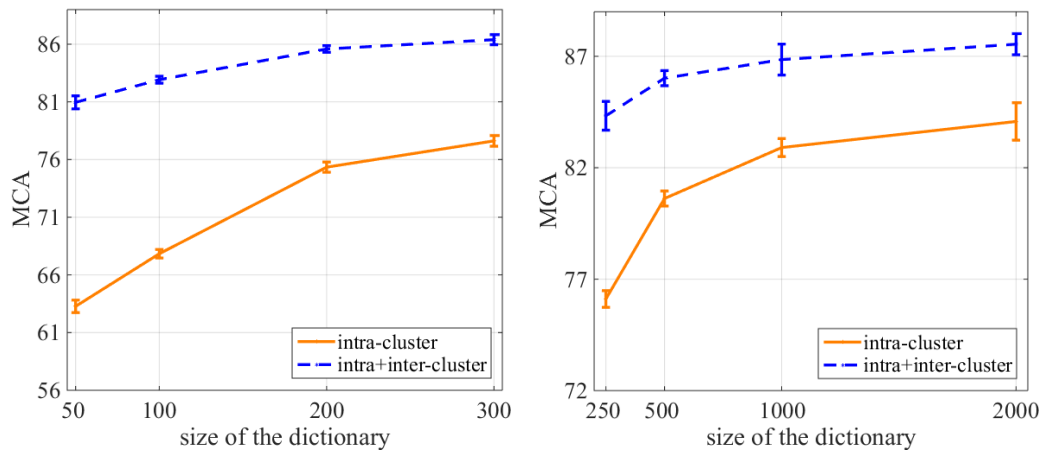
Figure 9.3: Effect of the inter-cluster features. $P = 0$ corresponds to intra-cluster feature (e.g. BOW), and $P > 0$ corresponds to inter-cluster feature plus intra-cluster feature. (a-b) BOW with co-occurrence, (c) VLAD with statistical cluster difference.

significantly increase the classification performance for both datasets (e.g. around 78% when $P = 0$ vs. 86% when $P > 0$ for ICPR dataset, and around 91% vs. 94% for IRMA dataset, both with dictionary size 200 and using intersection kernel). It also shows that the classification accuracy is not significantly different between selecting 10% (when $P = 10$) and all (when $P = 100$) cluster pairs, which indicates that only a small subset of cluster pairs are sufficient enough to capture the inter-cluster information. Figure 9.3(a)(b) also show that intersection kernel for intra-cluster feature cannot capture high-order information encoded in inter-cluster features, otherwise adding inter-cluster features would not improve the accuracy.

Similar findings have been confirmed when using VLAD as the intra-cluster feature and the VLAD-based inter-cluster statistics for the inter-cluster features (Figure 9.3(c)). By comparing the classification performance from Figures 9.3(a) and (c), it becomes

clear that, even using a smaller dictionary ($N = 32$) and a smaller subset of cluster pairs ($P = 10$ percent), VLAD plus VLAD-based inter-cluster features outperforms the corresponding BOW plus co-occurrences based inter-cluster features, i.e. 86.8% vs. 84.4% for ICPR dataset. Similar finding were found for IRMA dataset. This indicates that both VLAD intra-cluster feature and the VLAD-based inter-cluster feature captures richer statistical information than the BOW intra-cluster feature and the co-occurrence based inter-cluster feature.

To further confirm the effect of inter-cluster features, in Figure 9.4(a) the sizes of the dictionaries are varied and only 20% cluster pairs are chosen based on corresponding dictionaries. It shows a significant performance improvement when adding inter-cluster features, no matter what the dictionary size is.



(i) BOW + co-occurrence features for different dictionary sizes; the co-occurrence features were extracted from the corresponding dictionary.

(ii) BOW + co-occurrence features for different dictionary sizes; the co-occurrence features were extracted from a separate dictionary of size 100.

Figure 9.4: Classification performance (MCA + std) on ICPR dataset with BOW and co-occurrence based inter-cluster features using intersection kernel. See text for more details.

Since adding inter-cluster features for larger dictionaries tremendously increases the dimensionality of the final image representation, in another test, I capture inter-cluster features by considering only 20% pairs from a fixed small dictionary of size 100. Adding these fixed inter-cluster features to the traditional intra-cluster BOW features computed from any larger dictionary still increases the overall performance (Figure 9.4(b)). Notice that adding inter-cluster features from a fixed smaller dictionary not only increases the classification accuracy but also reduces the feature dimensionality.

9.3.2 Patch-based vs. image-based methods

This test is to compare the performance of patch-based with the image-based cluster pair selection for inter-cluster feature encoding on the IRMA dataset. For both methods, BOW was used as intra-cluster feature and co-occurrence of selected visual words as inter-cluster feature. The dictionary size was fixed to 200 and only 10% of pairs are selected to encode inter-cluster features. As expected, patch-based method gives the accuracy of 93.4%, much better than the accuracy 87.0% from image-based method (with standard deviation about 0.7%), supporting that patch-based method helps capture local structural information encoded in inter-cluster features.

9.3.3 LSA-based pair selection

In this section the LSA-based truncated term-term matrix is compared with the original co-occurrence matrix for pair selection. In this experiment a dataset containing radiographs of heads taken from four different angles collected from the IRMA dataset is considered. This dataset contains 50 images in each of the four classes. By keeping all the other factors (e.g. patch-based term-document construction and VLAD based inter-cluster feature encoding) unchanged, I found that when selecting a small subset ($P = 5$) of pairs for inter-cluster features, the pair selection based on the truncated term-term matrix performs significantly better than based on the original co-occurrence matrix (78.3% vs. 87.2%). This confirms the potential function of LSA-based pair selection in reducing noise and capturing high-order co-occurrence statistics.

9.3.4 Inter-cluster features for Fisher Vector

Some initial experiments with FV were also performed on the ICPR dataset to observe the effect of inter-cluster features for FV. Given an image, Fisher vector \mathbf{F}_i for each cluster C_i was computed based on soft-assignments (see [129] for details). The inter-cluster feature between any chosen cluster pair (i, j) was computed as $\|\mathbf{F}_i - \mathbf{F}_j\|^2$ (component-wise, as for VLAD). With a total of 16 clusters being used, accuracy of 85.2% was obtained by FV. In comparison, adding inter-cluster features ($P = 20$) to FV significantly improves the performance to 88.7%.

9.4 Conclusions and discussion

This chapter showed that adding inter-cluster features to the intra-cluster features significantly improves medical image classification. A new method was proposed to select a subset of cluster pairs to get the inter-cluster features. Experiments showed that adding rich inter-cluster statistics performs better than only considering the co-occurrence frequency information as the inter-cluster statistical feature. In future work I plan to select cluster pairs based on discriminative information (i.e. class labels) and add spatial information to final representation.

