COMPUTATIONAL SIGNAL ANALYSIS METHODS FOR INFORMATION EXTRACTION WITH APPLICATIONS IN BIOMEDICINE

Michael D. Vasilakakis

Department of Computer Science and Biomedical Informatics

University of Thessaly

A thesis submitted for the degree of

Doctor of Philosophy

March 2020

ABSTRACT

This doctoral thesis explores computational approaches for signal analysis and information extraction. In view of scientific challenges for developing innovative solutions with a broad social impact, it investigates various applications in biomedicine.

In this context, a novel signal analysis methodology, based on fuzzy logic, was proposed. This methodology, called Fuzzy Phrases, expresses signals or feature representations of signals using sets of words. Each of these words is represented by a fuzzy set. The words form phrases, which are obtained by the aggregation of the fuzzy sets. Experiments on publicly available datasets showed that it outperforms relevant state-of-the-art approaches. Advantages of the proposed signal representation approach include intuitiveness and tolerance to uncertainty from imprecise or missing information.

A significant part of this thesis is devoted to the analysis of endoscopic images, in the context of gastrointestinal tract examination. Endoscopic examination using a wireless camera (Wireless Capsule Endoscopy or WCE) has been established to study and diagnose gastrointestinal pathological conditions. The examination of the patient is performed by a swallowable camera that has the size of a large vitamin pill. The examination with WCE is a non-invasive screening procedure. During the examination, the wireless camera captures and transmits thousands of images of the gastrointestinal tract of the patient. Significant drawback of this examination is the large volume of images received; resulting in the capture of thousands of color images (> 100,000 images per patient). The clinical diagnosis of health conditions requires the review of this large number of collected images by medical specialists. Usually, the reviewers reach their human limits by trying to maintain their concentration undistracted in order to examine this large number of images within an average of 60-90min. This explains why this examination is prone to human errors and has low diagnostic accuracy.

To address this important problem, image analysis methods for information extraction to assist in the clinical diagnosis of pathological conditions, were investigated. The development of such methods aims to support medical decision making by reducing the required human effort and by providing a second opinion on the examined medical problem. In-depth research of state-of-the-art methods of image processing and analysis in the literature led to several contributions. One of the contributions is the detailed recording of the literature on technological developments in the field of wireless capsule endoscopy in the last five years 2013-2018.

Supervised classification methods were developed requiring only weak annotation of images by experts, *i.e.*, the experts do not need to annotate in detail the areas of abnormality. Experimental results proved the suitability of the proposed weakly supervised methods for the detection of images with pathological content. Subsequently, algorithms were developed to identify the areas of abnormalities within these images. As the results from weak annotation approaches were encouraging for the classification between abnormal and normal images, supervised methods were adopted also for the

semantic interpretation of the whole content of images using multi-label classification methods. The effort to interpret the whole content of endoscopic images proved to be reasonable, as the results for image classification and pathology finding were further improved compared to previous approaches.

An important aspect in image analysis is the detection of interest points. Such points are useful for the extraction of features from the areas around them. The development of these methods has the effect of limiting the sampling points of the images to these points only, and thus simplifying the computational cost of the analysis process. To this end, during this doctoral research the algorithms suggested by the literature for the detection of points of interest in images were studied and a novel detection algorithm for endoscopic images was proposed. The proposed algorithm was based on the color characteristics of the images. The experimental results were satisfactory with respect to locating points within pathological areas, and it is important to note that the algorithm was unsupervised, as it has no requirement of training, based on previous knowledge in order to detect points.

The work presented in this thesis provides the basis for further research on the topics studied. Particularly in the context of capsule endoscopy, it has contributed methods that can be practically used by physicians to detect various gastrointestinal abnormalities in endoscopic images.

ΠΕΡΙΛΗΨΗ

Η παρούσα διδακτορική διατριβή διερευνά πρωτότυπες υπολογιστικές προσεγγίσεις ανάλυσης σημάτων και εξόρυξης πληροφορίας. Λαμβάνοντας υπόψη τις επιστημονικές προκλήσεις για την ανάπτυξη καινοτόμων λύσεων με ευρύ κοινωνικό αντίκτυπο, διερευνά διάφορες εφαρμογές στη βιοϊατρική.

Αναπτύχθηκε μια πρωτότυπη μεθοδολογία βασισμένη στην ασαφή λογική, η οποία καλείται μέθοδος των Ασαφών Φράσεων (Fuzzy Phrases). Η μεθοδολογία αυτή εκφράζει τα σήματα ή τις χαρακτηριστικές αναπαραστάσεις τους με σύνολα λέξεων. Κάθε λέξη αναπαρίσταται από ένα ασαφές σύνολο. Οι λέξεις σχηματίζουν φράσεις, που λαμβάνονται από τη συνάθροιση των ασαφών συνόλων και αναπαριστούν το περιεχόμενο των σημάτων. Πειράματα σε δημοσίως διαθέσιμα δεδομένα έδειξαν ότι υπερτερεί σε αποτελεσματικότητα άλλων σύγχρονων συναφών προσεγγίσεων. Πλεονεκτήματα της συγκεκριμένης μεθόδου αποτελούν η διαισθητικότητα και η ανοχή στην αβεβαιότητα που προέρχεται από την ανακρίβεια ή την απώλεια πληροφορίας.

Η διδακτορική διατριβή εστιάζει στην ανάλυση εικόνων που λαμβάνονται στο πλαίσιο ενδοσκοπικών εξετάσεων του γαστρεντερικού συστήματος με τη χρήση ασύρματης κάμερας. Η ενδοσκόπηση με χρήση ασύρματης κάμερας (Wireless Capsule Endoscopy ή WCE) έχει καθιερωθεί ως ένας τρόπος για την μελέτη και διάγνωση παθολογικών καταστάσεων του γαστρεντερικού συστήματος. Πραγματοποιείται με την κατάποση κάμερας στο μέγεθος ενός χαπιού βιταμίνης και αποτελεί μια μη-επεμβατική διαδικασία εξέτασης. Κατά τη διάρκεια μιας εξέτασης η ασύρματη κάμερα επιστρέφει χιλιάδες εικόνες για κάθε ασθενή. Ένα από τα σημαντικότερα προβλήματα αυτής της εξέτασης είναι ο μεγάλος όγκος των προσλαμβανομένων εικόνων τα οποία συνιστούν χιλιάδες έγχρωμες εικόνες (> 100.000 εικόνες ανά ασθενή). Η λήψη αποφάσεων σχετικά με την υγεία ενός ασθενούς απαιτεί την εξέταση αυτού του μεγάλου αριθμού συλλεγμένων εικόνων από εξειδικευμένο ιατρικό προσωπικό. Συνήθως, οι εξεταστές αγγίζουν τα ανθρώπινα όριά τους προσπαθώντας να διατηρήσουν την συγκέντρωση τους αδιάκοπη, προκειμένου να εξετάσουν αυτό το μεγάλο αριθμό εικόνων μέσα σε 60-90 λεπτά κατά μέσο όρο. Αυτό εξηγεί γιατί η εξέταση αυτή είναι επιρρεπής σε ανθρώπινα σφάλματα και έχει χαμηλή διαγνωστική ακρίβεια.

Αναγνωρίζοντας τη σημαντικότητα αυτού του προβλήματος, διερευνώνται μέθοδοι ανάλυσης εικόνων για την εξόρυξη πληροφοριών προς την κλινική διάγνωση παθολογικών καταστάσεων. Σκοπός αυτών των μεθόδων είναι η αξιοποίησή τους στο πλαίσιο συστημάτων υποστήριξης ιατρικών αποφάσεων με στόχο κυρίως τη μείωση της ανθρώπινης προσπάθειας και την παροχή μιας δεύτερης γνώμης ως προς το υπό εξέταση ιατρικό πρόβλημα. Από την ενδελεχή έρευνα της βιβλιογραφίας, μελετήθηκαν μέθοδοι αιχμής επεξεργασίας και ανάλυσης εικόνων. Μια από τις συνεισφορές ήταν και η λεπτομερής καταγραφή της βιβλιογραφίας για τις τεχνολογικές εξελίξεις στον τομέα της ενδοσκόπησης με κάψουλα την πενταετία 2013-2018 και η δημοσίευση της.

Αναπτύχθηκαν καθοδηγούμενες μέθοδοι ταξινόμησης (supervised classification) που απαιτούν μόνο ασθενή επισημείωση (weak annotation) των εικόνων από τους ειδικούς, δηλαδή μόνο την ένδειξη αν στην εικόνα περιέχεται ανωμαλία. Τα πειραματικά αποτελέσματα κατέστησαν τις μεθόδους κατάλληλες για την εύρεση εικόνων με παθολογικό περιεχόμενο. Ακολούθως δημιουργήθηκαν αλγόριθμοι για τον εντοπισμό παθολογικών περιοχών/ανωμαλιών εντός των εικόνων αυτών. Καθώς τα αποτελέσματα από τις ασθενώς καθοδηγούμενες προσεγγίσεις για την ταξινόμηση εικόνων, σε παθολογικές ή μη εικόνες, ήταν ενθαρρυντικά, έδωσαν βήμα για τη σημασιολογική ερμηνεία του συνόλου του περιεχομένου των εικόνων χρησιμοποιώντας πολλαπλές κατηγορίες (multi-label weakly supervised methods). Η προσπάθεια για ερμηνεία ολόκληρου του περιεχομένου των ενδοσκοπικών εικόνων αποδείχθηκε βάσιμη, καθώς τα αποτελέσματα για την ταξινόμηση των εικόνων και την εύρεση των παθολογιών βελτιώθηκαν ακόμα περισσότερο συγκριτικά με την προηγούμενη προσέγγιση.

Ένα σημαντικό κομμάτι στην ανάλυση μεγάλου πλήθους εικόνων αποτελεί ο εντοπισμός σημείων ενδιαφέροντος και εξαγωγής χαρακτηριστικών από τις περιοχές που επιλέχθηκαν αυτά τα σημεία. Η ανάπτυξη των μεθόδων αυτών έχει ως αποτέλεσμα τον περιορισμό της δειγματοληψίας των εικόνων, μόνο στα σημεία αυτά, και κατά συνέπεια την απλοποίηση του υπολογιστικού κόστους για την ανάλυση των δειγμάτων. Για το σκοπό αυτό κατά τη διδακτορική αυτή έρευνα μελετήθηκαν οι προτεινόμενοι από τη βιβλιογραφία αλγόριθμοι για τον εντοπισμό σημείων ενδιαφέροντος σε εικόνες και προτάθηκε ένα νέος αλγόριθμος εντοπισμού σημείων για ενδοσκοπικές εικόνες. Ο προτεινόμενος αλγόριθμος βασίστηκε στα ιδιαίτερα χρωματικά χαρακτηριστικά των εικόνων. Τα πειραματικά αποτελέσματα ήταν ικανοποιητικά αναφορικά με τον εντοπισμό σημείων εντός παθολογικών περιοχών, ενώ είναι σημαντική η παρατήρηση πώς δεν απαιτήθηκε κάποια μορφή εκπαίδευσης ώστε ο αλγόριθμος να εντοπίζει σημεία.

Το έργο που παρουσιάστηκε στην παρούσα διατριβή παρέχει τις βάσεις για περαιτέρω έρευνα στα αντικείμενα που μελετήθηκαν. Ειδικότερα στο πλαίσιο της ενδοσκοπίας με κάψουλα, συνεισέφερε μεθόδους που μπορούν να χρησιμοποιηθούν στην πράξη από τους γιατρούς για την ανίχνευση διαφόρων γαστρεντερικών ανωμαλιών σε ενδοσκοπικές εικόνες.

This thesis is dedicated to my precious family and friends.

ACKNOWLEDGMENTS

The pursuit of a doctoral degree is a path that has many difficulties to overcome. Crossing this path with the right persons will not make it less difficult, but will make it more enjoyable.

The first person, that I would like to thank, is Dr. Dimitris Iakovidis for his valuable contribution and his fruitful discussions and advice throughout this doctoral dissertation. His support, his scientific knowledge, his example of hard work, and his optimism have all contributed significantly to the completion of this thesis.

I would like to express my gratitude to Dr. Anastasion Koulaouzidis for his financial support, cooperation and the provision of his medical knowledge. Without the support of Dr. Koulaouzidis, this thesis would have not probably completed.

I would like to thank Dr. Evaggelos Spyrou for his cooperation in many journal articles as well as his useful advices throughout the years of this thesis. Also, I would like to thank my co-authors Dr. Vassilios Plagiannakos and Dr. Spyros Georgakopoulos for their excellent cooperation.

I would like to thank my colleagues in the lab. Especially, I would like to thank Mr. George Dimas and Mr. Dimitris Diamantis, who are also good friends. Although there were difficulties to be encountered during this thesis, they were always there to help. The time we spent together will remain unforgettable.

Finally, I would like to thank my parents and my sisters for their support and contributions all these years, in order to achieve the goal of doctoral degree.

This work was supported by the project “Klearchos Koulaouzidis” Grant No. 5151

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

This chapter presents the aims of the performed research, as well as the novel contributions of this thesis. These contributions have been published in several international journals and international conferences.

1.1 Introduction

Signal analysis is important tool in many applications, including audio indexing and retrieval (Ghoraani & Krishnan 2011; Wilson et al. 1992), detection of various events from wearable sensors (Kalantarian et al. 2016), detection of problems in electrical power systems (Andrade et al. 2016), and many more.

This thesis addresses the analysis mainly of two-dimensional (2D) signals, *i.e.*, on images. Nowadays, as cameras have invaded to our everyday life, e.g., through smart phones, smart cars, and other smart applications, image processing and analysis is becoming more and more popular. Object detection and recognition are the most common tasks, generally posed as the problem of matching a representation of the target object with the available image features, while rejecting the background features (Lowe 2004; Oliva & Torralba 2007). Another application of image analysis is the segmentation for the detection of objects whose boundaries are not defined (Chan & Vese 2001; Ronneberger et al. 2015). For the application of image analysis in tasks such as classification and retrieval significant information is provided by the extraction of image features, which are able to describe the content of an image in a more abstract but meaningful way. A popular method for global image description based on the appropriate features is the Bag-of-Visual-Words (BoW) or Bag-of-Words model (Sivic & Zisserman 2008; Csurka et al. 2004). BoW is utilized for image classification and retrieval purposes (Wang et al. 2010; Sánchez et al. 2013).

The concepts of signal and image analysis have been widely used for the extraction of biomedical information in the context of many clinical procedures, including examination and diagnosis of various diseases. Recent practice of signal analysis can be found in health care systems for disease monitoring. Based on the type of the disease there are different monitoring approaches for recording one or 2D signals. For example, recording and analysis of one-dimensional (1D) signals through wearable smart devices (Athavale

& Krishnan 2017), which can provide recordings of Electrocardiography (ECG) or Electroencephalography (EEG). Analyzing an EEG signal for measuring the electrical activity of the brain, can provide clinical information on diagnosis patients with Alzheimer's disease, epilepsy, EEG seizure detection (Alotaiby et al. 2014), sleep disorders(Boostani et al. 2017). An ECG signal describes the electrical activity of the heart. From the analysis of ECG signals details can be derived concerning the heart rate, the diagnosis of heart disease, emotion recognition and biometric identification (Berkaya et al. 2018). In the recent study of (Koulaouzidis et al. 2016) telemonitoring (TM) has been introduced, aiming to detect patients at high risk of heart failure by the use of daily collected physiological data (blood pressure, heart rate, weight). Another example of image analysis for healthcare systems is coming from Gastrointestinal Endoscopy (GIE), which is a fundamental modality for the investigation of the gastrointestinal (GI) tract and the detection of luminal pathology. The most common GIE procedures are gastroscopy and colonoscopy. Another GIE procedure, which has become the prime choice for the examination of the small bowel, is the wireless capsule endoscopy (WCE) (Vasilakakis, Koulaouzidis, Yung, et al. 2019; Koulaouzidis et al. 2015). Image analysis covers the clinical needs related to the detection and localization of lesions suspicious for malignancy or of bleeding sources, and to provide a second opinion on the assessment of lesions that require a more thorough examination (Iakovidis & Koulaouzidis 2015; Vasilakakis, Koulaouzidis, Yung, et al. 2019; Vasilakakis, Koulaouzidis, Marlicz, et al. 2019). Another application of image analysis is the the detection of bone fractures in X-Ray images(Vasilakakis, Iosifidou, et al. 2019; Al-Ayyoub et al. 2013; Donnelley & Knowles 2005). Combining the knowledge around signal and image analysis Medical Decision Support Systems (MDSS) can provide assistance to physicians supporting them in correct clinical decisions, and consequently can contribute to the improvement of the quality of medical care.

The research of this thesis has been performed in the scope of the project “Klearchos Koulaouzidis”, investigating methodologies for endoscopic image analysis, especially in WCE images. The proposed methodologies of this thesis aim to the assistance of the medical experts during the review of the images captured in the endoscopic examination. Particularly, methods for the detection of images with pathologies of gastrointestinal tract have been studied to support the experts’ decisions as well as to increase the diagnostic yield of the examination.

1.2 Aims of this thesis

This doctoral research investigates novel approaches to signal analysis and their application to biomedical problem-solving with broader social impact. It focuses on the analysis of multichannel 2D signals received in the context of endoscopic examinations of the gastrointestinal system using a wireless camera. Specific aims of this thesis include investigation of:

- Methods for informative signal representation, considering aspects of real-world applications
- Methods for information detection and extraction from biomedical signals
- Applications of the signal analysis methods in the challenging domain of gastrointestinal endoscopy, towards effective approaches to computer-aided clinical diagnosis of pathological conditions.

1.3 Thesis Contributions

The effort that was made for the accomplishment of the aforementioned aims, led to the development of innovative methods, experimentally validated in comparison to relevant state-of-the-art methods. The novel contributions of this thesis can be summarized as follows:

- A methodology based on fuzzy logic for uncertainty-aware signal representation and classification.
- Detailed literature reviews on WCE technologies and endoscopic image analysis methods progress over the last five years 2013-2018
- Development of image analysis and machine learning methodologies for automatic lesion detection and the semantic interpretation of endoscopic images. These methodologies require only weakly annotated images by experts for supervised classification.
- An unsupervised approach to salient point and region of interest detection in wireless capsule images and video frames.

The research performed in the scope of this doctoral thesis and the contributions of this research in the above topics, they have been published in *five (5) international journals* and have been presented in *five (5) international conferences*. The list of publications is in the **Appendix A**.

It is worth noting that one of the published papers (Vasilakakis, Iakovidis, et al. 2018), was nominated as one of the best articles published in 2018 in the ‘Sensor, Signal, and Imaging Informatics’ subfield of medical informatics (Hsu et al. 2019) in the 2019 edition of the annual *International Medical Informatics Association (IMIA) Yearbook of Medical Informatics*.

1.4 Thesis Outline

The rest of this thesis is organized in five (5) chapters as:

- Chapter 2 provides a necessary theoretical background to the signal processing and analysis methods, with respect to the methods investigated
- Chapter 3 presents an original, generic framework for signal analysis based on fuzzy logic, developed in the context of this thesis. It includes experimental results on public datasets.
- Chapter 4 performs an extensive literature review on technologies and methods for gastrointestinal endoscopic image acquisition and analysis.
- Chapter 5 presents all the methods investigated and proposed in the context of this thesis for the extraction of semantic information from endoscopic images, and the detection of lesions indicating pathological conditions.
- Chapter 6 is the last chapter where the conclusions and future prospects for further research, are summarized.

CHAPTER 2

SIGNAL PROCESSING AND ANALYSIS

The main purpose this chapter is to provide the reader with the necessary background knowledge with respect to the signal processing and analysis methods considered in this thesis. This chapter covers methods for digital image processing and analysis, including image transformations, detection of interest points, feature extraction and image classification.

2.1 Introduction

Signal processing is an integral part of almost any digital media application, and it involves the transformation of signals to meet application requirement. For example, the most widely known processing operations are noise reduction and compression, which are common in audio and video transmission. The former one aims to improve the quality of the signals, whereas the second one to reduce the bandwidth required for the transmission.

Signal analysis is generally considered the logical next step of signal processing. The purpose of signal analysis is the use of the signal in its processed form for the interpretation of its content (Allen & Mills 2004). In this way signal analysis stands within the scope of machine learning.

Regardless the way and the purpose of signal collection, the digitization of signals involve sampling, quantization and coding. Once digitized signals are processed and analyzed, usually to detect meaningful patterns. The pattern discovery process involves feature extraction, i.e., the estimation of numerical quantities representing various signal characteristics. These quantities, called features, form numerical vectors, called feature vectors, and the process of the estimation of these features are known as feature extraction. Features may be extracted globally from the whole signal or locally from regions of interest in the signal. The result of feature extraction is a representation enabling the quantification and the semantic interpretation of the signal contents. This is usually performed using machine learning methods. Such methods classify the feature vectors into different classes, corresponding to semantics; thus, enabling computers to understand the contents of the signal.

2.2 Digital signal processing and analysis

2.2.1 Signals and Images

Signals can be categorized based on their form as one-dimensinal (1D), two_dimensinal(2D) (images), or multidimensional. An analog signal is defined as a function of a real variable

$$s_a : \mathbb{R} \rightarrow \mathbb{R} \tag{2.1}$$

where $\mathbb{R}$ is the set of real numbers, and $s_a(t)$ is the signal value at time *t*.

Some very simple elementary analog signals play pivotal roles in the theoretical development. The *Dirac delta* is one of these elementary signals and it is defined as:

$$\delta(t) = \begin{cases} 0, & t \neq 0 \\ 1, & t = 0 \end{cases} \tag{2.2}$$

The *unit step* signal finds use in chopping up analog signals. It is also a building block for signals that consist of rectangular shapes and square pulses.

$$u(t) = \begin{cases} 0, & t < 0 \\ 1, & t \geq 0 \end{cases} \tag{2.5}$$

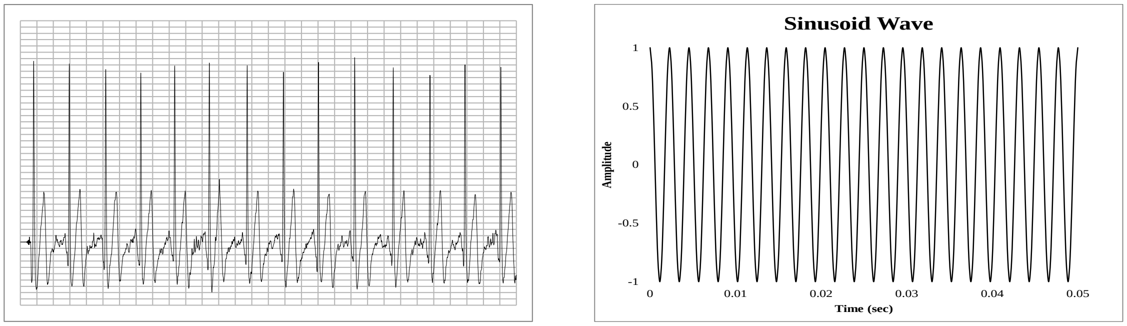
Periodic signals repeat their values over intervals. The interval over which a signal repeats itself is its period, and the reciprocal of its period is its frequency. The measure of frequency is Hertz (Hz), which represents the time period of the sinusoidal wave between to pikes per second.

An analog signal $s_a(t)$ is periodic if there is a T > 0 with

$$s_a(t+T) = s_a(t), \text{ T>0} \tag{2.6}$$

for all t.

Figure 2.1 provides two examples of real-world 1D signals.



(a) (b)

Figure 2.1 (a) Example of an ECG signal; (b) Example of a sinusoidal sound signal produced by a violin.

Electromagnetic waves can be considered as propagating sinusoidal waves with wavelength λ. They can be thought of as a stream of massless particles, that travel in a wavelike pattern and moving at the speed of light (Gonzalez & Woods 1992). Visible light is an electromagnetic wave within a certain portion of the electromagnetic spectrum. The electromagnetic spectrum is the range of frequencies of electromagnetic radiation and their respective wavelengths. Visible light is the portion of the spectrum that can be perceived by the human eye. The usual wavelengths of visible light have range between 400 nm and 700 nm. The electromagnetic spectrum can be expressed in terms of wavelength, frequency. Wavelength (λ) and frequency (v) are related by the expression

$$\lambda = c/v \tag{2.7}$$

where c is the speed of light (2.998×10^8 m/s). Electromagnetic waves can be visualized as propagating sinusoidal waves with wavelength λ, or they can be thought of as a stream of massless particles, each traveling in a wavelike pattern and moving at the speed of light(Gonzalez & Woods 1992).

Light is a particular type of electromagnetic radiation that can be seen and sensed by the human eye. The colors that humans perceive in an object are determined by the nature of the light reflected from the object. A body that reflects light and is relatively balanced in all visible wavelengths appears white to the observer. However, a body that favors reflectance in a limited range of the visible spectrum exhibits some shades of color. For example, green objects reflect light with wavelengths primarily in the 500 to 570 nm range while absorbing most of the energy at other wavelengths (**Figure 2.2)**. Light that is void of color is called achromatic or monochromatic light. The only attribute of such light is its intensity, or amount. The term gray level generally is used to describe monochromatic light intensity because it ranges from black, to grays, and finally to white.

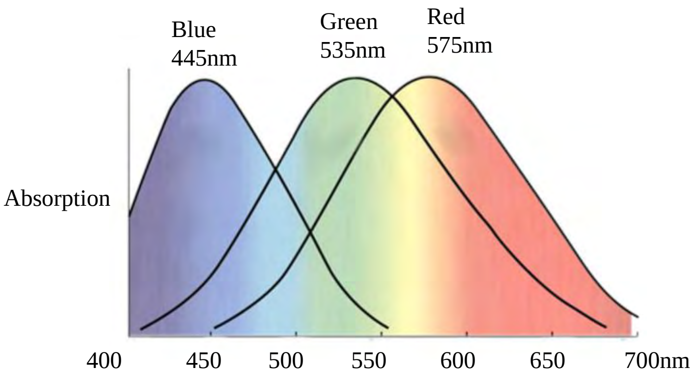


Figure 2.2 Human eye spectral sensitivity

Images are 2D signals generated by the combination of an “illumination” source and the reflection or absorption of energy from that source by the elements of the “scene” being imaged. An image can be represented as a 2D continues function (Gonzalez & Woods 1992)

$$I_g: \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+ \tag{2.8}$$

The amplitude of $I_g(x, y)$ at any pair of coordinates (x, y) is called the *intensity* or *gray level* of the image at that point. That is,

$$g = I_g(x, y) \tag{2.9}$$

The interval between the maximum and the minimum values of g is the gray scale. Common practice is to shift this interval numerically to the interval [0, G-1], where g=0 is considered black and g=G-1 is considered white on the gray scale. All intermediate values are shades of gray varying from black to white.

A photosensitive device that captures the illumination of a scene is called image sensor. An image sensor is arranged in the form of a 2-D array. Typically, the two main types of image sensors are CCD and CMOS sensors (El Gamal & Eltoukhy 2005) and are mainly used in digital cameras and other imaging devices like capsule endoscopes. CCD stands for Charged-Coupled Device and CMOS stands for Complementary Metal–Oxide–Semiconductor. Each sensor consists of cells, where each cell produces a single value independent of colour. To capture colour images, cells are organized in groups of four cells and a filter is placed on top of the group to allow only red light to hit one of the four cells, blue light to hit another and green light to hit the remaining two. The reasoning behind the two green cells is because the human eye is more sensitive to green light and it is more convenient to use a 4-pixel filter than a 3-pixel filter and can be compensated after an image capture with something called white balance.

Depending on the nature of the radiation source (Gonzalez & Woods 1992), which is transmitted through objects, different kinds of images are produced. Except of the electromagnetic radiation which is visible for the human eye, X-rays are also among the sources of electromagnetic radiation used for imaging. An example of X-rays is their use in medical diagnostics. An X-ray image is generated by the placement of a patient between an X-ray source and a film sensitive to X-ray radiation. The X-rays are absorbed as they pass through the patient's body, and the resulting radiation is capture by the film.

2.2.2 Signal digitization

Sensors capture the amplitude and spatial behavior of physical phenomena, like heart rate or illumination, and they produce as output a continuous waveform. The conversion of continue sensed data into digital form utilizes the procedures of sampling, quantization and digitization. Thus, after the acquisition, an analog signal is sampled and then is digitalized.

The sampling (Nyquist 1928) process keeps values of the analog signal at regular intervals. For one dimensional signal the sampling interval is the time between samples. For a time signal, the sampling frequency is measured in hertz (Hz) and it is the reciprocal of the sampling interval, measured in seconds (s). If $s_a(t)$ is an analog signal, then keeping samples in regular intervals from $s_a(t)$ can be expressed as

$$s_n(n) \;=\; s_a(nT) \tag{2.10}$$

where *T* is the time period and T> 0, the discrete value sampled signal s_n of an instant *n* is a real value function

$$s_n : \mathbb{Z} \to \mathbb{R} \tag{2.11}$$

The signal is digitized by quantizing the signal values(Allen & Mills 2004). Digital signals can take on only a finite number of output values in the dependent variable, as long as in computer processing only a finite number of bits can represent the value in binary form. Basically the quantization assigns to each signal value to a computer register. A digital signal is an integer valued function

$$s_d : \mathbb{Z} \to [-\mathrm{N}, \mathrm{N}] \tag{2.12}$$

where $N \in \mathbb{Z}$ and $N > 0$.

For images the result of sampling and quantization are illustrated in **Figure 2.3**. A continuous image $I_g(x, y)$ with continuous the x-coordinates, y-coordinates and amplitude have to be converted to digital form. In the case of images the sampling is the digitizing of the coordinate values and the quantization is the digitizing of the amplitude values (Gonzalez & Woods 1992). After the digitization of the image the coordinates x, y, and the amplitude values of I_g are all finite, discrete quantities. A digital image is composed of a finite number of elements, each of which has a particular location and value. These elements are referred to as picture elements or pixels. Pixel is the most widely used term to denote the elements of a digital image.

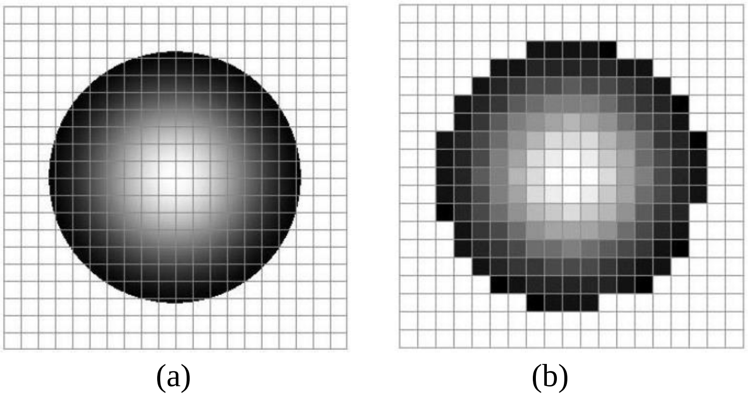


Figure 2.3 (a) a continues image before sampling and quantization (b) a sampled and quantized image

2.2.3 Discrete Fourier and Wavelet transformations

The signals can be viewed as functions of one variable, i.e. time, or of two variables, i.e. spatial x-coordinate and y-coordinate. So far, they have been studied without considering their frequency content. The information about the frequency content of a signal can be derived by transformations such as the Fourier or the wavelet transformations.

Fourier transformation

The Fourier Transform(Allen & Mills 2004) of a continuous signal s_a of a value variable, time *t*, is

$$S_a(\omega) = \int_{-\infty}^{\infty} s_a(t) e^{-j\omega t} dt \qquad (2.13)$$

where ω is a frequency variable and $j = \sqrt{-1}$. The *reverse* Fourier Transform is

$$s_a(t) = \int_{-\infty}^{\infty} S_a(\omega) e^{j\omega t} d\omega \qquad (2.14)$$

The definition of Discrete Fourier Transform (DFT) of a discrete $s_n(n)$ signal is

$$S_n(k) = \sum_{n=0}^{N-1} s_n(n) e^{\frac{-2\pi jnk}{N}} \tag{2.15}$$

where $0 \leq k \leq N - 1$ and $N > 0$.

In general, *S*$_n$ *(k)* is complex; the complex norm, | *S*$_n$ *(k)*|, and complex phase, *arg(S*$_n$ *(k))*, for $0 \leq k < N$, are called the discrete magnitude spectrum and discrete phase spectrum, respectively, of *s*$_n$ *(n)*.

There is an inversion theorem for the DFT. If *s*$_n$*(n)* is a discrete signal and *S*$_n$ *(k)* is the DFT of *s*$_n$*(n)* on [0, N−1]. Then

$$s_n(n) = \frac{1}{N} \sum_{k=0}^{N-1} S_n(k) e^{\frac{2\pi jnk}{N}} \tag{2.16}$$

Extension of one dimensional DFT to two dimensions is straightforward(Gonzalez & Woods 1992). The DFT of a two dimensional signal, image, $I_g(x, y)$ of size $M \times N$ is defined by the following equation

$$F_g(u, v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} I_g(x, y) e^{-2\pi j\left(\frac{ux}{M} + \frac{vy}{N}\right)} \tag{2.17}$$

where u and v variables are the frequency variables and x and y are the spatial image coordinate variable. In similar way the *inverse* DFT of an image is

$$I_g(u, v) = \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} F_g(u, v) e^{2\pi j\left(\frac{ux}{M} + \frac{vy}{N}\right)} \tag{2.18}$$

Wavelet transformation

The Fourier Transform was for many years the basis of signals transformation in order to obtain their spectral information. However, Fourier Transform was not able to provide temporal information about the frequency content of signals. The wavelet transform is a more recent transformation that is used to localize the frequency information of a signal *s*$_a$(*t*) while it keeps the temporal information of each frequency.

In one dimensional Discrete Wavelet Transform (DWT) of a signal, two functions mutually orthonormal are initially adopted: the scaling function φ and the mother wavelet *ψ*. Other wavelets are then produced by translations of the scaling function *φ* and dilations by the mother wavelet *ψ*, according to the equations (Mallat 1989):

$$\varphi_{j_0,k}(t) = 2^{j_0/2}\varphi\left(2^{j_0}t - k\right) \qquad (2.19)$$

$$\psi_{j,k}(t) = 2^{j/2}\psi(2^{j}t - k) \qquad (2.20)$$

for $j = j_0, j_0 + 1, \ldots$ and $j_0 \in \mathbb{Z}$. The scaling function *φ* defines a kernel function and the mother wavelet *ψ* results in an oscillation of the input signal. Families of scaling functions can act as suitable bases for ***L**²(**R**)* or for alternative spaces. The structure of a wavelet basis is deterministic in location and frequency due to translation and dilation respectively. A function $s \in \boldsymbol{L}^2(\boldsymbol{R})$ can be represented in a wavelet series, using a given basis, as:

$$s(t) = \sum_k c_{j_0,k}(t)\varphi_{j_0,k}(t) + \sum_{j=1}^{j_0}\sum_{k\in\mathbb{Z}} w_{j,k}(t)\psi_{j,k}(t) \qquad (2.21)$$

where j is the scale of the transform, j_0 is the "coarsest scale", $c_{j_0,k} = \langle f, \varphi_{j_0,k}\rangle$ and $w_{j,k} = \langle f, \psi_{j,k}\rangle$ are the wavelet coefficients and <., .> is the standard $\boldsymbol{L}^2$ inner product of two functions:

$$\langle s_1, s_2\rangle = \int_R \; s_1(t)s_2(t)dt \qquad (2.22)$$

The first term of (2.18) corresponds to a low resolution signal $L_{j_0} = \{c_{j_0,k}\}$, $k \in \mathbb{Z}$, that can be obtained by lowpass filtering. The coefficients $D_j = \{w_{j,k}\}$, $k \in \mathbb{Z}$, $1 \le j \le j_0$, constitute the detail signal at scale j, that can be obtained by highpass filtering. Together L_{j_0} and D_j are known as the *wavelet representation of depth j_0* of the signal *f*.

The expansion of DWT for two dimensional signals is straightforward considering that the wavelet transform is separable. The 2D DWT of a 2D signal can be calculated by first applying one dimensional DWT on its rows and then apply the one dimensional DWT on the columns of the resulted 2D-signal. More precisely a separable filterbank is applied to the original 2D signal L_0 according to the following recursive equations:

$$L_{j_0} = [H_x * (H_y * L_{j_0-1})_{\downarrow 2,1}]_{\downarrow 1,2} \qquad (2.23)$$

$$D_{j_1} = [H_x * (G_y * L_{j_0-1})_{\downarrow 2,1}]_{\downarrow 1,2} \qquad (2.24)$$

$$D_{j_2} = [G_x * (H_y * L_{j_0-1})_{\downarrow 2,1}]_{\downarrow 1,2} \qquad (2.25)$$

$$D_{j_3} = [G_x * (G_y * L_{j_0-1})_{\downarrow 2,1}]_{\downarrow 1,2} \qquad (2.26)$$

where $k \in \mathbb{Z}$,$1 \leq j \leq j_0, \dots ,j_3 \in \mathbb{Z}$, ↓2,1 and ↓1,2 denote the sub-sampling along the rows and columns respectively, * is the convolution operator, *H* is the lowpass filter and *G* is the highpass filter. As in the one dimensional DWT the coefficients $L_{j_0}, D_{j_1}, D_{j_2}, D_{j_3}$, $1 \leq j \leq j_0, \dots ,j_3$ are known as the *wavelet representation of depth j_0* of a two dimensional signal L_0.

2.3 Digital image processing and analysis

Every signal or image contains patterns that can be revealed through processing. Thus, for an image further details about the visual content of images can be extracted. Usually, these patterns are represented as numerical vectors and in literature are referred as features. The features are related to image properties such as the intensity, the color or the texture (Tuytelaars et al. 2008). Consequently, the effective feature extraction of an image is the key to obtain useful semantic information for the representation of the image, which can be exploited in image analysis tasks such as image classification.

A feature tends to describe either the global scenery of the image or a local image region, which is discriminated from its immediate neighborhood. However, features for global image representation are not efficient to distinguish foreground from background, resulting in the insufficient description of some characteristic of the image scenery, i.e. small objects. Thus, image description based on many features derived from different regions of the image is often more preferable than a feature that describes the global scenery of the image.

The effectiveness of image representation through multiple extracted features of different image regions is depended to the content of each region. In order to select the appropriate regions of the image to extract features, methods able to detect salient or interest points inside an image are often utilized. Salient or interest points are usually associated with a change of one or more image properties simultaneously (Tuytelaars et al. 2008), without necessarily the location of the detected interest point to be exactly on this change. The region around the interest point is a region of interest, which encloses the change of image properties.

2.3.1 Color spaces and color features

Color is one of the main characteristics of the surface of objects and the scenery of an image. Color is the most intuitive and obvious feature of an image and is robust to changes in noise, image size, orientation, and resolution. A simple holistic description of the image color is the distribution of the color in the image pixels. The color distribution can be depicted by a histogram, which measures the frequency with which each color appears among the image.

The purpose of a color space is to facilitate the specification of colors in some standard, generally accepted way. In essence, a color model is a specification of a coordinate system and a subspace within that system where each color is represented by a single point. A color system generally defines a set of axes along which certain properties of color are quantitatively expressed. As it was previously referred, due to the acquisition of image and the human visual perception, a single color is often characterized by a 3D vector. In this thesis, the elements of such vector are referred as the color components or color channels.

Since the advent of digital image processing, several color systems have been proposed, each with specific advantages. The RGB color space is aligned with the color channels of most electronic sensors and displays. The CIE-*Lab* color space is designed so that the same amount of numerical change in these values corresponds to roughly the same amount of visually perceived change.

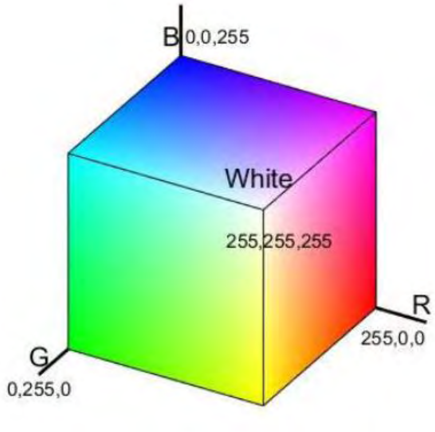


Figure 2.4 RGB color cube

In the RGB color system, colors are defined by the intensities of their spectral components around wavelengths of red, green and blue light. Hence, a color is represented by components R, G, B with $R, G, B \in [0{,}255]$, yielding a linear subspace with the form of a cube (see **Figure 2.4**). In this color cube, the points (255, 0, 0), (0, 255, 0) and (0, 0, 255) define the pure form of the colors red, green and blue, respectively. Black and white are represented by (0, 0, 0) and (255, 255, 255), respectively, and the line between these two points contains all grayscale values within the cube, i.e. all points with equal color components. Most hardware for capturing and displaying images employs separate red, green and blue components. For this reason, RGB is an attractive color system to work with, as the input signal requires no additional transformations.