CONCLUSION, DISCUSSION AND FUTURE WORK

In this chapter, I review and summarize the work presented in this thesis, highlighting the key contributions of my research. I also identify the limitations of my system, with their possible causes, before discussing possible extensions and future directions of my research.

10.1 Summary of the thesis

The main aim of this thesis is to build a state-of-the-art automated system to classify colonoscopy images into two classes, normal (healthy, containing no lesions) and abnormal (unhealthy, containing various lesions including polyps, cancers, bleeding, etc.). I mainly focussed on learning discriminative local features as well as computing discriminative image-level representations for classifying images into predefined classes. Our approach to accomplish this task uses data comprising images as well as annotations in the form of image-level labels. Since the proposed techniques are not specific to colonoscopy images, I also reported experiments based on other datasets (see Chapter 3 for the detail of the datasets used), including histology. The following sections summarize the thesis.

10.1.1 Novel feature learning approaches

Chapters 4, 5, 6 and 7 focussed on learning efficient image descriptors for image classification. A comparison with the proposed features and other widely used features

was given in Chapter 8.

Inspired by the success of the LBP descriptor and its variants, in Chapter 4 I proposed a generalisation of LBP called *generalised Local Ternary Patterns* (gLTP). LBP variants such as LTP and SILTP are designed to make LBP resilient to noise and illumination changes respectively, but neither are resilient to *both* noise *and* illumination changes. gLTP, on the other hand, was designed to make LBP resilient to *both* noise *and* illumination changes. I experimentally showed on two datasets (normal/abnormal colonoscopy and ICPR cell images) that the proposed gLTP descriptor gives competitive performance for image-level classification compared to the best performing descriptors (LBP, LTP or SILTP) in these datasets, confirming that gLTP is resilient to both noise and illumination simultaneously.

LBP and its variants, including gLTP, have some limitations, which can be summarised as follows. (1) Losing information due to the binarisation procedure involved in the feature extraction stage; (2) the dimensionality of the LBP-based histogram representation increases with the number of sampling points which is used to compute the LBP-based descriptor; (3) uniform-LBP patterns (see Chapter 5 for details) describe the commonly occurring patterns in the images such as edges, bright and dark spots, and reduce the dimensionality of the image-level representation. However, uniform patterns are identified based on some heuristics; identifying these patterns will be difficult, particularly when the number of sampling points is high.

To overcome these problems Chapter 5 proposed a novel descriptor called the *Multi-Resolution Local Patterns* (MRLP), and its extended version, the *Extended Multi-Resolution Local Patterns* (xMRLP). MRLP gives equal importance to different sampling points. On the other hand, xMRLP treats different sampling points (or neighbourhood pixels) differently by weighing them. Hence, in Chapter 5 I proposed an unsupervised approach to learn these weights, and showed improved performance over LBP-based descriptors.

In Chapter 6 I proposed a weakly-supervised approach to discriminatively learn the parameters of the xMRLP features. This weakly-supervised approach uses image-level labels to learn the local features. Although this approach uses the training labels to learn the feature parameters, the learned descriptor shows similar performance compared to the unsupervised learning approach proposed in Chapter 5.

The approach proposed in Chapter 6 uses the NBNN classifier [23] to learn the parameters of the xMRLP features. The NBNN classifier uses I2CD as the basic element for classification. I2CD can be easily affected by the noisy local features as well as the features from the image background (the features which are common to different classes). To overcome this problem, in Chapter 7 I proposed a feature learning approach based on weighted I2CD, where the I2CD calculated from different classes are weighted differently to learn the local features as well as an image-level classifier. This approach shows similar performance compared to the approaches proposed in Chapters 5 and 6 for the colonoscopy datasets, and improved performance for the cell dataset.

In Chapter 8 I showed that the proposed features perform better than recent, popular features in computer vision such as root-SIFT and Random Projections with BOW-based approaches. Since the parameters of the xMRLP features are learned based on different objectives defined in Chapters 5 and 7, computing an image representation from them could capture complementary information. I showed in Chapter 8 that the image representation obtained in this way outperforms the baseline features (e.g. SIFT, RP, LCH) and with reduced time complexity.

10.1.2 Experimental evaluation

In Chapter 8 I provided extensive comparative experiments using various state-of-the-art features as baselines. Also I proposed two automated systems to classify colonoscopy and cell images into predefined classes, and showed that my systems outperform the state-of-the-art. Unlike existing work reported for colonoscopy, to my best knowledge, I investigated different feature encoding approaches, and found that the BOW and SC often perform better than VLAD and FV with much reduced size of the image-level representations. For the cells dataset I investigated different components of the proposed system in detail (the discussion can be found in Chapter 8).

10.1.3 Inter-cluster statistics for feature encoding

The traditional feature encoding approaches, such as BOW, SC, VLAD and FV capture statistical information within each cluster of local features (intra-cluster features), and do not capture the inter-cluster statistics, such as how the visual words co-occur in

images or image regions. This information could be discriminative for classification. In the computer vision literature, this inter-cluster information is considered for BOW merely based on co-occurrence of visual words [180, 181]. In Chapter 9 I proposed a new approach to choose a subset of cluster pairs to capture the co-occurrence information, and proposed a new inter-cluster statistics which capture richer information than the traditional co-occurrence information. Unlike [180, 181], my approach can be easily extended to other feature encoding approaches such as SC, VLAD and FV, and can capture rich inter-cluster statistical information compared to the frequency-based information used in [180, 181]. I experimentally showed on two medical datasets (ICPR cells and IRMA radiology images) that explicitly encoding inter-cluster statistics in addition to intra-cluster statistics significantly improves the classification performance, and adding the rich inter-cluster statistics performs better than the frequency(co-occurrence)-based inter-cluster statistics. For example, adding inter-cluster statistics for FV representation improves the overall MCA by $\sim 3\%$ for the ICPR cells dataset.

10.2 Key contributions

In this thesis I contributed to the existing literature of colonoscopy image analysis by introducing a novel descriptor and learning algorithms to learn the parameters of the descriptor. I also proposed an approach to improve the traditional feature encoding approaches.

The main contribution of this thesis can be summarised as follows,

- A novel feature, the Generalised Local Ternary Patterns (gLTP), which is a generalised version of LBP and its variants such as LTP and SILTP (Chapter 4).
- Novel descriptors, the Multi-Resolution Local Pattern (MRLP), and its extended version, the Extended Multi-Resolution Local Pattern (xMRLP). An unsupervised (Chapter 5) and weakly-supervised learning algorithms (Chapters 6, and 7) to learn the parameters of xMRLP descriptors.
- Application of feature learning approaches to colonoscopy and histology image classification (Chapters 5, 6, 7, and 8).

- A novel approach to improve traditional feature encoding by adding rich inter-cluster statistical features (Chapter 9).

10.3 Limitations and analysis

In this section, I report the two main limitations of my system, and suggest the possible reasons for these weakness. The limitations are mainly due, arguably, to the characteristics of the colonoscopy datasets.