Due to the distribution of cones in the eye, the tristimulus values depend on the observer's field of view. To eliminate this variable, the International Commission on Illumination (CIE) defined a color-mapping function called the standard (colorimetric) observer, to represent an average human's chromatic response

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} 0.412 & 0.357 & 0.180 \\ 0.212 & 0.715 & 0.072 \\ 0.019 & 0.019 & 0.950 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix} \tag{2.27}$$

The CIE XYZ color space encompasses all color sensations that are visible to a person with average eyesight. That is why CIE XYZ (Tristimulus values) is a device-invariant representation of color. It serves as a standard reference against which many other color spaces are defined. The CIE model capitalizes on this fact by setting Y as luminance. Z is quasi-equal to blue, or the S cone response, and X is a mix of response curves chosen to be nonnegative. Setting Y as luminance has the useful result that for any given Y value, the XZ plane will contain all possible chromaticities at that luminance.

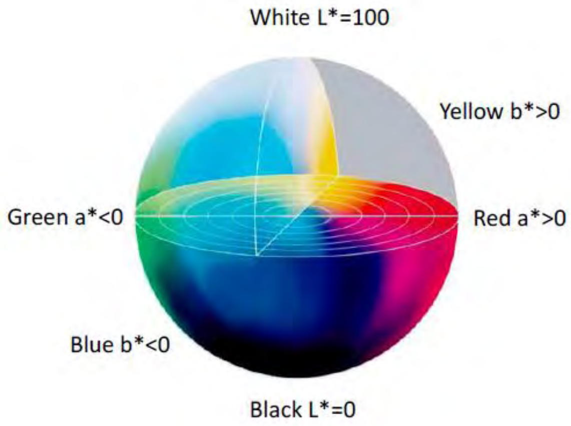


Figure 2.5 CIE-*Lab* color space represented as a sphere.

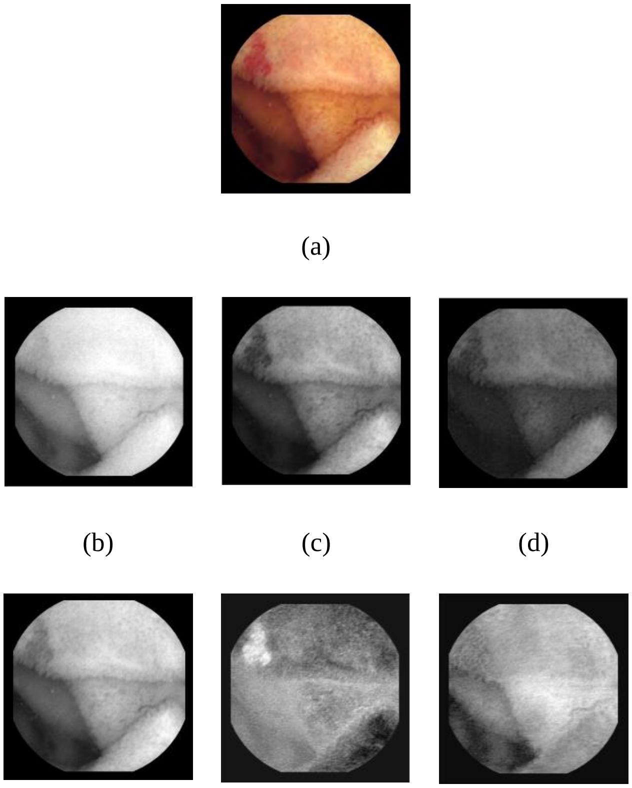
The CIELAB color space (Schwiegerling & others 2004) is an attempt at providing a perceptually uniform color space. It expresses color as three values

- L for the lightness with range [0,100], where black=0 and white=100
- a ranges from green, negative (−) values of a, to red, positive (+) values of a
- b ranges from blue, negative (−) values of b, to yellow, positive (+) values of b

CIELAB was designed so that the same amount of numerical change in these values corresponds to roughly the same amount of visually perceived change. In this color space, the distance between two points also approximately tells how different the colors are in luminance, chroma, and hue. The 1976 CIELAB coordinates (L, a, b) in this color space can be calculated from the tristimulus values XYZ with the following formulas. The subscript n denotes the values for the white point. CIELAB equations that define the color transformation from XYZ are:

$$\begin{aligned} L &= 116f\left(\frac{Y}{Y_n}\right) - 16 \\ a &= 500[f\left(\frac{X}{X_n}\right) - f\left(\frac{Y}{Y_n}\right)] \\ b &= 200[f\left(\frac{Y}{Y_n}\right) - f\left(\frac{Z}{Z_n}\right)] \end{aligned} \tag{2.28}$$

where $f(s)=s^{1/3}$ for $s>0.008856$ and $f(s)=7.7877s+16/116$ for $s<0.008856$ and X_n, Y_n and Z_n are the CIE XYZ tristimulus values of the reference white point.



(a)

(b) (c) (d)

(e) (f) (g)

Figure 2.6 (a) Original RGB image of vascular lesion; (b) R component of (a); (c) G component of (a); (d) B component of (a); (e) L component of CIE-*Lab* color space of (a); (f) a component of CIE-*Lab* color space of (a); (g) b component of CIE-*Lab* color space of (a);

Due to the high correlation between the components of RGB color space, CIE-*Lab* is usually preferred for medical images. In **Figure 2.6(a)** there is an example of a medical image of small bowel captured during endoscopic examination. In the upper left side of the image there is a vascular lesion, which can be distinguished from its red color. In the second row of **Figure 2.6** images (b)-(d) are respectively the R, G, B components, where there is no clear highlight of the lesion based on the intensity of image. On the other hand, in the last row of **Figure 2.6** in the image (f) the lesion can be discerned due to the high intensity on the red color.

Several studies used features extracted from color. Color histogram was one of the first approaches considered for color feature extraction from images (Swain & Ballard 1991). Being orderless(Grauman & Leibe 2011), the histogram offers invariance to viewing conditions and some tolerance for partial occlusions. This insensitivity makes it possible to use a small number of views to represent an object, assuming a closed-world pool with distinctly colored objects. However, color indexing fails when the incident illumination varies either spatially or spectrally. (Funt & Finlayson 1995) proposed a preprocessing phase with a color constancy algorithm using a histogram color ratios to overcome the limitation of illumination. On the domain of medical image processing (Bashar et al. 2010) proposed a methodology for removal of non-informative frames of an endoscopic video based on color histograms to discriminate frames that are highly contaminated by turbid fluids, faecal materials and/or residual foods.

2.3.2 Texture features

Texture is an essential image characteristic for describing the innate surface properties of a particular object and its relationship with the surrounding regions. Texture features have become an important basis for describing image characteristics because of the extensive use of image information.

Gray-Level Co-occurrence Matrix (GLCM)

The Gray-Level Co-occurrence Matrix (GLCM) (Haralick et al. 1973) is one well-known texture analysis method. The GLCM calculates how often the intensity, with value *i*, of a pixel within a grayscale image I_g occurs either horizontally, vertically, or diagonally to adjacent pixels with the value *j*. Practically, GLCM is able to capture local pixel differences providing a texture description of image contents based on the statistical properties of this matrix. The GLCM is expressed by the following equation:

$$C(i,j) = \sum_{x=1}^{M} \sum_{y=1}^{N} \begin{cases} 1, if\ I_g(x,y) = i\ and\ I_g(x+\Delta x, y+\Delta y) = j \\ 0, \quad otherwise \end{cases} \tag{2.29}$$

where C is the GLCM of a grayscale image I_g with size $M \times N$ pixels and Δx and Δy define the spatial relation for which this matrix is calculated. The GLCM method to extract textural features for different angles, such as 0°, 90°, 180° and 270. From the resulting co-occurrence matrices statistical features are extracted (Haralick et al. 1973). The description of selected relevant features is given in the following equations.

Energy :

$$\sum_i \sum_j C^2(i,j) \tag{2.30}$$

Entropy:

$$-\sum_i \sum_j C(i,j)\log C(i,j) \tag{2.31}$$

Contrast :

$$\sum_i \sum_j (i-j)^2 C(i,j) \tag{2.32}$$

Inverse Difference Moment:

$$\sum_i \sum_j \frac{C(i,j)}{|i-j|^2}, i \neq j \tag{2.33}$$

Correlation:

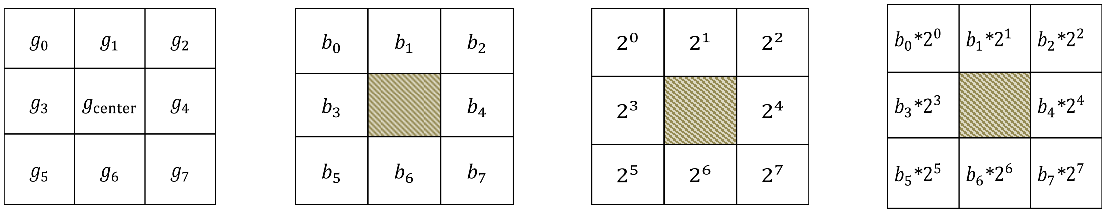
$$\frac{\sum_i \sum_j (i-\mu x)(j-\mu y) C_{(i,j)}}{\sigma_\chi \sigma_y} \tag{2.34}$$

In the above list, C denotes the normalized co-occurrence matrix, with C (i, j) referring to individual matrix elements. Parameters μ_x, μ_y and $\sigma_{x,}$ σ_y denote the statistical mean and standard deviation in the directions of i and j, respectively. The choice of the above six features is motivated by the visual features of the image signal: flatness (Homogeneity), contrast, signal variation (Energy), signal activity (Entropy), signal statistics (Correlation) and averaged local signal difference (Dissimilarity). Energy, also called Angular Second Moment and Uniformity, is a measure of textural uniformity of an image. Energy reaches its highest value when gray level distribution has either a constant or a periodic form. A homogenous image contains very few dominant gray tone transitions, and therefore the C matrix for this image will have fewer entries of larger magnitude resulting in large value for energy feature. In contrast, if the C matrix contains a large number of small entries, the energy feature will have smaller value. Entropy measures the disorder of an image and it achieves its largest value when all elements in C matrix are equal. When the image is not texturally uniform many GLCM elements have a very small value which implies that entropy is very large. Therefore, entropy is inversely proportional to GLCM energy. Contrast is a difference moment of the C and it measures the amount of local variations in an image. Inverse difference moment measures image homogeneity. This parameter achieves its largest value when most of the occurrences in GLCM are concentrated near the main diagonal Inverse different moment is inversely proportional to GLCM contrast film and digitizers (Partio et al. 2002). Correlation measures the linear dependence of gray levels on those neighboring pixels being examined.

Local binary pattern (LBP)

The Local binary pattern (LBP) operator was proposed by (Ojala et al. 1996; Ojala et al. 2002), as a non-parametric, grey-scale invariant texture representation model for the description of the local spatial structure of an image. The LBP operator represents the local texture around a pixel exploiting the information of its 3×3 neighborhood (**Figure 2.7**). More specifically, a neighborhood is represented by a set of nine pixels with intensities $G = \{g_c, g_0, g_1, \dots, g_7\}$, where g_c is the intensity of the pixel in center and $g_i (0 \leq i \leq 7)$ are the intensities of the pixels around the g_c (**Figure 2.7(a)**). The neighborhood is characterized by a set of binary values $b_i (0 \leq i \leq 7)$ based on the condition(**Figure 2.7(b)**)

$$b_i = \begin{cases} 0, if\ g_i < g_c \\ 1, if\ g_i \geq g_c \end{cases} \qquad (2.35)$$



g_0	g_1	g_2
g_3	g_{center}	g_4
g_5	g_6	g_7

(a)

b_0	b_1	b_2
b_3	▨	b_4
b_5	b_6	b_7

(b)

2^0	2^1	2^2
2^3	▨	2^4
2^5	2^6	2^7

(c)

b_0*2^0	b_1*2^1	b_2*2^2
b_3*2^3	▨	b_4*2^4
b_5*2^5	b_6*2^6	b_7*2^7

(d)

Figure 2.7 Local Binary Pattern computation steps

A unique LBP value is extracted for each center pixel of a neighborhood considering the following equation(**Figure 2.7(c)-(d)**):

$$LBP = \sum_{i=0}^{7} b_i \cdot 2^i \qquad (2.36)$$

This way, the local texture information around a pixel is described by a local binary pattern with code $LBP \in [0,255]$. Every pixel in an image generates a single LBP code.
The textural information of the image is described by a histogram that represents the occurrences of different LBP codes from all pixels. This histogram is the texture feature vector of the image.

Complete local binary pattern (CLBP)

An extension of LBP was Complete local binary pattern (CLBP) and proposed by (Guo et al. 2010). Despite the LBP, CLBP exploits the information of magnitude. A neighborhood is represented by a set of Q pixels with intensities$G = \{g_c, g_0, g_1, \dots, g_{Q-1}\}$, where g_c is the intensity of the pixel in center and $g_i(0 \leq i \leq Q - 1)$ are the intensities of the pixels around the g_c. In CLBP a local neighborhood of pixels is described by the local difference $b_i = g_i - g_c$ that is decomposed to sign-magnitude

$$b_i = g_i - g_c = s_i * m_i \tag{2.37}$$

$$s_i = sign(b_i) \text{ and } m_i = |b_i| \tag{2.38}$$

where s_i is the sign of b_i and m_i is the magnitude of b_i respectively. Then, two operators CLBP-Sign (CLBP_S) and CLBP-Magnitude (CLBP_M) are used to code them. Another operator the CLBP_Center (CLBP_C) is used for the coding of the central pixel of the neighborhood based on a global thresholding. All the three code maps produced by the CLBP_C, CLBP_S, and CLBP_M operators are in binary format. The CLBP feature for textural description of the image is the combination of all computed codes for the formation of a CLBP histogram.

2.3.3 Feature detectors-descriptors

The detection of regions of interests begins with the detection of salient/interest points. Preferably, these points have to be invariant and distinctive under various conditions as viewpoint changes, noise, translation or rotation of image. For the detection of invariant interest points (Lowe 2004) proposed Scale Invariant Feature Transform (SIFT) and (Bay et al. 2008) proposed Speeded-Up Robust Features (SURF).

Scale Invariant Feature Transform (SIFT)

In SIFT, the Difference-of-Gaussians detector (DoG) was utilized for the detection of interest point between different scales. In DoG there is a smoothing of the initial image multiple convolutions with a Gaussian mask. From every convolution a smoothed image is produced. The smoothed images are combined pairwise to compute a set of DoG points. More precisely, the scale space of an image is the product of the convolution of a variable-scale Gaussian, $G(x, y, \sigma)$:

$$G(x, y, \sigma) = \frac{e^{-(x^{2+}y^2)/2\sigma^2}}{2\pi\sigma^2} \tag{2.39}$$

with an input image $I_g(x, y)$ and scale variable σ. The DoG is the difference between two adjacent scales that are separated by a factor of k

$$D(x, y, \sigma) = \left(G(x, y, k\sigma) - G(x, y, \sigma)\right) * I_g(x, y) \qquad (2.40)$$

Every scale has at least s+3 (s an integer number) images, which compose an octave. These images are the different products of convolutions with Gaussian for different values of scale variant k. Then the initial image I_g is down-sampling by a factor 2 and the DoG for a new octave with the down-sampling image is computed. The minimum and the maximum extrema points are defined by comparing each point of $D(x, y, \sigma)$ with the immediate 8 neighbors and with the 9 closest neighbors in two adjacent scale levels. From the defined extrema point the final points of interest are determined after a threshold for the rejection of low contrast points and a threshold based on the ratio of principal curvatures.

The description of the region of interest around the detected interest point is based on the magnitude of a point coordinate at (x,y):

$$m(x, y) = \sqrt{\left(G(x+1, y, \sigma) * I_g(x, y) - G(x-1, y, \sigma) * I_g(x, y)\right)^2 + \left(G(x, y+1, \sigma) * I_g(x, y) - G(x, y-1, \sigma) * I_g(x, y)\right)^2} \qquad (2.41)$$

and the orientation of the same point

$$\theta(x, y) = \tan^{-1}\left(\left(G(x, y+1, \sigma) * I_g(x, y) - G(x, y-1, \sigma) * I_g(x, y)\right) / \left(G(x+1, y, \sigma) * I_g(x, y) - G(x-1, y, \sigma) * I_g(x, y)\right)\right) \qquad (2.42)$$

by encoding the image information of the region in a localized set of gradient orientation histograms. The region is a regular grid of 16 ×16 pixels that is divided in subgroups of 4 ×4. The content of 4×4 subgroup is then summarized in gradient orientation histogram with 8 orientation bins. The final feature vector of each region of interest is the concatenation of each subgroup’s histogram resulting in a feature vector with dimensionality of 128.

Speeded-Up Robust Features (SURF)

SURF(Bay et al. 2008) was proposed as a more efficient alternative of SIFT. It utilizes the integral images(Viola & Jones 2001) for fast approximation of the Hessian matrix $H(x, y, \sigma)$ of an image I_g at scale σ

$$H(x, y, \sigma) = \begin{bmatrix} L_{xx}(x, y, \sigma) & L_{yx}(x, y, \sigma) \\ L_{xy}(x, y, \sigma) & L_{yy}(x, y, \sigma) \end{bmatrix} \qquad (2.43)$$

where $L_{xx}(x,y,\sigma)$, $L_{xy}(x,y,\sigma)$, $L_{yy}(x,y,\sigma)$ are the convolutions of the Gaussian second order derivative with the image I_g in point *(x,y)*.

The sum of all pixels in a rectangular region formed by the origin $(0,0)$ and (x,y) of the input image I_g is preserved as index in the integral image $I_\Sigma(x,y)$:

$$I_\Sigma(x,y) = \sum_{i=0}^{i\leq x} \sum_{j=0}^{j\leq y} I_g(i,j) \tag{2.44}$$

After the computation of the integral image, four additions are needed to calculate the sum of the intensities over any upright, rectangular area, as shown in **Figure 2.9**.

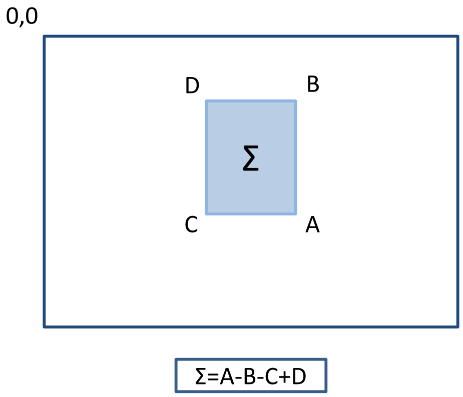


Figure 2.8 Using integral images, it takes only four operations to calculate the area of a rectangular region of any size

SURF doesn't rely on Gaussian derivatives as SIFT does. SURF utilizes simple 2D box filters ("Haar wavelets"). The box filters approximate determinant of the Hessian and are efficiently evaluated using integral images. Based on the approximated determinant of the Hessian in the image, the local maxima are detected using non-maximum suppression.

The SURF features are inspired by SIFT dividing the region of interest into a 4 × 4 grid. However, instead of building up a gradient orientation histogram, SURF computes the statistics $\sum d_x$, $\sum|d_x|$, $\sum d_y$, $\sum|d_y|$,where d_x, d_y are the wavelet responses vertically and horizontally.

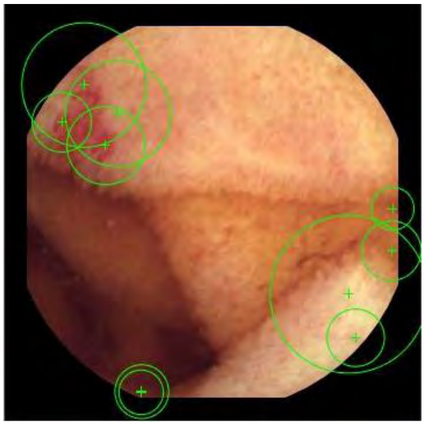


Figure 2.9 The detected point of interest in the a-channel of CIE-Lab color space of Figure 2.6 (a) using SURF.

In biomedical image processing, (Iakovidis & Koulaouzidis 2014b) proposed the use of SURF for the detection of lesion in endoscopic images, Figure 2.10. The proposed methodology has a preprocessing step of image transformation from RGB to CIE-*Lab* color space. SURF algorithm detects salient points in the *a* component of CIE-*Lab* color space.

2.3.4 Superpixel segmentation

Segmenting the image into group of pixels is an alternative way for feature extraction from arbitrarily shaped uniformly colored image regions image regions. Superpixels provide a convenient way for image representation, where the image segments typically have the right trade-off between locality and distinctiveness(Tuytelaars et al. 2008). The algorithm simple linear iterative clustering (SLIC) (Achanta et al. 2012) performs a local clustering of along with a distance measure that enforces compactness and regularity in the superpixel shapes, and seamlessly accommodates grayscale as well as color images. SLIC creates clusters of pixels defining regions of homogeneous color properties, called superpixels. SLIC is simple to implement and easily applied in practice the only parameter specifies the desired number of superpixels. This is done in the five-dimensional [L,a,b,x,y] space, where [L,a,b] is the pixel color vector in CIE-*Lab* color space, which is widely considered as perceptually uniform for small color distances, and x,y is the pixel position. While the maximum possible distance between two colors in the CIE-*Lab* space is limited, the spatial distance in the (x,y) plane depends on the image size. It is not possible to simply use the Euclidean distance in a 5D space without normalizing the spatial distances. In order to cluster pixels a distance measure is used that considers superpixel size. First the distance of pixels in CIE-Lab is defined as

$$d_{Lab} = \sqrt{(L_i - L_j)^2 + (a_i - a_j)^2 + (b_i - b_j)^2} \tag{2.45}$$

and then spatial distance of x,y

$$d_{xy} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{2.46}$$

The final distant measure is

$$D_s = d_{Lab} + \frac{m}{S} d_{xy} \tag{2.47}$$

where D_s is the sum of the L,a,b distance and the x,y plane distance normalized by the grid interval S. A variable m is introduced in D_s allowing us to control the compactness of a superpixel. The greater the value of m, the more spatial proximity is emphasized and the more compact the cluster.

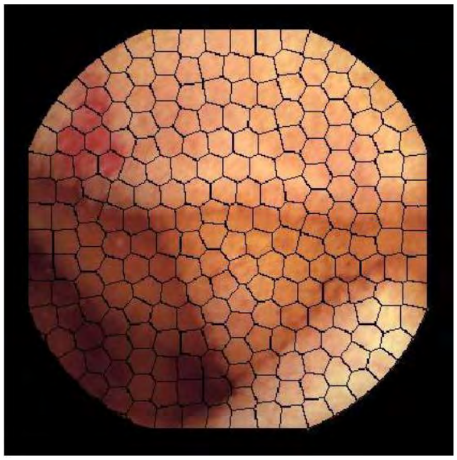


Figure 2.10 The superpixels with SLIC in **Figure 2.6(a)**

For medical images capture during endoscopic procedure, many recent methodologies have used SLIC in order to extract local features. In the approach proposed by (Iakovidis et al. 2015), the superpixels that contain at least one salient point are also characterized as salient. However, in that study the pixel-level saliency was disregarded, and the localization of abnormalities smaller than a superpixel was impossible. In the study of (Vasilakakis, Iakovidis, et al. 2018), the pixel-level saliency defined by DINOSARC algorithm is not superseded by the region-level saliency defined by the superpixels. Each DINOSARC salient region is defined by a superpixel that includes only a single, representative salient point. If the superpixel contains a cluster of salient points, then the cluster centroid is regarded as its corresponding salient point.

2.3.5 The model of Bag-of-Words for image representation

The Bag-of-Words or Bag-of-Visual Words (BoW/BoVW) (Csurka et al. 2004) is a widely used method to model generic categories in detection, classification and recognition problems (Sivic & Zisserman 2008). This method has been originally inspired by text document analysis techniques, and consists of calculating the frequency of appearances of a word. Bag-of-Visual-Words can be considered as a model built upon the notion of visual vocabularies. The BoW describes an image as a set of “words”, which capture its visual content.

A visual vocabulary can be defines as a quantization on the feature vector space of local image descriptors. That way, any novel descriptor vector can be coded in terms of the discretized region of feature space to which it belongs. The visual vocabulary consists of a collection of a large sample of features from representative images, and the quantization of the feature space. Given an adequately large dataset, a set of features is extracted from every image and typically quantized using a clustering approach, e.g., the k-means algorithm (Drake & Hamerly 2012). Upon clustering, the centroids (or in some approaches the medoids, which opposed to centroids are actual members of the dataset) that have been determined, are used as a “visual vocabulary” and are often referred to as “visual words.”

In that case, the visual “words” are the k cluster centers, and the size of the vocabulary k is a user-supplied parameter. Then a novel image’s features can be translated into words by determining which visual word they are nearest to in the feature space. This procedure involves a histogram construction, which describes the appearance frequency of every visual word within an image. Thus, this histogram is used to characterize the visual content of the image. Among the advantages of BoW, we should emphasize that it succeeds to reduce the problem of classifying a large number of high dimensional vectors from local point descriptors to a fixed-size, one dimensional vector without significant loss of visual information.

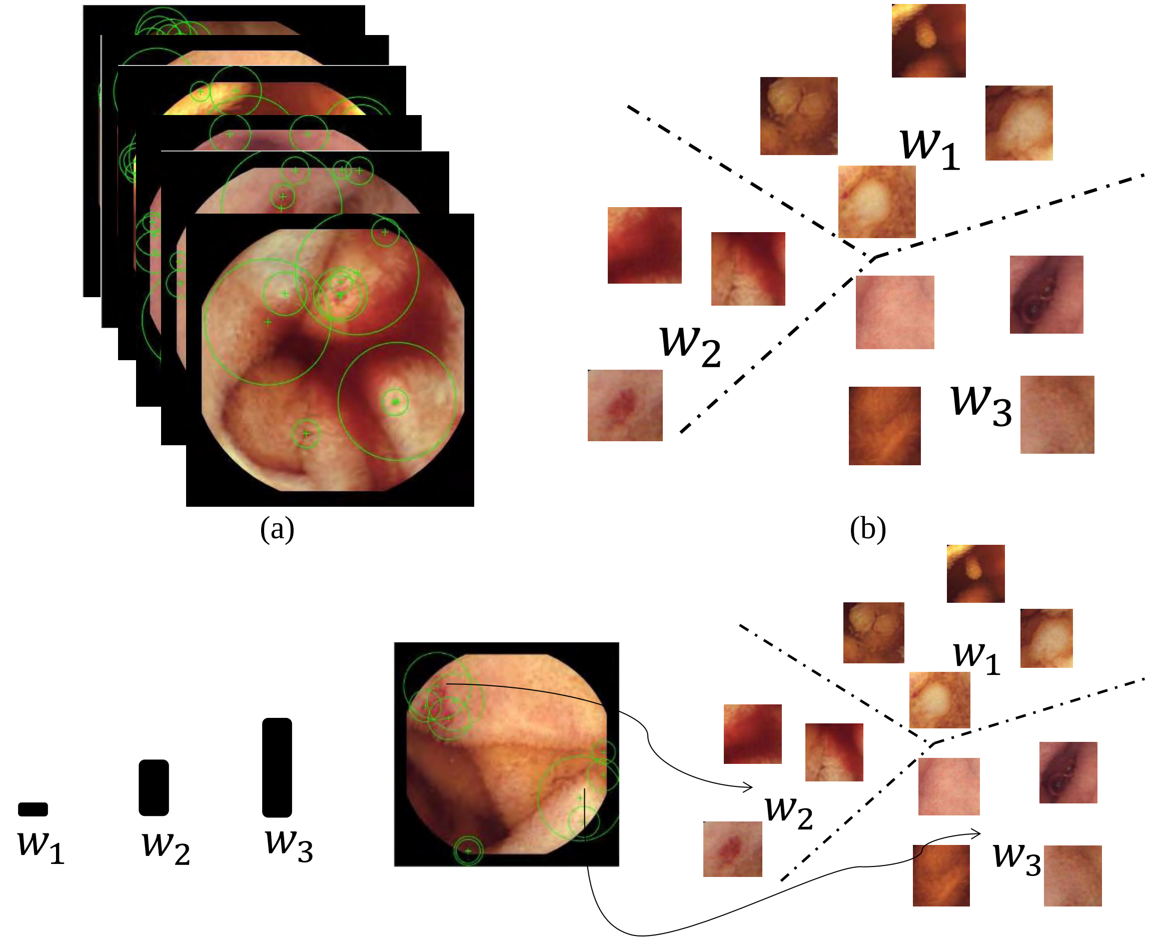


Figure 2.11 BoW Feature Model (a) features are extracted from a group of images (b) quantization on the feature vector space of image descriptors for the construction of visual vocabulary (c) histogram construction for the characterization of the visual content of the image based on the frequency that each word appears in the image

2.4 Data classification

Usually, the extracted features are labeled in relation to the content of the image or the content of the image region that these features represent. For example, features coming from an area of medical image with abnormality have the label “abnormal” and this label indicates the class that these features are categorized. Otherwise, feature from an area of a medical image without abnormality have the label “normal”.

Following the image processing, the extracted features are used for the analysis of the image content. The purpose of image analysis is to reveal the relations between features extracted from different images. In this way, image analysis stands within the scope of machine learning, as the image features are employed by machine learning algorithms.

Such algorithms are trained from the given features with known labels and classify the unknown features predicting their content.

In biomedicine, image classification is a task of great importance, as the classification result can improve the diagnostic accuracy of physicians. An example of image classification in medicine is the classification of endoscopic images, in order to detect a possible abnormality (Vasilakakis, Koulaouzidis, Yung, et al. 2019; D. K. Iakovidis et al. 2018).

2.4.1 Supervised learning

Let $X = R^d$ be the feature vector space derived from a set of images used to train the supervised classification system. Also, let Y denote the label space, where each label refers to a class, i.e. “normal” or “abnormal”. In the supervised learning an algorithm learns from a set of labeled examples a function $f: X \rightarrow Y$, by using the feature vectors $x \in X$, that have been extracted from a set of N images and are labeled by $y \in Y$. This algorithm learns from a set of labeled features that are used as training data, in order to make predictions for features with unknown label. The features with unknown are usually referred as test features $x^t \in X$ This is the most common scenario associated with classification problems (Mohri et al. 2012).

K-nearest neighbors (KNN) is a simple classification algorithm that stores all available cases and classifies new cases based on a similarity measure i.e. Euclidean distance. KNN is a non-parametric technique and it has been used in pattern recognition and feature classification from the beginning of 1970’s. Each feature can be classified based on the majority vote of its neighbors. This way, an unseen feature is assigned to the class to the class most common amongst its K nearest neighbors measured by a distance function that is used as a similarity measure.

KNN classification has two parameters that are needed to be set. The first is the number of neighbors K and the second is the similarity measure that used for determining the set of closest neighbors. For the case of K=1, the feature is assigned to the class of the neighbor that is nearest. Choosing the optimal value for K is best done by first inspecting the data. In general, a large K value is more precise as it reduces the overall noise. However, the increase in classification accuracy is no guarantee.

KNN is relatively slow during testing, although it can be trained fast. This happens because when a new sample is tested, it has to be compared to all training samples. Additionally, the deficiency of an underlying parametric model leads to a high memory demand, as all the training points have to be stored.

Another method for supervised classification is based on Support Vector Machine (SVM) (Sergios Theodoridis & Koutroumbas 2008). SVM has been proposed by Cortes and Vapnik (Cortes & Vapnik 1995). SVM performs classification by finding the hyperplane that maximizes the margin between the two classes. The hyperplane separates the labeled data of different classes and the features (vectors) that define the hyperplane are the support vectors. The region bounded by hyperplanes is called the margin.

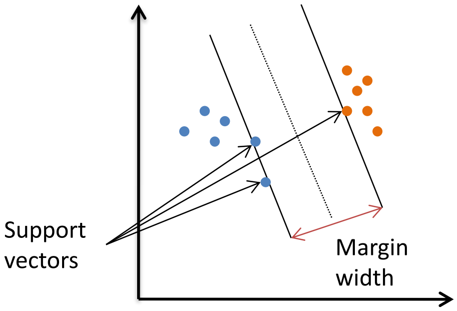


Figure 2.12 Example of linearly separable classes

To define an optimal hyperplane the maximization of the margin width w of label data is needed. Let y_i where $y \in \{-1,1\}$, denotes a class label of feature $x \in R^d$, SVM finds the hyperplane

$$\vec{w} \cdot \vec{x} + \vec{b} = 0 \tag{2.48}$$

where b is the offset of the hyperplane. The hyperplanes of each class can be described as

$$\vec{w} \cdot \vec{x} + \vec{b} \geq 1, if\ y = 1 \tag{2.49}$$

$$\vec{w} \cdot \vec{x} + \vec{b} \leq -1, if\ y = -1 \tag{2.50}$$

that divides the data points with label $y_i = -1$ from the data points with label $y_i = 1$ with maximum margin. These constraints state that each data point must lie on the correct side of the margin. This can be rewritten as

$$y_i(\vec{w} \cdot \vec{x}_i + \vec{b}) \geq 1, \forall x_i \tag{2.51}$$

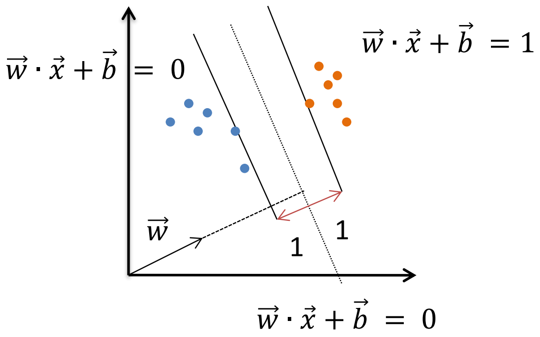

$\vec{w} \cdot \vec{x} + \vec{b} = 1$
$\vec{w} \cdot \vec{x} + \vec{b} = 0$
$\vec{w}$
1
1
$\vec{w} \cdot \vec{x} + \vec{b} = 0$

Figure 2.13 Linear support vector classifier

Based on the geometry the distance of two hyperplanes is $\frac{2}{\|w\|}$, so for the maximization of the distance between the planes, the minimization of $\|w\|$ is needed as well as to find w and b by solving the following objective function

$$min\|w\| \; subject \; to \; y_i(\vec{w} \cdot \vec{x}_i + \vec{b}) \geq 1, \forall x_i \qquad (2.52)$$

In SVM, if the data is linearly separable, there is a unique global minimum value. Ideally, SVM analysis should produce a hyperplane, where the feature vectors are completely separable into two non-overlapping classes. However, perfect separation may not be possible, or it may result in a model with many cases that the model does not classify correctly. In this situation SVM finds the hyperplane that maximizes the margin and minimizes the misclassifications

$$y_i\left(\vec{w} \cdot \vec{x}_i + \vec{b}\right) \geq 1 - \xi_i, \forall x_i \; and \; \xi_i > 0 \qquad (2.53)$$

where the slack variable ξ_i allows some instances to fall off the margin with a penalty.

Objective function penalizes for misclassified instances

$$\min \frac{1}{2} \|w\|^2 + C \sum_i \xi_i \; subject \; to \; y_i\left(\vec{w} \cdot \vec{x}_i + \vec{b}\right) \geq 1 - \xi_i, \forall x_i \; and \; \xi_i > 0 \qquad (2.54)$$

where C is a positive cost parameter.

The algorithm tries to maintain the slack variable ξ_i to zero while maximizing margin. However, it does not minimize the number of misclassifications, but the sum of distances from the margin hyperplanes.

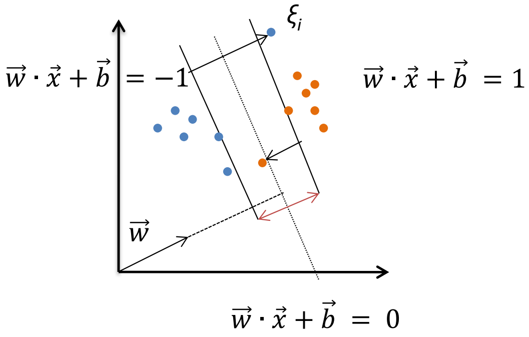

ξ_i
$\vec{w}\cdot\vec{x}+\vec{b} = -1$
$\vec{w}\cdot\vec{x}+\vec{b} = 1$
$\vec{w}$
$\vec{w}\cdot\vec{x}+\vec{b} = 0$

Figure2.14 Non-linear separable hyperlane

The simplest way to separate two groups of data is with a straight line, if they are 1D data, flat plane, if they are 2D data or an N-dimensional hyperplane for data of more than two dimensions. However, there are situations where a nonlinear region can separate the groups more efficiently.

SVM handles this by using a kernel function, as a way to map the data from the non-linear space into a different space with higher dimensionality, where a hyperplane can be used to separate the data. This mapping exploits the fact that the inner product of vectors of a higher dimensional space can be expressed as a function of the inner product of vectors of the initial space. The mapping $\varphi: R^d \to H$, where $H$is a higher dimensional space, is achieved with a kernel function $K(x_i, x_j) = \varphi(\vec{x}_i)\cdot\varphi(\vec{x}_j)$ that maps the inner product $\vec{x}_i \cdot \vec{x}_j \ \rightarrow\ \varphi(\vec{x}_i)\cdot\varphi(\vec{x}_j)$

This is called kernel trick which means the kernel function transform the data into a higher dimensional feature space to make it possible to perform the linear separation. Thus, the (2.31) can be rewritten as

$$\min\frac{1}{2}\|w\|^2 + C\sum_i \xi_i \ \ subject\ to\ y_i\left(\vec{w} * \varphi(\vec{x}_i) + \vec{b}\right) \geq\ 1-\xi_i, \forall x_i\ and\ \xi_i > 0 \qquad (2.55)$$

A kernel that is used in this thesis is the Gaussian Radial Basis function or RBF and it can be describe by the function below

$$K(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}} \qquad (2.56)$$

Artificial Neural Networks(ANNs)

Artificial Neural Networks(ANNs) (Sergios Theodoridis & Koutroumbas 2008) are inspired from the way that the human brain works. ANNs are composed by units named perceptron. Figure 2.14 illustrates the basic perceptron model. For a feature vector $x = [x_1 x_2 \dots x_n] \in R^N$ every x_i is multiplied with the respective weight $w = [w_1 w_2 \dots w_n] \in R^N$ which is called synaptic weight or synapses. Activation function applies step rule which converts the numerical value to 0 or 1 so that it will be easy for data set to classify. The activation functions that are usually use are the sigmoid

$$S(x) = \frac{1}{1+e^{-\beta x}} \tag{2.57}$$

$$S(x) = \tanh\frac{\beta x}{2} \tag{2.58}$$

where they have a smooth gradient for the values of parameter β.

Also, there is a bias value w_0 which is added to the weighted sum product $\vec{w} \cdot \vec{x}$ is an element that adjusts the boundary away from origin to move the activation function left, right, up or down.

Perceptron algorithms can be divided into two types they are single layer perceptron and multi-layer perceptron. In single-layer perceptron’s neurons are organized in one layer and they are used for linear classification problems. If there is more than one class, then more than one neuron are placed parallel in the same layer.

Multi-Layer perceptron is substantially formed from multiple layers of perceptron. Every single neuron present in the first layer will take the input signal and send a response to the neurons in the second layer and so on. The multilayer perceptron (MLP) are used for non-linear classification cases. In the first layer there are no neurons and each node represents each xi. In the output layer there is the same number of neurons as the number of classes. The hidden layers include a number of neurons that is usually defined experimentally.

In the first phase the activations are propagated forward from the input to the output layer and the error between the output and the desire value is calculated by

$$E = \sum_{k=1}^{M}(d_k - y_k)^2 \tag{2.59}$$

where y_k is the output of the *k*-neuron and d_k is the desire value.

The back propagation algorithm propagates backwards the error iteratively and tries to find the optimal values for weights, so as the error value to be minimized. To minimize the error back propagation algorithm calculates partial derivatives from the error function till each neuron's specific weight is defined. The basic iteration scheme for the computation of optimal weights is

$$w_j^r(new) = w_j^r(old) + \Delta w_j^r \tag{2.60}$$

$$\Delta w_j^r = -\eta \frac{\partial \mathrm{E}}{\partial w_j^r} \tag{2.61}$$

where is the partial derivatives $\frac{\partial \mathrm{E}}{\partial w_j^r}$ of the error function with respect to each weight *w* and η is the learning rate.

Convolutional Neural Networks

Within the last decades, Convolutional Neural Networks (CNNs)(LeCun et al. 1990) have shown high predictive capacity in the broader field of computer vision (Krizhevsky et al. 2012; Simonyan & Zisserman 2014; Szegedy et al. 2015) on single label datasets, such as ImageNet (Deng et al. 2009) and CIFAR (Krizhevsky et al. 2009). CNNs are contemporary extensions of the MPL characterized by a deep structure that enables feature extraction from raw input images through layers of adaptable filtering components (LeCun et al. 1998). This makes them independent from any hand-crafted feature extraction method "tailored" to specific diagnostic tasks (Greenspan et al. 2016).

The main layers of CNN architecture are the Convolution layer, the Pooling layer and the Fully Connected layer.

In the Convolution layer valuable features are extracted from an image. A convolution layer has a number of k filters that are convolved with the image. Let a $n \times m \times r$ image be the input of the convolutional layer, where n is the height and m is width of the image and r is the number of channels, i.e. RGB images have $r = 3$. Each filter has size $h \times w \times d$ where $h < n,\ w < m$. After the convolution of the image with k filters the result between input image and convolution layer is a feature map with $n' \times m' \times r'$ where $n' = (n - h + 1)$, $m' = (m - w + 1)$ and $r' = k$.

Pooling layer operates on each feature map independently. Its function is to progressively reduce the spatial size of the representation to reduce the amount of parameters and computation in the network. The most common approach used in pooling is max pooling, where selects the maximum value in a region $p \times p$ of the feature map, $p < n', m'$ and usually $p = 2$.

Fully connected layers are placed before the classification output of a CNN. Fully connected layers flatten the results of convolution and pooling layers before classification. This is similar to the output layer of an MLP.

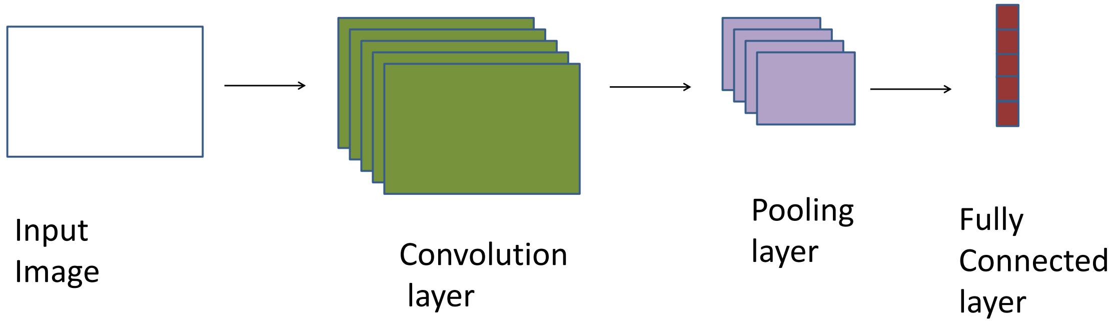


Figure 2.15 Basic CNN architecture

An approach of using a CNN in a classification task is through a pre-trained network used for feature extraction exploiting the convolution layer feature maps. Zhang et al. (Zhang et al. 2016) utilized a pre-trained network, and more specifically CaffeNet (Jia et al. 2014), to transfer learning by extracting features from intermediate convolution layers of the network and then use them to train an SVM classifier.

2.4.2 Unsupervised learning

In the previous section, the training features were available, and the classifier was designed by exploiting known information. However, this is not always the case, and there is another type of pattern recognition tasks for which training data, of known class labels, are not available. When the given training data are unlabeled and the learning algorithm has to predict for unseen data, this procedure is called unsupervised learning. Since in general no labeled example is available in that setting, it can be difficult to quantitatively evaluate the performance of a learner. Clustering and dimensionality reduction are example of unsupervised learning problems. The *k-means* algorithm (Drake & Hamerly 2012) is a well-known clustering algorithm.

2.4.3 Weakly-supervised learning

As it was previously mentioned, the image region from which the features are extracted is important, especially when the extracted features need to characterized a specific object in the foreground of the image without taking into account the background information. In object classification problems, the features are extracted from specific image regions, which contain the object and have been manually annotated. This specific human labeling about the object’s location inside the image limits the application of classification methods for two reasons. First, the manual annotation of the objects in a large collection of image is time consuming. Second, the manual annotation of the objects introduces arbitrary biases, as annotations of the same image from different people maybe differ.

In order to overcome the limitations of human annotation in the training data, the method of supervised learning on weakly annotated data was proposed(Hoai et al. 2014). In weakly supervised learning each label *y* refers to the semantic content of the whole image, since a given feature vector *x* describes the whole image rather than a specific region. The weakly annotation was useful especially for abnormality detection in medical images. A representative example is the gastrointestinal abnormality detection in capsule endoscopy images(Vasilakakis et al. 2016; D. K. Iakovidis et al. 2018).

2.4.4 Multi-label learning

Multi-label classification is a special case of data classification, where multiple labels may be assigned to a given instance. One may consider it as a generalization of multi-class classification, which enables the semantic characterization of data instances by labels that are not mutually exclusive.

One of the most exciting and continuously growing research areas in computer vision is the understanding of the visual image content. The majority of research efforts have turned to the automatic extraction of semantic annotations, aiming to imitate the way humans perceive and describe such content. This problem is often referred to as "bridging the semantic gap" (Smeulders et al. 2000) and consists of automatic extraction of high-level semantics from a given image, based on low-level features computed from raw data (pixels). To this goal, semantic concepts are formalized, learned, and ultimately linked to their linguistic representation. The semantic content of a video frame may not be completely characterized by a single annotation, as it may contain more than one semantic concept; therefore, the need for multi-label classification becomes evident. Moreover, semantic interpretation of images can become even more challenging by extending its application to video scale. State-of-the-art works on semantic video interpretation are mainly based on supervised machine learning algorithms capable of classifying the contents of the video frames into semantically relevant categories (H. Li et al. 2016).

Let $X = R^d$ be the feature space derived from a set of images used to train the supervised classification system. Also, let Y denote the label space. In the binary case the classification system aims to learn a function $f: X \to Y$, by using the feature vectors $x_i \in X$, $i = 1,2,\dots,N$, that have been extracted from a set of N images and are labeled by $y_j, \in Y$, $j = 1,2$ from a training set $\{(x_i, y_i \vee 1 \leq i \leq N, 1 \leq j \leq 2)\}$. In weakly supervised learning each label y_j refers to the semantic content of the whole image, since a given feature vector x_i describes the whole image rather than a specific region.

When tackling the problem of multiple-label classification, one may use a cascade of binary or multi-class classifiers on image regions. In the latter case, an image is labeled with a single label $y_j \in Y$, $y = 1,2,\dots,m$; m denoting the number of the available classes to describe image content. However, such an approach does not take into account that the visual content of a single image may be described with more than one, different labels. Therefore, taking this observation into account, a multi-label classification learning scheme will be more useful to be utilized (Tsoumakas & Katakis 2007; Zhang & Zhou 2013).

More specifically, let v be a vector of m multiple labels for each $x_i \in X$, $i = 1,2,\dots,N$, where $v_j \in Y, v_j = (y_1, y_2, \dots y_m)$, $j = 1,2,\dots,z$. Each label is a binary flag denoting the presence of different kinds of image contents. The purpose of training a multi-label classifier is to learn a function $h: Q \to 2^L$.

There are two main learning strategies, namely the algorithm adaptation, and the problem transformation strategies (Tsoumakas & Katakis 2007). Algorithm adaptation tackles multiple labels by adapting existing learning algorithms from single- to multi-label. Examples of algorithms implementing this strategy include an adaptation of the *k*-Nearest Neighbor (*k*-NN) (S Theodoridis & Koutroumbas 2008) a classifier for multi-label classification (MLkNN) (Zhang & Zhou 2007) and kernel methods, e.g., multi-label SVMs (Elisseeff & Weston 2002).

On the other hand, the problem transformation strategy deals with multi-label learning problem by reducing it into binary or multi-class categorization. This way, a traditional classifier, e.g., an SVM may be used. More specifically, the four basic problem transformation strategies are:

- The binary relevance (Tsoumakas & Katakis 2007) trains different binary classifiers, each of which classifies the video frames according to a single label
- The ranking and thresholding (Tsoumakas & Katakis 2007; Fürnkranz et al. 2008) aim to transform the task of multi-label learning into a multi-class problem. In ranking the task is to order the set of labels. A threshold function is constructed from multi-label data, so that the topmost labels are more related

with the new instance. A ranking of labels requires post-processing in order to give a set of labels, which is the proper output of a multi-label classifier.

- The pairwise classification (Fürnkranz et al. 2008; Mencιa & Furnkranz 2008) adopts the "one-*vs*-one" approach, where one classifier is associated with each pair of labels. This is contrary to the binary relevance approach of "one-*vs*-all" where one classifier is associated with the relevance of each label.
- The label combination (Read et al. 2008) transforms the task of multi-label learning into a standard, single-label, multi-class classification. It considers each different set of labels that exist in the multi-label data set as a single one. In this way, it treats every label combination in the training data as a unique class label in a binary label problem.

Artificial Neural Networks (ANNs) have been traditionally used in single-label binary classification tasks. Zhang and Zhou (Zhang & Zhou 2006) proposed an adaptation of the error function of a back-propagation learning algorithm for Multi-Layer Perceptron (MLP) architecture so as to account for multiple labels in the learning process.

2.5 Summary and conclusions

This chapter provided a brief overview about the needed background knowledge for signal processing and analysis methods that are used in a variety of applications, including applications in biomedicine. Furthermore, the discussion about 1D signals was extended to 2D signals.

This chapter presented various well-known image processing and analysis methods, including machine learning methods, which were considered in the context of this thesis, and they were used as building elements to develop novel approaches.

CHAPTER 3

FUZZY PHRASES

This chapter proposes a framework for the representation of data, enabling classification and feature selection based on fuzzy logic. As the representation of collected data may suffer for incompleteness and uncertainty, fuzzy logic can provide a solid mathematical background to overcome the representation problem. The proposed model is inspired by the bag-of-words (BoW) feature extraction, which follows an intuitive approach of describing data, using histograms of data granules, referred to as words.

3.1 Introduction

Real-world data are usually characterized by uncertainty and incompleteness. Since Zadeh (Lotfi Asker Zadeh 1975; Lotfi A Zadeh 1975a; Lotfi A Zadeh 1975b; Zadeh 1965; Zadeh 1988) established the foundations of the fuzzy sets theory, fuzzy logic has been effectively used as a basis for the foundation of various classification and pattern recognition models. Fuzzy logic is the intuition of reasoning which is more an approximate description rather than an exact description (Zadeh 1996). Thus, the theoretical framework of systems based on fuzzy set theory, enables the handling of the uncertainty and incompleteness of the input data. A fuzzy system can represent the input data and describe their relations with a non-linear manner. A fuzzy set can be seen as a set of input data forming a granule. Labeling each granule with a word, the data can be associated with expressions close to human perception (Zadeh 1999). A set of rules that have the form of antecedent and consequent can be utilized for the classification of input data. More specifically, these rules usually follow a form of a structure “IF x is a THEN z is b”.

Fuzzy logic has been adopted and applied in a variety of research domains. These include financial prediction(Chou et al. 2017), fault diagnosis of power systems (Peng et al. 2017), dig data processing (Wang et al. 2017), image processing (Ziólko et al. 2017), medical diagnosis(Vasilakakis, Iosifidou, et al. 2019; Pota et al. 2017; Vasilakakis & Iakovidis 2020)

The design of a fuzzy classification framework depends on various aspects (Pota et al. 2017), including the partitioning of the input data for the construction of the membership functions, the rules encoding the domain knowledge, and the inference process for the categorization of unknown data samples to a class.

Fuzzy classification has been investigated in several studies. Indicatively, in (Hu et al. 2018)the optimization of the granularization of the feature space has been considered in the context of pattern recognition. That study proposed a concept of fuzzy classifiers built upon a logic-based computing architecture utilizing t-norms and t-conorms. In (Hu et al. 2019) an Evolutionary Multi-Objective algorithm was proposed for the allocation of information granularity in the context of classification. In (Fu et al. 2019)a classification model is constructed by engaging a synergy of Fuzzy C-Means (FCM) clustering and the principle of justifiable granularity with weighted data. In (Duan et al. 2018)a time-series clustering method, called Linear Fuzzy Information Granule-based Dynamic Time Warping Hierarchial Clustering, was proposed by defining a new distance measure for hierarchical clustering.

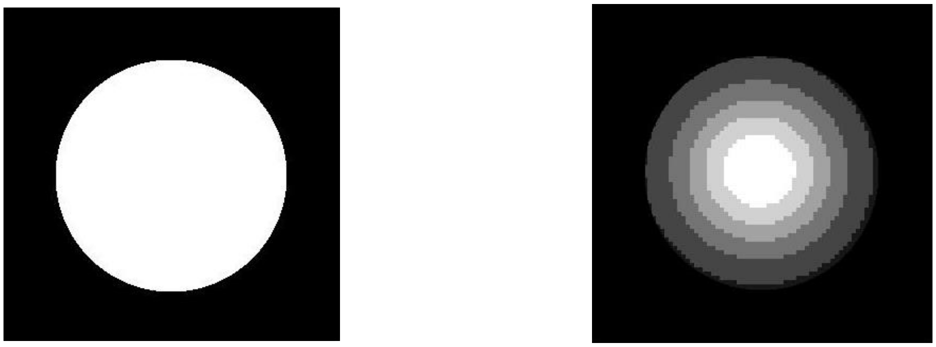
This chapter proposes a constructive fuzzy representation model that is called Fuzzy Phrases(Vasilakakis, Iosifidou, et al. 2019; Vasilakakis & Iakovidis 2020). The aim of Fuzzy Phrases is to enhance the expressivity of conventional feature spaces, and consequently to improve the classification of the respective data instances. The proposed model is inspired by the bag-of-words (BoW) method, which follows an intuitive approach for the description of data, that resembles the way humans use specific vocabularies of words for the description of real-world concepts. Unlike BoW, the proposed model considers that data can be represented by fuzzy phrases constructed by fuzzy words. Different words can be instantiated by different fuzzy sets derived from data. Besides its intuitiveness, this modeling approach is also simpler to implement than the respective state-of-the-art approaches. The general aspects of Fuzzy Phrases make the model able for handling like missing data values and feature selection.

A preliminary version of the proposed model was presented in (Vasilakakis, Iosifidou, et al. 2019)however, that study was focused on bone fracture detection in x-ray images, was based on different clustering approach and did not include feature selection.

3.2 Basic elements of fuzzy sets

Let V (**Figure 3.1** (a)) be a classical set of objects, called the of discourse s, whose generic elements are denoted o. Membership of a subset of V is often viewed as a characteristic function

$$\mu_V(o) = \begin{cases} 1, o \in V \\ 0, o \notin V \end{cases} \tag{3.1}$$



(a) (b)

Figure 3.1 (a) the set *V* where the membership of an objects has value one inside the set or the value zero outside the set (b) the fuzzy set *V'* where the membership of an objects value is in the interval [0,1], where the value set is closer to one as the objects is closer to the center of the set.

A membership function for a fuzzy set *V'*(**Figure 3.1** (b)) on the universe of discourse *V* is defined as $\mu_{V'}: V \rightarrow [0,1]$, where each object of V is mapped to a value between 0 and 1 (Zadeh 1965). This value, called membership value or degree of membership, quantifies the grade of membership of an object in *V* to the fuzzy set *V'*. The use of the interval [0, 1] allows a convenient representation of the gradation in membership (Dubois 1980). The more an object *o* belongs to *V'*, the value of the membership of *o* is closer to one.

The extension of union (∪) and intersection (∩)of ordinary sets to fuzzy sets proposed by (Zadeh 1965) .

Definition 3.1 Let *V'* and *Z'* be two fuzzy sets the union is define

$$\forall o \in V, \mu_{V' \cup Z'}(o) = \max\left(\mu_{V'}(o), \mu_{Z'}(o)\right) \qquad (3.2)$$

Definition 3.2 Let *V'* and *Z'* be two fuzzy sets the intersection is define

$$\forall o \in V, \mu_{V' \cap Z'}(o) = \min\left(\mu_{V'}(o), \mu_{Z'}(o)\right) \qquad (3.3)$$

3.3 The constructive fuzzy representation model: Fuzzy Phrases

The proposed Fuzzy Phrases methodology is a supervised method consisting of a training and a test phase.

3.3.1 Fuzzy Phrases training

The proposed Fuzzy Phrases methodology is illustrated in Figure. 3.2**.** Let K be the number of different classes of the classification problem under investigation, and N_K be the number of training feature vectors v_{ik} from each class in the training phase. Every feature vector is composed of different features. Let $v_{ik}(f_1^{v_{ik}}, f_2^{v_{ik}}, \dots, f_L^{v_{ik}})$ be an L-dimensional feature vector extracted from a training sample, with features $f_l^{v_{ik}}$, l=1,2,…, L, k=1,2,…,K, where K is the number of classes, and i=1,2…,N_k, where N_k is the number of training samples per class k.

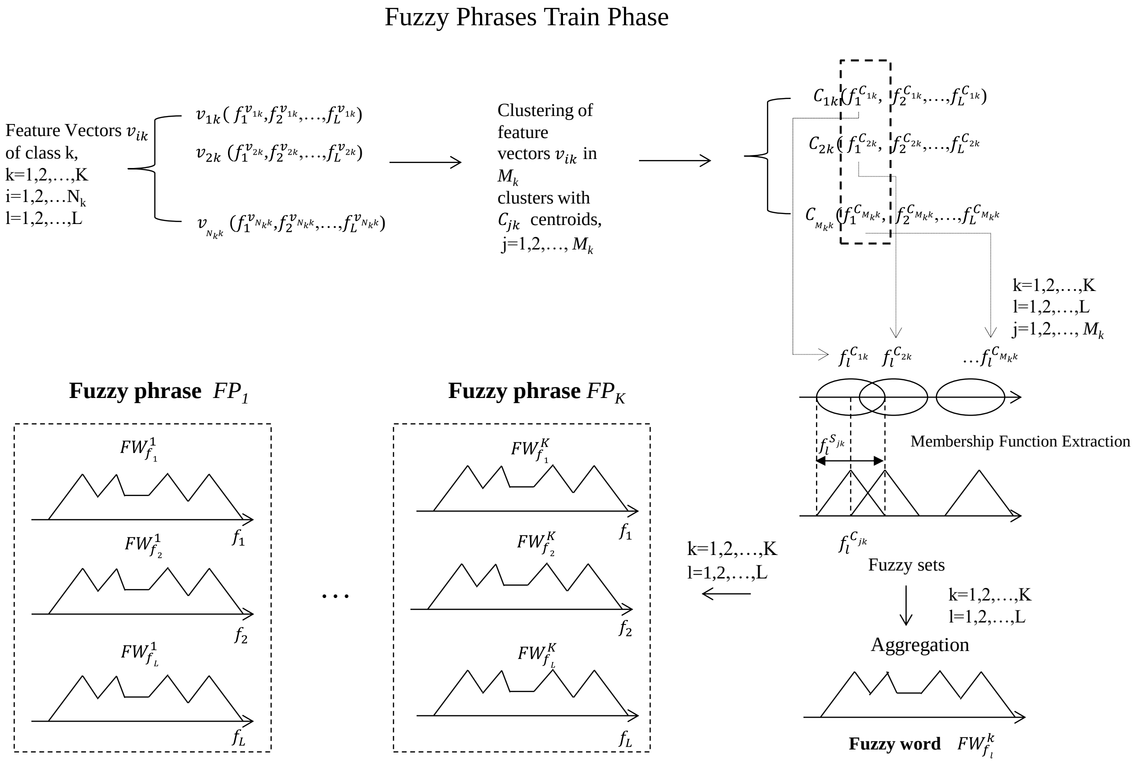

Fuzzy Phrases Train Phase
Feature Vectors v_{ik} of class k, k=1,2,…,K i=1,2,…N$_k$ l=1,2,…,L
$v_{1k}(f_1^{v_{1k}}, f_2^{v_{1k}}, \dots, f_L^{v_{1k}})$
$v_{2k}(f_1^{v_{2k}}, f_2^{v_{2k}}, \dots, f_L^{v_{2k}})$
$v_{N_k k}(f_1^{v_{N_k k}}, f_2^{v_{N_k k}}, \dots, f_L^{v_{N_k k}})$
Clustering of feature vectors v_{ik} in M_k clusters with C_{jk} centroids, j=1,2,…, M_k
$C_{1k}(f_1^{C_{1k}}, f_2^{C_{1k}}, \dots, f_L^{C_{1k}})$
$C_{2k}(f_1^{C_{2k}}, f_2^{C_{2k}}, \dots, f_L^{C_{2k}})$
$C_{M_k k}(f_1^{C_{M_k k}}, f_2^{C_{M_k k}}, \dots, f_L^{C_{M_k k}})$
k=1,2,…,K l=1,2,…,L j=1,2,…, M_k
$f_l^{C_{1k}}$ $f_l^{C_{2k}}$ …$f_l^{C_{M_k k}}$
$f_l^{S_{jk}}$
Membership Function Extraction
$f_l^{C_{jk}}$
Fuzzy sets
k=1,2,…,K l=1,2,…,L
Aggregation
Fuzzy word $FW_{f_l}^k$
k=1,2,…,K l=1,2,…,L
Fuzzy phrase FP_1
$FW_{f_1}^1$
f_1
$FW_{f_2}^1$
f_2
$FW_{f_L}^1$
f_L
…
Fuzzy phrase FP_K
$FW_{f_1}^K$
f_1
$FW_{f_2}^K$
f_2
$FW_{f_L}^K$
f_L

Figure 3.2 The Fuzzy Phrases method schematically. In the training phase the features of each feature vector are used for the extraction of the fuzzy phrases.

Fuzzy Phrases applies a clustering algorithm to cluster the feature vectors $v_{ik}, i = 1,2..N_k$, into a set of $M_K < N_K$ clusters. Every cluster has a centroid C_{jk}, which has the form $C_{jk}(f_1^{C_{jk}}, f_2^{C_{jk}}, \dots, f_L^{C_{jk}}), j = 1,2,\dots, M_K$. Each feature $f_l^{C_{jk}}$, l=1,2,…, L of the centroid C_{jk} of the jth cluster represents a centroid of the features $f_l^{v_{ik}}$ in l^{th} dimension with a respective standard deviation $f_l^{S_{jk}}$ of the features $f_l^{v_{ik}}$. After the computation of the centroid coordinates $f_l^{C_{jk}}$ and their standard deviations $f_l^{S_{jk}}$ of the features $f_l^{v_{ik}}$ in l^{th} dimension, a fuzzy set can be defined with a respective membership function having the form $\mu_l^{\left(f_l^{C_{jk}}\right)}\left(f_l^{v_{ik}}\right)$. For instance, if the membership functions are Gaussians, then

$$\mu_l^{\left(f_l^{C_{jk}}\right)}\left(f_l^{v_{ik}}\right) = e^{\frac{-(f_l^{v_{ik}} - f_l^{C_{jk}})^2}{2*p*(f_l^{S_{jk}})^2}} \qquad (1)$$

where p is a parameter given by the user.

Definition 3.3 The fuzzy sets of a class k, which are defined according to the aforementioned procedure, are aggregated using the union for the fuzzy sets. The new fuzzy sets defined by this aggregation operation are considered as *fuzzy words*

$$FW_{f_l}^k = \cup_{j=1}^{M_k} \mu_l^{\left(f_l^{C_{jk}}\right)}\left(f_l^{v_{ik}}\right) \qquad (3.2)$$

for l=1,2,…, L, k=1,2,…, K, and i=1,2…,N_k.

Let $\mu_l^{FW_{f_l}^k}$ be the aggregated membership function of each fuzzy word $FW_{f_l}^k$. A feature $f_l^{v_{ik}}$ of a feature vector v_{ik} is a member of fuzzy word $FW_{f_l}^k$ if

$$0 < \mu_l^{FW_{f_l}^k}\left(f_l^{v_{ik}}\right) \leq 1 \qquad (3.3)$$

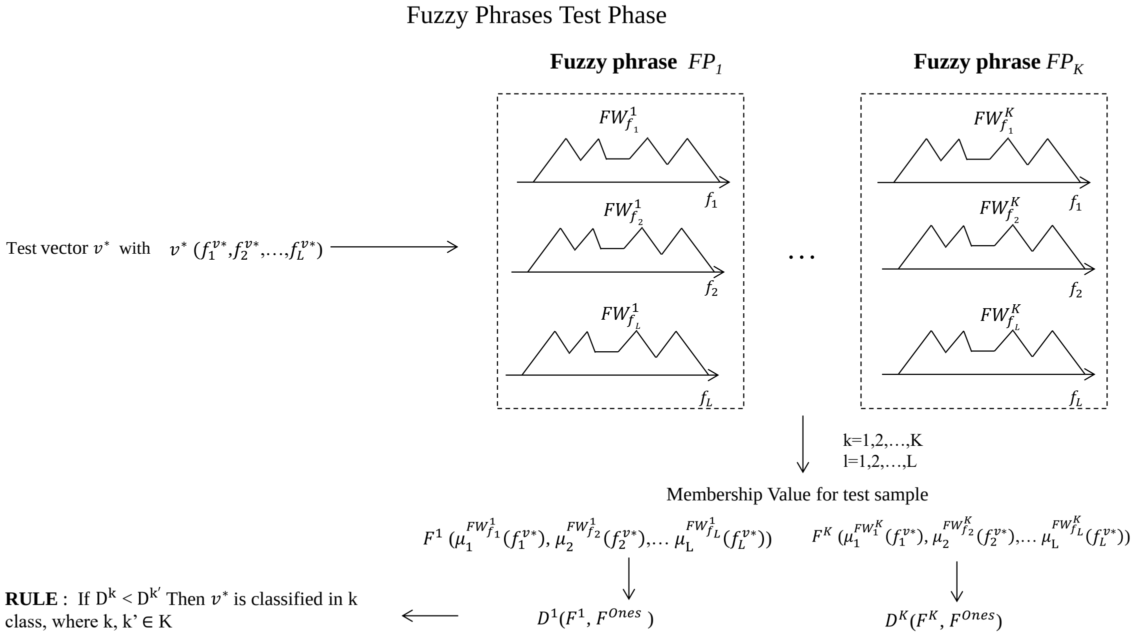
for l=1,2,…, L, k=1,2,…, K, and i=1,2…,N_k.

Definition 3.4 A set of fuzzy words $FW_{f_l}^k$ defines a *fuzzy phrase*, which is representative for class k

$$FP_k = \{FW_{f_1}^k, FW_{f_2}^k, \dots, FW_{f_L}^k\} \qquad (3.4)$$

3.3.2 Fuzzy Phrases testing

The test phase of the proposed Fuzzy Phrases classification model is illustrated in Fig. 3.3. During the test phase, let v^* be an unknown sample that is represented with the feature vector $v^*\left(f_1^{v^*}, f_2^{v^*}, \dots, f_L^{v^*}\right)$. For each feature $f_l^{v^*}$, l=1,2,…, L, the respective membership to the fuzzy sets $FW_{f_l}^k$, k=1,2,...,K. is computed.



Fuzzy Phrases Test Phase
Fuzzy phrase FP_1
Fuzzy phrase FP_K
$FW_{f_1}^1$
f_1
$FW_{f_2}^1$
f_2
$FW_{f_L}^1$
f_L
$FW_{f_1}^K$
f_1
$FW_{f_2}^K$
f_2
$FW_{f_L}^K$
f_L
Test vector v^* with $v^*\,(f_1^{v*}, f_2^{v*}, \dots, f_L^{v*})$
…
k=1,2,…,K
l=1,2,…,L
Membership Value for test sample
$F^1\,(\mu_1^{FW_{f_1}^1}(f_1^{v*}), \mu_2^{FW_{f_2}^1}(f_2^{v*}), \dots \mu_L^{FW_{f_L}^1}(f_L^{v*}))$
$F^K\,(\mu_1^{FW_1^K}(f_1^{v*}), \mu_2^{FW_{f_2}^K}(f_2^{v*}), \dots \mu_L^{FW_{f_L}^K}(f_L^{v*}))$
$D^1(F^1, F^{Ones})$
$D^K(F^K, F^{Ones})$
RULE : If $D^k < D^{k'}$ Then v^* is classified in k class, where k, k' ∈ K

Figure 3.3 The Fuzzy Phrases method schematically. In the test phase the features of the unknown test sample are described based on the membership calculated from its features to the fuzzy words. Finally, the test sample is classified based on the adopted decision rule

The overall membership of the feature vector v^* to a class k is represented by a membership vector $F^k(\mu_1^{FW_{f_1}^k}, \mu_2^{FW_{f_2}^k}, \dots \mu_L^{FW_{f_L}^k})$ where each feature $\mu_l^{FW_{f_l}^k}$ represents a membership to the fuzzy word $FW_{f_l}^k$.

Consequently, the test sample with feature vector $v^*\left(f_1^{v^*}, f_2^{v^*}, \dots, f_L^{v^*}\right)$ is described by F^k, k=1,2,...,K membership vectors, one for each class.

In order to classify the test sample with feature vector v^* to a class, a rule is adopted based on the distance of each membership vector $F^k(\mu_1^{FW_{f_1}^k}, \mu_2^{FW_{f_2}^k}, \dots \mu_L^{FW_{f_L}^k})$ k=1,2,...,K from a membership vector with components equal to one. More specifically, considering the fact that the maximum value of a membership is equal to one and that the minimum is equal to

zero, the ideal case of a sample from the class k is to have all the features of the membership vector F^k equal to one. Thus, a $F^{ones}(f_1^{ones}, f_2^{ones}, \dots, f_L^{ones})$ membership vector is defined to represent the ideal case, where each feature $f_l^{ones}, l = 1 \dots L,$ is $f_l^{ones} = 1$.
Let $D^k(F^k, F^{ones})$ be the distances of F^k, k=1,2,..., K from F^{ones}, and $D^{k\prime}(F^{k\prime}, F^{ones})$ be the distances of $F^{k\prime}$ and F^{ones}, where $k, k' \in K$ and $k \neq k'$, The following rule can be used for classification of v^* in class k

$$RULE\!: If\ D^k < D^{k\prime}\ \ then\ v^* classified\ in\ class\ k$$

3.3.3 Fuzzy Phrases for missing values

It can be noticed that such a classification approach can be applied even if the test data have missing values. For example, given an unknown sample vector $v^*\left(f_1^{v^*}, f_2^{v^*}, \dots, f_L^{v^*}\right)$ represented by the membership values $\mu_l^{FW_{f_l}^k}\left(f_l^{v^*}\right)$ of each feature $f_l^{v^*}, l = 1 \dots L$. Consequently, for features with missing values the membership value will be $\mu_l^{FW_{f_l}^k}\left(f_l^{v^*}\right) = 0$ for all *k=1…K*. Thus, Fuzzy Phrases achieves to exploit information for the unknown test sample relying on the rest of its features.

3.3.4 Fuzzy phrases for feature selection

The proposed model, Fuzzy Phrases, can be used for feature selection, in order to reduce the complexity of the classification task and identify the most informative features within a dataset. The Fuzzy Phrases Feature Selection (Fuzzy Phrases -FS) is based on the information, derived from the fuzzy words FW_l^k, defined during the training phase.

More specifically, Fig. 3.4 illustrates that each class k is represented by a fuzzy phrase FP_k, which is a set of fuzzy words FW_l^k. These fuzzy words are fuzzy sets that have been produced by clustering of the feature vectors v_{ik}. The presence of overlap between two fuzzy words indicates that these words carry redundant information; therefore, the respective feature $f_l^{v_{ik}}$ can be considered as less important. For example, the fuzzy word illustrated in Fig.3.4(b) is expected to be less informative than the fuzzy word illustrated in Fig. 3(a). The ovelap h of the fuzzy words FW_l^k and $FW_l^{k\prime}$ of the classes k and *k'* respectively, is estimated by the intersection divided by the union of these fuzzy words. However, h can be considered only as a weak indicator of redundancy, since it results from a non-deterministic clustering procedure (e.g., clustering algorithms, such as the k-means, usually depend on a random initialization and arbitrarily determined parameters, such as the number of clusters). A stronger redundancy indicator can be obtained by aggregation of multiple overlap observations from multiple executions of the clustering algorithm. Based on this approach, a feature is selected as informative if and only if all overlap observations of the respective words, are low.

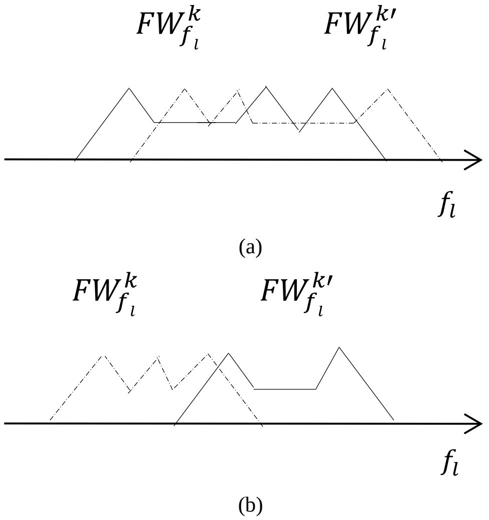


Figure 3.4(a) Fuzzy words $FW_{f_l}^{k}$ and $FW_{f_l}^{k'}$ of two different classes k and k' with high overlap. (b) Fuzzy words $FW_{f_l}^{k}$ and $FW_{f_l}^{k'}$ of two different classes k and k' that have lower overlap.

3.4 Results

Several experiments were performed in order to specify the representation of input data utilizing the Fuzzy Phrases model. The input data were from different data collections and were related with various real world problems.

3.4.1 Real-word data classification

The UCI (Blake & Merz 1998) and KEEL (Alcalá-Fdez et al. 2011) are two of the mostly used dataset repositories for machine learning. 16 real-life datasets available from the UCI and KEEL dataset repositories were used for the experimentally evaluation of classification performance of the proposed Fuzzy Phrase constructive fuzzy representation model. **Table 3.1** provides the main characteristics about these adopted datasets. The asterisk (*) above the name of some datasets indicated the existence of missing values in the instances of the dataset.

Table 3.1 Real-World Datasets

Data	Samples	Dimensionality
Australian	690	14
Balance	625	4
Sonar	208	60
Diabetes	768	8
Musk	6598	166
Spectheart	267	44
Wine	178	13
Glass	214	9
Haberman	306	3
Ionosphere	351	33
Seismic	2584	18
Yeast	1484	8
Liver	345	6
Wpbc	198	33
Breast Cancer*	699	10
Votes*	435	16
Mammographic mass*	961	5
Monk2	432	6
Parkisons	19	23
Heart(Statlog)	270	13

Since the Fuzzy Phrases algorithm was proposed for representation and classification model based on fuzzy logic, Fuzzy Phrase was compared with 14 state-of-the-art fuzzy classification approaches as well as with the SVM classifiers. **Table 3.2** presents the comparisons with the proposed fuzzy classifiers(FC) HID-TSK of (Zhang et al. 2017) , FRODT(Cai et al. 2019) and the fuzzy classifier of (Fu et al. 2019)**.** The comparison results were better or comparable with the results of the other classifiers. The usual classification tactic of the other classifiers was to omit the sample with missing values, as it was happened for datasets Breast Cancer and Vote. However, the Fuzzy Phrases was able to use all the samples and extract valuable information from the sample with missing values.

Table 3.2 Comparative results FP with fuzzy classifiers(FC) HID-TSK of (Zhang et al. 2017) , FRODT(Cai et al. 2019) and the fuzzy classifier of (Fu et al. 2019)

Data	Study		Study		Study		Study	
	Fuzzy Phrases		HID-TSK (Zhang et al. 2017)		FRODT (Cai et al. 2019)		(Fu et al. 2019)	
	Accuracy	Std	Accuracy	Std	Accuracy	Std	Accuracy	Std
Australian	**0.8735**	0.0196	0.8337	0.0187	0.8550	0.0100	0.8619	0.0188
Balance	0.7725	0.0384	0.8416	0.0155	N/A	N/A	**0.8811**	0.0168
Sonar	**0.7793**	0.0264	0.6585	0.0351	0.7458	0.0247	N/A	N/A
Diabetes	**0.7556**	0.0393	0.7182	0.0156	0.7496	0.0105	0.7513	0.0280
Musk	**0.8849**	0.0180	0.8763	0.0153	N/A	N/A	N/A	N/A
Spectheart	**0.8322**	0.0572	0.7520	0.0344	N/A	N/A	N/A	N/A
Wine	**0.9610**	0.0257	N/A	N/A	0.9468	0.0101	0.9472	0.0273
Haberman	0.7258	0.0457	N/A	N/A	**0.7457**	0.0970	0.7420	0.0380
Ionosphere	0.8717	0.0443	N/A	N/A	**0.8906**	0.0094	N/A	N/A
Seismic	0.8959	0.0279	**0.9321**	0.0009	N/A	N/A	N/A	N/A
Liver	0.6379	0.0607	0.6563	0.0212	**0.6658**	0.0259	N/A	N/A
Wpbc	**0.7928**	0.0268	0.7745	0.0029	0.7521	0.0181	N/A	N/A
Breast Cancer*	**0.9785**	0.0113	0.9512	0.0081	0.9713	0.0240	N/A	N/A
Votes*	**0.9497**	0.0282	0.9113	0.0121	N/A	N/A	N/A	N/A
Monk2	**0.7922**	0.0257	0.6494	0.0012	N/A	N/A	N/A	N/A
Heart(Statlog)	**0.8215**	0.0124	N/A	N/A	0.8156	0.0141	N/A	N/A

The classification results of Fuzzy Phrases in comparison to the classifiers PSO-FR (Chen et al. 2016)**,** NEWFM(Lee 2015) and FCCI-TSK(Wang et al. 2019) are presented in **Table 3.3.** The reported results of Fuzzy Phrases were better or comparable with the other studies, as in four out of eight dataset the performance of Fuzzy Phrases were better.

Table 3.3 Comparative results FP with PSO-FR (Chen et al. 2016), NEWFM(Lee 2015) and FCCI-TSK(Wang et al. 2019)

Data	Study		Study		Study		Study	
	Fuzzy Phrases		PSO-FR (Chen et al. 2016)		NEWFM (Lee 2015)		FCCI-TSK(Wang et al. 2019)	
	Accuracy	Std	Accuracy	Std	Accuracy	Std	Accuracy	Std
Diabetes	**0.7556**	0.0393	0.6759	0.186	N/A	N/A	N/A	N/A
Glass	**0.5990**	0.0521	0.5425	0.0441	N/A	N/A	N/A	N/A
Haberman	0.7258	0.0457	**0.7365**	0.0139	N/A	N/A	0.7609	N/A
Ionosphere	**0.8717**	0.0443	N/A	N/A	0.8488	N/A	0.7360	N/A
Yeast	0.4789	0.0147	**0.5270**	0.0116	N/A	N/A	N/A	N/A
Liver	0.6379	0.0607	0.5814	0.0267	N/A	N/A	**0.7067**	N/A
Parkisons	**0.8928**	0.0592	0.8147	0.0245	0.8308	N/A	N/A	N/A
Heart(Statlog)	**0.8215**	0.01245	N/A	N/A	0.8112	N/A	N/A	N/A

The comparison results of Fuzzy Phrases with the linear SVM classifier, as well as with the fuzzy classifiers FS-FCSVM and zero order TSK-FC (Zhang et al. 2017) are presened in **Table 3.4**. The overall classification ability of Fuzzy Phrases model exceeds the other classifiers, as in nine dataset comparison the Fuzzy Phrases has better results.

Table 3.4 Comparative results FP with linear SVM, FS-FCSVM and zero order TSK-FC (Zhang et al. 2017)

Data	Study		Study		Study		Study	
	Fuzzy Phrases		linear SVM		FS-FCSVM (Zhang et al. 2017)		zero order TSK-FC (Zhang et al. 2017)	
	Accuracy	Std	Accuracy	Std	Accuracy	Std	Accuracy	Std
Australian	**0.8735**	0.0196	0.7291	0.0243	0.7688	0.0501	0.7298	0.0399
Balance	0.7725	0.0384	**0.8566**	0.0278	0.6861	0.0803	0.7907	0.0416
Sonar	**0.7793**	0.0264	0.4824	0.0413	0.5572	0.0394	0.4930	0.0540
Diabetes	**0.7556**	0.0393	0.6975	0.0121	0.6605	0.0168	0.6687	0.0327
Musk	**0.8849**	0.0180	0.8458	0.0013	0.8460	0.0019	0.8685	0.0013
Spectheart	**0.8322**	0.0572	0.7492	0.0463	0.7317	0.0396	0.7516	0.0346
Seismic	0.8959	0.0279	**0.9340**	0.0019	0.9023	0.0011	0.9320	0.0016
Liver	0.6379	0.0607	**0.6458**	0.0275	0.5540	0.0502	0.5523	0.0573
Wpbc	**0.7928**	0.0268	0.7584	0.0126	0.7697	0.0138	0.7607	0.0094
Breast Cancer*	**0.9785**	0.0113	0.9171	0.0459	0.9498	0.0142	0.8846	0.0432
Votes*	**0.9497**	0.0282	0.8885	0.0374	0.8573	0.0237	0.8645	0.0087
Monk2	**0.7922**	0.0257	0.6270	0.0459	0.6473	0.0163	0.6322	0.0517

The details about the classification result of Fuzzy Phrases compared to the Bayesian TSK classifier (Gu et al. 2016), NCGMANF(Gao et al. 2019) and SVM with RBF kernel(Zhang et al. 2017) are provided in **Table 3.5**.