10.3.1 Classifying colon images from new videos

For the cell image dataset I experimentally showed that the proposed system can work well even for images taken from unseen specimens. However, the colonoscopy datasets used in this thesis are small compared to the cell dataset; they were collected from short video segments (<1 min. long) as well as images obtained from the internet. The images in this dataset do not have video or patient ID information. Without patient characterisation and consequent stratification, my system is trained on randomly sampled set of the colon datasets, which guarantees similar characteristics of the images from the test set present in the training set. It proved impossible, in the course of the work, to evaluate the performance of the proposed colonoscopy image classification system on additional and well-characterised datasets.

10.3.2 Multiple annotations

It is not uncommon that inter-observer variability of annotations on medical images is high. This is due to various factors, including the clinician's expertise and experience. Due to this variability, it is well known that an automated system has to be trained and evaluated on a dataset annotated by multiple human experts. However, obtaining such annotations particularly for medical images is a difficult and time consuming task. Our colonoscopy datasets were annotated by only one clinician. Hence the results reported in this thesis may be biased towards the expert who annotated the datasets. This limitation was caused by clinician availability, and well beyond the author's control.

10.4 Future work

In this thesis, I mainly focussed on computing discriminative local features, and image-level representations for individual frame-level classification. This section suggests some possible future directions for colonoscopy image analysis.

10.4.1 Incorporating temporal information for classification

Our approach and the existing approaches for classifying colonoscopy images assume that the frames are independent of each other, and label each frame independently, without considering the temporal consistency between adjacent frames. Temporal consistency, however, can improve the classification accuracy in the presence of unclear/uncertain images.

There are reasons to expect that temporal consistency will improve the classification. First, some frames are genuinely ambiguous, and a single view will not be sufficient for reliable classification even for experts, whose decisions are based on multiple observations generated by moving the scope. Second, the colonic wall may not be clearly visible in specific frames due to poor illumination, blur due to fast camera movements, and surgical smoke. Third, the appearance of lesions (e.g. scale, orientation) varies in different frames. Fourth, frame-level representations for classification are often obtained by aggregating the statistics of the local features extracted from that frame (e.g. bag-of-visual-words). Such representations may not capture small lesions sufficiently well, vis-à-vis the volume and appearance of background features (extracted from normal tissue).

Therefore, future systems should investigate ways to incorporate temporal consistency in frame classification; this can be either done at the (1) classifier level, where the final decision of a classifier will not only rely on the classification score of a particular frame, but also rely on the scores of adjacent frames; or, at the (2) feature level, where the image-level representation could not only capture the statistics of the local features from a particular frame, but also capture some temporal context information, i.e. the statistics of the local features from adjacent frames. In turn, this could be done by a weighted pooling approach, where the pooled features (e.g. BOW

histograms) from a particular frame, and the pooled features from the adjacent frames are weighted to get the final representation of an image, on which the classification will be done. Again, this system will heavily depend on the characteristics of the training dataset, especially whether consisting of a set of videos or of consecutive frames. I carried out an initial investigation in collaboration with the University of Texas at Austin, and results are being published at MICCAI CARE 2015. These results were not included in this thesis as they are not part of the central objective.

10.4.2 Discriminative inter-cluster features for classification

In Chapter 9 I proposed an approach to capture inter-cluster statistical information in addition to the intra-cluster statistics which is often used by the traditional feature encoding approaches. This information is very useful, particularly for medical images, e.g. the distribution of different cells in cancer as well as in healthy regions may be different. This region-level co-occurrence information cannot be captured by traditional feature encoding approaches, but it can by my approach. However, the latter selects the most frequently co-occurring cluster pairs and encodes their statistics. Although I showed improved performance based on this approach, the most co-occurring cluster pairs, may not be discriminative for classification and may come from background regions. Therefore selecting the cluster pairs which discriminate different classes, and computing statistics from them to get the final image-level representation could improve the classification performance compared to my proposed approach (Chapter 9).

10.4.3 I2CD prototype learning

In Chapters 7 and 6 I proposed weakly-supervised approaches to learn discriminative local features. Our approaches use I2CD as the key element for feature learning. However, as argued in Chapter 7, I2CD can be easily affected by noisy local features, as well as the features from the image background (the features which are common to different classes, and carry no discriminative information). I2CD are also computationally expensive as they require NN search. For this reason, I used only few images (e.g. 70 images from each class for colonoscopy) to learn the local features. This limits the use of the I2CD. On the other hand, a set of discriminative prototypes which describe the discriminative local features from the training images could be learned in

a weakly-supervised manner. Prototype learning approaches based on the supervised training set as well as discriminative dictionary learning approaches are well explored in the computer vision literature (e.g. [30, 143]). Since I learn only a few prototypes, the learned prototypes could dramatically reduce the computational time required for I2CD calculations, and could be robust to noisy local features as well as the features from the image backgrounds.

10.4.4 Lesion localisation and multiple instance learning

Supervised approaches for lesion detection require region-level annotations. e.g. detecting abnormal regions in colonoscopy [91]. Multiple instance learning (MIL) approaches are becoming popular to detect objects [166] or lesions [93] in a weakly-supervised manner. MIL uses image-level labels for lesion detection, hence reduces the region-level annotation efforts needed. To my best knowledge, MIL has not been explored yet for colonoscopy image analysis. Our feature learning approaches can be integrated within MIL settings, where the features together with a MIL classifier can be learned from the weakly-labelled data for lesion detection.

BIBLIOGRAPHY

- [1] “Bowel cancer,” <http://www.nhs.uk/conditions/Cancer-of-the-colon-rectum-or-bowel/Pages/Introduction.aspx>, accessed on 20-Sept-2015.
- [2] “Cancer Research UK,” <http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer>, accessed on 20-Sept-2015.
- [3] “CODIR - Colon Disease Investigation by Robotic Hydro-colonoscopy,” www.codir.org, accessed on 20-Sept-2015.
- [4] “Colonoscopy,” <http://www.webmd.com/colorectal-cancer/guide/colonoscopy-what-you-need-know>, accessed on 20-Sept-2015.
- [5] “Confocal laser endomicroscopy,” <http://clinicaltrials.gov/ct2/show/NCT00561938>, accessed on 20-Sept-2015.
- [6] “Gastrolab-the gastrointestinal site,” <http://www.gastrolab.net/>, accessed on 20-Sept-2015.
- [7] “Importance of colorectal cancer screening,” <http://www.cancer.org/cancer/colonandrectumcancer/>, accessed on 20-Sept-2015.
- [8] “Tests to detect colorectal cancer and polyps,” <http://www.cancer.gov/cancertopics/factsheet/detection/colorectal-screening>, accessed on 20-Sept-2015.
- [9] A. Alahi, R. Ortiz, and P. Vandergheynst, “FREAK: Fast Retina Keypoint,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.
- [10] S. Ameling, S. Wirth, D. Paulus, G. Lacey, and F. Vilarino, “Texture based polyp detection in colonoscopy,” in *Bildverarbeitung für die Medizin*, 2009.
- [11] B. Andre, T. Vercauteren, A. Perchant, A. Buchner, M. Wallace, and N. Ayache, “Endomicroscopic image retrieval and classification using invariant visual features,” in *IEEE International Symposium on Biomedical Imaging*, 2009.
- [12] —, “Introducing space and time in local feature-based endomicroscopic image retrieval,” in *Medical Content-Based Retrieval for Clinical Decision Support*, ser. Lecture Notes in CS, 2010, vol. 5853, pp. 18–30.