Table 3.5 Comparative results FP with Bayesian TSK classifier (Gu et al. 2016), NCGMANF(Gao et al. 2019) and SVM with RBF kernel(Zhang et al. 2017)

Data	Study		Study		Study		Study	
	Fuzzy Phrases		(Gu et al. 2016)		NCGMANF(Gao et al. 2019)		**RBF SVM**	
	Accuracy	Std	Accuracy	Std	Accuracy	Std	Accuracy	Std
Australian	**0.8735**	0.0196	0.8688	0.0354	N/A	N/A	0.7314	0.0243
Balance	0.7725	0.0384	0.9505(2vs3)	0.0120	0.8168	N/A	**0.8875**	0.0294
Sonar	**0.7793**	0.0264	N/A	N/A	0.7173	N/A	0.5294	0.0053
Diabetes	**0.7556**	0.0393	0.7722	0.0156	0.7375	N/A	0.7099	0.0274
Musk	**0.8849**	0.0180	N/A	N/A	N/A	N/A	0.8466	0.0013
Spectheart	**0.8322**	0.0572	N/A	N/A	N/A	N/A	0.7517	0.0432
Glass	**0.5990**	0.0521	N/A	N/A	0.5327	N/A	N/A	N/A
Haberman	0.7258	0.0457	**0.7587**	0.0293	N/A	N/A	N/A	N/A
Seismic	0.8959	0.0279	N/A	N/A	N/A	N/A	**0.9343**	0.0011
Liver	0.6379	0.0607	**0.6671**	0.0405	0.6430	N/A	0.6535	0.0281
Wpbc	**0.7928**	0.0268	N/A	N/A	N/A	N/A	0.7652	0.0134
Breast Cancer*	**0.9785**	0.0113	0.9600(no missing-values)	0.0160	0.6794	N/A	0.9062	0.0371
Votes*	**0.9497**	0.0282	N/A	N/A	N/A	N/A	0.9049	0.0111
Mammographic mass*	0.8233	0.0418	**0.8313**	0.0210	N/A	N/A	N/A	N/A
Monk2	**0.7922**	0.0257	N/A	N/A	N/A	N/A	0.6343	0.0294
Heart(Statlog)	0.8215	0.01245	**0.8562**	0.0363	N/A	N/A	N/A	N/A

Finally, **Table 3.6** concentrates the last comparisons between the proposed Fuzzy Phrases and the fuzzy classifiers FC (Hu et al. 2018), FCO (Hu et al. 2018) and (Pota et al. 2017).

Table 3.6 Comparative results FP with FC (Hu et al. 2018), FCO (Hu et al. 2018) and (Pota et al. 2017)

Data	Study		Study		Study		Study	
	Fuzzy Phrases		(Hu et al. 2018) **FC**		(Hu et al. 2018) **FCO**		(Pota et al. 2017)	
	Accuracy	Std	Accuracy	Std	Accuracy	Std	Accuracy	Std
Diabetes	0.7556	0.0393	0.7662	0.0372	**0.7727**	0.0471	0.7734	N/A
Spectheart	0.8322	0.0572	0.8667	0.0435	**0.8778**	0.0699	N/A	N/A
Wine	0.9610	0.0457	0.9727	0.0614	0.9727	0.0614	N/A	N/A
Haberman	0.7258	0.0457	N/A	N/A	N/A	N/A	0.7614	N/A
Ionosphere	0.8717	0.0443	0.9405	0.0388	**0.9491**	0.0414	N/A	N/A
Breast Cancer*	**0.9785**	0.0113	N/A	N/A	N/A	N/A	0.9757	N/A
Mammographic mass*	**0.8233**	0.0418	0.8096	0.0491	0.8096	0.0424	N/A	N/A
Parkisons	**0.8928**	0.0592	0.8792	0.0860	0.8897	0.0513	N/A	N/A
Heart	0.8215	0.01245	0.7519	0.0553	0.7519	0.0553	**0.8284**	N/A

3.4.2 Bone fracture detection

Fuzzy Phrases was utilized for the problem of bone fracture detection(Vasilakakis, Iosifidou, et al. 2019). The database used in the experiment consisted of 790 x-ray bone images of upper and lower extremity collected from Public General Hospital in digital Portable Network Graphics (PNG) format. The method used to create images with x-radiation, is to pass an x-ray beam through the body section, which is needed to be examined. The characteristic parameters of the x-ray beam are Peak kilovoltage (kV) and Milliampere-seconds (mAs). Based on the type of bones the parameters of the x-ray beams are varied. For example, x-rays images of small bones of the body (wrist, ankle etc) need kV=43 and mas=5, while long bones (Humerus, femur etc) need KV=60 and mas=18. 300 x-ray bone images were randomly selected from this database, 200 normal and 100 abnormal (fractured) x-ray bone images. The images where sampled using 32×32 pixel sub-images, in order to evaluate the local texture of the bones. The sub-image size was determined so that the number of pixels belonging in the fracture to be approx. 20-30% of the total number of pixels in the sample. Examples are illustrated in **Figure 3.5**. It can be noticed that the samples with the bone fractures are characterized by directional textural patterns. This justifies the texture analysis approach considered in this study.

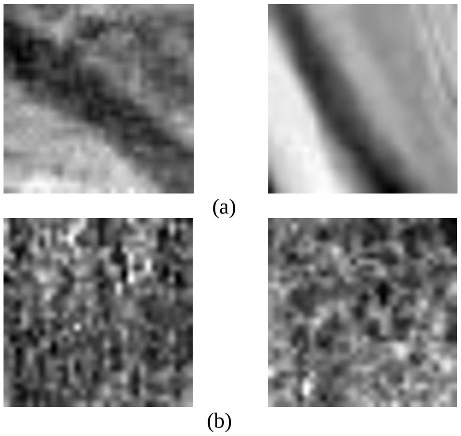


Figure 3.5 Samples obtained from different x-ray images of the database used (a) Samples with a bone fracture. (b) Samples without a bone fracture.

Several experiments were performed in order to evaluate the proposed classification methodology for the analysis of the bone x-ray images. In the case of Fuzzy Phrases classification method, k-means (Drake & Hamerly 2012)was used to cluster the different attributes of normal (without fractures) and abnormal (with fractures) classes. Different numbers of clusters for attributes belonging to vectors of normal and abnormal classes were examined. The number of clusters tested was in the range from 1 to 5 clusters. Also, Gaussian membership functions used, for p ranging from 1 to 5; Slightly better performance was observed using 3 clusters with p = 2.

For comparison purposes GLCM-based (Umadevi & Geethalakshmi 2012), the Hough Transformed-based (Donnelley & Knowles 2005) and the DWT-based (Al-Ayyoub et al. 2013)feature extraction approaches were used. The classification of the feature vectors obtained by these methods was implemented by an SVM classifier. A linear, polynomial and Radial Basis Function (RBF) kernels, and followed the grid-search approach (Chang & Lin 2011)to determine its optimal parameters were tested. The RBF kernel provided the best results, for a minimum cost parameter c = 10.

Table 3.7 Bone Fracture Classification Results

Classifier	**Features**	***Accuracy***	***Sensitivity***	***Specificity***
SVM	GLCM	0.71	0.38	**0.89**
SVM	HT	0.79	0.53	**0.94**
SVM	DWT	0.80	0.59	**0.92**
FP	GLCM	**0.77**	**0.81**	0.75
FP	HT	**0.82**	**0.73**	0.86
FP	**DWT**	**0.84**	**0.81**	0.86

The classification performance was thoroughly investigated using the accuracy as well as the sensitivity and specificity of the classification of the test images. Experiments were performed using the 10-fold cross validation evaluation scheme. The results with regards to the bone fracture detection are summarized in **Table 3.1**. It can be noticed that the Fuzzy Phrases method achieved higher overall results both in terms of accuracy and sensitivity than the SVM classifier. The best results were performed by the Wavelet transformed features **Figure 3.6**.

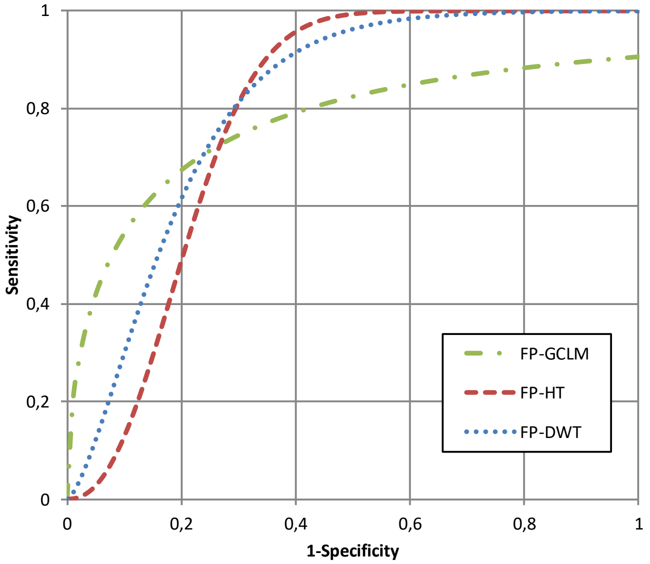


Figure 3.6 The ROCs achieved by the proposed FP classification method.

The average training and testing time of the DWT-FP was 0.36 seconds and 0.038 seconds respectively, while the training and testing time of SVM was 0.57seconds and 0.047 seconds respectively.

3.4.3 Fuzzy Phrases for Heart Disease detection and Heart Failure prediction results

The experimental validation of Fuzzy Phrases classification model was performed using two datasets from the cardiology domain. These datasets have been used in recent studies and have been considered to enable comparisons with the state-of-the-art. Among the studies investigating the detection of Heart Disease (HD) with fuzzy computational approaches, a recent one(Lee 2015), was based on a supervised feature selection method, considering the bounded sum of weighted fuzzy membership functions. Also, in a previous study was presented (Koulaouzidis et al. 2016), a methodology for prediction of patients at high risk of Heart Failure (HF), using features extracted from daily collected physiological data, based on Multi-Resolution Analysis (MRA) using the Discrete Wavelet Transform (DWT).

The first dataset is Statlog Heart dataset is from the Cleveland Clinic Foundation, contains 270 samples. The dataset is publicly available from UCI repository (Blake &

Merz 1998). The classification problem under study with respect to this dataset is the absence (150 samples) or presence (120 samples) of heart disease, based on 13 features, which are presented in **Table 3.2**.

The second dataset is the Heart Failure dataset based on retrospective telemonitoring (TM) data from 308 patients in Kingston-upon-Hull(Koulaouzidis et al. 2016). The data, which are presented in **Table 3.3**, have been collected using the Motiva telemonitoring system (Domingo et al. 2012)as part of the daily HF service between 2010 and 2013. An admission as HF Hospitalization (HFH) was defined based on the first diagnosis, which should be ‘Heart Failure’. The study is limited on the cases for which death has been reported (6.5%) and considers only the information extracted from the monitored physiological signals heart rate (HR), systolic blood pressure (SBP), diastolic blood pressure (DBP), and body weight (BW). The MRA features extracted from Heart Failure dataset correspond to time-intervals of 4-day patient monitoring. The dataset is highly imbalanced, with only 0.2% of the vectors representing Worsening HF (WHF) precursor patterns, i.e., vectors corresponding to HFH. This makes the classification task even more challenging. Another challenge in this dataset is that in one fourth of the patients' data there are missing values during consecutive 4 days prior to HFH, respectively. This is due to low compliance of the patients to the use of TM.

In both case studies the k-means was used for clustering, with k ranging from 1 to 12 clusters. Also, the Gaussian membership function was utilized for p ranging from 1 to 2. Regarding the Fuzzy Phrases-FS the number of clustering executions was set to 5.

The classification performance was thoroughly investigated using the accuracy, the sensitivity and specificity of the classification of the test data as well as the Area Under the receiving operating Characteristic (AUC). Experiments were performed using the 10-fold cross validation evaluation scheme in order to limit the bias in the selection of the vectors used for training and testing the classifier.

Table 3.8 13 Features of Statlog Heart dataset

Feature type	**Feature Description**		**Feature value**
Numerical	Age	(f_{01})	[29, 77]
Binary	Sex	(f_{02})	[0, 1]
Nominal	Chest pain type	(f_{03})	[1, 4]
Numerical	Resting blood pressure	(f_{04})	[94, 200]
Numerical	Serum cholesterol in mg/dl	(f_{05})	[126, 564]
Binary	Fasting blood sugar > 120 mg/dl	(f_{06})	[0, 1]
Nominal	Resting electrocardiographic results	(f_{07})	[0, 2]
Numerical	Maximum heart rate achieved	(f_{08})	[71, 202]

Binary	Exercise (f_{09})	[0, 1]
Numerical	Oldpeak = ST depression induced by exercise relative to rest (f_{10})	[0.0, 6.2]
Nominal	The slope of the peak exercise ST segment (f_{11})	[1, 3]
Nominal	Number of major vessels (f_{12})	[0, 3]
Nominal	Thal: Defect type (f_{13})	[3, 7]

Table 3.916 features of Heart Failure dataset

Feature type	**Feature Description** **(day$_1$, day$_2$, day$_3$, day$_4$)**
MRA Numerical	Heart rate (HR) (f_{01}, f_{05}, f_{09}, f_{13})
MRA Numerical	Systolic blood pressure (SBP) (f_{02}, f_{06}, f_{10}, f_{14})
MRA Numerical	Diastolic blood pressure (DBP) (f_{03}, f_{07}, f_{11}, f_{15})
MRA Numerical	Body weight (BW) (f_{04}, f_{08}, f_{12}, f_{16})

Experiments were conducted with and without feature selection using the proposed model. The results are summarized in **Table 3.4**, in comparison with the results obtained from the application of state-of-the-art methods on the same datasets.

It can be noticed that the overall performance of Fuzzy Phrases is comparable or higher to the performance of the state-of-the-art methods, whereas the performance of Fuzzy Phrases-FS is higher in all compared cases. The subset of features selected by Fuzzy Phrases-FS is {f_{03}, f_{04}, f_{05}, f_{08}, f_{09}, f_{11}, f_{12}, f_{13}}. The respective Receiver Operating Characteristics (ROCs) (Fawcett 2006) are presented in **Fig. 3.6**.

Table 3.10 Statlog Heart dataset Comparative Results

Method	**Evaluation metrics**			
	AUC	***Accuracy***	***Sensitivity***	***Specificity***
FP	0.81	0.82	**0.80**	0.87
FP-FS	**0.88**	**0.85**	0.71	**0.94**
Lee (Lee 2015)	N/A	0.82	N/A	N/A
Hu *et al.* (Hu et al. 2018)	0.74	0.75	N/A	N/A

Table 3.11 Heart Failure dataset Comparative Results

Method	Evaluation metrics			
	AUC	*Accuracy*	*Sensitivity*	*Specificity*
FP	**0.80**	**0.93**	**0.50**	**0.94**
FP-FS	0.78	0.88	0.38	0.88
MRA (Koulaouzidis et al. 2016)	0.76	N/A	0.47	0.96
MRA BW (Koulaouzidis et al. 2016)	0.75	N/A	0.38	0.98
MRA BW, DBP (Koulaouzidis et al. 2016)	0.77	N//A	0.48	0.96

The proposed Fuzzy Phrases framework was compared with the state of the art study of Koulaouzidis et al. (Koulaouzidis et al. 2016). In (Koulaouzidis et al. 2016) a Naïve Bayes classifier was used to predict a possible HFH event for the Heart Failure dataset, i.e., to classify the MRA vectors into two classes, namely the normal and HFH classes. The results are summarized in Table IV, which shows that Fuzzy Phrases outperforms the state-of-the-art methods. On the other hand, the performance of Fuzzy Phrases -FS is lower than Fuzzy Phrases. This could be attributed to the fact that Heart Failure dataset is particularly imbalanced, which can affect the clustering process applied for the formation of the fuzzy words. The features selected using Fuzzy Phrases -FS include {f_{04}, f_{08}, f_{12}, f_{16}}, which represent the BW, in agreement with the results of relevant medical studies. The respective ROCs are presented in **Fig. 3.6**

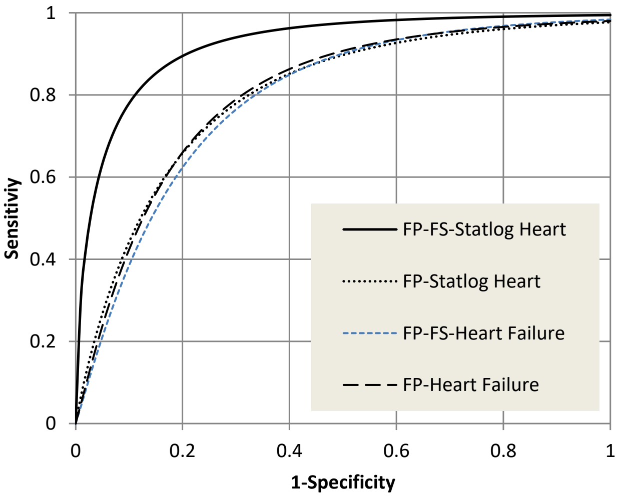


Figure 3.7 The ROCs achieved by the proposed Fuzzy Phrases (FP) for Statlog Heart and Heart Failure datasets respectively.

3.5 Conclusion

This chapter investigated a novel, generic model for data classification, named Fuzzy Phrases. This model manages to enhance the expressivity of crisp feature spaces, such as those considered in the experimental study of the paper. The proposed model resembles the human way of expression using multiple different words to describe and make decisions about real-world concepts. Furthermore, Fuzzy Phrases was a generic model that was efficiently extended to a novel feature selection methodology, named Fuzzy Phrases-FS.

The performance of both Fuzzy Phrases and Fuzzy Phrases -FS were investigated and evaluated. Both approaches resulted in a better or comparable performance from the previously reported methodologies for detection of HD and prediction of HF. The case study for HD showed that Fuzzy Phrases-FS can result in a better classification performance than Fuzzy Phrases using a smaller number of features. Also, of particular interest for medical applications, is the inherent tolerance of Fuzzy Phrases in missing data.

The Fuzzy Phrases classification method is still in an early stage. Further investigation is required to fully explore all the potentials of this method. Potential areas for further research include alternatives rules for decision making, and systematic evaluation of its robustness to noise and the presence of missing values.

CHAPTER 4

GASTROINTESTINAL IMAGE ANALYSIS: BACKGROUND AND STATE-OF-THE-ART

This chapter summarizes background and the state-of-the-art of CE as a multidisciplinary field and aims to provide an overview of both medical and technological advances, including concept and prototype capsule endoscopes, and software methodologies for enhanced visualization, abnormality detection and capsule localization.

4.1 Introduction

4.1.1 The beginning of Wireless Capsule Endoscopy

The imaging of gastrointestinal (GI) tract for diagnosis and treatment of diseases is a matter of great significance, because the examination of the whole GI tract due to its large size is still remaining a challenging task. The beginning of for the imaging of GI tract was back in back in 1868, when the flexible endoscopes were first used by Wolf and Schindler (Sliker & Ciuti 2014). From then, the endoscopes have evolved to an efficient and reliable tool for the screening through different segments of the GI tract, such as esophagus, stomach, large bowel or colon and part of the small bowel. Other imaging techniques of GI tract are magnetic resonance enterography (MRE), or computed tomography enterography (CTE).

Although, the conventional endoscopes can efficiently diagnose pathologies of GI tract, they still have safety issues as they are able to traumatize the patient. The discomfort of the patient during the endoscopic procedures, such as colonoscopy, must be considered. Also, the main drawback of conventional endoscopes is still their inability to reach and examine all the areas of the GI tract, such as most of the small bowel.

The beginning of the 21st century was also the beginning of a new technology keen to overcome the discomfort and the pain during the examination of GI tract. A capsule endoscope that has the size of large vitamin pill with the ability to capture and transmit wirelessly video frames was first proposed by (Iddan et al. 2000). After that, Wireless Capsule Endoscopy (WCE) or Capsule Endoscopy (CE) has become widely adopted into conventional clinical practice since the introduction of the first commercial model in 2001(Iakovidis & Koulaouzidis 2015; Vasilakakis, Koulaouzidis, Yung, et al. 2019), because it was able to overcome the drawbacks of the conventional endoscopes. Due to the ability of the capsule to be swallowed, CE can be characterized to be a minimal invasive procedure, which is a user-friendlier examination method, than the conventional endoscopes. The biggest advantage of CE is the visualization of all segments of GI tract and especially the small bowel. For this reason, CE has known great success in the field of small bowel examination as well as monitoring of bowel diseases.

4.1.2 Diagnostic Yield in CE

The common significant pathologies that can be found in the GI tract are classified into a few main groups: vascular lesions, neoplasms (i.e. polyps/tumours), and inflammatory lesions – idiopathic inflammatory bowel disease (IBD) and pharmacogenetic or infectious inflammation. CE has proven valuable for clinical problems such as small bowel bleeding and the investigation or monitoring of inflammatory bowel disease as well as the ability to directly visualize the small bowel mucosa. The data so far show that small bowel (SB) CE has an overall diagnostic yield (DY) of about 50% (Koulaouzidis et al. 2012; Koulaouzidis, Rondonotti, et al. 2013). There are two main aims of the CE for the improvement of DY. The first one is the identification and selection of patients who are likely to have relevant findings. The second one is optimization of images obtained. Indication is the key in the selection of patients who are likely to benefit from CE; as detailed previously, CE is of diagnostic value in certain well-established indication groups.

A group of patients who benefit from SBCE are those with known or suspected inflammatory bowel disease (IBD). Crohn’s Disease (CD) affects only the SB in up to a third of patients, in which case CE is valuable for both diagnosis and monitoring of treatment response (Kopylov & Seidman 2014), and can play a complementary role compared to MRI enterography (Enns et al. 2017). Emerging data now show that CE has comparable accuracy to conventional methods for the investigation of SB pathologies, such as ileocolonoscopy, enterography and push or device assisted-enteroscopy (Enns et al. 2017; Kopylov et al. 2017), whilst offering the advantages of minimal invasiveness and direct mucosal visualization.

However, age is another significant factor which has been shown to affect DY (Diana E Yung, Rondonotti, Giannakou, et al. 2017). Elderly patients have a higher overall DY, especially in the setting of suspected small bowel bleeding, and the most common findings are angioectasias and bleeding. Furthermore, even a negative CE examination is of value in guiding further investigation and management, as pooled data have shown that patients with no significant findings on CE have a much lower rate of re-bleeding (19%) compared to those with a positive capsule (40%)(Diana E Yung, Koulaouzidis, et al. 2017). This is especially important in the significant subgroup of patients with iron deficiency anaemia (IDA), who are often referred for repeated capsule investigations with poor diagnostic yield (DY)(Sonnenberg 2015; Woodward et al. 2016). CE is useful for monitoring or detection of these pathologies of the small bowel, such as CD or bleeding, because the small bowel is predominantly inaccessible to regular endoscopes due to its long and convoluted structure.

Younger patients are more likely to be referred for the investigation of known or suspected IBD, with ulcers the more common finding in this group; however, notably, young patients referred with IDA are more likely to have significant SB findings such as SB malignancies. Timing of CE examination in the setting of SB bleeding is another important factor, with studies showing that carrying out a capsule closer to the onset of symptoms improves DY(Singh et al. 2013). Some studies have even investigated the use of CE in the acute to semi-acute setting as a method of triage or to guide further management of such patients(Chandran et al. 2013; Gralnek et al. 2013; Gutkin et al. 2013; Meltzer et al. 2013; Sung et al. 2016).

Although there remain significant technological limitations to the quality of images which can be obtained by current models of capsule endoscopes, clinical methods to optimize the images, which are obtained have been investigated. Chiefly, the use of simethicone has been shown in meta-analysis to improve image clarity by reducing interference from gas and bubbles in the small bowel(Koulaouzidis, Giannakou, et al. 2013). The need for bowel preparation with laxatives remains a controversial topic, with conflicting results from various studies and meta-analyses(Diana E Yung, Rondonotti, Sykes, et al. 2017).

Finally, the reporting of a CE examination is hindered by long reading times and reviewer’s concentration span (Kim et al. 2018). Long reading times in SBCE can be reduced by optimizing reading settings and by using “QuickView” mode when appropriate; however, it may result in missed findings, thus the use of this function is recommended mostly when panenteric pathology is expected(Mitselos & Christodoulou 2018). Another approach to counteract reader’s stress and reducing attention span is by pausing and replaying CE video segments, thus allowing the reader to rest. However, this way the reading time may eventually increase, thus affecting the overall productivity of the reviewer. In summary, although CE has come a long way since its introduction to clinical practice, there remain several key limitations yet to be addressed in order to optimize its utilization.

The significant advantages of CE have led to the further development of capsules for colonic, oesophageal and gastric investigation (gastric capsule endoscopes are not yet available; technological advances aim at capsule locomotion control and tissue sampling (Hale et al. 2015)).

In general, the appeal of CE as a minimally-invasive, more comfortable alternative to conventional endoscopy can be extended to indications outwit the SB. Colon CE(CCE) has been shown in several studies to have a comparable DY for colonic pathology such as polyps and colitis(Spada et al. 2015). Therefore, although not as widely used, it is recommended as a viable option for colon investigation in patients who failed or unwilling to undergo conventional colonoscopy. Oesophageal CE (OCE) has also been investigated in a smaller number of studies and may offer benefit to patients with coagulation disorders who require repeated oesophageal surveillance for varices (Parker et al. 2015). At this time, esophageal capsules do not demonstrate to confer any diagnostic advantage against oesophagogastroduodenoscopy (OGD) (Park et al. 2018).

4.1.3 Motivations for further improvement

Advances in CE technology have tended to concentrate on the areas of image quality, battery life and processing software. Since its inception, the main hardware constraint has been the size and volume of the capsule, estimated approximately at 2 cm^3, which limits both the quality and quantity of its components. Therefore, image quality remains inferior to that of conventional endoscopy, with lack of capacity for image enhancement technologies, such as high definition and narrow band imaging. Moreover, there is at present no reliable technology for localization or steering, and current capsules are unable to collect and procure tissue samples.

From the point of view of several comprehensive review papers on CE (Iakovidis & Koulaouzidis 2015; Vasilakakis, Koulaouzidis, Yung, et al. 2019; Koulaouzidis et al. 2015; Fisher & Hasler 2012), it is clear that CE is a rapidly evolving, multidisciplinary field, and the annual tally of relevant research contributions is high. More specifically, since 2009 the annual number of publications tagged with “capsule endoscopy” as a keyword, across all fields, is in the range of 500 per year. The majority of these publications are medical, although a reducing from 92% in 2009 to 79% in 2017. This reflects the interest levels of non-medical (mainly biomedical engineering and information technology) scientific communities, to address the challenges posed by the still-open practical issues of CE affecting its DY.

4.2 State-of-the-art Commercial Wireless Capsule Endoscopes

In the first 18 years of commercial availability of CE the main players in the CE market are the five companies. The names of these companies are Medtronic, Intromedic, Olympus, JINSHAN Science &Technology Co. Ltd, and CapsoVision Inc. Recently, a number of other companies from China have also claimed a share of the market. These new companies are ANKON Technologies Co., Ltd and Shangxian Minimal Invasive Inc. They provide diagnostic technology, especially for SB investigation, but also products for the non-invasive exploration of the oesophagus and the colon.

The commercially-available capsule endoscopes share common manufacture principles i.e. a pill-shaped device with, usually, 1-2 domed cameras on either end. They contain Complementary Metal-Oxide Semiconductor (CMOS) cameras and light-emitting diodes (LEDs) in various numbers and configurations, miniaturized batteries and components for data transmission or storage. The average weight of a capsule endoscope is approximately 3-4g, and most capsule models require an external data receiver and recorder system. Each model comes with its own proprietary reading software, most of which offer some form of image enhancement technology. **Table 4.1** contains a comparative summary of existing capsule models on the market.

Table 4.1 Specifications of current commercial capsule endoscopes

Company	**Capsule endoscope**	**Dimensions (mm)**	**LED lights**	**FPS**	**Weight (g)**	**Resolution (pixels)**	**Battery life (h)**	**Reviewing software**	**Optical enhance-ments**	**Image Sensor**
Medtronic	Pillcam® SB3	11.4×26.2	4	2-6	3.0±0.1	340×340	≥11.5	RAPID 9.0	Blue mode; FICE	CMOS
Medtronic	Pillcam® Crohn’s Capsule	11.6×32.3	8	4-35	2.9 + 0.1	* n/a	≥10	RAPID 9.0	Blue mode; FICE	CMOS
Medtronic	Pillcam® COLON 2	11.6×31.5	8	4-35	2.9±0.03	* n/a	≥10	RAPID 9.0	Blue mode; FICE	CMOS
Medtronic	Pillcam® UGI	11.6×32.3	4	18-35	2.9 ± 0.1	* n/a	90 min	RAPID 9.0	Blue mode; FICE	CMOS
IntroMedic Co Ltd	Mirocam® 1200	10.8×24.5	6	3	3.25	320×320	≥10	MiroView™ U 4.0	n/a*	CMOS
IntroMedic Co Ltd	MiroCam 2000	10.8×31.1	12	6	3.5±0.1	320 ×320	12	MiroView™ U 4.0	n/a	CMOS
Olympus Corporation	EndoCapsule® EC-S10	11×26	4	2	3.3	512×512	12	Endocapsule Software 10 Server	Image adjustment function	CMOS
Capsovision	CapsoCam® Plus	11×31	16	20	4	221,884	15	CapsoView®	Advanced Color Enhancement (ACE)	CMOS

Jinshan Science and Technology	OMOM	13×27.9	6	2	6	640×480	~12	VUE™	RGB Imaging Color Enhancement	CMOS
Jinshan Science and Technology	OMOM Capsule 2	11×25.4	6	2	4.5	256×256	> 10 hours	VUE™	RGB Imaging Color Enhancement	CMOS
Ankon Technologies Co., Ltd.	Ankon NaviCam®	11.8× 27	* n/a	* n/a	5	480×480	>8hours	* n/a	* n/a	CMOS
Shangxian Minimal Invasive Inc.	Capsubot	11×28	* n/a	* n/a	3.9	* n/a	* n/a	* n/a	* n/a	* n/a

* n/a (not available).

4.2.1 PillCam™ CE

The PillCam™SB 3 (PillCam™ 2016) is the latest commercial SB capsule from Medtronic. The image capturing software offers adaptive variable frame rate, which changes from 2 to 6 frames per second (fps) based on capsule speed as it is propelled through the small bowel. Images captured are transmitted by a radiofrequency (RF) transmitter to a receiver belt worn around the patient’s waist, and then stored in an external data recorder which has real-time viewing and image capture/thumbnailing capability. A study conducted by Medtronic concluded that PillCam™SB3 had an approximately 40% increase in DY for relevant pathology compared to the earlier model (Dunn et al. 2014).

The proprietary reading software is the RAPID™Reader 9.0 version. The software provides additional diagnostic features and study reviewing aids. It contains enhanced controls similar to the ribbon-type toolbar concept used in Microsoft® products, including the Lewis Score (LS) (Gralnek et al. 2008) calculator, the Fujinon Intelligent Colour Enhancement (FICE) (D. E. Yung et al. 2017) the suspected blood indicator(Kopylov et al. 2017), QuickView (QV) (Koulaouzidis et al. 2015), a thumbnail comparison feature, backward compatibility with studies from previous RAPID™ software versions and an improved progress indicator/localization guide.

A further iteration of the PillCam is the Crohn’s capsule, which has two heads located on both ends of the capsule in order to increase mucosal coverage. It has been developed specifically for the investigation and monitoring of patients with Crohn’s disease, who often require serial visualization of the digestive tract. The accompanying software reflects this, with features to facilitate LS calculation and comparison of capsules in patients undergoing repeat examinations.

In addition, Medtronic has developed the Pillcam™COLON2 and Pillcam™Upper GI (UGI) tract. Pillcam™COLON2 is designed for colon investigation, while Pillcam™UGI enables direct visualization of the upper GI tract, sacrificing recording time for a greatly increased image capture rate of up to 35 fps as a reflection of the much faster transit time through the oesophagus.

4.2.2 MiroCam®

The MiroCam (http://www.intromedic.com/eng/main/ 2018) (which stands for Micro Intelligent Robotic Object Camera) was developed by the intelligent Microsystem center established by the Korea Ministry of Science & Technology in Seoul, South Korea, renamed IntroMedic Co Ltd in 2006. The standard model is MC1000-W®. The newer MC2000 capsule has two CMOS cameras, located at both ends of the capsule, and provides a bidirectional 12-hour view of the SB. Instead of using RF to transmit video images to a data recorder, the MiroCam utilizes electric field propagation to transmit data, which is described as human body communication (HBC). HBC uses the capsule itself to generate an electrical field and the human body as a conductive medium for data transmission, with less power consumption compared to RF transmission. Images are transmitted to a sensor array which must be kept in direct contact with the patient’s body, and stored on an external data recorder. The data recorder has real-time viewing and image capture capabilities, whereas older models without a LCD display allow real-time viewing through USB connection to a notebook or wirelessly to a smartphone or tablet.

The MiroCam®Navi (Rahman et al. 2014) magnetic capsule is marketed for visualization of the stomach, and as such has limited steering capabilities using magnetic force. It can be controlled with an external hand held magnet, allowing clinicians to position the capsule to direct gastric views via real time imaging.

4.2.3 Endocapsule 10

The Endocapsule (EC-S10) (https://www.olympus-europa.com/medical/en/Products-and-Solutions/Products/Product/ENDOCAPSULE-10-System.html 2018) is the newest version of the small bowel CE manufactured by Olympus Corporation, Japan and approved for use by the Food and Drug Association (FDA). Images are captured by an eight-sensor array inserted into a belt worn around the waist. The data recorder has a battery life of 12 h and has real-time viewing capabilities allowing capture of images and playback.

The proprietary reading software for this capsule is the Endocapsule System 10 software. The system tracks the capsule's journey through the small intestine utilizing the receiving signal of the belt-style antenna. The 3D Track function allows the system to display in 3D the capsule track. This functionality provides an estimation of the location of captured images and in this way assists the detection of abnormalities location. It is also supports Omni-selected Mode (Hosoe et al. 2016), which skips over images that overlap with previous ones and can recognize similar images when captured from different angles. The Adjust mode also changes playback speed depending on the differences detected in the images.

4.2.4 CapsoCam Plus®

The CapsoCam® Plus (http://www.capsovision.com/physicians/product-specifications 2018) (Capsovision, Saratoga, CA) capsule is unique in having 4 CMOS cameras placed around the body of the capsule at 90° angles, giving a 360° side-on panoramic field of view. Instead of transmitting images like other CE devices, CapsoCam®Plus uses a large-capacity onboard storage system and does not require external receiver equipment. The capsule itself must be retrieved, following which images are downloaded directly from the capsule to a workstation for review.

The proprietary reading software is CapsoView 3.4. It includes an automated Red Detection system that highlights suspected images of bleeding and Advanced Color Enhancement (ACE) technology which uses computed spectral sequences based on the in vivo image data to enhance tissue characterization.

4.2.5 OMOM®

Jinshan Science and Technology, Chongqing, China has developed the OMOM Capsule (http://english.jinshangroup.com/capsuleendoscopy.html 2018). The OMOM Capsule is still not available for use in the USA. A unique and notable difference of the OMOM capsule is that the physician can view the captured images in real time and send a signal to the capsule to change the frame rate from 0.5 to 1 or 2 fps, in order to optimize visualization. The latest version OMOM capsule 2 has a wider field of view and it is slightly smaller. Images are sent via RF transmission to a receiver belt and external data recorder.

The proprietary review software is VUE™. VUE offers bleeding detection, which highlights portions of the recording that appear red and quick view to find areas of interest in short time. For Image Enhancement (ICE) RGB mode separates the different RGB spectrums in the images, in order to achieve a more detailed view of mucosa and capillaries.

4.2.6 Ankon NaviCam®

Ankon Technologies Co., Ltd. (http://www.ankoninc.com.cn 2018) established in 2008 in China and it has developed Ankon NaviCam® capsule. Ankon NaviCam® is a magnetic guidance robot. It has field of view of 140° and battery life more than 8 hours. Ankon NaviCam® has a CMOS camera capturing 2 frames per second.

4.2.7 Capsubot

Shangxian Minimal Invasive Inc.(http://www.shangxianinc.com/en/ 2018) has developed Capsubot capsule endoscope. The Capsubot capsule is available for use in China.

4.2.8 Drawbacks of commercial capsules

Despite their simplicity in patient examination and high DY commercial capsules have several drawbacks. Traversing the GI tract in a passive manner is perhaps the most important of them; physicians are not able to interfere in the movement and/or the speed of the capsule as it moves in the lumen of the GI tract propelled by contractions. In other words, it is not possible to stop or navigate the capsule towards an area of interest, for a more thorough review. Other drawbacks are the limited battery life, which may result in incomplete SB examination, and their lack of capability for treatment, such as drug delivery, or biopsy of suspicious areas.

4.3 State-of-the-art research capsule endoscopes

New capsule modes have been designed by several research groups aiming to improve on existing capsule designs and add novel functions. Various solutions have been proposed to cope with the limitations of existing CE systems. The first category of developmental capsule endoscopes aims to provide capabilities for enhanced diagnosis, and the second group aims to provide capabilities that aid therapeutic interventions.

Also, the research for the enhancement of diagnostic ability of CE devices leads towards Therapeutic CE (TCE) devices. Research for TCE has been limited because of drawbacks, such as the inaccurate capsule localization that makes difficult the drug delivery in specific regions of GI tract. As the number of capsule endoscopy related trials currently exceeds 150 records in ClinicalTrials.gov registry, so the demand for new capsule based technologies and solutions is also growing(clinicaltrials.gov 2019)

Table 4.2 presents a summary of representative state-of-the-art research prototypes or concepts of capsule endoscopes.

Table 4.2 Research (P)rototypes or (C)oncept capsule endoscopes

Study (year)	**Project**	**Status**	**Multimodal imaging**	**Active actuation**	**Magnetic propulsion**	**Drug delivery**	**Specific treatment**	**Biopsy capabilities**
(Jang et al. 2018)	4-Camera High-Resolution and -Throughput Capsule Endoscope	P	yes	no	no	no	no	no
(Fontana et al. 2017)	Wireless Spherical Endoscopic Capsule	P	no	no	yes	no	no	no
(Son et al. 2017)	Magnetically Actuated Soft Capsule Endoscope for Fine-Needle Aspiration Biopsy	P	no	no	yes	no	no	yes
(Fu et al. 2017)	Magnetically Actuated Microrobotic Capsule with Hybrid Motion	C	no	yes	yes	no	no	no
(Winstone et al. 2017)	Bio-Inspired Tactile Sensing Capsule Endoscopy for Detection of Sub-mucosal Tumors	C	yes	no	no	no	no	no
(Leung et al. 2017)	A Capsule for haemostasis utilizing an inflated	P	no	no	no	no	yes	no

	balloon							
(Stewart et al. 2017)	SonoCAIT	C	no	no	no	yes	no	no
(Guo et al. 2017)	Development of a novel wireless spiral capsule robot with modular structure	P	no	no	yes	no	no	no
(Le et al. 2016)	A soft-magnet-based drug-delivery module for ALICE system	P	no	no	yes	yes	no	no
(Demosthenous et al. 2016)	Infrared Fluorescence-Based Cancer Screening Capsule for the Small Intestine	P	yes	no	no	no	no	no
(Z. Li et al. 2016)	Blue Light Therapy Capsule for Helicobacter pylori	C	no	no	no	no	yes	no
(Tortora et al. 2016)	A blue and red light Capsule for the Photodynamic Therapy of Helicobacter Pylori	P	no	no	no	no	yes	no
(Woods & Constandinou 2016)	A wireless capsule endoscope with holding mechanism for medication release	C	no	no	no	yes	no	no
(Gao et al. 2016)	Motor-based Capsule Robot Powered by Wireless Power Transmission	P	no	no	yes	no	no	no
(Lee et al. 2015)	Active Locomotive Intestinal Capsule Endoscope (ALICE) System	P	no	no	yes	no	no	no

(Gu et al. 2015)	6-Camera Endoscopic Capsule	P	yes	no	no	no	no	no
(Beccani et al. 2015)	A capsule with a coil actuation mechanism for drug release	P	no	no	no	yes	no	no
(Yu et al. 2015)	A drug delivery capsule in specified location by magnet	P	no	no	no	yes	no	no
(Zhong et al. 2015)	Tadpole Endoscopic Capsule	P	no	yes	no	no	no	no
(Shi et al. 2015)	Worm-Inspired Capsule Endoscope with Arc-Shaped Spiral Legs and Wireless Power Transmission	P	no	yes	no	no	no	no
(Yim et al. 2014)	Magnetically Actuated Soft Capsule Endoscope with Untethered Microgrippers for Biopsy	P	no	no	yes	no	no	yes
(Yim & Jeon 2014)	Ring-Shaped Magnetic Capsule Robot	P	no	no	yes	no	no	no
(Sun et al. 2014)	Legged Capsule Robot Actuated Wirelessly by Magnetic Torque	P	no	yes	yes	no	no	no
(Chen et al. 2014)	Wireless Autonomous Endoscope with Micro-Jaw Forceps for Biopsy	P	no	yes	no	no	no	yes

4.3.1 Capsule Locomotion and Navigation

A large proportion of work addresses capsule locomotion and navigation, as the ability to control capsule movement would improve mucosal visualization, as well as enabling biopsy. Two main approaches are considered: active actuation and external magnetic guidance. In the former approach the capsule is equipped with actuators, such as robotic legs, enabling autonomous motion of the capsule within the GI tract. In the latter approach, the motion of the capsule is based on externally applied magnetic fields.