- [13] R. Arandjelović and A. Zisserman, “Three things everyone should know to improve object retrieval,” in *IEEE Computer Vision and Pattern Recognition*, 2012.
- [14] S. Baheerathan, F. Albrechtsen, and H. Danielsen, “New texture features based on the complexity curve,” *Pattern Recognition*, vol. 32, no. 4, pp. 605 – 618, 1999.
- [15] M. Baker, L. Bogoni *et al.*, “Computer-aided detection of colorectal polyps: can it improve sensitivity of less-experienced readers? preliminary findings,” *Journal of Radiology*, vol. 245, pp. 140–149, 2007.
- [16] G. Balathasan, X. Yuan, J. Liu, B. Bill, J. Oh, and S. J. Tang, “Bleeding detection from capsule endoscopy videos,” in *IEEE Conference on Engineering in Medicine and Biology Society*, 2008.
- [17] R. Banerjee and D. N. Reddy, “Advances in endoscopic imaging: Advantages and limitations,” *Journal of Digestive Endoscopy*, vol. 3, pp. 7–12, 2012.
- [18] R. Barclay, J. Vicari, A. Doughty, J. Johanson, and R. Greenlaw, “Colonoscopic withdrawal times and adenoma detection during screening colonoscopy,” *The New England Journal of Medicine*, vol. 355, pp. 2533–2541, 2006.
- [19] C. Becker, R. Rigamonti, V. Lepetit, and P. Fua, “Supervised feature learning for curvilinear structure segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2013, vol. 8149.
- [20] S. Bejakovic, R. Kumar, T. Dassopoulos, and G. H. Gerard Mullin, “Analysis of crohn’s disease lesions in capsule endoscopy images,” in *IEEE international Conference on Robotics and Automation*, 2009.
- [21] A. Bhattacharyya, “On a measure of divergence between two statistical populations defined by their probability distributions,” *Bulletin of the Calcutta Mathematical Society*, vol. 35, no. 1, pp. 99–109, 1943.
- [22] E. Bingham and H. Mannila, “Random projection in dimensionality reduction: applications to image and text data,” in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2001, pp. 245–250.
- [23] O. Boiman, E. Shechtman, and M. Irani, “In defense of nearest-neighbor based image classification,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.
- [24] Y. Boureau, N. Le Roux, F. Bach, J. Ponce, and Y. LeCun, “Ask the locals: multi-way local pooling for image recognition,” in *IEEE International Conference on Computer Vision*, 2011.
- [25] B. Bressler, L. F. Paszat, Z. Chen, D. M. Rothwell, C. Vinden, and L. Rabeneck, “Rates of new or missed colorectal cancers after colonoscopy and their risk factors: A population-based analysis,” *Journal of Gastroenterology*, vol. 132, no. 1, pp. 96–102, 2007.

- [26] M. Brown, G. Hua, and S. Winder, "Discriminative learning of local image descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 43–57, 2011.
- [27] L. Cao, R. Ji, Y. Gao, Y. Yang, and Q. Tian, "Weakly supervised sparse coding with geometric consistency pooling," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.
- [28] T. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma, "PCANet: A simple deep learning baseline for image classification," *IEEE Transactions on Image Processing*, vol. 24, pp. 5017–5032, 2014.
- [29] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.
- [30] F. Chang, C. Lin, and C. Lu, "Adaptive prototype learning algorithms: Theoretical and experimental studies," *Journal of Machine Learning Research*, vol. 7, pp. 2125–2148, 2006.
- [31] T. Chen, K.-H. Yap, and L.-P. Chau, "From universal bag-of-words to adaptive bag-of-phrases for mobile scene recognition," in *International Conference on Image Processing*, 2011, pp. 825–828.
- [32] E. Ciaccio, G. Bhagat, C. Tennyson, S. Lewis, L. Hernandez, and P. Green, "Quantitative assessment of endoscopic images for degree of villous atrophy in celiac disease," *Journal of Digestive Diseases and Sciences*, vol. 56, pp. 805–811, 2011.
- [33] A. Coates and A. Y. Ng, "Learning feature representations with k-means." in *Neural Networks: Tricks of the Trade (2nd ed.)*. Springer, 2012, vol. 7700, pp. 561–580.
- [34] A. Coates, A. Y. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 2011.
- [35] F. Cohen, Z. Fan, and M. Patel, "Classification of rotated and scaled textured images using gaussian markov random field models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 13, no. 2, pp. 192–202, 1991.
- [36] M. Coimbra, P. Campos, and J. Cunha, "Extracting clinical information from endoscopic capsule exams using MPEG-7 visual descriptors," in *The 2nd European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology*, 2005, pp. 105–110.
- [37] —, "Topographic segmentation and transit time estimation for endoscopic capsule exams," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, 2006, pp. II–II.
- [38] M. Coimbra and J. Cunha, "MPEG-7 visual descriptors contributions for automated feature extraction in capsule endoscopy," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 5, pp. 628–637, 2006.

- [39] L. Cui, C. Hu, Y. Zou, and Meng, "Bleeding detection in wireless capsule endoscopy images by support vector classifier," in *IEEE International Conference on Information and Automation*, 2010.
- [40] F. DA, M. JT, and e. a. Ben-Menachem T, "Complications of colonoscopy," *Journal of Gastrointestinal Endoscopy*, vol. 74, pp. 745–752, 2011.
- [41] J. P. da Silva Cunha, M. T. Coimbra, P. Campos, and J. M. Soares, "Automated topographic segmentation and transit time estimation in endoscopic capsule exams." *IEEE Transactions on Medical Imaging*, vol. 27, no. 1, pp. 19–27, 2008.
- [42] D. de Ridder and R. P. Duin, "Locally linear embedding for classification," Pattern Recognition Group, Dept. of Imaging Science and Technology, Delft University of Technology, Delft, The Netherlands, Tech. Rep., 2002.
- [43] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of The American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.
- [44] S. Engelhardt, S. Ameling, D. Paulus, and S. Wirth, "Features for classification of polyps in colonoscopy," in *CEUR Workshop Proceedings*, 2010.
- [45] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.
- [46] B. Fernando, I. Fromont, and T. Tuytelaars, "Effective use of frequent itemset mining for image classification." in *European Conference on Computer Vision*, 2012.
- [47] H. Goh, "Learning Deep Visual Representations," Ph.D. dissertation, University Pierre et Marie Curie - Paris, 2013.
- [48] D. Gragnaniello, C. Sansone, and L. Verdoliva, "Biologically-inspired dense local descriptor for indirect immunofluorescence image classification," in *Pattern Recognition Techniques for Indirect Immunofluorescence Images (I3A), 1st Workshop on*, 2014, pp. 1–5.
- [49] A. Hafiane, G. Seetharaman, and B. Zavidovique, "Median binary pattern for textures classification." in *ICIAR*, ser. Lecture Notes in Computer Science, Springer, vol. 4633, 2007.
- [50] R. M. Haralick, "Statistical and structural approaches to texture," in *Proceeding of IEEE*, vol. 67, 1979.
- [51] C. Harris and M. Stephens, "A combined corner and edge detector," in *In Proceedings of Fourth Alvey Vision Conference*, 1988.
- [52] D. C. He and L. Wang, "Texture unit, texture spectrum, and texture analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 28, pp. 509–512, 1990.