Experimental or prototype capsules with active locomotion can be further grouped based on their method of propulsion. The first group uses magnetic forces. A preliminary model of a legged capsule robot with active locomotion has been presented by (Sun et al. 2014). The four legs move the capsule robot forward or backward. The activation of the legged capsule to move or to stop is achieved by magnetic torque without the dependence of a battery for power supply. The actuation of two external permanent magnets forces an internal permanent magnet to rotate. The internal permanent magnet is integrated in the capsule robot and rotates relative to the body of the robot. An internal mechanism transmits the magnetic torque of the rotating internal permanent magnet in order to activate or stop the legs. The group of (Yim & Jeon 2014) has presented a capsule robot able to move in a fluid-filled tube. The capsule is designed with three different parts. The first part is the rotating frontal part of the robot, which consists of a permanent magnet that has the shape of a ring. The second part is the linearly moving clamper that is also equipped with a permanent magnet, a clamper with ribs to support the directional movement of the capsule, and a slider. The third part is the capsule body, which consists of a stroke limiter and a small motor of a diameter with a size 6 mm. This mechanic arrangement of the parts provides the locomotion of the capsule. The motor rotates the frontal part. The magnetic force of the magnet of the frontal part pulls or pushes the magnet of the second part. Thus, the second part is passively moving linear and forces the movement of the capsule robot. This mechanism enables the capsule robot to achieve enough speed for the generation of the appropriate propulsion inside the liquid tube. Several mobility mechanisms for the capsule endoscope have been developed by researchers without achieving adequate degrees of freedom or sufficiently-diverse capsule motions. The group of (Lee et al. 2015)have developed the Active locomotion intestinal capsule endoscope (ALICE) system with diverse mobility consists of an Electro-Magnetic Actuation (EMA) system and capsule endoscope. The EMA system has three pairs of orthogonal uniform magnetic coils for 3-D alignment, and two pairs of gradient magnetic coils for propulsion and it achieves to realize complex motions of the capsule. The capsule has a tubular shape and moves through 5 degrees of freedom, enabling complex movements e.g., helical movements to closely scan the inner wall of the GI tract. The designed capsule endoscope has a diameter of 8 mm and length of 20 mm. The motion of ALICE system was evaluated in ex-vivo experiments, which verified the feasibility of the ALICE system for investigation of GI tract.

A second group of capsules draws on movements seen in the natural world. Inspired by the movements seen in natural world the group of (Zhong et al. 2015) have proposed the Tadpole Endoscope, which adopts a thunniform swimming technique of tadpoles and can propel itself through the GI tract. The Tadpole endoscope follows the traditional capsule endoscope design equipped with a soft tail and contains the driving unit, camera, control and an application specific integrated circuit transmitter, antenna and batteries. The driving unit drives the tail to flap and generates a propulsion force. The body of the capsule has a diameter of 13 mm and a length of 38 mm, where the 9 mm is the length of the tail. The total weight is only 6 g and one button battery of 1.5 V is used to drive the circuit board and the magnetic coil. Motion instructions are sent to the circuit board via infrared communication. The maximum swimming speed is 12.5 mm/s and the minimum turning radius is 25 mm. The tadpole endoscope’s motion was tested ex-vivo in a pig stomach, where it was showed that the movement of Tadpole endoscope and the change of direction when Tadpole endoscope reached the stomach wall. Relying on GI peristalsis, the WCE cannot actively move and treat effectively, necessitating further research into intestinal robotic devices. In order to support the locomotion in the intestinal tract capsule endoscopes are equipped with legs or paddles. However, the sharp ends of the legs have the potential to damage the intestine of the patient while navigating through it. The group of (Shi et al. 2015) proposed a capsule robot that mimics the movement of a worm equipped with elastic spiral legs. These arc-shaped spiral legs slide along the intestinal wall without hurting the intestine of the patient. In the body of the capsule a CMOS camera and a transmitting circuit are integrated for image acquisition. To avoid the limitation of insufficient battery duration a wireless power transmission system is used for power supply. The power supply system consists of receiving coils, a rectifier circuit and a voltage regulator circuit. The ring-like receiving coils are designed to surround the shell and to protect the inner space of the robot. The capsule has diameter of 16mm and length of 31mm. The locomotion and the wireless power transmission where tested in ex-vivo experiments. Sufficient power supply is important to support the functionalities of a capsule endoscope such as the video recording, video wireless transmission, biopsy devices etc. However, commercially-available button batteries are not sufficient to power these functionalities for long durations. The group of (Gao et al. 2016)has proposed a motor-based capsule robot utilizing an inchworm mechanism consisting of two expanding devices at both ends and a middle extensor for active locomotion. The capsule robot is powered by Wireless Power Transmission (WPT). WPT is based on near-field inductive coupling could supply several hundred milliwatts to the moving capsule robot and has been considered as a promising solution. The WPT system normally consists of a 1D (1-D) transmitting coil that excites alternating magnetic field and a three-dimensional (3-D) receiving coil with an inserted ferrite core that induces electromotive force (EMF). The motor-based capsule robot has a diameter of 13 mm and length of 42 mm for exploring the intestinal tract. The frame rate of the propose capsule is 30 frames per second with resolution of 320×240 dpi. Ex-vivo evaluation of proposed capsule robot was performed proving that the capsule is able to navigate in a collapsed porcine intestine. The group of (Fu et al. 2017) has proposed a conceptual, magnetically actuated, hybrid micro-robot with hybrid motion, that include screw jet, paddling and fin motion. It is driven by an electromagnetic actuation system, which generates a rotational magnetic field and alternate magnetic field. The experimental results from simulation indicated that the micro-robot realized the flexible motion in the pipe, by adjusting the changing magnetic frequency. The physician obtains the ability to move the microrobot in order to accomplish functions such as endoscopy examination or drug delivery, after taking into consideration information like the position and the posture of the microrobot. The proposed dimensions of the microrobot are a diameter of 13 mm and length of 30mm.

The application of external magnetic fields for the motion maneuvering of the capsule endoscopes has also been a matter of study. In the framework purposed by the group of (Fontana et al. 2017) a capsule with spherical shape was presented. The capsule has 26 mm diameter and weighs 12.70 g. The capsule transmits images of 320×320 resolution at 1.5 fps. The battery of the capsule can be recharged up to 6 times through a proposed recharging circuit. A power-on circuit and a localization module are embedded in the capsule. In order to reduce the friction during the examination of the colon that leads the camera in insufficient angles of view a structure combining an outer and an inner shell has been constructed. These shells have a 360° degree of orientation and the simultaneously movement maintaining the camera in the correct direction. The actuation of the capsule is based on the interaction of the integrated permanent magnet in the capsule and an external electromagnet. The capsule has been tested in an in-vitro colon simulation for image capturing. Detection and treatment of GI pathologies are two functionalities, which need different hardware integration, and the co-existence of both, in the same capsule is limited by its size. Thus, the group of (Guo et al. 2017) has proposed a wireless spiral capsule robot with modular structure driven by the external magnetic field. The capsule robot consists of two robots with modular structure with a mechanism to combine and separate them. It consists of a guide robot and an auxiliary robot with helical diversion grooves. The guide robot has a length of 39.4mm and a diameter of 12.5mm. The auxiliary robot has a length of 33.8mm and a diameter of 12.5mm. In the two robots, permanent magnets have been placed in the center and magnetized in the radial direction. Under the same external magnetic field, generated by the three-axis Helmholtz coils, the motion of each capsule is relative to the other. After the inspection of the guide module robot, the treatment module robot can be swallowed and then dock with the guide module robot and save the long time to reposition the target location.

4.3.2 Capsules with Biopsy Capabilities

A more limited amount of work has been done attempting to enhance capsule endoscopes with biopsy capabilities. Two types of micro-jaw mechanisms for biopsy, the slide-jaw and the sleeve-jaw, for a wirelessly powered capsule endoscope have been designed by (Chen et al. 2014). The capsule robots have the forceps in the front part of their body. The actuator of the capsule robot utilizes the lead-screw mechanism for the implementation of the stretching out and withdrawing motions of the micro-jaw. During the biopsy there are three movements that have to be executed. First one is the extension of the micro-jaws to the sampling area. Second the micro-jaws bite and cut off tissue. Third the micro-jaws withdraw back in the capsule. The collected samples are stored inside the capsule. The proposed two types of micro-jaw aim for supporting biopsy in surgical tasks. The slide-jaw extends along a guiding slot and pull back to cut off a tissue sample. The tissue sample is kept squeezed inside the two parts of jaw. The adjustment of the forward distance of jaw manages the sampling amount of tissue. The volume of the sampling tissues is around 1 mm^3 to 1.5 mm^3. The sleeve-jaw forceps are inspired by the traditional biopsy tools. The sleeve-jaw cuts off the tissue samples utilizing the pressure from the compressed spring. The sleeve-jaw is used for the retrieval and storage of smaller amount of tissue samples about 0.5 mm^3 to 1 mm^3 than slide-jaw. The cutting ability of the forceps was evaluated ex-vivo using small intestine of a pig. Another group (Yim et al. 2014) has developed a capsule endoscope system combining two devices for biopsy. Firstly, the system exploits the extra axial degree of freedom of the MASCE device for the functionalities of remote actuation, controlled navigation and drug release. Secondly, the system utilizes the microgrippers which can be actuated autonomously at the body temperature and self-fold collecting tissue samples in a highly parallel manner. The microgrippers have been optimized in order not to fold earlier than the pass of a 10 min period. This time threshold for microgrippers to close prevents them to fold before their release from the capsule. The constructed capsule has a diameter of 18mm and a length of 31.5mm. Ex-vivo experiments were conducted and showed that the proposed capsule is capable for the retrieval of multiple tissue samples. The extraction of samples from suspicious tissue lesions through a thin and hollow needle is a method called fine-needle aspiration biopsy. The group of (Son et al. 2017) has proposed a magnetically actuated soft capsule endoscope (B-MASCE) that utilizes the fine-needle aspiration biopsy functionality. The fine needle can penetrate deep inside the mass of a lesion, even in the case of sub-mucosal tumors and it can improve the diagnostic yield. The B-MASCE has been manufactured to enable the rolling locomotion on the surface of a stomach and the motion of axial jabbing of the needle. A magnetic field is used for the controlling and torqueing of the magnet inside the capsule endoscope. Four legs assist the guidance of the needle for the penetration in the area of the lesion. These four legs are made from soft material, thus the capsule endoscope is called “soft”. The capsule is designed for the examination of the upper GI tract and it has diameter of 12mm and length of 30mm. The fine needle has a length of 15 mm and penetration depth of 10 mm. In-vitro experiments for sampling were performed using pork fat as GI tumor. Samples were successfully captured from the proposed capsule with fine-needle aspiration biopsy functionality.

4.3.3 Multimodal Imaging

Various approaches have been proposed to enhance the multimodal imaging capabilities of current capsule endoscopes and add other novel lesion detection techniques apart from conventional white-light imaging.

The random movement of the capsule and the limited visual field in GI tract results in a high miss rate of significant findings. For this reason, the group of (Gu et al. 2015) has implemented a Multiple Cameras Endoscopic Capsule (MCEC) with smart control to reduce the miss rate. Multiple cameras are employed to create a larger visual field. Instead of wirelessly transmitting the image data, stores them in a flash memory thus enabling a high image acquisition rate. Moreover, a low-complexity image compression algorithm is proposed achieving a decrease in power consumption. A smart image capture control strategy based on motion information is used to control frame rate upon random movements within the digestive tract. This prototype endoscopic capsule has six cameras and captures frames with resolution of 480×480 pixels. Another multi-camera capsule, proposed by the group of (Jang et al. 2018) has developed a capsule consisting of 4 Video Graphics Array (VGA) cameras. The capsule has a diameter of 12mm and length of 32mm, while it weighs < 4g. The field of view for every single camera is 120° enabling the capsule for capturing 360° images of 640×480 pixels at a frame rate of 4 frames per second (fps). The capsule consists of a transmitter providing the image transmission up to 80Mb/s to an external receiver with power consumption at 0.8mW. The two batteries inside the capsule can last for >12 hours. Also, the capsule can provide images with information of the location inside the GI tract achieving sub-cm accuracy. The location of the capsule is specified measuring the signal power between the capsule and the external receiver that has 8-nodes and hub implanted on a tight vest. The location is determined to be near the node with the less channel attenuation and then specified using an adaptive selection of the nearest four nodes. In-vitro experimental results in pig intestine proved the sub-cm accuracy.

Moving away from white-light imaging, cancer detection in early-stage through infrared fluorescent-labelling is a well-known technique. The group of (Demosthenous et al. 2016) has developed a screening capsule prototype for the detection of fluorescence emitted by very low concentrations of Indocyanine Green (ICG) fluorophores. The capsule has a diameter of 13mm and length 25mm. It is able to detect and record fluorescence levels for about 9 hours via a variable sampling rate methodology that reduces the amount of redundant data collected. The fluorescence levels are stored in the internal memory of the capsule. Therefore, this capsule provides a viable general screening method for small-bowel cancer. Analysis of 8 or more hours of video for each patient is not required, since the physicians can examine whether the detected fluorescence levels exceeded a predefined threshold when plotted on a chart. The proposed near infrared-based fluorometric capsule is the first of its kind, in that without external body-worn hardware and labour intensive video analysis, early stage cancers can be detected cost-efficiently and reliably. Ex-vivo experiments of the proposed capsule were conducted using ICG-impregnated swine intestine showing the detection and screening ability of the system at different ICG concentrations.

Finally, a totally different approach to support diagnosis in CE has been proposed, based on tactile sensing technology. Inspired from human finger sensing anatomy (Winstone et al. 2017) have developed a biomimetic tactile fingertip sensor, named Tactip. Tactip uses remote palpation to stimulate a tactile sensing surface that deforms when is pressed against soft or hard lumps on the surface of the GI tract. This new diagnostic method enables the enhancement of visual investigation of lesions and it could provide further information about the structure of the lesions. A system able to classify abnormalities based on the shape, size and softness was utilized to test the sensor. Thus, information of the characteristics in different locations inside the surface of bowel without relying on the vision alone can be provided by the system. Tactip has embedded an artificial cast silicone skin, optically clear flesh like gel, camera with 720×1200 resolution and internal illumination using LEDs.

4.3.4 Therapeutic Capsule Endoscopes

Following the aforementioned categories capsule endoscopes, it is clear that the focus of capsule functionalities has to extend further than the diagnostic capabilities. In this subsection the potential capsule endoscopes capable for therapeutic functionalities are described. Therapeutic capsules can be further grouped based on their application. One group is the drug delivery capsules and the second group is the treatment capsules for specific pathologies as bleeding in the GI tract or the Helicobacter Pylori.

For the first group of therapeutic capsule, a smart capsule for location-specific drug release in the GI tract has been presented by the group of (Yu et al. 2015). The proposed capsule has length of 26mm and diameter of 9mm. The specific location for the releasing of the drug is determined by an implanted or externally worn permanent magnet. The capsule is activated when it is closed to the magnet. Then, the reed switch closes and the capacitor is discharged by the nichrome wire. The nylon fuse is melting while the cap is opening and the drug is released. The drug delivery procedure was tested in-vitro showing the locomotion of the capsule next to the permanent magnet. Another group of (Beccani et al. 2015) has designed a prototype Magnetic Drug Delivery Capsule (MDDC). A coil actuation mechanism is used by the capsule for the release of the drug. The capsule consists of a coil, a magnet and a drug chamber, in which matching magnets are placed. The drug chamber remains attached to the body of the capsule by the generated attraction between the magnets. The coil is able to produce the appropriate force for the magnets to be repulsed. Then, the chamber opens and the drug is released in the specific area of the GI tract. The diameter of the capsule is 13 mm and the length is 30 mm. The capsule has a weight of 12 g and the drug chamber can store to 2.4 ml of drug.

The group of (Woods & Constandinou 2016) has presented a concept of a microrobot with a medication release and a drug infusion mechanism for targeted drug delivery. The microrobot has been designed with resistance to natural peristalsis deploying a holding mechanism enabling the microrobot to localize a pathological area of interest in the GI tract. Then, a needle is positioned in this area and delivers a 1 ml dose of medication. The needle has the ability to be placed in a 360°scale, while simultaneously maintaining a diametrically opposite relationship with the holding mechanism. This feature guarantees the penetration of the GI tract wall by the needle. The holding mechanism utilizes a single micromotor to open and close two legs. The legs stretch and hold the microrobot in the region of interest inside the GI tract. A module for drug delivery that is combined with the ALICE (Lee et al. 2015)system has been presented by the group of Le et al (Le et al. 2016). The drug delivery module consists of two ring-type soft magnets and a simple plastic hinge. The ring-type magnets are axially magnetized attracting to each other keeping the drug enclosed inside the module. The ALICE system provides controlled navigation of the integrated ALICE with drug delivery module to investigate and accurately infuse the drug to the lesion area. The drug-delivery module is opened by the repulsive force between the two radially magnetized soft-magnetic rings, when the axial magnetization of the rings stops. The rings are demagnetized and a strong pulsating magnetic field in a radial direction is applied and the enclosed into the module drug is released. Then, the two rings are axially magnetized again, attracting to each other and thanks to the plastic hinge the drug-delivery module is returned to its initial shape. The integrated drug delivery module with ALICE has a diameter of 12 mm and a length of 33 mm. The proposed drug delivery module with ALICE was tested in-vitro.

For an ultrasound (US)-mediated targeted drug delivery (UmTDD) proof-of-concept capsule named SonoCAIT has been developed by the group of (Stewart et al. 2017). The prototype capsule is able for drug delivery in a specific location, utilizing US to release drugs and/or to increase drug uptake through sonoporation. SonoCAIT has a pill-like shape with the dimensions of 10mm in diameter and 30mm in length. An US transducer, a drug delivery channel, a vision module and the multi-channel external tether are integrated into the capsule. The vision module consists of a CMOS camera and circuit board with four white LEDs lights. The camera is cylindrical with resolution of 220×224 pixels. The aim of SonoCAIT is to deliver drugs to the wall of the GI tract. One example of a therapeutic preparation is drug-filled microbubbles (MBs). When these reach the target zone, they must be released in close proximity to the wall where the drugs can then be released by US. That means the US focus and MBs have to be directed towards the same target. In-vitro experiments showed that SonoCAIT achieved enhancement of drug uptake.

For the second group of capsule that are focus for specific treatment a blue light emission capsule for the therapy of Helicobacter pylori has been proposed by the group of (Z. Li et al. 2016). A module for pH sensing and measuring is used to differentiate locations and evaluate the digestive function by monitoring the pH values of the GI tract. The optical source of capsule consists of eight blue LEDs and emits blue light for treatment according to the preset range of pH values. Also, the capsule consists of a low power-consumption microcontroller for the processing of the pH signal and a wireless communication module for the transmission of the measured pH values to an external receiver. The proposed capsule has a diameter of 11.5mm and a length of 22mm. The group of (Tortora et al. 2016) has also investigated the efficiency of the LEDs in specific wavelengths needed for the therapy of the infection from Helicobacter Pylori. Based on their measurements, two easy-to-swallow capsule devices have been developed. The first is a capsule for research purposes that has a more performant battery that is not permitted in clinical examination. The second is the preindustrial capsule that integrates certified modules for clinical examination. The capsules integrate 8 LEDs placed in an electronic board along with a magnetic switch and a battery. There have been constructed two versions of capsule based on the emitting wavelength. The one emits red light only at 625nm and the other blue light at 405 nm. The capsules have a diameter of 14mm and length of 27 mm.

Finally, an inflatable prototype capsule for haemostasis in the GI tract based on balloon tamponade effect has been proposed by the group of (Leung et al. 2017). The capsule consists of three segments linked with flexible joints. These segments are the gas generation chamber with a length of 13mm, the acid injector with a length of 35mm and the circuit box with a length of 12mm. The capsule has a diameter of 14mm and it is enclosed into a silicone balloon. The balloon inflates at a bleeding lesion and achieves haemostasis by the tamponade effect. The inflation of the balloon is achieved by an acid injection into a gas generation chamber filled with base powder. The amount of infused acid controls the pressure and the volume of the silicone balloon in order to suite the variation of the diameter and texture of the intestine. The inflation of the balloon is capable to achieve the appropriate pressure to the bowel wall to anchor the capsule steadily in the position of the bleeding. Ex-vivo experiments for the evaluation of the appropriate pressure and in-vivo experiment of bleeding in the small intestine of a pig were conducted, showing that the proposed capsule is able to achieve haemorrhage control in the lower GI.

4.4 Capsule endoscope Localization

During the passive movement of capsule endoscopes through the GI tract, the accurate localization of the capsule is of great importance. Accurately identifying the location of the capsule can determine the exact position of possible abnormalities detected, and can therefore guide further management such as surgery or local drug delivery.

Several methods have been proposed for capsule localization (Iakovidis & Koulaouzidis 2015; Vasilakakis, Koulaouzidis, Yung, et al. 2019; Than et al. 2012). A summary of such studies performed during the last five years is provided in **Table 4.3**. The main approaches include electromagnetic or Radio-Frequency signal localization (RF) and magnetic localization (M); other techniques include Computed Tomography (CT) for patency capsule localization (Omori et al. 2015), and a recent approach based on Positron Emission Tomography (PET) (Than et al. 2017). However, the latter techniques involve radiation, which may have adverse health effects (Than et al. 2012).

Radiofrequency based localization techniques include Time-Of-Arrival (TOA), Time-Difference-Of-Arrival (TDOA), direction-of-arrival (DOA), and Received Signal Strength (RSS). The transmitting signal from the capsule endoscope is measured by the installed sensors around the patient’s abdomen and the sensors compute the received signal strength. However, RSS localization varies depending on the unique characteristics of each patient’s body and suffers from signal attenuation due to the complex non-homogeneous environment. Several studies experiment with different ways of computing this path loss signal propagation in order to compute the location of capsule(Hany &

Akter 2017a; Hany & Akter 2017b; Hany & Akter 2018a; Hany & Akter 2018b; Ye et al. 2014; Hany et al. 2017). In (Nafchi et al. 2014) the performance of TOA/DOA and TDOA/DOA measurements was investigated and compared for the capsule localization, while in (Ito et al. 2016) a hybrid TOA and RSS estimation method was proposed.

Current magnetic localization methods exploit a permanent magnet inside the capsule (He et al. 2015; Mahoney & Abbott 2016; Pham & Aziz 2014; Song et al. 2014) and they use of an external array of magnetic sensors to localize the capsule in the 3D abdominal space. In (Islam & Fleming 2014) a sensing coil was used inside the capsule and an alternating magnetic field is generated from outside the body. The voltages that are induced into the sensing coil could be used for the determination of capsule's position. In (Song et al. 2016) a multiple objects positioning and identification method has been constructed aiming to the localization of more than one different magnetic targets, such as capsules. In (Umay & Fidan 2016) a hybrid scheme of a permanent magnet and RF was used to achieve better localization result. In first place the main study of the proposed works is around the required number, arrangement, position, and array of the magnet and magnetic sensors.

Computer Vision (CV) algorithms can assist the localization process of the capsule endoscope by exploiting the visual content of the raw CE video frames (Dimas, Iakovidis, Ciuti, et al. 2017; Geng & Pahlavan 2016; Iakovidis et al. 2016; Mehmet Turan et al. 2018). Thus, the need for any other equipment, such as external sensors, can be bypassed. Recently published studies have demonstrated very promising results towards this direction. In (Dimas, Iakovidis, Karargyris, et al. 2017)an Artificial Neural Network (ANN) is used to automatically calculate the distance traveled by the endoscope based on visual cues. Its advantage over previous approaches (Spyrou et al. 2015; Spyrou & Iakovidis 2014) is that it does not require any prior knowledge about the geometric model of the capsule endoscope camera and its intrinsic parameters, such as its focal length. In (Dimas, Spyrou, et al. 2017) that work was extended with the use of color information to enhance the localization performance. More recently, a less parametric, so called "deep visual measurement" approach was proposed for visual localization of capsule endoscopes with even higher accuracy (Dimitris K Iakovidis et al. 2018). In addition to localization, this methodology provided a means for contactless size measurement of lesions. A hybrid localization was proposed in(Turan et al. 2017), combining both computer vision and magnetic localization methods. Another hybrid localization was proposed in (Bao et al. 2015), where computer vision and radio frequency localization methods were combined. A thorough, but more technical, review on capsule localization methods has been performed by Mateen et al. (Mateen et al. 2017).

Table 4.3 provides a summary of the state-of-the-art capsule localization methods and respective results. There is at present no common evaluation metric for assessing localization. The metrics reported in the reviewed studies mainly express the error in the estimation of the capsule's location. These include the Root Mean Square Error (RMSE), Localization Error (LE), Average Localization Error (ALE), Mean Absolute Error (MAE), Mean Error (ME), Translational Error (TE), Position Error (PE), Mean Distance Error (MDE), Average Position Error (APE). More details on how exactly these metrics are estimated can be found in the respective studies. Most of the results reported in the reviewed studies are not comparable to each other, not only due to the use of the different metrics, but also due to the different experimental setups used. Most studies on RF and magnetic capsule localization have reported errors of sub-centimeter ranges, whereas studies based on CV have reported errors of a few centimeters. However, it should be noted that the former address the localization of the capsule in the 3D abdominal space, whereas the latter ones address the localization of the capsule within the intestinal lumen, which can be directly exploited for therapeutic interventions, e.g., localized drug delivery. This is a major difference, which can explain the differences in the error levels. Even promising perspectives arise from the hybrid approaches combining CV and sensor based localization, as in (Geng & Pahlavan 2016), where the localization error reported was very low (47mm).

Although the developments are significant, the experimental assessment of the respective methods was mainly based on simulation/emulation models, fewer were based on ex vivo setups (Mehmet Turan et al. 2018), whereas only some of them were performed in vivo. Limitations for performing in vivo experiments are posed by the higher costs and the legal aspects to be treated.

Table 4.3 State-of-the-art methods for capsule localization

Study (Year)	Experiments	CV	RF	M	Results	
(Hany & Akter 2018b)	Simulation	N	Y	N	ALE, RMSE	18.66mm, 24.53mm
(Mehmet Turan et al. 2018)	Ex vivo (pig stomach), Simulation	Y	N	N	TE	6%
(Hany & Akter 2018a)	Simulation	N	Y	N	RMSE	6.2mm
(Than et al. 2017)	Simulation	N	N	N	PE	0.39mm ±0.22mm
(Dimas, Iakovidis, Ciuti, et al. 2017)	In vitro	Y	N	N	MAE	7.9±5.1mm
(Umay & Fidan 2016)	Simulation	N	Y	Y	ALE	0.25mm
(Hany et al. 2017)	Simulation	N	Y	N	ALE	3.8mm
(Turan et al. 2017)	Ex vivo (5 pig stomachs)	Y	N	Y	TE	>=2%
(Dimas, Iakovidis, Karargyris, et al. 2017)	In vitro	Y	N	N	MAE	1.14± 0.75cm
(Dimas, Spyrou, et al. 2017)	In vitro	Y	N	N	MAE	2.7±1.62cm
(Hany & Akter 2017a)	Simulation	N	Y	N	ALE, RMSE	7.28mm, 10.43mm
(Hany & Akter 2017b)	Simulation	N	Y	N	ALE, RMSE	4.53mm 5.14mm
(Mahoney & Abbott 2016)	Simulation	N	N	Y	ALE	2.1mm
(Iakovidis et al. 2016)	In vitro	Y	N	N	MAE	1.4 ± 0.8 cm
(Geng & Pahlavan 2016)	Simulation	Y	Y	N	RMSE	47mm

(Ito et al. 2016)	Simulation	N	Y	N	RMS	1.3mm
(Song et al. 2016)	Simulation	N	N	Y	MDE	3.5mm-4.0mm
(He et al. 2015)	Simulation	N	N	Y	ALE	0.76mm
(Bao et al. 2015)	Emulation	Y	Y	N	ALE	23mm
(Nafchi et al. 2014)	Simulation	N	Y	N	RMSE	≥ 1cm (per axis)
(Song et al. 2014)	Simulation	N	N	Y	APE	0.003 mm
(Pham & Aziz 2014)	In vivo	N	N	Y	LE	5mm
(Ye et al. 2014)	Simulation	N	Y	N	ALE	5cm (RSS) 1.5cm (TOA)
(Islam & Fleming 2014)	Simulation	N	N	Y	ME	6mm

4.5 Image Enhancement

Enhancing the visualization of CE video streams can affect diagnostic yield. Such enhancements mainly include approaches for faster reviewing of the CE video, as well as more complete and accurate display (Iakovidis & Koulaouzidis 2015; Vasilakakis, Koulaouzidis, Yung, et al. 2019) . As described in section 4.2, the commercially available capsule endoscopes provide software solutions for this purpose. Research towards enhanced visualization is active but still not sufficiently explored. In the following section, some of the latest and most representative works are presented.

A cyber physical system for simultaneous RF experimentation and 3D imaging inside the small intestine has been proposed in (Pahlavan et al. 2015). 3D reconstruction was based on a hybrid localization and mapping technique. This technique uses the RF signal received from body mounted sensors and similarities among consecutive images from the VCE to construct 3D model of the small intestine. The path reconstruction algorithm was validated with clinical experimentations using a 3D X-Ray procedure. Another 3D mapping approach has been proposed in the context of a system for navigation of either active locomotion or magnetically propelled capsules, by a haptic user interface (Mura et al. 2016). This was achieved by using a robotic manipulator coupled with a computer vision module able to infer the 3D structure of the environment on a frame-by-frame basis. Based on the user input and the estimated scene structure, the control system was gently able to generate forces guiding the user along the centerline of the GI tract. Another 3D reconstruction method for visualization is a has been proposed in (M. Turan et al. 2018). That method was based solely on color images.

Another approach aiming to time-efficient visualization of CE videos was based on the elimination of redundant video frames (Chen et al. 2015). The identification of such frames was based on their temporal correlation and their color and texture features. The selective elimination of the redundant frames, e.g., by keeping only representative frames from the CE video, results in a video summary that can be examined faster by the CE reviewers (Iakovidis et al. 2008). Such video summarization approaches have been investigated in the studies(Ben Ismail & Bchir 2016; Chen et al. 2017; Mehmood et al. 2014a; Mohammed et al. 2017). In particular in(Mehmood et al. 2014a), a Mobile-Cloud-Assisted Tele-Endoscopic System (MCATS) was proposed, aiming to provide ubiquitous access to CE videos through a flexible framework capable of adaptively performing video summarization. Through that framework CE videos can be visualized and shared through smartphones. Addressing the demands of growing CE data volume another video summarization framework was proposed for efficient management and analysis of CE data obtained from a tele-endoscopy system (Mehmood et al. 2014b). In that framework, a smartphone collects frame sequences and performs video summarization to generate keyframes. In parallel, the smartphone also transmits the generated keyframes to the corresponding medical specialists for analysis

4.6 State-of-the-art abnormality detection software

Several efforts have been made to develop computer-based medical systems capable of analyzing CE image sequences for the detection and recognition of abnormalities during the last five years. In such systems the images undergo transformations that enhance features significant for diagnosis, such as color, texture and shape. Based on these features, which are numerically represented, the systems are able to discriminate different kinds of tissues. The discrimination is performed by algorithms capable of classifying the tissue images, based on their features, into different categories, including normal, abnormal, or categories representing specific types of abnormalities, e.g., polyp, blood, angiectasia, ulcer, etc. The automatic detection of abnormalities can contribute in the reduction of the number of false negative diagnoses and, indirectly, it could contribute in the reduction of the time it takes to review WCE videos.

Usually, supervised classification algorithms, such as Artificial Neural Networks (ANNs), and Support Vector Machines (SVMs) (Sergios Theodoridis & Koutroumbas 2008) are employed for image classification and/or segmentation. The results are presented in terms of average accuracy (ACC), representing the number of correctly detected abnormal samples divided by the total number of samples, and/or the average sensitivity (SN), which represents the true positive detection rate, and the specificity (SP), which represents the true negative detection rate. In few studies precision is provided instead of SP, the former representing the proportion of true positives over all positives. Also, in some studies, the Receiver Operating Characteristic (ROC) and the Area Under the ROC (AUC) are used as a more reliable metric for abnormality detection. Interestingly, most of these studies have focused on the detection of one category of pathologies. Only a few of them have focused on the detection of suspicious CE video frames, regardless of the pathology.

The following paragraphs of this thesis present the state-of-the-art abnormality detection methods for computer-based medical software systems, per abnormality type.

4.6.1 Blood detection methods

Computer-based abnormality detection systems have been proved especially useful in detecting and locating the origin of obscure gastrointestinal bleeding (OGIB), which is defined as chronic bleeding from a source not found after traditional (wired) endoscopy. Most of the proposed systems concentrate on features associated with the color of the image, as blood has a distinct red hue. Color is the most important aspect that differentiates bleeding and non-bleeding region. However, in some cases edge pixels, such as those found in intestinal folds, and bleeding pixels share similar dark hues; which lead traditional algorithms to often mistake edge pixels for bleeding pixels. (Fu et al. 2014) and (Usman et al. 2016), used a technique to remove these edge pixels aiming to enhance bleeding detection. (Usman et al. 2016) used the transformation of an image to HSV (Hue Saturation Value) color space to extract image features, while in (Fu et al. 2014), the pixels are grouped adaptively into uniformly-colored segments based on color and location, with a procedure called superpixel segmentation. The feature is used from each superpixel for bleeding detection is the ratio of red to green, blue, or the sum of red, green and blue intensities in RGB (Red, Green, Blue) color space.

(Yuan et al. 2016a), considered quantized color histograms as features for blood detection, using the technique of Bag-of-Words (BoW). In second a stage, the bleeding regions are localized using a saliency map that indicates regions of importance within the image, estimated based on color information from various color spaces. A classifier fusion algorithm to detect the bleeding frames and localize the bleeding area was proposed by (Deeba et al. 2018). It combines the results of two classifiers trained using first-order statistical features extracted from RGB and HSV color spaces, that include mean, standard deviation, entropy, skew and energy. Furthermore, (Ghosh et al. 2018) noted that the blue (B) component of RGB does not carry any valuable information for the discrimination between bleeding and non-bleeding zones. Instead, a composite color component obtained by dividing the green (G) with the red (R) components, i.e., G/R was found to be more informative. First-order statistical features were extracted from G/R and were used as input to an SVM classifier for the detection of bleeding frames.

Aside color features, texture and shape features have been also considered for the discrimination of bleeding frames. In the study conducted by (Hu et al. 2016), a geometric image feature, called local-contrast- enhanced higher-order local auto-correlation (LCEHLAC), was utilized along with an image pre-processing method for a non-linear conversion model of the HSV color space. A methodology combining color histogram features and automatically extracted features, was proposed by (X. Jia & Meng 2017) . The former ones are extracted from CIE-Lab color space, and the later ones are extracted using a deep (multi-layer) neural network architecture, called Convolutional Neural Network (CNN).

A summary of the afore-mentioned blood detection methods along with their results in various datasets is provided in **Fig. 4.1**. It can be noticed that in all cases the performance metrics are high, exceeding 90%. The different methods cannot be directly compared to each other, since they have been tested on different datasets. It is notable that the highest overall performance was reported by (Hu et al. 2016), with an AUC of 99%, using one of the largest datasets, with 11,118 image frames, as compared to the other reviewed studies.

4.6.2 Polyps and tumor detection methods

Polyps are growing protrusions of mucosa inside the intestine due to excessive proliferation of tissue and inflammation or deep-seated malformations. Polyps are mainly discriminated by their shape and texture (Hu et al. 2016). (Mamonov et al. 2014) started by accepting polyps as protrusions that are mostly round in shape. Thus, best fit ball radius was used as a decision parameter of a classifier. (Yuan et al. 2016b) use the Scale-Invariant Feature Transform (SIFT) to detect salient points in images that may correspond to polyps. From the neighborhoods of these points, texture features, called Complete Local Binary Pattern (CLBP) are extracted. A more complex methodology was proposed by (Liu et al. 2016) for small bowel tumor detection. This methodology is based on both texture and color analysis. Multi-scale texture analysis is performed (i.e. the analysis is performed at different image resolutions), by means of the curvelet transformation and fractal encoding. Color information is captured by means of higher order moments between different color channels. The extracted features classified by an optimally selected SVM classifier. In another study, (Alizadeh et al. 2017), proposed an adaptive neuro-fuzzy inference system aiming to classify CE video frames containing polyps. The system extracts 32 features including four statistical measures (namely contrast, correlation, homogeneity and energy) calculated from co-occurrence matrices. Mutual information (a measure of the information shared between pairs of features) was used to select a subset of more informative features for the discrimination of polyps from normal tissues.

The results of the polyp/tumor detection methods reviewed in this subsection are presented in **Fig.4.1**. The highest results were reported by (Liu et al. 2016) that achieved accuracy 97.3%. However, the dataset used is smaller than the one used by (Hu et al. 2016) that achieved accuracy 92.4%. The lower accuracy of the latter can be mainly attributed to its lower specificity.

4.6.3 Ulcer detection methods

Ulcer is one of the most common pathologic outcomes of several diseases affecting the GI tract. Small-bowel ulcers, of variable severity and activity, are often challenging in terms of detection by traditional imaging techniques. Hence, CE is increasingly being used in ulcer diagnosis and management. Several solutions have been proposed for the detection and follow up of different forms of CD or NSAID-induced ulcerations. In this context, the majority of recent studies were based on combinations of texture and color features to discriminate ulcers from normal tissues. In the study of (Suman et al. 2017), the color components of 7 different color spaces were analyzed in order to find an optimal combination of their color components that better discriminates ulcers from normal tissues. The selected components were Cr (which represents the difference of red from a reference value) from YCbCr color space, the yellow (Y) component from CMYK color space, and the blue (B) component from RGB. A two-staged fully-automated computer-aided detection system is proposed by (Yuan et al. 2015) to detect ulcers in CE images. In the first stage, the image is segmented using superixels of different sizes. Color and texture features are extracted from each superpixel. Then, the extracted features are fused to form a saliency map per image. In the second stage the BoW technique uses the obtained saliency map to better characterize the images. A system capable of using weakly annotated images was proposed by (M. Vasilakakis et al. 2017). Instead of annotating the images in detail, i.e., pixel-by-pixel, the images can be annotated at image level. In this way, a binary semantic label is assigned per image to indicate whether its content is normal or abnormal, e.g., a keyword “abnormal” if the image contains an abnormality. This system offers the convenience to robustly detect which images contain possible ulcers. A weakly image annotation method based on a CNN architecture was presented by (Georgakopoulos et al. 2016). The CNN receives a single CE image as input to process and it makes the decision regarding the presence of an inflammatory lesion in the input image (i.e. two-neuron output layer).