- [53] J. He, A. H. Tan, C. L. Tan, and S.-Y. Sung, *On Quantitative Evaluation of Clustering Systems*. Kluwer Academic Publishers, 2003.
- [54] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, pp. 504–507, 2006.
- [55] G. Hua, M. Brown, and S. Winder, "Discriminant embedding for local image descriptors," in *IEEE International Conference on Computer Vision*, 2007.
- [56] D. Huang, C. Shan, M. Ardabilian, Y. Wang, and L. Chen, "Local binary patterns and its application to facial image analysis: A survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 41, no. 6, pp. 765–781, 2011.
- [57] S. Hwang, J. H. Oh, W. Tavanpong, J. wong, and P. C. de Groen, "Polyp detection in colonoscopy video using elliptical shape feature," in *IEEE International Conference on Image Processing*, 2007.
- [58] A. K. Jain and F. Farrokhnia, "Unsupervised texture segmentation using Gabor filters," *Pattern Recognition*, vol. 24, no. 12, pp. 1167–1186, 1991.
- [59] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "What is the best multi-stage architecture for object recognition?" in *IEEE International Conference on Computer Vision*, 2009.
- [60] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *IEEE Computer Vision and Pattern Recognition*, 2010.
- [61] Y. Jia, C. Huang, and T. Darrell, "Beyond spatial pyramids: Receptive field learning for pooled image features," in *IEEE Computer Vision and Pattern Recognition*, 2012.
- [62] H. Jin, Q. Liu, H. Lu, and X. Tong, "Face detection using improved LBP under bayesian framework," in *IEEE First Symposium on Multi-Agent Security and Survivability*, 2004, pp. 306–309.
- [63] Y. S. Jung, Y. H. Kim, D. H. Lee, and J. H. Kim, "Active blood detection in a high resolution capsule endoscopy using colour spectrum transformation," in *IEEE Conference on BioMedical Engineering and Informatics*, 2008.
- [64] A. Karargyris and N. Bourbakis, "A methodology for detecting blood-based abnormalities in wireless capsule endoscopy videos," in *IEEE BioInformatics and BioEngineering*, 2008.
- [65] A. Karargyris and N. G. Bourbakis, "Detection of small bowel polyps and ulcers in wireless capsule endoscopy videos," *IEEE Transactions on Biomedical Engineering*, vol. 58, pp. 2777–2786, 2011.
- [66] S. Karkanis, K. Galousi, and D. Maroulis, "Classification of endoscopic images based on texture spectrum," in *Workshop on Machine Learning in Medical Applications, Advance Course in AI*, 1999.

- [67] S. A. Karkanis, D. K. Iakovvidis, D. E. Maroulis, D. A. Karras, and M. Tzivras, "Computer aided tumour detection in endoscopic video using colour wavelet features," *IEEE transactions on IT in biomedicine*, vol. 7, pp. 141–152, 2003.
- [68] Y. Ke and R. Sukthankar, "PCA-SIFT: A more distinctive representation for local image descriptors," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2004.
- [69] D. Kersten, *Statistical limits to image understanding*. Cambridge University Press, 1990, pp. 32–43.
- [70] P. C. Khun, Z. Zhuo, L. Z. Yang, L. Liyuan, and L. Jiang, "Feature selection and classification for wireless capsule endoscopic frames," in *International Conference on Biomedical and Pharmaceutical Engineering*, 2009.
- [71] S. Kim, X. Jin, and J. Han, "DisIClass: discriminative frequent pattern-based image classification," in *International Workshop on Multimedia Data Mining*, 2010.
- [72] V. Kodogiannis and M. Boulougoura, "Neural network-based approach for the classification of wireless-capsule endoscopic images," in *IEEE International Joint Conference on Neural Networks*, vol. 4, 2005.
- [73] A. Kontostathis and W. M. Pottenger, "A framework for understanding latent semantic indexing performance," *Journal of Information Processing and Management*, vol. 42, no. 1, pp. 56–73, 2006.
- [74] S. Krishnan, X. Yang, K. Chan, S. Kumar, and P. Goh, "Intestinal abnormality detection from endoscopic images," in *IEEE Conference on Engineering in Medicine and Biology Society*, 1998.
- [75] R. Kumar, P. Rajan, S. Bejakovic, S. Seshamani, G. Mullin, T. Dassopoulos, and G. Hager, "Learning disease severity for capsule endoscopy images," in *IEEE International Symposium on Biomedical Imaging*, 2009.
- [76] R. Kumar, Q. Zhao, S. Seshamani, G. Mullin, G. Hanger, and T. Dassopoulos, "Assessment of crohn's disease lesions in wireless capsule endoscopy images," *Journal of Biomedical Engineering Online*, vol. 59, pp. 352–362, 2012.
- [77] P. Y. Lau and P. Correia, "Detection of bleeding patterns in WCE video using multiple features," in *IEEE Conference on Engineering in Medicine and Biology Society*, 2007.
- [78] Q. V. Le, M. Ranzato, R. Monga, M. Devin, G. Corrado, K. Chen, J. Dean, and A. Y. Ng, "Building high-level features using large scale unsupervised learning," in *Proceedings of the 29th International Conference on Machine Learning*, 2012.
- [79] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, 1998.
- [80] Y.-G. Lee and G. Yoon, "Bleeding detection algorithm for capsule endoscopy," in *World academy of Science, Engineering and Technology*, 2011.