Fig. 4.1 gathers the results from the previously presented studies, which focus on the ulcer detection. (Suman et al. 2017) reported the best results compared to the other methods achieving accuracy 97.9%. However, the results cannot be directly compared to the other methods, since different datasets were used. In the method of (Georgakopoulos et al. 2016) where a CNN architecture is utilized, the performance is noticeably high with an accuracy of 90.2%. Although, the dataset in the study of (Georgakopoulos et al. 2016)consists of weakly annotated images, automatically extracted features of the proposed method are able to characterize the images and provide a classification system independent from handcrafted features and pixel-level image annotation.

4.6.4 Hookworm detection methods

Hookworms are a leading cause of maternal and child morbidity, and their detection often is a challenging task. An automatic hookworm detection system for CE images was proposed by (Wu et al. 2016). The system is based on the basic characteristics of hookworms, which are their tubular body structure, the parallel edges in the shape of their body and their body color features. The first step is the processing of the image for the enhancement and the detection of locations with tubular structure. The second step is a parallel region detection method identifying the potential regions having hookworm bodies. The last step for the detection of hookworms is the features based on their intensity, which discriminate the hookworms from different components of gastrointestinal, i.e. bubbles. A histogram of average intensity is proposed to represent their properties. A deep hookworm detection system based on CNN was proposed by (He et al. 2018) for CE images. The system consists of two CNN networks, which enable the simultaneously model of visual appearances and tubular patterns of hookworms.

The results of the proposed methods are in the **Fig.4.1**. The experiments performed in the same dataset for all the aforementioned hookworm detection methods. It can been noticed that the method of (He et al. 2018) for the hookworm detection achieved higher classification performance than the method of (Wu et al. 2016). This can be attributed to the better enhancement of the visual pattern of the hookworms achieved by the CNN.

4.6.5 Multiple lesion detection methods

The detection of different lesions is one of the most challenging tasks in the CE reviewing process. To this end, the majority of research studies focus on the detection of one kind of abnormality, probably, because it is easier for an algorithm to distinguish one abnormality each time, i.e. bleeding detection is mainly based on the characteristic red hue of blood. (Nawarathna et al. 2014) proposed a texture analysis method was proposed. Different texture features, i.e. LBP, of an image extracted and the distribution of these various texture features is captured by a histogram to characterize the content the each image. An automated technique was proposed (Sekuboyina et al. 2017) for abnormality detection in CE images. Every image is divided into several blocks and from each block the color information is extracted using a CNN to overcome the drawbacks of handcrafted features. In (Y. Yuan et al. 2017), the SIFT features are extracted from images in HSV color space to obtain visual words that represent the three lesion categories (bleeding, polyps, ulcer) and the normal images. Then, these four types of visual words are combined together to composite the representative visual words for classifying the CE images. It is known that the semantics of the normal content include mucosal tissues, the hole of the lumen, bubbles, and debris. Thus, an investigation of such a semantic interpretation of the CE images may yield an improvement in lesion detection. In (M. D. Vasilakakis et al. 2017), a system for the semantic interpretation of the whole CE content was proposed, where each semantic content category, i.e. the hole of lumen, it represents a different label. The system consists of a salient point detection algorithm to detect points of interest in weakly annotated images and it extracts color features from the area around them. The representation of the CE images is based on the BoW image representation technique. Then, multi-label SVM classifiers are utilized to discriminate the labels that exist in a CE image. The study of (M. D. Vasilakakis et al. 2017)was further extended in (Vasilakakis, Diamantis, et al. 2018), where a convolutional neural network architecture enabling multi-scale feature extraction (MM-CNN) was proposed to detect the existing labels in the CE images. In (D. K. Iakovidis et al. 2018), a three phase methodology based on a CNN architecture for automatic detection and localization of lesions in CE images was presented. In the first phase, the proposed CNN architecture automatically extracts features from weakly annotated images. In the second phase, it suggests the possible locations of lesions in the detected CE images. In the third phase, a new algorithm is proposed to localize the lesions in the detected CE images. This algorithm uses the automatically extracted features from CNN. Then, the algorithm detects from the suggested locations from the second phase, which belong to lesions.

A summary of previously presented methods for the detection of different kind abnormalities is on **Fig.4.1**. It can be noticed that the method proposed by (Vasilakakis, Diamantis, et al. 2018) has considerably higher results with AUC 90.0%, which indicates that the semantic interpretation of the content of a CE images can provide valuable information to assist the detection of different lesions.

Table 4.4 State-of-the-art abnormality detection algorithms

Study & year	Detection	Dataset Images /Patients	Features	Best results (%)			
				ACCURACY	SENSITIVITY	SPECIFICITY	AUC
(Ghosh et al. 2018)	Bleeding	2350	C	94	94.78	93.58	
(Deeba et al. 2018)	Bleeding	Database1 (D1) 1224 Database2 (D2) 7648	C	D1: 97.22 D2: 94.5	D1: 98.13 D2: 92.32	D1: 96.52, D2: 95.07	
(He et al. 2018)	Hookworm	440K/11	S	88.5	84.6	88.6	89.5
(D. K. Iakovidis et al. 2018)	Bleeding, Polyps, Inflammatory Lesions	2352	A				81
(Vasilakakis, Diamantis, et al. 2018)	Bleeding Polyps Inflammations, Intestinal content	2352	A				90
(Suman et al. 2017)	Ulcers	48,000	C	97.89	96.22	95.09	
(Alizadeh et al. 2017)	Tumors	315	T	94	94.16	96.27	
Y. Yuan, Li, & Meng, 2017)	Bleeding, Polyps, Ulcers	1650	C,T	88.61			
(X. Jia & Meng 2017)	Bleeding	1500/80	A,C		91	94.79 (precision)	
(Sekuboyina et al. 2017)	Inflammatory Lesions, Vascular lesions, Lymphangiectasias , Polypoid Lesions	137	C,T				Aphthae: 78.81 Bleeding: 64.08 Chylous Cysts: 87.85

(M. D. Vasilakakis et al. 2017)	Bleeding, Polyps Inflammations, Intestinal Content	2352 images	C				0.88
(Wu et al. 2016)	Hookworm	440K/11	C	78.2	77.2	77.9	
(Hu et al. 2016)	Bleeding ,Tumors	Normal: 5642, Bleeding: 5476 Tumor: 1164 /28 patients	C,T	Bleeding: 98.54 Tumor: 92.44	Bleeding: 99.49 Tumor:96.74	Bleeding: 97.59 Tumor: 88.14	Bleeding: 99 Tumor: 97
(Yuan et al. 2016b)	Polyps	Normal: 2000normal/500 polyps	T	93.2	90.88	94.54	
(Usman et al. 2016)	Bleeding	3000bleeding/5500normal	C	92	94	91	
(Yuan et al. 2016a)	Bleeding	400bleeding/2000normal	C	95.75	92	96.5	97.71
(Liu et al. 2016)	Tumors	15patients/900tumor+900normal	T	97.3	97.8	96.7	
(Georgakopoulos et al. 2016)	Inflammatory Lesions, Ulcer	826	A	90.2	92.6	88.9	
(Vasilakakis et al. 2016)	Multiple Inflammations	1557	C				81
(Yuan et al. 2015)	Ulcers	20patients/170 ulcer and 170normal images	C,T	92.65± 1.2	94.12±2.4	91.18±0.9	
(Nawarathna et al. 2014)	Erythema, Bleeding, Polyps, Ulcers	500 images	C,S	91.8		91.88	
(Mamonov et al. 2014)	Polyps	18968/5	C,S		81	90	
(Fu et al. 2014)	Bleeding	5000	C	95	99	94	

C: Color features, S: Shape features, T: Texture features, A: Automatic features using Artificial Neural Networks

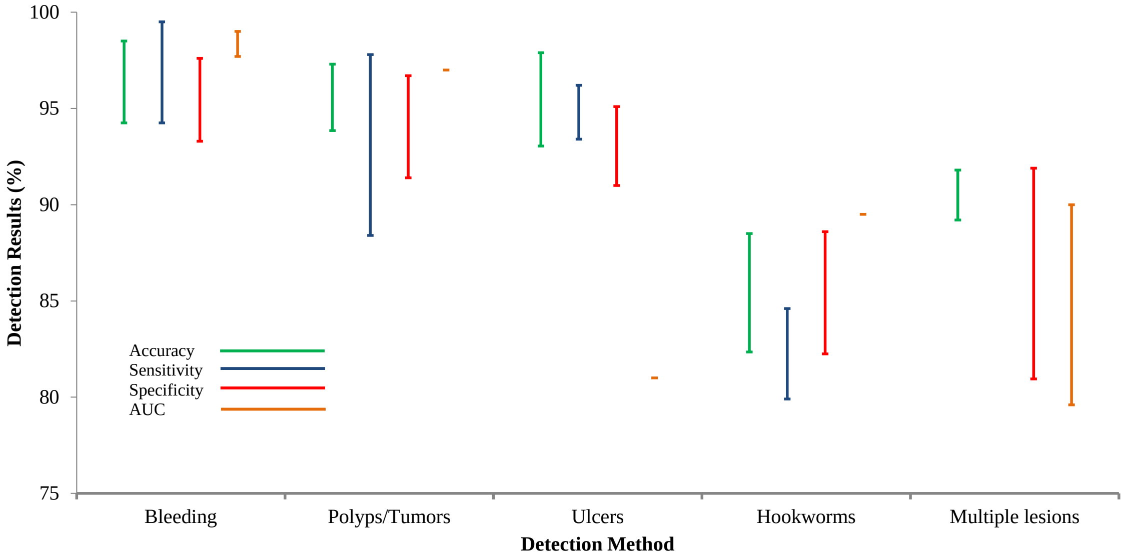


Figure 4.1.(Vasilakakis, Koulaouzidis, Marlicz, et al. 2019) Graphical representation and comparison of the aforementioned detection methods for the different kind of abnormalities during the last five years (2014-20181). The vertical bars indicate overall ranges of accuracy, sensitivity and specificity obtained from various abnormality detection methods. Single points indicate no variance. Missing bars indicate unavailable data due to inconsistent reporting of results across studies.

4.7 KID dataset

The aforementioned studies have reported rather high performances with respect to abnormality detection or discrimination of normal intestinal content. However, the use of different datasets and more importantly the unavailability of these datasets, prohibit direct comparisons of the results. Also, every study uses different evaluation method. Recently Koulaouzidis and Iakovidis proposed the KID ('Κάψουλα' – i.e., Greek for 'Capsule'-Interactive Database) (Koulaouzidis et al. 2017), which is a publicly available database of annotated WCE images and videos and can be used as a reference dataset. Another visible drawback is the small size of the datasets, so it is difficult to know the computational efficiency and performance accuracy in a real time examination of entire endoscopic videos consisted from thousand number of frames. Also, more attention should be given in the concept of automatic feature extraction techniques, especially regarding the relevance of the automatically extracted features with the problem under investigation. Future approaches could be enhanced if they incorporate medical knowledge. Also, the current "black box" classification approaches, could become more acceptable for use in clinical practice, if they would be capable of providing a clinically-relevant reasoning for diagnosis.

4.8 Conclusion

In this chapter a complete review of the background and the state-of-the-art methods in GI image analysis was presented. Based on this review chapter a key weakness in clinical management of capsule endoscopy is that is still demanding with respect to required review time and effort per examination. Beyond the clinical protocols, training, and guidelines that could contribute towards this direction, several computer-based methods have been proposed for faster, more effective visualization, and artificially intelligent approaches have devised for automated lesion detection and localization. However, although higher and higher results are being reported, only small steps have been done towards their adoption from commercially available CE platforms. Essential progress towards the translation of the research accomplishments to the market and finally to the clinical practice can be driven by increasing the availability of CE data to the wider research community. Such datasets can contribute in setting standard measurable objectives and triggering the competitive spirit in algorithm development. The legal framework for data sharing can be prohibiting, but this can be considered as another challenge to be taken. Recent computational methods, such as (D. K. Iakovidis et al. 2018), include algorithms that require minimum effort from the experts for the generation of large annotated capsule endoscopy datasets.

Ultimately, tomorrow’s CE platforms should be able to support clinicians by providing both diagnostic and therapeutic features. However, capabilities such as locomotion, biopsy excision and drug delivery, requires actuation mechanisms, such as mechanisms to manipulate forceps and needles, which are energy-demanding. Research towards the development of energy-efficient CE platforms, e.g., by exploiting energy harvesting techniques, could pave the way for navigable robotic capsule endoscopes with sufficient autonomy to perform diagnostic and therapeutic interventions. It is envisaged that wireless power transfer will become a viable alternative for future capsules with enhanced diagnostic and therapeutic capabilities. As this technology has already demonstrated its feasibility, it is a prospective direction towards research outcomes of higher technology readiness levels (TRL), i.e., closer to the market. A significant amount of work has been done over the last couple of decades, albeit mostly in modular fashion, but the main structure and function of the available capsules remain unaltered. In order to remove barriers in the development and clinical adoption of new capsules, several workshops have been developed to allow consolidation of intricate and multidisciplinary networks between clinicians, engineers and other stakeholders in the field.

The **key points** of this review chapter can be summarized as follows:

- Different kinds of capsule endoscopes have been proposed as solutions for image quality enhancement, navigation, biopsy, wireless power supply
- Therapeutic capsule endoscopes targeting to treat specific lesions have been investigated
- Clinical need of therapeutic capsule endoscopes is the therapy of pathologies, such as gastrointestinal bleeding, Crohn’s disease and small intestinal tumors
- Most capsule endoscopes intend to improve the examination and therapy of the whole gastrointestinal tract
- Different drug release and delivery mechanisms for capsule endoscopes have been presented in this review
- Following the advances of capsule endoscopes in control and navigation inside the GI tract contributes to the progress of capsule endoscopes targeting the treatment
- Most of the capsule endoscopes are concepts or prototypes without approval for in vivo clinical experiments
- Some capsule endoscopes have diameter and/or length that make swallowing difficult
- At this time there is no commercial or research capsule endoscope enable to put through all the aforementioned functionalities
- The development of high resolution, high definition image should be seen as the cornerstone for further active promotion and wider clinical adoption of capsules
- Automatic lesion detection and reporting and development of an accurate lesion localization system remain priority software challenge of our days
- The establishment of standard, sufficiently large datasets for CE for lesion detection experiments is necessary
- The lesion detection software is the lack of public, large and diverse datasets
- The lesion detection methods that were presented have to comparatively evaluate using a standard data base of CE images/videos
- The lack of regional controllability over the drug delivery in the target areas, in order to prevent the spreading of medication over other areas of GI tract
- Treatment capsule endoscopes must been able to navigate either internally or externally and anchor at the target for therapy area inside the GI tract
- Insufficient power backup
- The size of the aforementioned experimental capsules has to adjust, in order to ease the swallowing for the examination and/or treatment of the patient
- Miniaturization of the drug chamber, in order to facilitate the ingestible of therapeutic capsules due to theirs limited size

CHAPTER 5

GASTROINTESTINAL ABNORMALITY DETECTION

This chapter investigates the approaches for the detection of gastrointestinal anomalies. These approaches can be categorized in two types. The first type includes the approaches that examine and classify video frames as normal or as abnormal. The second type includes approaches for localization of lesions inside the endoscopic video frames.

In Chapter 5 the approached for the detection of suspicious for lesion in video frames are focused on weakly-supervised learning. The purpose for this type of machine learning is to cope with the resource-demanding issue of detailed image (video frame) annotation. Thus, in this chapter as a first step is the investigation of machine learning approaches that are train to classify frame content based on a simple semantic label or more than one labels per video frame.

After the examination of the weakly-supervised learning methodologies, as a solution to overcome the difficulty of detailed annotation of images, this chapter also examines the approaches for the localization of possible lesions. More specifically this chapter examines a weakly supervised methodology for lesion localization, as well as, an unsupervised lesion detection algorithm, which is result of a chromatic image analysis.

5.1 Introduction to Medical Decision Support Systems

Gastrointestinal Endoscopy (GIE) is a fundamental modality for the investigation of the gastrointestinal (GI) tract and the detection of luminal pathology. The most common GIE procedures are gastroscopy and colonoscopy. Another GIE procedure, which has become the prime choice for the examination of the small bowel, is wireless capsule endoscopy (WCE). WCE is performed with a swallowable, untethered capsule equipped with a camera that captures color images during its journey along the GI tract as it was described in the previous chapter. The amount of color images that are captured during any GIE procedure is significantly large and with a diverse content, making the detection of GI anomalies a challenging task for image-based Medical Decision Support Systems (MDSS).

The first MDSS for automated detection of GI anomalies in GIE video sequences appeared in the early 2000’s (Karkanis et al. 2003). Since then, a variety of such systems has been proposed, aiming to reduce the number of the lesions missed during GIE (Liedlgruber & Uhl 2011). These mainly include supervised approaches based on CNNs addressing the detection of only a single or a few kinds of GI anomalies, including polyps (Zhang et al. 2016; Ribeiro et al. 2016; Tajbakhsh et al. 2015; Bernal et al. 2017), both ulcers and polyps (Karargyris & Bourbakis 2011), celiac disease(Wimmer et al. 2016), inflammatory lesions(Georgakopoulos et al. 2016), and bleeding (Yuan et al. 2016a; Jia & Meng 2016; Xiao Jia & Meng 2017). Some recent approaches are more general in the sense that they address the detection of various kinds of anomalies(D. K. Iakovidis et al. 2018).

MDSS for GIE appeared primarily to cover clinical needs related to the detection and localization of lesions suspicious for malignancy or of bleeding sources, and to provide a second opinion on the assessment of lesions that require a more thorough examination(Iakovidis & Koulaouzidis 2015; Vasilakakis, Koulaouzidis, Yung, et al. 2019; Karkanis et al. 2003). This way, the application of such systems can contribute in speeding up the flexible endoscopy procedures, which are uncomfortable for a lot of patients, and can also enable less experienced personnel to perform it, including physicians' extenders or specialty nurses. Therefore, a consequent increase in clinical productivity, and an overall cost reduction for healthcare systems is possible (Koulaouzidis et al. 2015). The immense clinical need for such systems is more apparent in WCE.

5.1.1 MDSS for WCE abnormality detection

A major issue that is still unresolved in WCE is that it requires a lot of human effort for manually reviewing of the produced videos. During a WCE video review, WCE readers usually reach their human limits. Reviewing a WCE examination video is a very time-consuming task, which can often be a burden to the everyday clinical practice, especially so in units with high turnaround of WCE procedures. The reviewer's concentration should remain undistracted for a careful inspection of the output video in order to examine approximately 50,000-120,000 images within an average of 45-90min (Vasilakakis, Koulaouzidis, Yung, et al. 2019). Such a tiring procedure is prone to human errors; a fact with serious consequences in the diagnostic yield, which is alarmingly low the diagnostic accuracy of WCE(Zheng et al. 2012).

In order to cope with this problem, automated lesion detection methods based on computer vision algorithms have been proposed(Iakovidis & Koulaouzidis 2015; Vasilakakis, Koulaouzidis, Marlicz, et al. 2019). Most of these methods exploit supervised machine learning methodologies, capable of learning what is defined as normal and what is defined as an abnormal finding within the WCE video.

The majority of current MDSS for GIE are based on supervised machine learning algorithms aiming to detect/diagnose possibly abnormal conditions in the medical images. These algorithms are called supervised because they are trainable with annotated images. The annotated images include information about the presence, the location and the pathology of their contents, as assessed, usually, by a group of experts. Typically, the training images are annotated by experts at pixel-level, i.e., the experts indicate which pixels correspond to anomalies.

The generation of datasets for training the learning machines requires that experts indicate which pixels correspond to normal or abnormal tissues within the WCE images. Considering that the videos produced by a WCE examination are composed of thousands of frames, such a pixel-wise annotation task can prove very time-consuming and discouraging for annotation of large datasets by the experts. In this doctoral research thesis novel automated lesion detection methods based on computer vision algorithms have been proposed in order to assist MDSS systems.

Firstly, a supervised learning of weakly labeled images for automated video analysis in GIE is investigated. This type of machine learning was presented in Chapter 2 and was investigated as an alternative to cope with the resource-demanding issue of detailed image annotation(Hoai et al. 2014; Blaschko et al. 2010; Manivannan & Trucco 2015). It involves annotation of the training images only at image-level, using a semantic tag indicating whether the image contains anomalies or not; thus, omitting the details that can be specified by pixel-level annotation. This way, a binary semantic label is assigned per video frame indicating whether its content is normal or abnormal.

Acknowledging the significance of incorporating an image-level instead of pixel-level annotation process in the development of training datasets for lesion detection systems in WCE, in this thesis BoW-based supervised learning of weakly labeled images approach (Vasilakakis et al. 2016; Georgakopoulos et al. 2016; D. K. Iakovidis et al. 2018) is utilized using the state-of-the-art features that have been proposed in (Iakovidis & Koulaouzidis 2014a). These features represent colour information both at pixel and region level in CIE-Lab colour space, and despite their simplicity they have been proved very effective in the detection of a diverse set of abnormalities (Iakovidis & Koulaouzidis 2014b).The system can robustly detect which frames contain possible lesions, which is significantly important for video reviewers, since the localization of the lesion becoming easier. However, the drawback of the lesion localization still remains a challenge.

It is true that the normal content of images is diverse, while it includes various kinds of tissues and artifacts. Multiple semantic categories co-exist inside the GI tract. Considering that the image features of these contents are usually different (e.g., bubbles include white reflections, debris has green/yellow hues) approach to identify them as members of separate classes aiming to simplify lesion detection has been examined.

Thus, for each video frame a more complete description is provided using multiple semantic identifiers (labels). Semantic interpretation of endoscopy video using multi-label classification techniques has not been previously proposed. A system (M. D. Vasilakakis et al. 2017) for the semantic interpretation of the whole CE content was examined, where each semantic content category, i.e. the hole of lumen, it represents a different label. The system consists of a salient point detection algorithm to detect points of interest in weakly annotated images and it extracts color features from the area around them. The representation of the CE images is based on the BoW image representation technique. Then, multi-label SVM classifiers are utilized to discriminate the labels that exist in a CE image. This study (M. D. Vasilakakis et al. 2017) was further extended in (Vasilakakis, Diamantis, et al. 2018), where a convolutional neural network architecture enabling multi-scale feature extraction (MM-CNN) was proposed to detect the existing labels in the CE images. Thus, for each video frame a more complete description is provided using multiple labels corresponding to the different categories of its content.

5.1.2 WCE abnormality localization

After the detection of GI images/frames with a possible lesion, the next step for a complete MDSS is the information about the possible lesion. For this reason, Iterative Cluster Unification (ICU) (D. K. Iakovidis et al. 2018) was proposed as a novel localization algorithm for the abnormal areas. The novelty of ICU algorithm is the ability to localize the anomalies within the video frames, based on Pointwise Cross-Feature-Map (PCFM) Weakly Supervised CNN (WCNN) features from weakly labeled images. These features are automatically extracted from the salient points detected by a Deep Saliency Detection (DSD) algorithm, enabling the detection of salient points relevant to GI anomalies in endoscopic video frames. The proposed ICU has been applied in the GIE domain, using publicly available datasets that include a diverse set of anomalies and normal video frames from various parts of the GI tract.

A second attempt to localize abnormalities based on a unsupervised methodology, after an analysis on chromatic components of the GI images/frames. Distances on selective aggregation of chromatic image components (DINOSARC)(Vasilakakis, Iakovidis, et al. 2018) has been proposed as a methodology for the detection of salient points in GI images. This methodology includes an unsupervised salient point and region detection algorithm, and the estimation of local and global image descriptors enabling the characterization of various abnormalities both at a regional and at an image level. It consists of several novel components, including a color-based salient point detector, a salient region detector defining salient superpixels, and a method to derive a vectorial representation of the color of the salient superpixels by taking into account both point and region-level information. This enables more accurate the localization and characterization of even very small abnormalities. Besides the local image descriptors derived, global image descriptors are derived for supervised abnormality detection based on a BoW model.

The remainder of this chapter is structured as follows. Section 5.2 introduces the supervised model using of weakly labeled images for the lesion detection in endoscopic video frames. Section 5.3 extends the classification of two classes to multiple classes for the semantic interpretation of the whole video frame. Section 5.4 discusses the point localization in video frames aiming for the localization of lesions. Section 5.5 presents the DINOSARC algorithm for detection of salient point in endoscopic frames without previous knowledge. Finally, section 5.6 provides the experimental results of the proposed methodologies.

5.2 Supervised Lesion Detection in Video Capsule Endoscopy based on a Bag-of-Colour Features Model using Weakly Labeled Images

Contrary to conventional supervised learning, weakly-supervised learning (Hoai et al. 2014) does not require explicit and detailed annotation. Instead, only video frame-level annotation of the semantics of the video frame is required. Thus, a given video frame may be annotated, e.g., with the semantic concept “abnormal”, if it contains an abnormality, or with the semantic concept “normal”, if it does not contain an abnormality. An approach towards supervised learning of weakly labeled images is the previously described BoW model. BoW has been shown to be an effective strategy to cope with the demand for annotated GIE video frames (Vasilakakis et al. 2016). The vocabulary used for image/video frame representation is typically based on hand-crafted features.

The model of BoW was presented in Chapter 2 in the subsection 2.3.4 as a method of image representation. From the description of BoW model can be considered as a model for supervised learning of weakly labeled images, as it is able to describe the image content relying on salient points that detected by detection algorithms such as SURF. SURF, which was also described in Chapter 2, is a powerful and fast descriptor scheme and has been successfully applied to a plethora of computer vision problems. It has been shown to achieve comparable repeatability and performance to other, more sophisticated schemes, at a lower computational cost. This way the specific human annotation is not needed in the BoW model.

In the domain of WCE for the detection of polypoids, Hwang (2011) applied BoW using SURF in order to detect interest points and extract descriptions from patches around them. In (Yu et al. 2012) the performance of BoW was investigated using SIFT and LBP for ulcer detection. A more complex feature extraction scheme for the construction required in BoW was proposed in(Yuan et al. 2016b). This scheme was applied for polyp detection and includes extraction of SIFT, LBP, CLBP and histogram of oriented gradients (HoG) features from neighbourhoods of salient points detected using the SIFT key-point detector. In the context of bleeding detection, colour histograms extracted from various colour spaces were considered(Yuan et al. 2016a). Colour along with textural information has also been exploited in (S. Wang et al. 2016) for detection of gastric and oesophageal cancer, gastritis, and oesophagitis. In (S. Wang et al. 2016) superpixel segmentation was exploited for estimation of image descriptors from homogeneous regions. In (S. Wang et al. 2016) the descriptors considered include colour histograms from various colour spaces as well as LBP-based textural signatures. Most of the aforementioned approaches are based on SVM classifiers for the classification of abnormal images based on the BoW image description.

This section investigates two different sampling methods of image regions. The first method utilizes the SURF detector and the second utilizes the dense sampling. Also, two different features are investigated, the SURF features and the color features of (Iakovidis & Koulaouzidis 2014b; Iakovidis & Koulaouzidis 2014a). The extracted features from the image regions are incorporated in the framework of BoW for image description.

More specifically, **Figure5.1** shows an initial WCE image of lymphangiectasia.

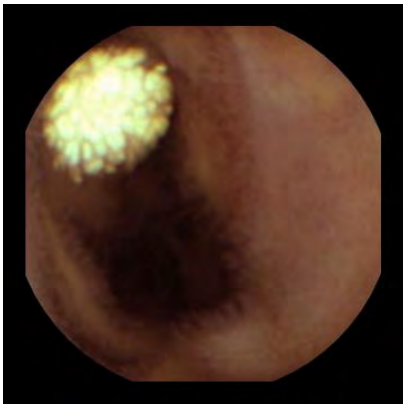


Figure 5.1: A raw WCE image depicting lymphangiectasia

Figuge. 5.2 illustrates the set of the SURF interest points extracted from the **Fig.5.1**, combined with SURF regions.

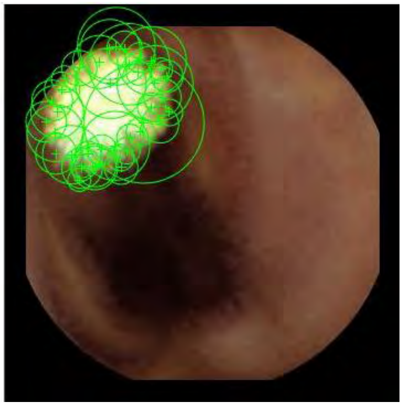


Figure 5.2: SURF regions

However, except of the use of SURF for the detection of interest points and the sampling of image, another “naïve” approach for interest point selection that is known as “dense sampling”, is used. Following this approach, all pixels are sampled using a regular grid (i.e., one with equal horizontal and vertical inter-pixel distances), which are then used as interest points. Although these points cannot be matched accurately, when compared to the SURF interest points, they carry valuable information regarding image content interpretation (Tuytelaars 2010). **Figure 5.3** illustrates the set of the dense interest points, also combined with SURF regions.

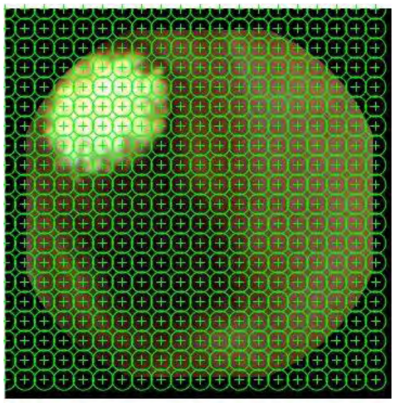


Figure 5.3: dense SURF regions

Except of the use of SURF visual descriptor, in (Vasilakakis et al. 2016) a visual descriptor proposed by (Iakovidis & Koulaouzidis 2014b; Iakovidis & Koulaouzidis 2014a) was also utilized. The colour-based features of (Iakovidis & Koulaouzidis 2014b; Iakovidis & Koulaouzidis 2014a) are extracted from patched around the interest points that detected from SURF detector or from the whole image based on the dense sampling. During the procedure of the visual descriptor the images are first transformed to the CIE-*Lab* colour space. Following the procedure colour information is extracted from a square region centered at each point. The colour information uses the *Lab* values of each interest point, as well as, the minimal and maximal values of each chromatic component. This results to a visual description vector consisting of 9 values. **Figure 5.4** illustrates the set of interest points detected with SURF detector and the square region, where the colour information is extracted.

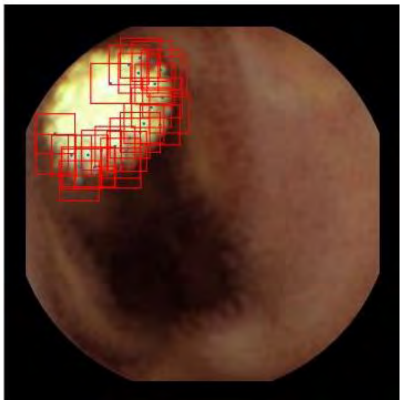


Figure 5.4 Lab regions

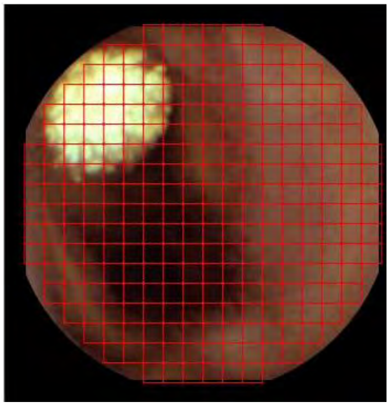


Figure 5.5 illustrates dense interest points.

One may easily observe that SURF points do not cover the visual properties of the whole image. Yet, the latter is achieved by the dense features.

5.3 Semantic Interpretation of Gastrointestinal Image/Frame

5.3.1 Multi-label classification in WCE video frames

In subsection 2.4.4 the multi-label classification was presented as a special case of data classification, where multiple labels may be assigned to a given data. In this section the application of multi-label classification is applied on the domain of gastrointestinal (GI) video endoscopy and the focus of investigation is on Wireless Capsule Endoscopy (WCE) as a more challenging application domain (Iakovidis & Koulaouzidis 2015). The video frames obtained from such an endoscopic examination are generally characterized as normal or abnormal depending on whether they contain abnormalities, such as lesions or blood. However, normal frames may include content belonging to a variety of semantic categories such as normal mucosa, bubbles, and debris. Also, the content of the abnormal video frames may include one or more kinds of abnormalities, as well as normal content.

Following the BoW feature extraction process, the content of the WCE video frames needs to be classified into semantic categories. Usually the classification of the endoscopic video frame content is performed into two categories, corresponding to normal and abnormal tissues, using binary classifiers, e.g., SVMs (S Theodoridis & Koutroumbas 2008) . However such approaches only provide an abstract categorization of the video frame content. This happens due to the initial assumption that every video frame belongs in exactly one of the aforementioned categories. This assumption does not consider the multiple semantics that may co-occur within a given video frame. For example, the semantics of a normal video frame besides mucosal tissues may include normal intestinal content such as bubbles and debris, and the lumen hole. More specifically, based on the theoretic background of the Chapter 2, here a total of 5 labels are considered, indicating the presence of normal (l_1), abnormal (l_2), debris (l_3), bubbles (l_4), and lumen hole (l_5).

This section applies the multi-label classification in the context of endoscopic video frame analysis the problem transformation methods that were previously described in the subsection 2.4.4 (Tsoumakas & Katakis 2007). Here each of these multi-label classification methods are customized to fit in the context of the endoscopic video frames.

In the binary relevance method (Tsoumakas & Katakis 2007) binary classifiers are used to determine the existence of each of the five categories of content considered, e.g., the existence of abnormalities or not, the existence of debris or not, etc. However, this methodology does not consider possible dependencies between labels.

In the label combination method (Read et al. 2008) apart from the five classes that were referred earlier, there are also “classes” that derive from their combinations, e.g., a “new” class may be considered to be the set of video frames are labeled both as normal and debris. This artificial class is then denoted as the normal-debris class.

The pairwise classification (Fürnkranz et al. 2008; Mencιa & Furnkranz 2008) instead of five binary problems, ten binary problems are formed, because there exist ten different pairs of labels. Typically, each pairwise problem is constructed from examples with which either labels (but not both) are associated, thus forming a decision boundary for these two labels.

As it was mentioned in subsection 2.4.4, MLP architecture was proposed (Zhang & Zhou 2006) for multi-label classification. Recently the usage of CNNs has been extended into multi-label classification problems. A common method to extend a CNN to multi-label classification is to transform it into multiple single-label classification problems by using one output neuron per label, as e.g., in the work of Gong et al. (Gong et al. 2013), who explored various multi-label loss functions on a network similar to the one proposed by Krizhevsky et al. (2012). While in a typical multi-class classification problem a common practice is to use softmax activations on the output neurons, in multi-label classification problems this does not apply since the softmax function forces the output neurons to express the selected class as a probability, which depends on the rest of the classes. In multi-label classification the usage of sigmoid neurons is typically employed with a cross-entropy loss (Guillaumin et al. 2009), which expresses the probability of a given class as a Bernoulli distribution. Other approaches have also been proposed, such as the one of (J. Wang et al. 2016) which utilized a combination of a Recurrent Neural Network (RNN) and a CNN to leverage the label dependencies that exist in natural images.

In (Vasilakakis, Diamantis, et al. 2018) a Multi-scale and Multi-label CNN (MM-CNN) was proposed. The overall MM-CNN architecture consists of five layers of multiscale convolutional blocks. In order to perform the task of multi-label classification, the network has 4 output sigmoid neurons; i.e., one for each label.

5.3.2 Weakly Multi-labeled Classification

At next level, further semantic concepts may be also added to the annotation process. In the context of lesion detection in WCE, different “normal” concepts can be associated with different normal intestinal content, e.g., “debris”, “bubbles”, whereas different “abnormal” concepts may include GI lesions, e.g., “inflammatory lesions”, “vascular lesions”, etc.

In the context of reviewing of large WCE videos, such an approach could significantly reduce the amount of effort required by the video reviewer, since it detects frames that possibly contain lesions. Since such frames are usually a rather small subset of the entire WCE video, the reviewers’ task may be limited to localization of abnormalities within this subset, which is a less tiring task.

5.3.3 BoW Weakly Multi-labeled classification based on Difference of Maxima Salient point detection

One of the challenges within the problem of semantic description of WCE videos is the lack of standardized interpretation methods. Therefore, many research works begin by constructing a saliency map. Given such a map, they are then able to select a subset of a given image/video frame, i.e., regions that would be examined for potential existence of abnormalities. For example, (Yuan et al. 2016a), proposed a saliency map extraction method for the detection of bleeding frames in WCE videos, by creating two saliency maps and by fusing color information of the a and the M channel of the CIELab and the CMYK color spaces, respectively, as well as heuristic properties of the "reddish" colors. Superpixel-based segmentation has been investigated by several research efforts among others, also for the detection of bleeding regions. (Yixuan Yuan et al. 2017)fused contrast-based and object center-based saliency maps and used strong classifiers. Also, (Bernal et al. 2015) proposed the use of energy maps that indicated the likelihood of polyp presence. The problem of detection of multiple abnormalities within the same image/video frame has recently gained the attention of the research community. (Y. Yuan et al. 2017) aimed to detect bleeding, polyp, ulcer and normal video frames. To this goal they calculated color SIFT (Lowe 2004) features from each semantic category, separately extracted visual words from each and combined all words to obtain a visual dictionary. Video frames were also encoded by a novel adaptive saliency algorithm.

The BoW approach can be based on a set of extracted patches surrounding dense points that result from a sampling process using a regular grid (i.e., one with equal horizontal and vertical inter-pixel distances) as it was described previously. In the context of endoscopic video frame analysis the application of SURF on channel *a* of the CIE-*Lab* color space (*a*-SURF) resulted in salient points on all the abnormalities included in that study. Also, the results of a preliminary study (Vasilakakis et al. 2016) showed that dense sampling may be more time-consuming, but it can result in higher abnormality detection rates.

The dense sampling process using regular grid, extracts a large number of feature vectors. These feature vectors are not easily separable by a clustering algorithm. For this reason, there is a need to select some video frame points to extract fewer feature vectors without significant loss of information. A way to reduce these points in a video frame represented in CIE-*Lab* color space is to get points only from the video frame regions where a significant color change is observed(Vasilakakis, Diamantis, et al. 2018). The purpose behind this idea is to discriminate and sample video frame regions, where a discontinuity in their color description appears. The discontinuity in color of channel *a*, indicates the region as a region of interest. In that sense these regions can be considered as "salient" points. In order to detect such salient points, the difference between two maximum values in *a*-channel around the densely sampled points are considered.

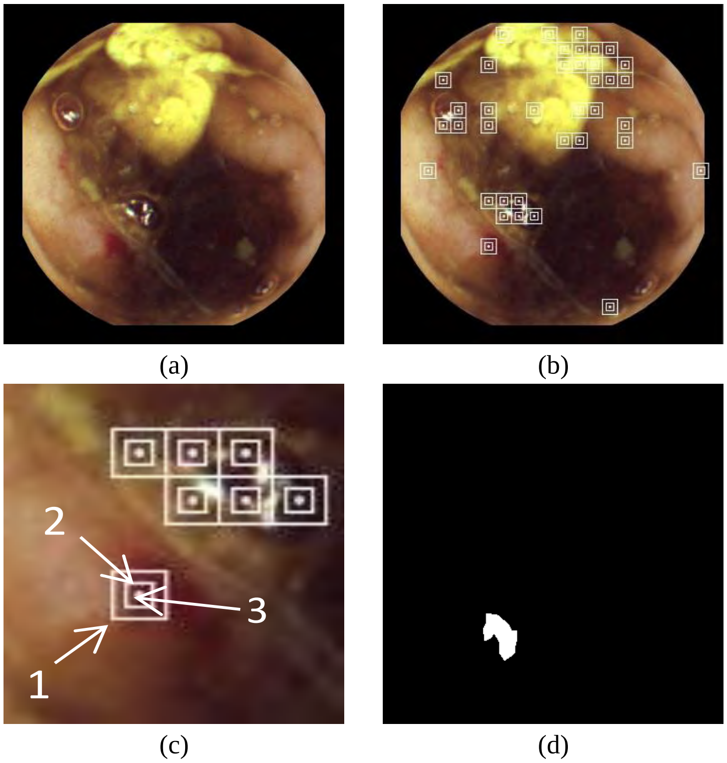


Figure 5.6: DoM salient point detection. (a) Original video frame. (b) The remaining points after dense sampling, around which the Euclidean distances are estimated. (c) A magnified region of **Figure 5.6** (b), clearly indicating the outer window (1), the inner window (2) where the maxima are calculated, and the central point (3) of these windows. (d) Detailed graphical image annotation of **Figure 5.6** (a).

In this section Difference of Maxima (DoM) Algorithm 5.1 (**Figure 5.6**) for salient point detection is proposed. The steps of the DoM algorithm are:

Algorithm 5.1 Difference of Maxima (DoM)

Input: Video Frame $I(M \times N)$,
Output: List of salient points $SP[]$

```
// Initialize
$i \leftarrow 0$;
$d[] \leftarrow$ **null**;
$I[] \leftarrow$ **null**;
// Dense Sampling Video Frame $I$ using window of size $X$ and evaluate saliency
**For each** $I(x, y) \neq 0$, $(x, y) \in [M, N]$ **do**
// $I(x, y)$ is the point 3 in Fig. 5.6(c)
// $X \times X$ neighborhood is point 1 in Fig 5.6(c)
// $X/2 \times X/2$ neighborhood is point 1 in Fig 5.6(c)
    $temp^{large}[] \leftarrow X \times X$ neighborhood centered at $I(x, y)$;
    $temp^{small}[] \leftarrow X/2 \times X/2$ neighborhood centered at $I(x, y)$;
    $M^{l\arg e} \leftarrow \max\left(temp^{l\arg e}[]\right));$
    $m^{l\arg e} \leftarrow \min\left(temp^{l\arg e}[]\right);$
    $M^{small} \leftarrow \max\left(temp^{small}[]\right);$
    $m^{small} \leftarrow \min\left(temp^{small}[]\right);$
    $i \leftarrow i + 1$;
    $d[i] \leftarrow \sqrt{(M^{l\arg e} - M^{small})^2 + (m^{l\arg e} - m^{small})^2}$
    $SP[i] \leftarrow I(x, y);$
**End For**
// Filter salient points upon their proximity
**For** $i = 1$ **to** length($d[]$) **do**
    **If** $d[i] \leq$ average($d[]$) **then**
            remove($d[i]$);
        remove($SP[i]$);
    **End If**
**End For**
```

The BoW approach is adopted and it is based on a set of extracted patches surrounding points that result from the salient point detection process of DoM algorithm. For the low-level description of the video frame patches, a set of color-based features is adopted, which has been previously applied to the problem of lesion detection in section 5.2 and yielded superior results compared to the state-of-the-art approaches (Iakovidis & Koulaouzidis 2014b; Iakovidis & Koulaouzidis 2014a).

5.4 Gastrointestinal Lesion Localization

5.4.1 CNN and CNN features

CNNs have been utilized in a variety of medical imaging domains both as conventional supervised classifiers, trained using image patches (Sekuboyina et al. 2017; Zhang et al. 2016; Ribeiro et al. 2016; Tajbakhsh et al. 2015; Bernal et al. 2017; Zhu et al. 2015; Van Grinsven et al. 2016; Anthimopoulos et al. 2016) and as weakly-supervised classifiers trained using weakly-annotated images (Georgakopoulos et al. 2016; D. K. Iakovidis et al. 2018; Zhang et al. 2016; Carneiro et al. 2017). Considering that image patches are sampled from known locations within the images, patch-based methods enable both the detection and the localization of possible anomalies; however, they require training with images annotated at pixel-level. In one of the most recent patch-based CNN approaches addressing the detection of various kinds of GI anomalies (Sekuboyina et al. 2017), input images were represented in CIE-Lab color space, and the CNN had a relatively low number of filters.

A preliminary study utilizing a CNN in a weakly-supervised framework (WCNN) was performed by (Georgakopoulos et al. 2016), aiming at the detection of inflammatory lesions. Recently, weakly supervised CNN-based approaches have been proposed in the context of GIE. In (Zhang et al. 2016), accurate detection of polyps in white-light and narrow-band imaging endoscopy, was reported using a pre-trained CNN only as a feature extractor. The pre-training was performed with non-medical images from the ImageNet dataset. A standard SVM was used for the classification of the CNN feature vectors.

5.4.2 Localization methodologies

A drawback of most weakly-supervised approaches, over the patch-based ones, is that they do not provide information about the location of anomalies within an image. Only a few approaches have been proposed to this direction.

State-of-the-art generic weakly supervised CNN-based methodologies with localization capabilities have been proposed mainly in the context of classification and segmentation of real-world objects. The methodology proposed in (Papandreou et al. 2015) uses weakly-labeled images or sub-images as bounding boxes of the objects of interest. In (Papandreou et al. 2015), it was shown that the use of weak annotations solely at the image-level is insufficient to train a high-quality segmentation model, and that the segmentation results become sufficient only when bounding boxes are used. The results improved with the use of pixel-level annotations from a subset of training images in a semi-supervised context.

In the context of GIE image analysis, abnormality localization has been based mainly on unsupervised image segmentation approaches, applicable on images with identified anomalies. As in the case of current abnormality detection methods most of them address the segmentation of only specific kinds of anomalies, such as polypoid lesions (Karkanis et al. 2003; Ganz et al. 2012; Jia 2015), and bleeding regions (Xiao Jia & Meng 2017). Recently, an application for a localized region-based active contour model for the unsupervised segmentation of various kinds of lesions was investigated (Koulaouzidis et al. 2017), and its utility was highlighted for the measurement of lesion sizes. Lesion localization, as considered in the current study, aims to attract the attention of the video reviewer at specific points within an image, where anomalies are possibly located. The specification of a few points instead of the segmentation of whole image regions provides more targeted cues about the location of the anomalies, while it usually involves fewer computations; therefore, it is preferable in terms of time-efficiency for application on GIE video frame sequences.

5.4.3 Iterative Cluster Unification Algorithm (ICU)

The proposed approach exploits WCNN architecture (D. K. Iakovidis et al. 2018) to detect and describe salient points within GIE images, **Figure 5.7**. In contrast to the current CNN-based image descriptors, which are mainly global (Zhang et al. 2016; Qian et al. 2016; Rui et al. 2018), PCFM pixel-level descriptors are extracted from each salient point. At this thesis a novel Iterative Cluster Unification (ICU) algorithm (D. K. Iakovidis et al. 2018) that exploits these descriptors to discern pixels that correspond to suspicious image regions without any detailed, pixel-level annotation. Unlike state-of-the-art approaches its application is not limited to specific GI anomalies, and it is investigated across different GIE modalities, including WCE and gastroscopy.

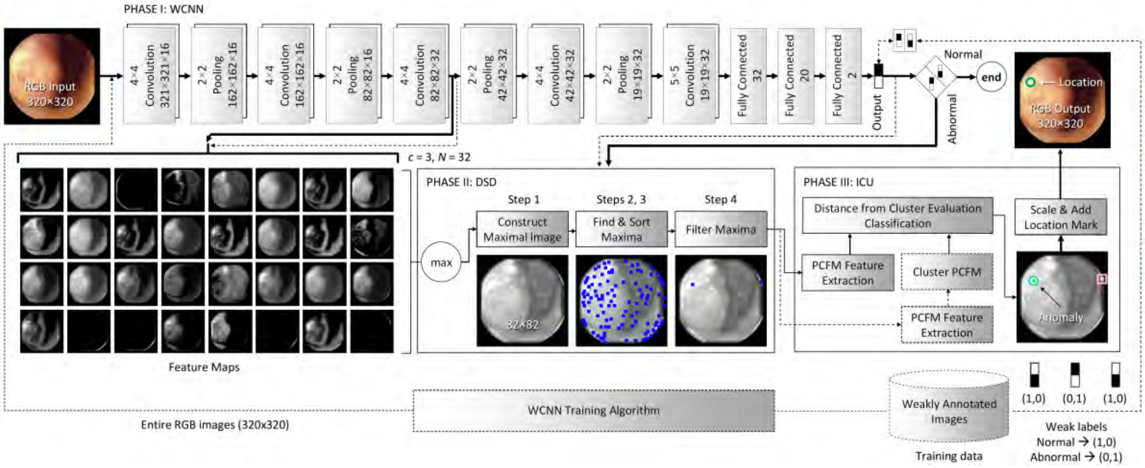


Figure 5.7 Proposed WCNN for the localization of abnormalities in weakly annotated image. In Phase I, an endoscopic image is semantically characterized as abnormal or

normal by the WCNN. Abnormal images are further analyzed by DSD salient point detection algorithm (Phase II). The salient points are classified by ICU algorithm to identify and localize possible anomalies (Phase III). The dashed lines are used to indicate the workflow of the training process.

The use of ICU algorithm is based on two preliminary phases. The first phrase a deep WCNN architecture classifies the GIE images as abnormal (having GI anomalies) or normal. In the second phase a Deed Saliency Detection algorithm is applied on the abnormal images to detect salient points in the input images using information extracted from the feature maps of a deeper WCNN convolutional layer. In the third phase ICU is now applied to identify a subset of salient points that possibly belong to GI anomalies. The coordinates of these points are then transformed (linearly scaled up) to match the spatial resolution of the input endoscopic image, on which they are superimposed to indicate the possible locations of the anomalies. Precisely, the phase of the ICU aims to determine which salient points of the abnormal images discovered in the first phase belong to GI anomalies. The motivation behind this idea is the assumption that abnormal images contain both abnormal and normal regions, whereas normal images contain only normal regions. Each salient point, detected using the DSD algorithm on these images, is represented by a feature vector composed of the values of each of the feature-maps derived from convolutional layer *c* of the WCNN at this point. The dimensionality of the feature vector is the equal to the number of feature maps of convolutional layer *c* (for *c*=3 the dimensionality is 32). The derived pointwise cross-feature-map (PCFM) features are used as input to the ICU classifier.

The ICU algorithm (Algorithm 5.2) is based on clustering to classify the salient points detected by the DSD algorithm in a weakly supervised way. It involves a training and a testing phase. During training it receives both abnormal and normal training images, and clusters their salient points upon their vector representations (which are unlabeled because the images are only weakly annotated). It considers that in the abnormal images some salient points may fall into normal regions as well. Iteratively, ICU unifies the clusters of the salient points of the abnormal images that are more similar to the clusters of the normal images. In the testing phase ICU receives an image classified as abnormal by WCNN, and the detected salient points are classified into possibly abnormal or normal upon the *K*-nearest neighbor (*K*-NN) clusters in the unified cluster space.

For simplicity, the clustering algorithm used in this study is the well-known *k*-means algorithm (S Theodoridis & Koutroumbas 2008). Preliminary investigation using other clustering algorithms, including fuzzy c-means (S Theodoridis & Koutroumbas 2008) and the recent random direction divisive clustering algorithm (Tasoulis et al. 2013), did not lead to any significant classification performance improvement. To cope with the fact that the result of this algorithm depends significantly on its initialization, the clustering algorithm is performed for T iterations with different initializations. Thus, a richer and more representative clustered vector space is generated by selecting $T > 1$. For the estimation of the distances between the clusters the Euclidean distance metric between the centroids of the clusters was used.

Algorithm 5.2 *Iterative Cluster Unification (ICU)*

```
Training phase
1. Let $I_n$ and $I_a$ be the training sets of normal and abnormal images respectively;
2. Let $Z_n$ and $Z_a$ be the sets of salient points extracted using DSD algorithm from $I_n$ and $I_a$;
3. For $i$ = 1 to $T$ do:
        For each normal image in $I_n$ do:
            Extract PCFM representations of $Z_n$;
        Cluster the PCFM representations of $Z_n$ into $Q$ clusters $N_q$, $q = 1,2,\dots,Q$;
        For each abnormal image in $I_a$ do:
            Extract PCFM representations of $Z_a$;
        Cluster the PCFM representations of $Z_a$ into $R$ clusters $A_r$, $r = 1,2,\dots,R$;
4. Set $Q = Q \cdot T$, $R = R \cdot T$;
5. For each abnormal cluster $A_r$, $r = 1,2,...R$ do:
        Calculate all distances $d_{rq}(A_r, N_q)$, $q = 1,2,...Q$ between $A_r$ and $N_q$;
        Sort distances $d_{rq}$ in ascending order;
        Calculate the normalized distance $d_{rq12} = d_{rq1}/d_{rq2}$,
    where $d_{rq1}$ and $d_{rq2}$ represent the distances of $A_r$ to its closest neighboring clusters $N_{q1}$ and $N_{q2}$;
6. Estimate the mean normalized distance $d_{q12}$ from all $d_{rq12}, r = 1,2,...R$ calculated in step 5;
7. For each abnormal cluster $A_r$, $r = 1,2,...R$ do:
            If $d_{rq12} < d_{q12}$ then unify $A_r$ with normal clusters:
                $Q$=$Q$+1; $N_Q = A_r$; $A_r = \varnothing$;

Test phase
1. Let $I_i^a$ be a new input image, characterized as abnormal by the WCNN classifier;
2. For each salient point $s$ in $I_i^a$ do:
    Extract a PCFM representation of point $s$;
    Calculate the distances of $s$ from all clusters in $A_r \cup N_q$;
        Classify $s$ as normal or abnormal based on its $K$ nearest neighbors by majority voting;
```

5.5 DINOSARC: Gastrointestinal Image/Frame Color Analysis

5.5.1 Introduction to chromatic analysis of Gastrointestinal Images

Many WCE image analysis approaches begin by detecting salient points to possible regions of abnormalities and construct saliency map, as it was stated earlier in this chapter. This section investigates a salient point detection method that is based on chromatic components of WCE. Based on this analysis a novel methodology for the detection of salient point, as well as, the definition of salient regions based on superpixel segmentation and color feature extraction is proposed. From each detected salient point and region a feature vector is calculated to describe the local color properties of the image that differentiate the abnormal from the normal tissues.

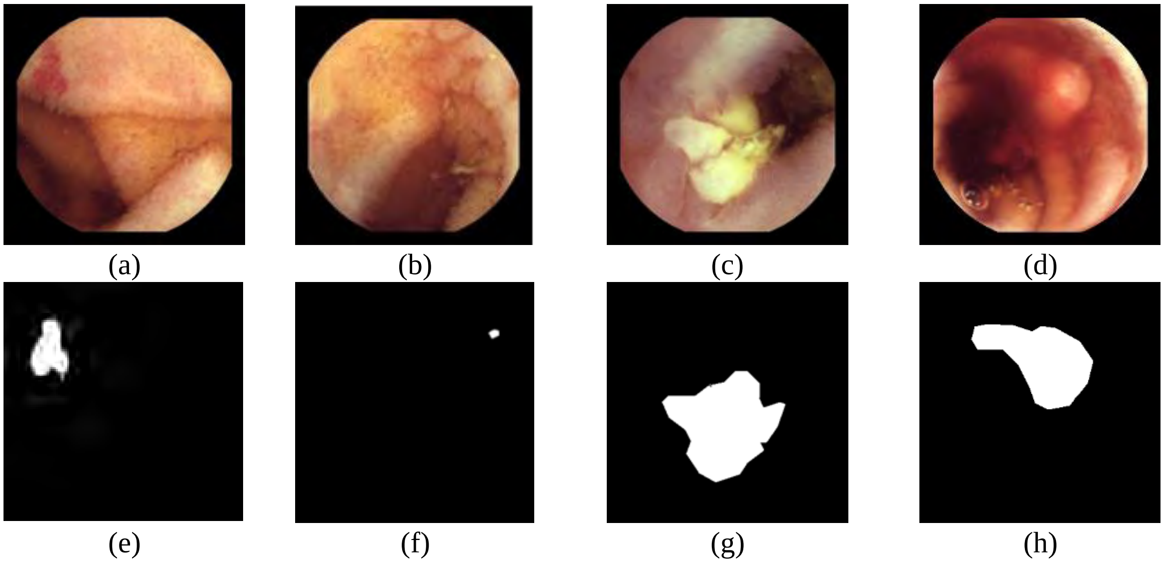


Figure 5.8 : Representative images from KID dataset (Koulaouzidis et al. 2017) (a) Vascular lesion. (b) Small inflammatory lesion. (c) Large inflammatory lesion. (d) Polypoid lesion. (e-h) Detailed graphic annotations

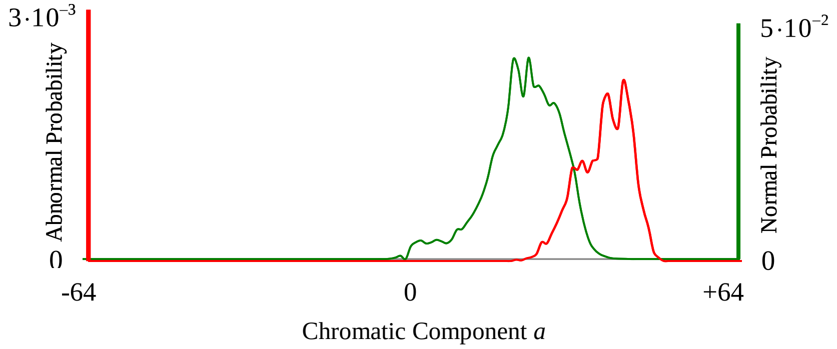
The saliency is defined respectively to *color differences* that appear in the abnormal regions. The proposed approach is based on the observation that within WCE images, the appearance of abnormalities may be described within a relatively small color range that is usually located on the margins of the overall color range of an image. In many cases this range is non-overlapping with the color range of the normal image content. By examining each WCE image separately, one may observe that the color ranges are different for each image, even for the same kind of abnormalities (Figure 5.8). Also, given a diverse set of WCE images, the color ranges of both the normal and abnormal content are completely overlapping. Therefore, it is not straightforward to specify a standard color range discriminating the abnormalities from normal content.

The rationale of the proposed saliency detection algorithm can be explained by the respective color histograms of WCE images. WCE images are represented in the CIE-*Lab* color space, which describes color with approximately decorrelated components (Iakovidis & Koulaouzidis 2014b). The components of this space represent lightness (*L*), the quantity of red (a>0) or the quantity of green (-a>0), the quantity of yellow (b>0) or the quantity of blue (-b>0) of a pixel. This way, color can be examined separately from lightness, which in our case varies significantly depending on the distance and the angle of the endoscope from the tissue surface. Thus, by only using the chromatic components *a* and *b*, the color information can be isolated and an approximately illumination-invariant description of the image content may be obtained.

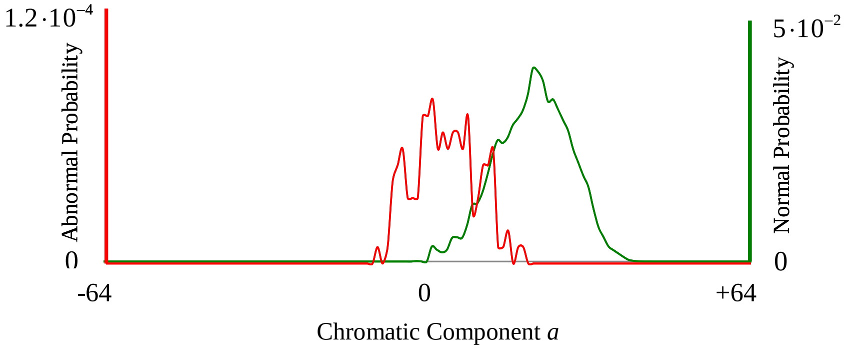
Let H^A and H^N be the normalized histograms (probability distributions) of a WCE image for abnormal and normal regions, respectively. For the images of Figure 5.8 the respective histograms are illustrated in Figure 5.9. H^A is represented by a red line, and H^N is represented by a green line. The histograms is provided for the chromatic components i.e., *a* or *b*, of the images where a non-overlapping range between the two histograms can be observed. For example, in Figure 5.9(a) a non-overlapping region between H^A and H^N can be observed only in component *a* (in the chromatic region $a \in [-9,-1]$); this the respective histograms of component *b* are omitted. Similarly, Figure s. 5.9 (b)-(d) present the non-overlapping histograms of the images illustrated in Figure s. 5.8(b)-(d). In Figure 5.10 the normalized histograms estimated over all images of KID dataset (section 5.6) are provided, where a total overlap between the abnormal and normal chromatic ranges can be observed.

5.5.2 Salient point detection

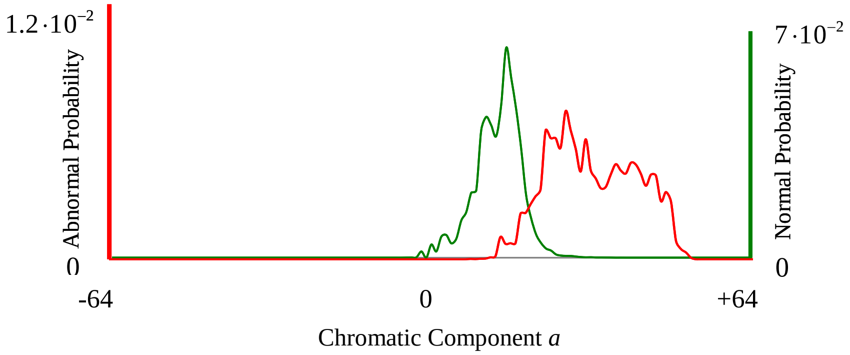
A novel algorithm named SARC (Algorithm 5.3) is proposed, which performs a Selective Aggregation of chRomatic image Components after an automatic segmentation process. This is based on the observation that abnormal image regions are usually characterized by higher positive or negative values of the *a* and *b* chromatic components. SARC produces saliency maps which emphasize on the regions that correspond to possible abnormalities.


$3\cdot10^{-3}$
Abnormal Probability
0
$5\cdot10^{-2}$
Normal Probability
0
-64
0
+64
Chromatic Component *a*

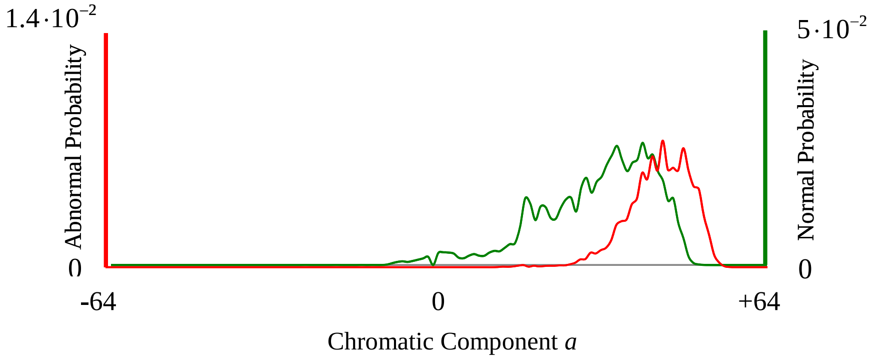
(a)


$1.2\cdot10^{-4}$
Abnormal Probability
0
$5\cdot10^{-2}$
Normal Probability
0
-64
0
+64
Chromatic Component *a*

(b)


$1.2\cdot10^{-2}$
Abnormal Probability
0
$7\cdot10^{-2}$
Normal Probability
0
-64
0
+64
Chromatic Component *a*

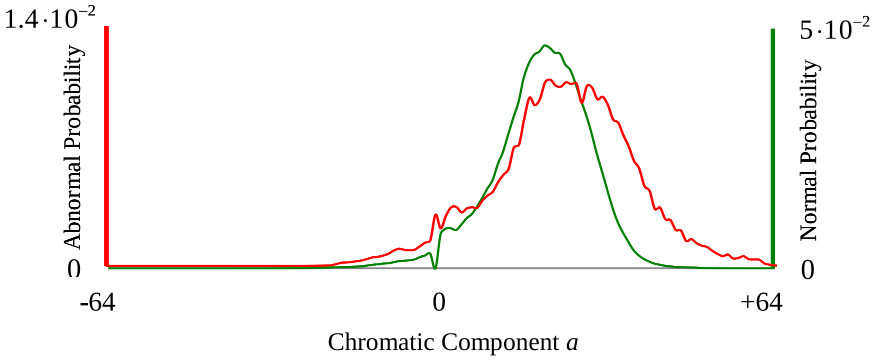
(c)


$1.4\cdot10^{-2}$
Abnormal Probability
0
$5\cdot10^{-2}$
Normal Probability
0
-64
0
+64
Chromatic Component *a*

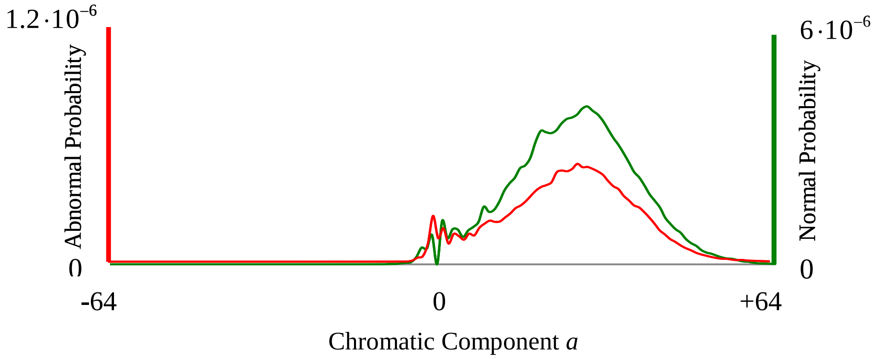
(d)

Figure 5.9 Normalized chromatic histograms of WCE images of Fig. 5.8 (for chromatic components that have non-overlapping regions). Histogram H^A which is estimated from abnormal image regions is represented with a solid red line, and H^N which is estimated from normal image regions is represented with a dashed green line. (a) Vascular lesion. (b) Small inflammatory lesion. (c) Large inflammatory lesion. (d) Polypoid lesion.

Let I_{Lab} be a M×N-pixel CIE-*Lab* input image, and I_a and I_b be the grayscale images representing a and b components of I_{Lab}. This algorithm uses the histogram H_c of image I_c, c=a, b, to determine optimal thresholds maintaining the image regions that have a higher probability to include an abnormality. It calculates the first (r_c) and second derivatives (R_c) of the positive (+) and negative (-) axes of H_c, and determines the maximum of each of the second derivatives as an optimal image threshold for maintaining as much as discriminating information about the possible abnormal regions within the chromatic components of the image as possible. This process determines the value of the chromatic component (*a* or *b*) where the rate of the first derivative changes. Considering that the chromatic ranges of the abnormal regions are located at the margins of the histograms, this value will be the most probable one for the abnormality. Figure 5.11 illustrates this concept. It can be noticed that the maximum of the second derivative corresponds approximately to the value of the chromatic component *a* with the maximum probability of the abnormality.



(a)



(b)

Figure 5.10 Normalized chromatic histograms estimated from all WCE images in KID dataset (Koulaouzidis et al. 2017). Histogram H$_A$ which is estimated from abnormal image regions is represented with a solid red line, and H_N which is estimated from normal image regions is represented with a dashed green line. (a) Chromatic component *a*. (b) Chromatic component *b*.

By applying the determined thresholds on the respective chromatic image components, four images are obtained. Indicative examples of such images for the cases of Fig.5.8(a) and Fig.5.8(b) are illustrated in the first and in the second row of Fig. 5.12, respectively. These images are subsequently filtered using a sliding window of $n \times n$ pixels, which aims to discard local non-maxima. The final step of SARC algorithm is the aggregation of the four filtered images using the sum operator.

Algorithm 5.3. Selective Aggregation of chRomatic image Components *(SARC)*

Input: Images $I_c(M \times N)$, $c=a, b$

Output: Image I_{SARC}

```
Compute histogram $H_c$ of images $I_c$,
$L$ = length($H_c$);
// Calculate the first and second derivatives of $H_c$
// $r_c^+$ first rate
$r_c^+ \leftarrow dH_c(i)/di, \;\; i = L/2-1, L/2-2, \ldots, 0;$
$r_c^- \leftarrow dH_c(i)/di, \;\; i = -L/2+1, -L/2+2, \ldots, 0;$
$R_c^+(i) \leftarrow r_c^+(i)/di, \;\; i = L/2-2, L/2-3, \ldots 0;$
$R_c^-(i) \leftarrow r_c^-(i)/di, \;\; i = -L/2+2, -L/2+3, \ldots 0;$
// denote the max values of value field
$T_c^{+/-} \leftarrow \arg\max\left(R_c^{+/-}\right);$
For each $(x, y) \in [M, N]$ do
    If $I_c^+(x, y) < T_c^+$ then $I_c^+(x, y) \leftarrow 0;$
    If $I_c^-(x, y) > T_c^-$ then $I_c^-(x, y) \leftarrow 0;$
End For
```

```
$I_c^{+/-} \leftarrow$ Normalize $\left|I_c^{+/-}\right|$; // Enhance contrast
// Non-maxima filtering
For each $(x, y) \in [M, N]$ in $I_c^{+,-}$ do
    $temp[] \leftarrow n\times n$ neighborhood centered at $I_c^{+,-}(x, y)$
    For $i=0$ to $n\times n$ do
        If $temp[i] < p(x, y)$ then $temp[i]=0$;
    End For
//temp is matrix with odd number of rows and columns
    If $\sum_{i\in[0,n\times n]} temp[i] > 0$ then $temp\lceil(n\times n/2)+n/2\rceil = 0$;    $I_c^{+,-}(x, y) \leftarrow temp[]$;
End For
$I_{\text{SARC}} = \sum_{c=a,b} I_c^{+} + \sum_{c=a,b} I_c^{-}$
```

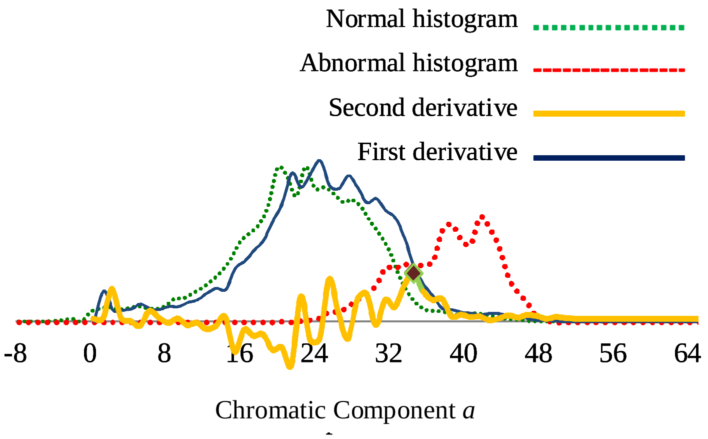


Figure 5.11 Determination of the optimal threshold T_a^+ (black point) based on the second derivative R_a^+ of the histogram estimated from chromatic component a of the WCE image of Fig. 5.8(a). The application of this threshold on a results in Fig. 5.12(a).

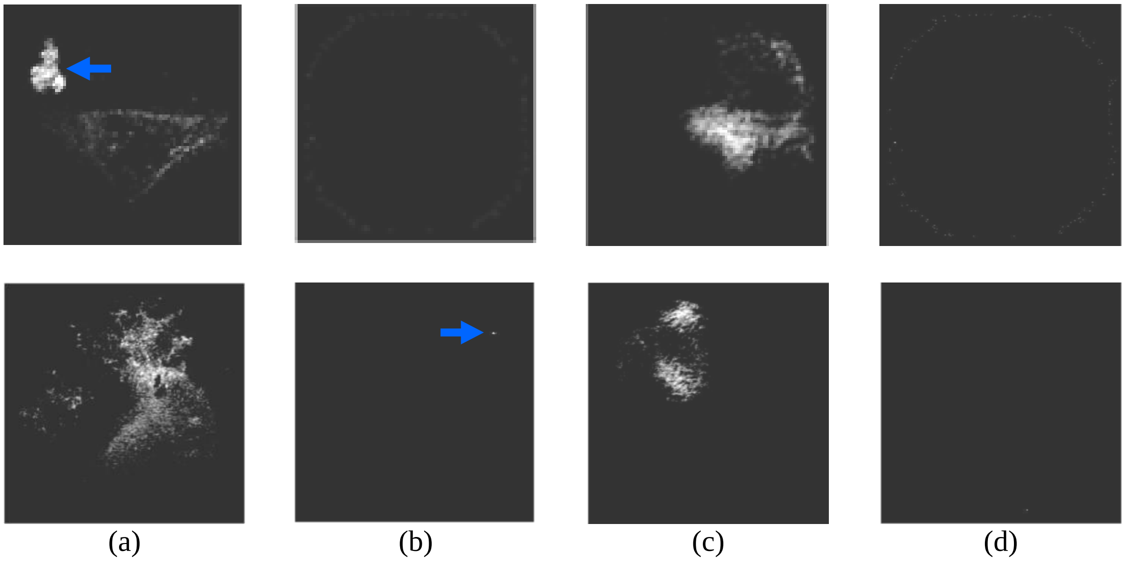


Figure 5.12 Representative output images obtained from the application of Algorithm 1 on the WCE images illustrated in Figure 5.8 (a) (first row) and Figure 5.8 (b) (second row) respectively. (a) I_a positive. (b) I_a negative. (c) I_b positive. (d) I_b negative. The arrows indicate the locations of the lesions. The vascular lesion of Fig. 1(a) is clearly discriminated in the respective I_a positive image. Also the small inflammatory lesion of Fig. 1(b) is clearly discriminated in the respective I_a negative image.

In the next step the DInstances On SARC (Algorithm 5.4) uses I_{SARC} to detect regions with significant changes in the chromatic components of CIE-*Lab* color space. Initially, I_{SARC} is sampled using concentric square windows of s×s and $s/2$×$s/2$ pixels respectively at each pixel of I_{SARC} with non-zero value. These pixels are more considered as points of more interest, since they have a maximum value within their neighborhood. From these points as salient are defined those ones that are characterized by a significant change in the chromatic values of their local neighborhood. This change is calculated by the distance between the maxima and the minima extracted from the concentric square windows. This definition of saliency is inspired by the fact that such chromatic changes are usual in the neighborhoods of most abnormalities.

Algorithm 5.4. DIstaNces On SARC (DINOSARC) Salient Point Detection

Input: Images $I_c\,(M{\times}N)$, $c{=}a, b$; $I_{SARC}\,(M{\times}N)$

Output: List of salient points $I[]$

```
// Initialize
$i \leftarrow 0$;
$d[] \leftarrow$ null;
$I[] \leftarrow$ null;
// Sample $I_{\mathrm{SARC}}$ and evaluate saliency
For each $I_{\mathrm{SARC}}(x, y) \neq 0,\ (x, y) \in [M, N]$ do
    $temp^{large}[] \leftarrow s{\times}s$ neighborhood centered at $I_c(x, y)$;
    $temp^{small}[] \leftarrow s/2{\times}s/2$ neighborhood centered at $I_c(x, y)$;
    $M^{l\arg e} \leftarrow \max(temp^{l\arg e}[]))$;
    $m^{l\arg e} \leftarrow \min(temp^{l\arg e}[])$;
    $M^{small} \leftarrow \max(temp^{small}[])$;
    $m^{small} \leftarrow \min(temp^{small}[])$;
    $i \leftarrow i + 1$;
    $d[i] \leftarrow \sqrt{(M^{l\arg e} - M^{small})^2 + (m^{l\arg e} - m^{small})^2}$
    $I[i] \leftarrow I_c(x, y)$;
End For
// Filter salient points upon their proximity
For $i = 1$ to length($d[]$) do
    If $d[i] \leq$ average($d[]$) then
        remove($d[i]$);
        remove($I[i]$);
    End If
End For
```

5.5.3 Salient region detection

The salient point detection process is followed by sampling image regions from their neighborhoods in order to estimate relevant descriptors. Instead of sampling square-shaped neigborhoods, as in (Iakovidis & Koulaouzidis 2014b; Iakovidis & Koulaouzidis 2014a), the DINOSARC descriptors are extracted from arbitrary-shaped neighborhoods. To this end, the input images are segmented using the simple iterative linear clustering (SLIC) algorithm (Achanta et al. 2012). SLIC creates clusters of pixels defining regions of homogeneous color properties, called superpixels (Figure 5.13). Considering the approach proposed by (Iakovidis et al. 2015), the superpixels that contain at least one salient point are also characterized as salient. However, in that study the pixel-level saliency was disregarded, and the localization of abnormalities smaller than a superpixel was impossible. In this study, the pixel-level saliency defined by DINOSARC algorithm is not superseded by the region-level saliency defined by the superpixels. Each DINOSARC salient region is defined by a superpixel that includes only a single, representative salient point. If the superpixel contains a cluster of salient points, then the cluster centroid is regarded as its corresponding salient point.

5.5.4 Local and Global Color Image Descriptors

Another novel contribution of this work is that both DINOSARC salient regions and points are represented by a local color feature vector. The local feature vectors are subsequently used for the formation of feature vectors globally representing the WCE images. The feature extraction process presented is an extension of the approach (Iakovidis & Koulaouzidis 2014b) for only local representation of square WCE image patches along with their central point.

The proposed, extended approach forms a 9-dimentional feature vector from the color components (L, a, b) of the CIE-*Lab* representation of a salient point, as well as the minimum and maximum values of each of the L, a, and b components within the DINOSARC salient region, i.e., min(L), max(L), min(a), max(a), and min(b), max(b). This is inspired by the way the WCE video reviewers empirically assess the image regions for the detection of abnormalities, which, takes into account regional color differentiations (Iakovidis & Koulaouzidis 2014a). By only including the minimum and maximum values from the salient regions (which are also determined by salient points derived from color differences) such differentiations can be captured.

The local image representation approach is extended by adopting the BoVW model (Csurka et al. 2004) for the extraction of global features from the WCE images. This model considers that an entire image can be represented by a visual vocabulary. Such a vocabulary may be seen as a set of “exemplar” image patches (visual words), in terms of which any given image may be described. The vocabulary may be seen as a means of quantization of the feature space derived from the local feature vectors. Then, any previously unseen descriptor may be easily quantized to its nearest visual word. Thus the DINOSARC feature vectors are used to form histograms of visual words for the representation of entire WCE images.

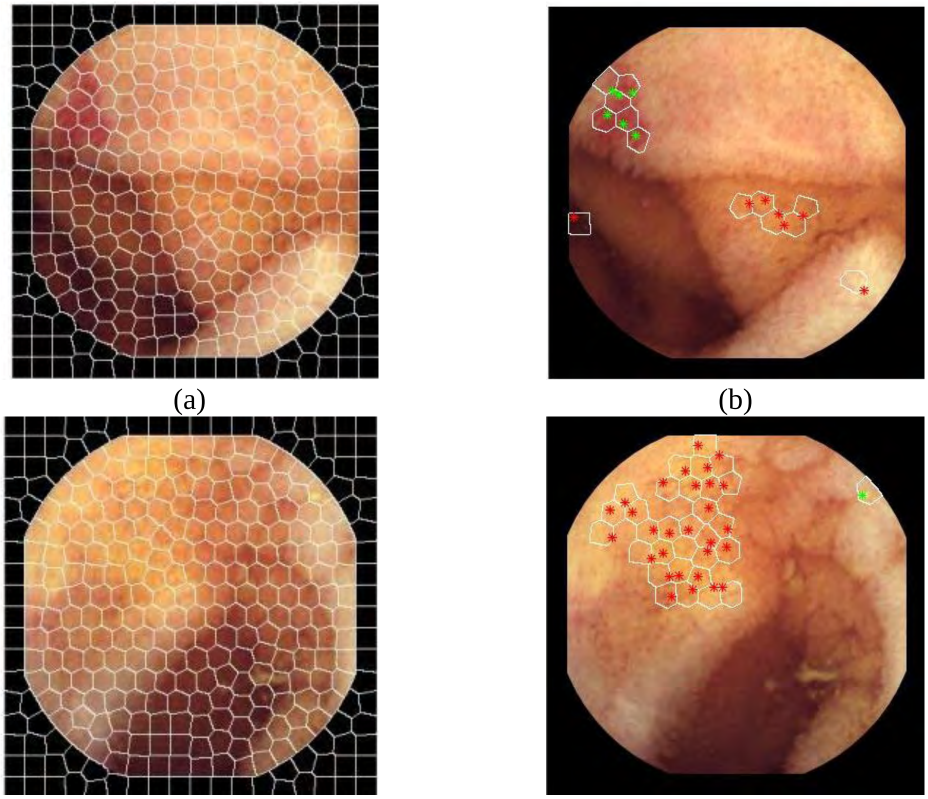


Figure 5.13 Result of SLIC algorithm on an endoscopy image (left), superpixels with DINOSARC points (right). The images (c) and (d) show the effectiveness of saliency detection for small lesions.

5.6 Experimental Results

5.6.1 Datasets

In order to enable reproducibility of the experiments and comparisons with current and future studies, two publicly available image datasets were used for the evaluation of the proposed methodologies of this chapter. These datasets have been acquired with different endoscopic imaging modalities. They have been selected primarily for their diversity, as they include different kinds of anomalies and normal images.

The first dataset is composed of images obtained from gastroscopies. It was released for the purposes of a challenge that took place in MICCAI 2015 (Deep sparse feature selection for computer aided endoscopy diagnosis). The task in that challenge was to correctly classify the gastroscopic images and to detect abnormalities. In the whole chapter, the same dataset is used for the detection of abnormal images using only semantically annotated training images.

The gastroscopy challenge dataset was derived from a total of 10,000 images, obtained from 544 healthy volunteers and from 519 volunteers with various lesions, such as gastritis, cancer, bleeding and gastric ulcer. The original image resolution was 768×576 pixels. The images were anonymized by cropping the image regions containing sensitive patient information. The size of the derived images is 489×409 pixels (Deep sparse feature selection for computer aided endoscopy diagnosis).