- [81] Z. Lei and S. Li, "Learning discriminant face descriptor for face recognition," in *Asian Conference on Computer Vision*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013, vol. 7725.
- [82] Z. Lei, M. Pietikainen, and S. Z. Li, "Learning discriminant face descriptor," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 2, pp. 289–302, 2014.
- [83] F. Leung, J. Leung, R. Siao-Salera, and S. Mann, "The water method significantly enhances proximal diminutive adenoma detection rate in unsedated patients," *Journal of interventional gastroenterology*, vol. 1, pp. 8–13, 2011.
- [84] F. Leung, J. Leung, R. Siao-Salera, S. Mann, and G. Jackson, "The water method significantly enhances detection of diminutive lesions (adenoma and hyperplastic polyp combined) in the proximal colon in screening colonoscopy - data derived from two RCT in US veterans," *Journal of interventional gastroenterology*, vol. 1, pp. 48–52, 2011.
- [85] J. Leung, S. Mann, R. Siao-Salera, C. Ngo, R. McCreery, W. Canete, and F. Leung, "Indigocarmine added to the water exchange method enhances adenoma detection - a RCT," *Journal of Interventional Gastroenterology*, vol. 2, pp. 106–111, 2012.
- [86] J. Leung, S. Mann, R. Siao-Salera, K. Ransibrahmanakul, B. Lim, W. Canete, L. Samson, R. Gutierrez, and F. W. Leung, "A randomized, controlled trial to confirm the beneficial effects of the water method on U.S. veterans undergoing colonoscopy with the option of on-demand sedation," *Clinical Endoscopy*, vol. 73, pp. 103–10, 2010.
- [87] S. Leutenegger, M. Chli, and R. Siegwart, "Brisk: Binary robust invariant scalable keypoints," in *IEEE International Conference on Computer Vision*, 2011.
- [88] B. Li and M. Meng, "Computer aided detection of bleeding in capsule endoscopy images," in *Canadian Conference on Electrical and Computer Engineering*, 2008.
- [89] —, "Computer-aided detection of bleeding regions for capsule endoscopy images," *IEEE Transactions on Biomedical Engineering*, vol. 56, pp. 1032–1039, 2009.
- [90] M. Li, "Detection of bleeding patterns in WCE video using TV-Retinex," *Journal of Biomedical Science and Engineering*, vol. 3, pp. 1143–1145, 2010.
- [91] P. Li, K. L. Chan, and S. Krishnan, "Learning a multi-size patch-based hybrid kernel machine ensemble for abnormal region detection in colonoscopic images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.
- [92] P. Li, K. L. Chan, S. Krishnan, and Y. Gao, "Detecting abnormal regions in colonoscopic images by patch-based classifier ensemble," in *IEEE International Conference on Pattern Recognition*, 2004.

- [93] W. Li, J. Zhang, W. Zheng, M. Coats, F. A. Carey, and S. J. McKenna, "Learning from partially annotated OPT images by contextual relevance ranking," in *International Conference on Medical Image Computing and Computer Assisted Intervention*, 2013.
- [94] S. Liao, G. Zhao, V. Kellokumpu, M. Pietikainen, and S. Li, "Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.
- [95] S. Liao, X. Zhu, Z. Lei, L. Zhang, and S. Z. Li, "Learning multi-scale block local binary patterns for face recognition," in *Proceedings of the 2007 International Conference on Advances in Biometrics*. Springer-Verlag, 2007, pp. 828–837.
- [96] S. Liao, X. Zhu, Z. Lei, L. Zhang, and S. Li, "Learning multi-scale block local binary patterns for face recognition," in *Advances in Biometrics*, ser. Lecture Notes in CS, 2007, vol. 4642, pp. 828–837.
- [97] M. Liedlgruber and A. Uhl, "Computer-aided decision support systems for endoscopy in the gastrointestinal tract: A review," *IEEE Reviews in Biomedical Engineering*, vol. 4, pp. 73–88, 2011.
- [98] C. Lima, D. Barbosa, A. Ramos, A. Tavares, L. Montero, and L. Carvalho, "Classification of endoscopic capsule images by using colour wavelet features, higher order statistics and radial basis functions," in *IEEE Conference on Engineering in Medicine and Biology Society*, 2008.
- [99] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
- [100] L. Liu and P. Fieguth, "Texture classification from random features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 3, pp. 574–586, 2012.
- [101] X. Liu, D. Tosun, M. W. Weiner, and N. Schuff, "Locally linear embedding (lle) for mri based alzheimer's disease classification." *Journal of NeuroImage*, vol. 83, pp. 148–157, 2013.
- [102] B. C. Lovell, G. Percannella, M. Vento, and A. Wiliem, "Performance evaluation of indirect immunofluorescence image analysis systems," *International Conference on Pattern Recognition Workshop*, Tech. Rep., 2014.
- [103] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [104] D. Lowe, "Object recognition from local scale-invariant features," in *IEEE International Conference on Computer Vision*, 1999.
- [105] A. Luis, Alexandre, J. Casteleiro, and N. Nobre, "Polyp detection in endoscopic video using svms," in *Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases*, 2007.

- [106] M. Mackiewicz, M. Fisher, and C. Jamieson, "Bleeding detection in wireless capsule endoscopy using adaptive colour histogram model and support vector classification," *SPIE Medical Imaging*, vol. 6914, pp. 69 140R–69 140R–12, 2008.
- [107] M. Madhoun and W. Tierney, "The impact of video recording colonoscopy on adenoma detection rates," *Journal of Gastrointestinal Endoscopy*, vol. 75, pp. 127–133, 2012.
- [108] J. Mairal, J. Ponce, G. Sapiro, A. Zisserman, and F. R. Bach, "Supervised dictionary learning," in *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2009, pp. 1033–1040.
- [109] P. Majewski and W. Jedruch, "Endoscopy images classification with kernel based learning algorithms," in *Innovations in Applied AI*, ser. Lecture Notes in CS, 2005, vol. 3533.
- [110] S. Maneewongvatana and D. M. Mount, "Analysis of approximate nearest neighbor searching with clustered point sets," *Data Structures, Near Neighbor Searches, and Methodology*, vol. 59, pp. 105–123, 2002.
- [111] S. Maneewongvatana and D. Mount, "On the efficiency of nearest neighbor searching with data clustered in lower dimensions," ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2001, vol. 2073, pp. 842–851.
- [112] T. Mang, G. Hermosillo, M. Wolf, L. Bogoni, M. Salganicoff, V. Raykar, H. Ringl, M. Weber, C. Mueller-Mang, and A. Graser, "Time-efficient CT colonography interpretation using an advanced image-gallery-based, computer-aided "first-reader" workflow for the detection of colorectal adenomas," *European Society of Radiology*, vol. 22, pp. 2768–2779, 2012.
- [113] S. Manivannan, R. Wang, E. Trucco, and A. Hood, "Automatic normal-abnormal video frame classification for colonoscopy," in *IEEE International Symposium on Biomedical Imaging*, 2013.
- [114] S. Manivannan and E. Trucco, "Learning discriminative local features from image-level labelled data for colonoscopy image classification," in *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*, 2015.
- [115] S. Manivannan, W. Li, S. Akbar, R. Wang, J. Zhang, and S. J. McKenna, "Hep-2 cell classification using multi-resolution local patterns and ensemble svms," in *I3A 1st workshop on Pattern Recognition Techniques for Indirect Immunofluorescence Images, IEEE International Conference on Pattern Recognition*, 2014.
- [116] —, "Hep-2 specimen classification using multi-resolution local patterns and svm," in *I3A 1st workshop on Pattern Recognition Techniques for Indirect Immunofluorescence Images, ICPR (invited paper)*, 2014.
- [117] —, "An automated pattern recognition system for classifying indirect immunofluorescence images of hep-2 cells and specimens pattern recognition," *Pattern Recognition*, 2016, (accepted).

- [118] S. Manivannan, H. Shen, W. Li, R. Annunziata, H. Hamad, R. Wang, and J. Zhang, “Brain tumour region segmentation using local co-occurrence features and conditional random fields,” *CVIP, School of Computing, University of Dundee, Tech. Rep.*, 2014.
- [119] S. Manivannan, R. Wang, and E. Trucco, “Extended gaussian-filtered local binary patterns for colonoscopy image classification,” in *IEEE International Conference on Computer Vision Workshops*, 2013.
- [120] —, “Inter-cluster features for medical image classification,” in *International Conference on Medical Image Computing and Computer Assisted Interventions*, 2014.
- [121] D. E. Maroulis, D. K. Iakovidis, S. A. Karkanis, and D. A. Karras, “CoLD: a versatile detection system for colorectal lesions in endoscopy video-frames,” *Computer Methods and Programs in Biomedicine*, vol. 70, pp. 151–66, 2003.
- [122] Z. MatthewD and F. Rob, “Visualizing and understanding convolutional networks,” in *European Conference on Computer Vision*, 2014.
- [123] T. Mäenpää and M. Pietikäinen, “Multi-scale binary patterns for texture analysis,” in *Image Analysis*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2003, vol. 2749, pp. 885–892.
- [124] M. Mirmehdi, X. Xie, and J. Suri, *Handbook of Texture Analysis*. London, UK, UK: Imperial College Press, 2009.
- [125] E. Nowak, F. Jurie, and B. Triggs, “Sampling strategies for bag-of-features image classification,” in *European Conference on Computer Vision*, 2006.
- [126] T. Ojala, M. Pietikainen, and D. Harwood, “Performance evaluation of texture measures with classification based on Kullback discrimination of distributions,” in *International Conference on Pattern Recognition*, 1994.
- [127] T. Ojala, M. Pietikainen, and T. Maenpaa, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” *IEEE Transactions on Pattern Analysis Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.
- [128] A. Oliva and A. Torralba, “Modeling the shape of the scene: A holistic representation of the spatial envelope,” *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.
- [129] F. Perronnin and C. R. Dance, “Fisher kernels on visual vocabularies for image categorization,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.
- [130] F. Perronnin, J. Sánchez, and T. Mensink, “Improving the fisher kernel for large-scale image classification,” in *European Conference on Computer Vision*, 2010.

- [131] F. Perronnin, J. Sánchez, and T. Mensink, “Improving the fisher kernel for large-scale image classification,” in *European Conference on Computer Vision*, 2010, pp. 143–156.
- [132] J. Philbin, M. Isard, J. Sivic, and A. Zisserman, “Descriptor learning for efficient retrieval,” in *European Conference on Computer Vision*, 2010.
- [133] M. Pietikainen, A. Hadid, G. Zhao, and T. Ahonen, *Computer Vision Using Local Binary Patterns*. Springer, 2011.
- [134] J. C. Platt, “Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods,” in *Advances in large margin classifiers*. MIT Press, 2000, pp. 61–74.
- [135] C. K. Poh, T. M. Htwe, L. Li, W. Shen, J. Liu, J. H. Lim, K. L. Chan, and P. C. Tan, “Multi-level local feature classification for bleeding detection in wireless capsule endoscopy images,” in *IEEE Conference on Cybernetics and Intelligent Systems*, 2010.
- [136] D. Regge, C. Hassan, P. Pickhardt, A. Laghi, A. Zullo, D. Kim, F. Iafrate, and S. Morini, “Impact of computer-aided detection on the cost-effectiveness of CT colonography,” *Journal of Radiology*, vol. 250, pp. 488–497, 2009.
- [137] F. Riaz, M. Areia, F. Silva, M. Dinis-Ribeiro, P. Nunes, and M. Coimbra, “Gabor textons for classification of gastroenterology images,” in *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*, 2011, pp. 117–120.
- [138] F. Riaz, M. Dinis-Ribeiro, P. P. N. Nunes, and M. Coimbra, “A dft based rotation and scale invariant Gabor texture descriptor and its application to gastroenterology,” in *IEEE International Conf. on Image Processing*, 2013.
- [139] F. Riaz, M. Ribeiro, P. Pimentel-Nunes, and M. Tavares Coimbra, “A DFT based rotation and scale invariant Gabor texture descriptor and its application to gastroenterology,” in *20th IEEE International Conference on Image Processing*, 2013, pp. 1443–1446.
- [140] F. Riaz, F. Silva, M. Ribeiro, and M. Coimbra, “Invariant Gabor texture descriptors for classification of gastroenterology images,” *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 10, pp. 2893–2904, 2012.
- [141] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio, “Contractive auto-encoders: Explicit invariance during feature extraction,” in *28th International Conference on Machine Learning*, 2011, pp. 833–840.
- [142] R. Rigamonti, M. A. Brown, and V. Lepetit, “Are sparse representations really relevant for image classification?” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.
- [143] P. M. Roth, M. Hirzer, M. Köstinger, C. Beleznai, and H. Bischof, “Mahalanobis distance learning for person re-identification,” in *Person Re-Identification*, ser. *Advances in Computer Vision and Pattern Recognition*. Springer, 2014, pp. 247–267.

- [144] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323–2326, 2000.
- [145] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Neurocomputing: Foundations of Research*. MIT Press, 1988, pp. 673–695.
- [146] C. Schmid, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *IEEE Computer Vision and Pattern Recognition*, 2006.
- [147] C. Schüldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local svm approach," in *International Conference on Pattern Recognition*, 2004.
- [148] J. Shotton, M. Johnson, and R. Cipolla, "Semantic texton forests for image categorization and segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.
- [149] J. S. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, "Towards embedded detection of polyps in wce images for early diagnosis of colorectal cancer," *International Journal of Computer Assisted Radiology and Surgery*, vol. 9, no. 2, pp. 283–293, 2014.
- [150] K. Simonyan, A. Vedaldi, and A. Zisserman, "Descriptor learning using convex optimisation," in *European Conference on Computer Vision*, 2012.
- [151] —, "Learning local feature descriptors using convex optimisation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.
- [152] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *IEEE International Conference on Computer Vision*, 2003.
- [153] Z. Sobri and H. A. M. Sakim, "Texture colour fusion based features extraction for endoscopic gastritis images classification," *International Journal of Computer and Electrical Engineering*, vol. 4, pp. 674–678, 2012.
- [154] A. Sousa, M. Dinis-Ribeiro, M. Areia, and M. Coimbra, "Identifying cancer regions in vital-stained magnification endoscopy images using adapted colour histograms," in *Proceedings of the 16th IEEE International Conference on Image Processing*, 2009.
- [155] M. Swain and D. Ballard, "Indexing via colour histograms," in *Third International Conference on Computer Vision*, 1990.
- [156] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1635–1650, June 2010.
- [157] The National Bowel Cancer Screening Program Quality Working Group, "Improving technical services in australia," Australian Government Department of Health and Ageing, Canberra, Tech. Rep., 2009.

- [158] M. P. Tjoa and S. Krishnan, "Feature extraction for the analysis of colon status from the endoscopic images," *Biomedical Engineering Online*, vol. 2, 2003.
- [159] E. Tola, V. Lepetit, and P. Fua, "DAISY: An efficient dense descriptor applied to wide-baseline stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 5, pp. 815–830, 2010.
- [160] T. Trzcinski, C. M. Christoudias, and V. Lepetit, "Learning Image Descriptors with Boosting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.
- [161] T. Trzcinski, C. M. Christoudias, V. Lepetit, and P. Fua, "Learning Image Descriptors with the Boosting-Trick," in *International Conference on Neural Information Processing Systems*, 2012.
- [162] T. Tuytelaars, M. Fritz, K. Saenko, and T. Darrell, "The NBNN kernel," in *IEEE International Conference on Computer Vision*, 2011, pp. 1824–1831.
- [163] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," <http://www.vlfeat.org/>, 2008.
- [164] A. Vedaldi and A. Zisserman, "Efficient additive kernels via explicit feature maps," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 3, pp. 480–492, 2011.
- [165] R. L. P. Viana, Y. Iwahori, K. Funahashi, and K. Kasugai, "Automated polyp detection from endoscope images," in *Proceeding of SCIS - ISIS*, 2012.
- [166] P. Viola, J. C. Platt, and C. Zhang, "Multiple instance boosting for object detection," in *Advances in Neural Information Processing Systems*, 2007.
- [167] M. B. Wallace, "Improving colorectal adenoma detection: technology or technique?" *Journal of Gastroenterology*, vol. 132, pp. 1221–1223, 2007.
- [168] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *IEEE Computer Vision and Pattern Recognition*, 2010.
- [169] P. Wang, , S. Krishnan, C. Kugean, and M. Tjoa, "Classification of endoscopic images based on texture and neural network," in *IEEE Conference on Engineering in Medicine and Biology Society*, vol. 4, 2001.
- [170] S. Wang and S. Wang, "A robust CBIR approach using local colour histograms," university of Alberta, Tech. Rep., 2001.
- [171] Z. Wang, Y. Hu, and L.-T. Chia, "Image-to-class distance metric learning for image classification," in *European Conference on Computer Vision*, 2010.
- [172] A. Wiliem, C. Sanderson, Y. Wong, P. Hobson, R. F. Minchin, and B. C. Lovell, "Automatic classification of human epithelial type 2 cell indirect immunofluorescence images using cell pyramid matching," *Transactions on Pattern Recognition*, vol. 47, no. 7, pp. 2315–2324, 2014.

- [173] A. Wiliem, Y. Wong, C. Sanderson, P. Hobson, S. Chen, and B. C. Lovell, "Classification of human epithelial type 2 cell indirect immunofluorescence images via codebook based descriptors," in *IEEE Workshop on Applications of Computer Vision*, 2013, pp. 95–102.
- [174] S. Winder, G. Hua, and M. Brown, "Picking the best daisy," in *IEEE Computer Vision and Pattern Recognition*, 2009.
- [175] S. A. J. Winder and M. Brown, "Learning local image descriptors," in *IEEE Computer Vision and Pattern Recognition*, 2007.
- [176] S. Xia, W. Mo, Z. Zhang, S. Xia, W. Mo, and Z. Zhang, "A content-based retrieval system for endoscopic images," 2005.
- [177] X. Xie and M. Mirmehdi, *A Galaxy of Texture Features*. Imperial College Press, 2008, pp. 375–406.
- [178] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
- [179] J. Yang and M.-H. Yang, "Learning hierarchical image representation with sparsity, saliency and locality," in *British Machine Vision Conference*, 2011.
- [180] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proceedings of the 18th International Conference on Advances in Geographic Information Systems*. ACM, 2010, pp. 270–279.
- [181] —, "Spatial pyramid co-occurrence for image classification." in *IEEE International Conference on Computer Vision*, 2011, pp. 1465–1472.
- [182] L. Yu, P. Yuen, and J. Lai, "Ulcer detection in wireless capsule endoscopy images," in *IEEE International Conference on Pattern Recognition*, 2012.
- [183] Y. Yuan, B. Li, and M.-H. Meng, "Improved bag of feature for automatic polyp detection in wireless capsule endoscopy images," *IEEE Transactions on Automation Science and Engineering*, vol. PP, no. 99, pp. 1–7, 2015.
- [184] Y. Yuan, J. Wang, B. Li, and M.-H. Meng, "Saliency based ulcer detection for wireless capsule endoscopy diagnosis," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 2046–2057, 2015.
- [185] Y. Yuan and M. Q. Meng, "Polyp classification based on bag of features and saliency in wireless capsule endoscopy," in *2014 IEEE International Conference on Robotics and Automation*, 2014, pp. 3930–3935.
- [186] S. Zelikovitz and H. Hirsh, "Using LSI for text classification in the presence of background text," in *10th ACM International Conference on Information and Knowledge Management*, 2001, pp. 113–118.

- [187] J. Zhang, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: a comprehensive study," *International Journal of Computer Vision*, vol. 73, no. 2, pp. 213–238, 2007.
- [188] J. Zhang, T. Tan, and L. Ma, "Invariant texture segmentation via circular Gabor filters," in *16th International Conference on Pattern Recognition, 2002*, vol. 2, 2002, pp. 901–904 vol.2.
- [189] J. Zhang and T. Tan, "Brief review of invariant texture analysis methods," *Pattern Recognition*, vol. 35, no. 3, pp. 735 – 747, 2002.
- [190] Q. Zhao and M. QH.Meng, "Polyp detection in wireless capsule endoscopy images using novel colour texture features," in *World Congress on Intelligent Control and Automation*, 2011.
- [191] X. Zhen, L. Shao, and F. Zheng, "Discriminative embedding via image-to-class distances," in *British Machine Vision Conference*, 2014.
- [192] Z. Zuo, G. Wang, B. Shuai, L. Zhao, Q. Yang, and X. Jiang, "Learning discriminative and shareable features for scene classification," in *European Conference on Computer Vision 2014*, 2014.