For the purposes of the MICCAI challenge, a subset of images was selected and separated, into two approximately balanced sets of training and a set of testing images. The training set consists of 205 normal and 260 abnormal images, and the test set consists of 104 normal and 129 abnormal images.

The second dataset is composed of WCE and has been recently released by (Koulaouzidis et al. 2017). KID dataset is a publicly available database of annotated WCE images and videos (including pixel-level annotations), which can be used as a reference for both training and evaluation of such systems (Iakovidis & Koulaouzidis 2015; Koulaouzidis et al. 2017). It contains WCE images obtained from the whole GI tract using a MiroCam capsule endoscope with a resolution of 360×360 pixels. These include 303 images of vascular anomalies (small bowel angiectasias, lymphangiectasias, and blood in the lumen), 44 images of polypoid anomalies (lymphoid nodular hyperplasia, lymphoma, Peutz-Jeghers polyps), 227 images of inflammatory anomalies (ulcers, aphthae, mucosal breaks with surrounding erythema, cobblestone mucosa, luminal stenoses and/or fibrotic strictures, and mucosal/villous oedema), and 1,778 normal images obtained from the esophagus, the stomach, the small bowel and the colon.

5.6.2 Preliminary evaluation of Classification on inflammatory lesion of KID database Results

The proof of concept for the use of color BoW on WCE, a preliminary evaluation of the proposed supervised BoW using weakly labeled images approach(Vasilakakis et al. 2016) was conducted. The experiments were performed using a subset of KID database. This dataset displays a variety of different kinds of abnormalities. More precisely, the selected subset consists of 227 images of most common inflammatory lesions including ulcers, aphthae, mucosal breaks with surrounding erythema, cobblestone mucosa, stenoses and/or fibrotic strictures, and significant mucosal/villous oedema. It also includes a set of 1327 normal images derived from the small bowel (728 images) and the stomach (599 images).

Considering that supervised methods use whole images for training and testing, a more thorough evaluation, using 6-fold Cross Validation (CV) was computationally feasible, using KID dataset. This enables a less biased evaluation with respect to the selection of the training and testing sets, by randomly splitting the dataset into 6 non-overlapping parts. Out of the 6 parts, 1 was used for training and one for testing, repeatedly, until each part was used for testing once. The classification performance was investigated using Receiver Operating Characteristic (ROC) curves. An ROC curve depicts relative tradeoffs between benefits (correct decisions about abnormal cases, characterized as True Positives, TPs) and costs (false decisions about normal cases, characterized as False Positives, FPs) (Fawcett 2006). From a medical viewpoint, the ROC yields a pure measure of diagnostic accuracy, independent of the diagnostic criterion and of the frequencies of the alternative conditions under study (Swets 1979). With respect to the abnormality detection problem, the respective conditions are defined by the presence of an abnormal tissue within an endoscopic image or not. With respect to the localization problem, the respective conditions are defined as whether a point diagnosed as abnormal is located within an abnormal image area or not. The Area Under the ROC (AUC) is an overall summary measure of diagnostic accuracy (Alemayehu & Zou 2012). In order to enable comparisons between the ROC curves, the AUC was used as a classification performance measure which, unlike accuracy, is relatively robust for datasets with imbalanced class distributions (Provost et al. 1997), as in the case of KID dataset.

In order to investigate whether BoW could be used as a reliable classification approach, its performance is examined in five different experiments. These differentiate on the method for the selection of interest points, the description of patches around the aforementioned points; and the colour space used. For the latter case the greyscale images and also transformations of CIE-Lab are used (using standard illuminant D65), where L and b channels had been discarded, keeping only the colour information of a. We shall refer to the latter as the “Lab images”. More specifically, the performed experiments are as follows:

i) SURF points and features on the greyscale image;
ii) dense points and SURF features on the Lab images;
iii) SURF points and colour features of (Iakovidis & Koulaouzidis 2014a);
iv) dense points and colour features of (Iakovidis & Koulaouzidis 2014a);
v) the state-of–the-art method of (Yuan et al. 2016a), where image description is based on the combination of SIFT and compound local binary pattern features (CLBP).

In each case, interest points are extracted, then their descriptions for the visual vocabulary, which is use for image BoW description and finally train SVM classifiers.

Feature Extraction	Feature Description	Window Size	Vocabulary Size	AUC
dense (18)	*Lab* (Iakovidis & Koulaouzidis 2014a)	18×18	500	0.80
dense (4)	SURF (g)	N/A	800	0.70
dense (36)	*Lab* (Iakovidis & Koulaouzidis 2014a)	36×36	700	0.79
dense (18)	SURF (g)	18×18	800	0.69
dense (10)	*Lab* (Iakovidis & Koulaouzidis 2014a)	18×18	700	**0.81**
SURF (*a*)	*Lab* (Iakovidis & Koulaouzidis 2014a)	N/A	700	0.77
SURF (g)	SURF (g)	N/A	500	0.59

Table 5.1: Experimental Results; in dense(x), x denotes the step, in SURF(y), y denotes the color space (g: greyscale, a: a channel of Lab). Note that in case of SURF feature description, image patches are selected by the algorithm, thus is marked herein as “N/A”

The visual vocabulary size ranged from 300 to 1200 words. For the experiments with dense SURF, multi-scale feature extraction with scale step 1.6, starting from scale 1.6, up to scale 6.4 is used. Experiments with various sizes of square regions, for the extraction of the colour features are done. Square areas of size 18×18 and 36×36 are used. For dense feature extraction grid steps of 4, 10, 18 and 36 pixels, both horizontally and vertically are used. For the method of (Yuan et al. 2016a) CLBP of patch size 4×4 and 8×8 is used. For the classification an SVM with RBF kernel is used.

Most notable results are summarized in Table 5.1. In the table it is observed that best performance was achieved for the case of dense Lab features using a window size of 18×18 pixels and a visual vocabulary of 700 words. The best performance of standard SURF features (i.e., applied on grayscale images) was achieved using dense extraction and a vocabulary size of 800 words. However, this advantageous performance comes at cost of efficiency, since the number of samples obtained by dense SURF is higher (due to the regular sampling process). In addition, the proposed approach had better results in comparison with of the state-of-the-art method of (Yuan et al. 2016a). In any case the application of SURF on the *a* channel of CIE-Lab leads to an increase of AUC.

5.6.3 Performance Evaluation of supervised classification using weakly labeled images

The results of the previous paragraph have been extended to the whole KID and MICCAI datasets. As in the previous paragraph, the supervised methods use whole images for training and testing, a more thorough evaluation, using 10-fold CV was computationally feasible, using MICCAI and KID datasets.

The color BoW was compared (D. K. Iakovidis et al. 2018)with four state-of-the-art supervised approaches. These include the CNN-based approaches of (Zhang et al. 2016) , and of Jia and Meng (Jia & Meng 2016), and the BoW-based approaches of Yuan et al (Yuan et al. 2016b) using SVM as a classification scheme and the optimal parameters suggested in the respective studies. The average results obtained over the CV evaluation are summarized in **Table 5.2** and the standard deviation of the measurements was of the order of 10^{-2}. Overall, focusing on the AUC measures, color BoW performs better than the compared supervised schemes, with a significant advantage over Yuan's et al in the classification of KID. The classification performance of color BoW is almost equivalent to that of Zhang's et al method on MICCAI, while it outperforms the other state-of-the-art methods on KID dataset.

Table 5.2: 10-Fold CV Classification Results of the WCNN and State-of-the-Art Supervised Methods

Measure	(Zhang et al. 2016)		(Jia & Meng 2016)		(Yuan et al. 2016b)		(Vasilakakis et al. 2016)	
	MICCAI	**KID**	**MICCAI**	**KID**	**MICCAI**	**KID**	**MICCAI**	**KID**
AUC	0.951	0.773	0.902	0.705	0.940	0.709	0.946	**0.802**
Accuracy	0.851	0.760	0.827	0.690	0.867	0.696	**0.892**	**0.768**
Sensitivity	0.930	0.537	0.806	0.602	0.876	0.432	0.911	0.454
Specificity	0.779	0.836	0.857	0.785	0.854	0.820	**0.872**	**0.886**

5.6.4 Performance evaluation of Multi-label Classification

Several experiments were performed to evaluate multi-label classification as a means for semantic interpretation of endoscopy video frames. The saliency-enabled BoW methodology was evaluated in comparison to state-of-the-art approaches.

In the case of the saliency-enabled BoW methodology, for each video frame, features have been extracted using the proposed DoM salient point detection method and the "naive" approach of dense feature extraction. For the proposed DoM salient point detection method image samples of 24×24 pixels were used. The BoW model was constructed with a visual vocabulary with sizes in the range from 500 to 2000 words using the *k-means* clustering algorithm (Drake & Hamerly 2012). The classification of the feature vectors obtained using the BoW method, was implemented by an SVM classifier. Linear, polynomial and Radial Basis Function (RBF) kernels have been tested, and followed the grid-search approach (Chang & Lin 2011) to determine its optimal parameters. The RBF kernel provided the best results, for a minimum cost parameter c=10 and γ=2^{-8}.

To compare the classification performance of saliency-enabled BoW with the transfer learning approach in multi-label classification of WCE gastrointestinal tract images, the implemented methodology by (Zhang et al. 2016) was followed. More specifically, for the feature extraction the same procedure as presented by the authors was followed, while for the classification of the extracted features, multi-label "one-*vs*-all" SVM with $c = 2^{-9}$and polynomial kernel was followed. The parameters of the SVM where selected after a series of experiments in order to determine the optimal values for the domain.

The classification performance was thoroughly investigated using Receiver Operating Characteristic (ROC) analysis and the area under the ROC (AUC). Experiments were performed using the 10-fold cross validation evaluation scheme, using SVMs as a binary classifier. Multi-label classification was implemented using a derivative of WEKA library (Witten et al. 2016) called MEKA(Read et al. 2016).

Initially, the case of binary classification of the video frames into normal and abnormal classes was examined. The performance of BoW method was investigated using the proposed DoM for salient point detection in comparison with the state of the art methodologies, of (Yuan et al. 2016b), who used SIFT algorithm for the detection of interest points and a concatenated feature vector of SIFT and LBP, or SIFT and CLBP(Yuan et al. 2016b), for the description of video frame regions. The comparison also includes the method of (Vasilakakis et al. 2016), who used SURF algorithm for salient point detection in the a-channel (SURF(a)) of CIE-Lab and the dense BoW approach and the CNN(Zhang et al. 2016). The results with regards to lesion detection are presented in **Table 5.3**. It can be noticed that the use of the proposed DoM algorithm increases the binary classification performance to an AUC of 0.81. All methods provide a low sensitivity. This indicates the difficulty of the lesion detection problem. The BoW method using DoM provided significantly higher specificity (less false positives) than all other methods. The higher sensitivity was obtained by CNN, at the cost of a higher false positive rate.

Table 5.3: Binary classification results for lesion detection using various supervised BoW methods and CNN with an SVM classifier. The confusion matrix (True Positives – TP, False Negatives – FN, False Positives – FP, and True Negatives – TN), the sensitivity and the specificity are provided along with the AUC for each method

Methods	TP	FN	FP	TN	Sensitivity	Specificity	AUC
BoW+SURF(a)	232	342	214	1564	0.40	0.87	0.78
BoW+SIFT+LBP	181	393	210	1568	0.31	0.88	0.72
BoW+SIFT+CLBP	207	367	200	1578	0.36	0.88	0.78
BoW+Dense	240	334	196	1582	0.42	0.88	0.8
CNN	299	275	259	1519	0.52	0.85	0.78
BoW+DoM	252	322	176	1602	0.44	0.90	**0.81**

Multi-label classification was performed using the following labels: abnormal, debris, bubbles, and lumen hole. The use of DoM for multi-label classification, results in an even higher classification performance than the conventional binary classification scheme. Best results were obtained using the Multi-Layer Perceptron (MLP) multi-label classification method with 100 hidden layer neurons, a learning rate of 0.1, trained with the features extracted from BoW model. The obtained AUC reached up to 0.83 using a vocabulary of 800 visual words. The results for all methods using BoW features are presented in **Figure 5.14**. The basic methods for multi-label classification, which used, were Binary Relevance (BR), Label Combination (LC), Ranking and Thresholding (RT) and Pairwise Classification (PC). For all multi label methods the same SVM with Radial Basis Function kernel (RBF) with c=10 was used. As in the binary classification experiments, these parameters were determined using the afore-mentioned kernels and grid-search approach. Also, **Figure 5.14** includes the results of CNN (Zhang et al. 2016) for multi-label classification It can be noticed that saliency-enabled BoW provided the highest performance compare to all the other approaches and achieved an AUC equal to 0.83.

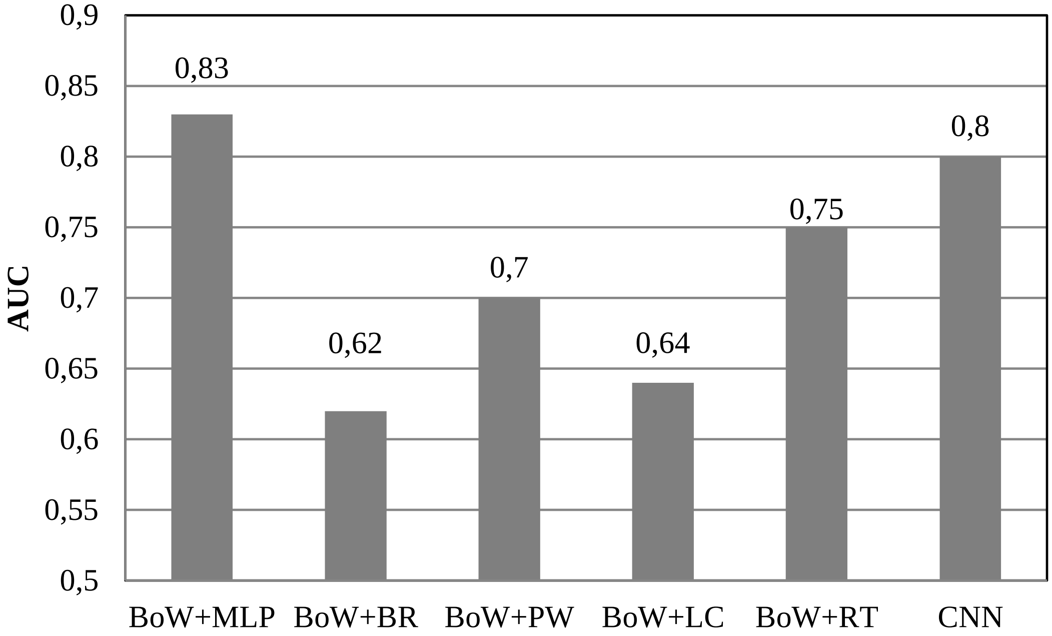


Figure. 5.14: Comparative multi-label lesion detection results for each multi-label method tested.

The classification results per semantic category are presented in **Figure 5.15**. It can be noticed that the result for debris are significantly higher than the results of bubbles and lumen hole. The reason is that the most video frames in KID dataset had debris as content compare to the number of video frames that had bubbles and/or lumen hole.

It can also be noticed that the classification performance of the CNN is not always better than the BoW-based approaches, although it has been proved effective in the context of endoscopy (Zhang et al. 2016). This could be explained by the diversity of the KID dataset, which includes several different kinds of lesions, whereas the dataset used by (Zhang et al. 2016) included only colorectal polyps.

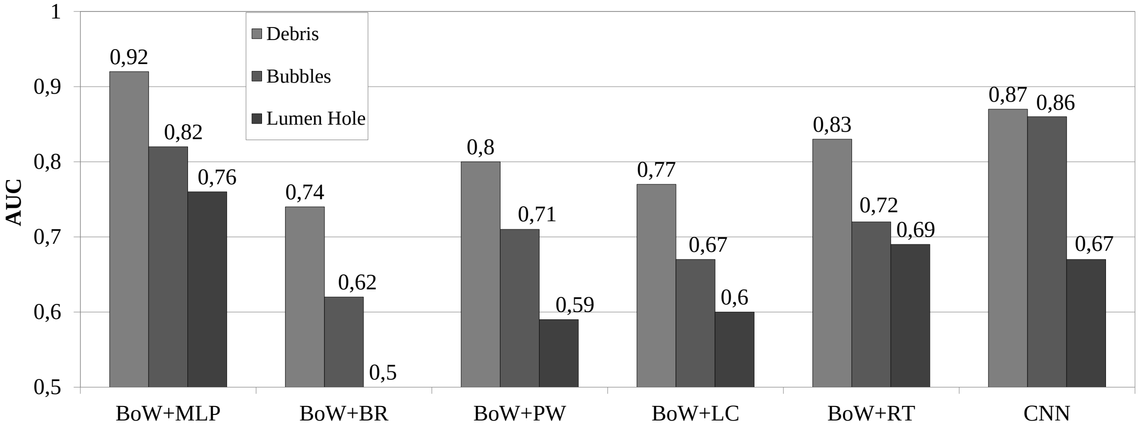


Fig. 5.15: Classification results for each semantic label in KID dataset 2 for each multi label method

5.6.5 Performance evaluation on Abnormality Localization

The performance of the proposed methodology (D. K. Iakovidis et al. 2018) in the localization of GI anomalies was evaluated on both MICCAI and KID dataset, by extending the 10-fold CV scheme. The ICU algorithm (Algorithm 5.2) filters the salient points detected using the DSD algorithm (D. K. Iakovidis et al. 2018), by classification, and outputs a number of points that indicate possible locations of GI anomalies within abnormal images. The results obtained using the proposed PCFM features are compared to the results obtained using standard color features. To this end, a feature vector composed of the mean values of the respective CIE-Lab color space components (a, b) is used. The means are estimated over a 5×5 pixel neighborhood centered at the salient points. The choice of the window size used was determined as best, based on preliminary experimentation among window sizes of 1×1 to 16×16 pixels.

The number of clusters Q and R tested in the k-means algorithm varied from 2 to 10, and the number of k-means executions was T=10. The number of nearest neighbors tested was K=1,3,5,7 and the best performance was achieved by K=1. The results obtained from the output of ICU, in terms of AUC are illustrated in **Figure 5.16**. This figure shows that best results in the two datasets are achieved for the least number of clusters (k=2). For higher values of k the two classes are more difficult to discriminate. Using the PCFM features the best localization performance achieved in MICCAI dataset is 0.848, and in KID dataset it is 0.877. Using CIE-Lab features the respective performances were 0.801 and 0.852. Overall, PCFM features perform better than CIE-Lab features, especially in the case of the larger dataset (KID).

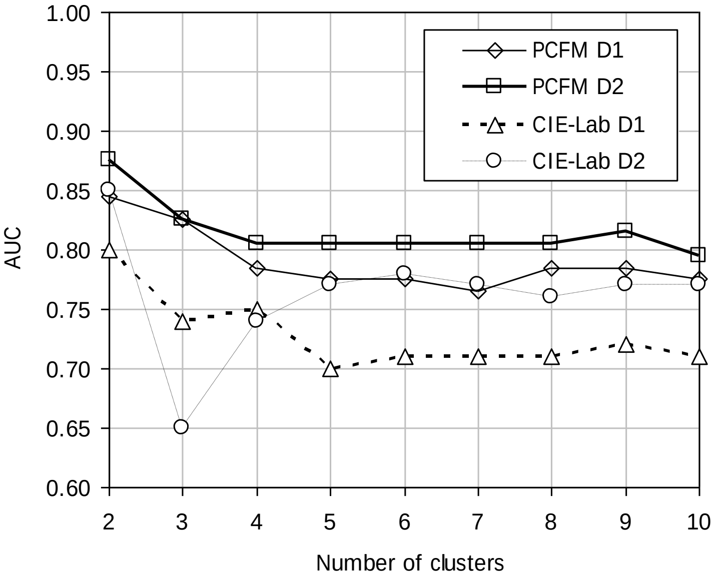
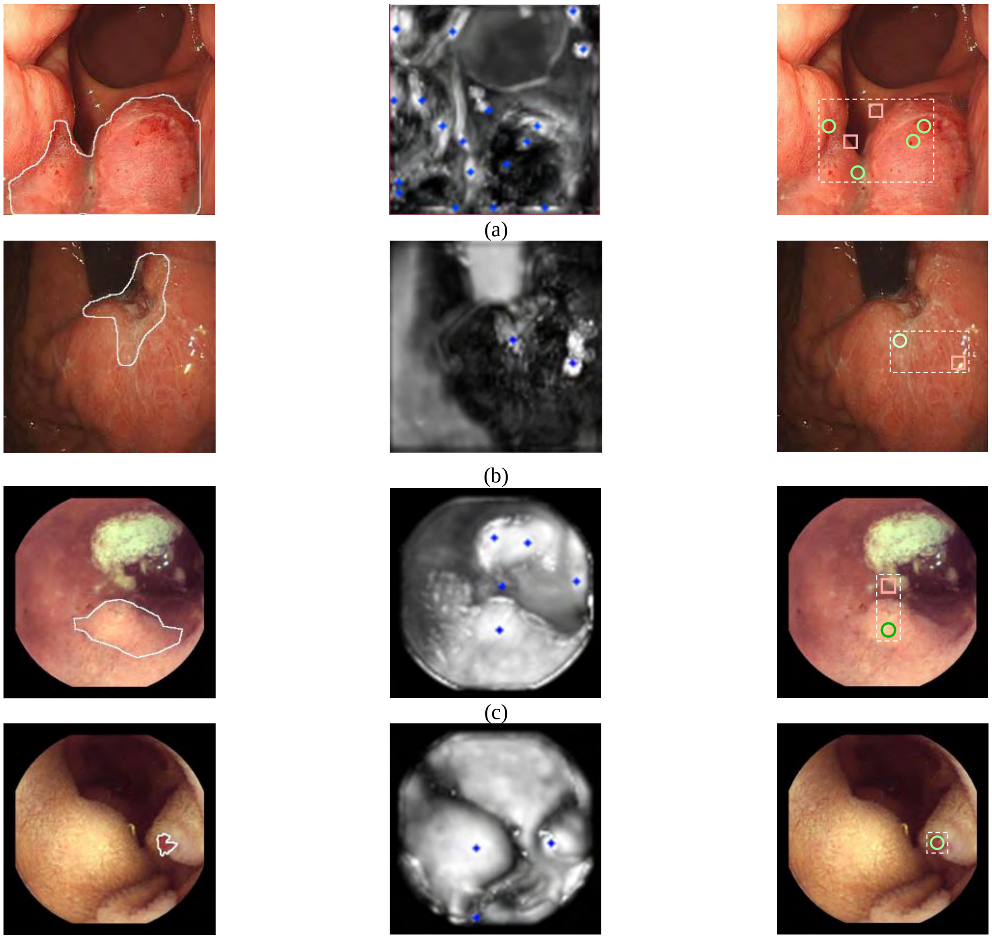


Figure 5.16: Abnormality localization performance of the proposed method using PCFM vs. CIE-Lab features on MICCAI and KID datasets in terms of AUC, per target number of clusters *k*.

Apart from these overall results, it is important to investigate the localization performance achieved at an image level. To this end the results per image were analyzed. This analysis showed that the average number of TP output points per image (i.e., points characterized as abnormal by the system that fall within the ground truth regions of the anomalies) in MICCAI dataset was 1.78, ranging from 1 to a maximum of 7 points. The respective number of FPs (i.e., points that fall outside the ground truth regions of the anomalies but are characterized as abnormal by the system) was 0.74, ranging from 0 to 5. In KID dataset the average number of TPs was 1.14, ranging from 1 to 6. The average number of FPs was 0.69, ranging from 0 to 5 per image. Representative examples of results produced by the proposed ICU are illustrated in the last column of Figure 5.17. The output of ICU is a subset of the salient points detected by DSD, classified as abnormal. The dashed frame indicates the bounds of the region where these suspicious points are located. Points classified as normal are rejected. For example the result of ICU in **Figure 5.17** (a) includes only two FP points (indicated with red squares), and four TP points. The FPs can be attributed to the lower illumination present in these regions. Each of the Figures 5.17 (b) and (c) has only one FP, and one TP. The FP of **Figure 5.17** (b) corresponds to a reflection, and the FP of **Figure 5.17** (c) corresponds to a point of under-illuminated debris. **Figure 5.17** (d) includes a TP and it does not include any FP.



(a)

(b)

(c)

(d)

Figure 5.17: Lesion localization results on four abnormal images. The ground truth lesion areas are outlined on the original images presented in the left column. The maximal images are illustrated (scaled up) in the middle column, along with the respective salient points detected by DSD. The localized lesions after the application of ICU in Phase III are presented in the right column. All points indicated with red square and green circular marks comprise the output of ICU. The green circular marks indicate the TPs and the red square marks indicate the FPs. (a-b) Images from MICCAI. (c-d) images from KID.

A summary of the localization results after the application of ICU using PCFM features, over all images of the available datasets is provided in Table 5.4. The output of this algorithm is a set of points (a subset of those detected by DSD) that are possibly abnormal (positive). This table lists the percentages of the images for which ICU produced 0, 1, 2,..., 7 TP and FP points (the maximum number of points per image was 7). For example, the percentage of images with 0 detected FP points per image (i.e., without any FP) was 54.3% in Dataset 1 and 52.9% in Dataset 2; the percentage of images with one TP point, was 59.9% in Dataset 1 and 92.9% in Dataset 2; and the percentage of images which had one FP point, was 28.0% in Dataset 1 and 31.0% in

Dataset 2. It is notable that TPs were identified in all the abnormal images (the percentage of images with 0 detected TP points is 0.0%), and that the number of TPs or FPs per image did not exceed 7 in any case.

Table 5.4: Abnormality Localization Results over All Images of MICCAI and KID datasets Using ICU With PCFM Features

Detected Points per Image	MICCAI		KID	
	TP (%)	FP (%)	TP (%)	FP (%)
0	0.0	**54.3**	0.0	**52.9**
1	**59.9**	28.0	**92.9**	31.0
2	22.4	11.2	5.3	9.6
3	11.2	3.7	0.6	3.9
4	4.7	1.9	0.0	2.2
5	0.9	0.9	0.6	0.4
6	0.0	0.0	0.6	0.0
7	0.9	0.0	0.0	0.0

The performance of ICU was compared with the performance of a related state-of-the-art approach, which is based on the creation of energy maps (Bernal et al. 2017). According to that approach, the salient points detected by DSD are considered as ’fixations’ or votes, and energy maps are created from this set of discrete fixations/votes. These fixation points are interpolated by a Gaussian function to build up the final energy map, from which the location of the global maximum of the saliency map is selected as the final output. After several experiments using Gaussian functions with different standard deviation values (σ = 16, 32, 64), the best result, considering as a priority not to miss any anomalies, was obtained for σ = 32. The respective percentages of images with TP and FP points are summarized in Table 5.5. It can be observed that the application of the energy-maps approach results in a significantly lower number of points per image; however, there are several images without any TP points detected (37.0% in Dataset 1 and 77.3% in Dataset 2). Thus, ICU is preferable.

Table 5.5: Abnormality Localization Results over All Images of MICCAI and KID datasets using Energy Maps

Detected Points per Image	MICCAI		KID	
	TP (%)	FP (%)	TP (%)	FP (%)
0	37.0	63.0	77.3	22.7
1	63.0	35.1	21.4	57.2
2	0.0	1.9	1.3	20.1
3	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0
5	0.0	0.0	0.0	0.0
6	0.0	0.0	0.0	0.0
7	0.0	0.0	0.0	0.0

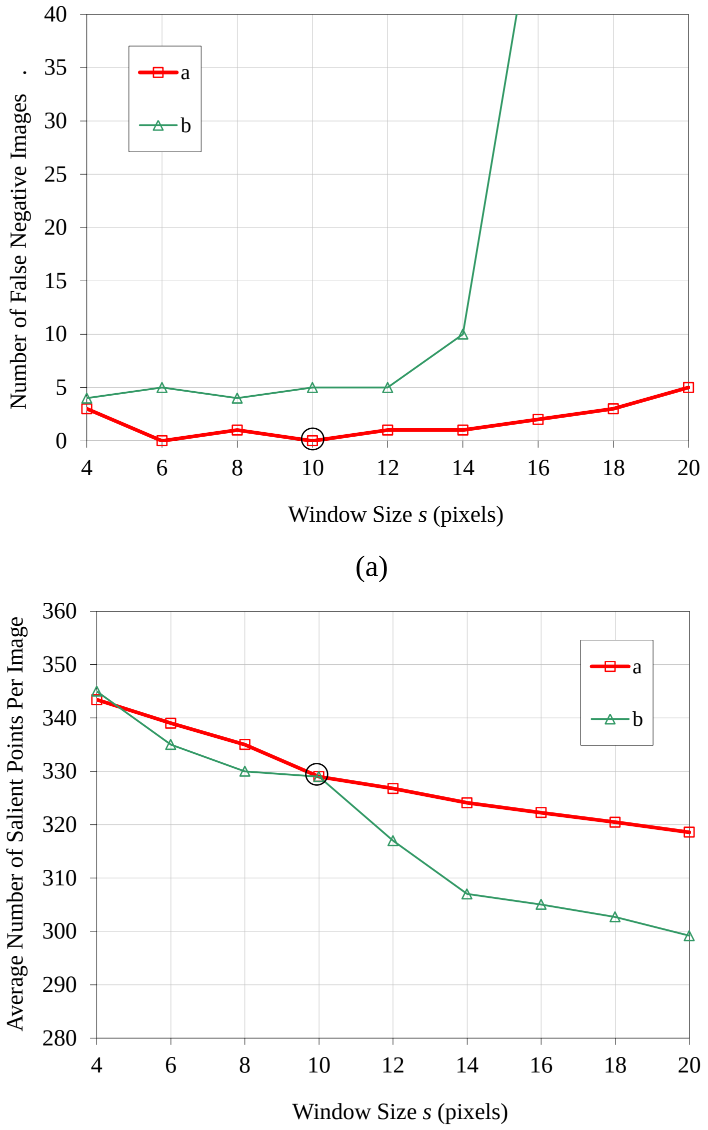
Aiming to a further reduction of the FPs produced by ICU, the experiments were repeated with the energy maps used as a post-processing step that could possibly refine its output. However, although the FPs were reduced, the reduction of the TPs was unacceptable, as the percentage of images without any TP reached 43.7% in Dataset 1 and 77.2% in Dataset 2.

5.6.6 DINOSARC: Salient point detection

Experiments on the WCE images were performed to evaluate the proposed DINOSARC feature extraction methodology in comparison to the state-of-the-art using the publicly available data described in section 5.5.

Prior to the application of DINOSARC algorithm a series of experiments were performed to determine its optimal parameters. The criterion considered for this tuning process was the number of false negative images, i.e., the number of images that were actually containing abnormalities, but no salient points were detected on these abnormalities. Since the salient point detection process is considered as the first step in the analysis of the WCE images, it is important to be able to detect points on abnormalities, in as many as possible (ideally in all) abnormal images.

To this end, the salient point detection performance of DINOSARC algorithm was investigated using various window sizes s×s between 4×4 and 20×20 pixels in each component of CIE-*Lab* color space. The results are illustrated in Figure 5.18. By using window sizes of 6×6 and 10×10 pixels in component a, at least one salient point was detected within the abnormalities. Among these choices the 10×10 pixel window is considered preferable because it results in less salient points per image (Figure 5.18b).



(a)

(b)

Figure 5.18: Salient point detection results for different (square) window sizes *s*×*s*. (a) Average number of salient points detected per image. (b) Number of images in which salient points have not been detected within abnormal regions (false negative).

The DINOSARC salient point detector was compared with the standard SIFT (Lowe 2004)(SIFT-*L*) and SURF (Bay et al. 2008) (SURF-*L*) algorithms, as they are typically applied on the luminance component (*L*) of images. Also, it was compared with the SURF-*a* color salient point detection method proposed in(Iakovidis & Koulaouzidis 2014b), where SURF was applied on component *a* of CIE-*Lab* color space. For completeness, SIFT was also tested on that color component (SIFT-*a*). The evaluation criterion for every detector was the minimum number of salient points needed in order to have the zero false negative images **(Figure 5.19**).

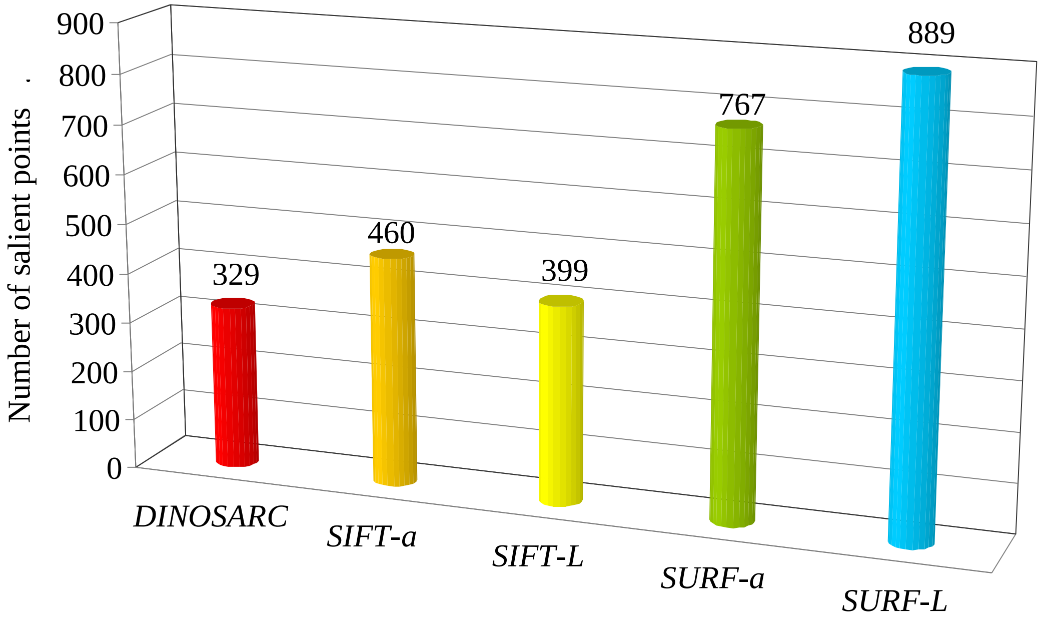


Figure 5.19: Number of salient points detected within abnormal regions over abnormal images using different methods.

Further reduction of the DINOSARC salient points is achieved by the salient region detection process (5.5), which results in only a single salient point per salient region.

In the evaluation of DINOSARC detection algorithm we also computed the percentage of the salient points falling on the abnormal regions of the images. The percentages of these, true positive points, over the total number of detected points in the image are presented in Table 5.6. It can be noticed that the proposed salient point detection algorithm results in more true positive points in every abnormal image than the other algorithms.

Table 5.6: True positive points for each image

Algorithm	**Salient points (%)**
DINOSARC	20
SIFT-*L*	14
SURF-*L*	11
SIFT-*a*	15
SURF-*a*	12

5.6.7 DINOSARC: Salient region discrimination using local descriptors

For the discrimination of abnormal from normal salient regions classification experiments were performed using various local image descriptors. As a baseline for the comparison of DINOSARC descriptor the hue histogram of the area around each salient point is considered. Since this descriptor is not associated with a particular salient point detection algorithm, the DINOSARC salient point detector was used. The hue histogram was quantized into 15 bins, which was the best performing one among histograms of $15 \cdot i$, i=1,..,24 bins. Also, for comparison purposes we selected three state-of-the-art methodologies. The methodology of (Yuan et al. 2016b), which is very recent, the methodology of Li et al. (Li & Meng 2012), and the methodology of (Iakovidis & Koulaouzidis 2014b), which is a predecessor of the proposed approach. The experiments were performed using the 10-fold cross validation evaluation scheme, and a Support Vector Machine (SVM) with Radial Basis Function (RBF) kernel, as a standard classifier. The classification performance was thoroughly investigated using Receiver Operating Characteristic (ROC) analysis and the Area Under the ROC (AUC) was estimated to be able to compare the classification performances using a single measure. The classification accuracy, the sensitivity and the specificity are provided as additional measures facilitating comparisons.

The results are presented in Table 5.7. The best AUC was obtained with the proposed methodology.

Table 5.7: Classification results of salient regions using local image descriptors.

	Hue Histogram	(Iakovidis & Koulaouzidis 2014b)	(Yuan et al. 2016b)	(Li & Meng 2012)	**DINOSARC**
AUC	0.584	0.774	0.606	0.718	**0.813**
Accuracy	0.671	0.772	**0.874**	0.698	0.809
Sensitivity	0.833	0.699	0.142	0.432	0.680
Specificity	0.232	0.782	**0.974**	0.829	0.814

The respective ROCs are illustrated in **Figure 5.20**. It can be noticed that methodology of (Yuan et al. 2016b) provides a higher accuracy; however, this is due to the high specificity, whereas the sensitivity is very low, i.e., its capability to detect positive image regions is low. Also low was the performance of the hue histogram descriptor. The second best performance was obtained by the method of (Iakovidis & Koulaouzidis 2014b).

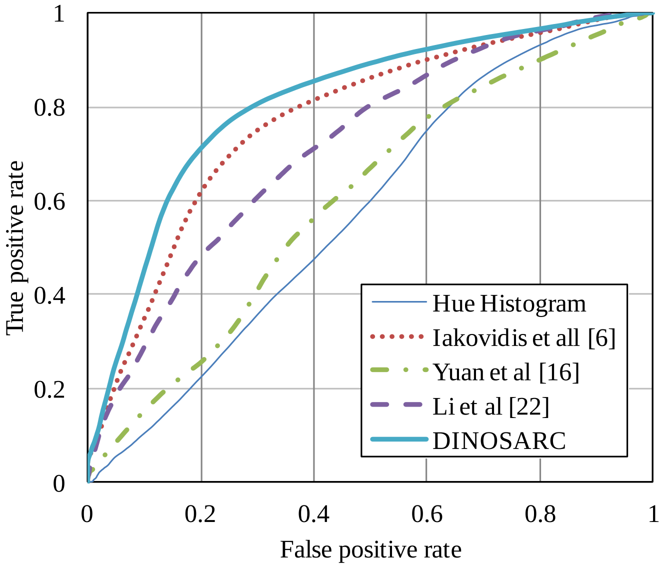


Figure 5.20: The ROCs corresponding to the AUCs reported in Table 5.7 for the classification of salient regions using local image descriptors.

An interpretation of these results can be based on the physical meaning of the respective descriptors. The descriptors proposed by (Yuan et al. 2016b) and (Li & Meng 2012) encode the texture of an image area (both SIFT/CLBP and ULBP/wavelet transform are textural descriptors), and hue histograms encode its colors as they are perceived by humans (Wyszecki & Stiles 1982). The best performing approaches are also based on color; however, the regional minima and maxima of the opponent color components tend to provide more discriminative information about the abnormalities. The approach of (Yuan et al. 2016b)was originally proposed for the detection of polyps, and the approach of (Li & Meng 2012) was proposed for the detection of tumors, including adenomas and adenocarcinomas. Texture has been a discriminative feature of polyps and tumors in several studies with flexible endoscopy images (Karkanis et al. 2003; Liedlgruber & Uhl 2011). However, the significantly lower resolution of WCE images limits the visibility of texture, and consequently, texture becomes less discriminative. More importantly, the database used in our experiments contains polyps but also several other kinds of abnormalities, for which texture may not be as discriminative as color, e.g., vascular lesions.

5.6.8 DINOSARC: Abnormal image detection using global descriptors

Experiments were performed for the investigation of the classification performance of entire WCE images using the DINOSARC features. For image representation global features were extracted using the BoVW model. The BoVW model was constructed with a range of visual vocabulary sizes in the range from 500 to 700 words. The experiments were performed using the 10-fold cross validation evaluation scheme, and an RBF-SVM classifier. Table 5.8 summarizes the results obtained. The proposed DINOSARC features achieved better results from the other methods.

Table 5.8: Classification results of WCE images using global image descriptors.

	Hue Histogram	(Iakovidis & Koulaouzidis 2014b)	(Yuan et al. 2016b)	(Li & Meng 2012)	**DINOSARC**
AUC	0.684	0.774	0.701	0.754	**0.815**
Accuracy	0.730	0.786	0.746	0.751	**0.818**
Sensitivity	0.391	0.496	0.406	0.358	**0.512**
Specificity	0.871	0.890	0.884	0.870	**0.908**

The respective ROCs are illustrated in **Figure 5.21**. Considering the AUCs, the method of (Iakovidis & Koulaouzidis 2014b) was ranked second, the method of (Li & Meng 2012) was ranked third, the method of (Yuan et al. 2016b) was ranked fourth, and the lowest classification performance was obtained by the hue histograms.

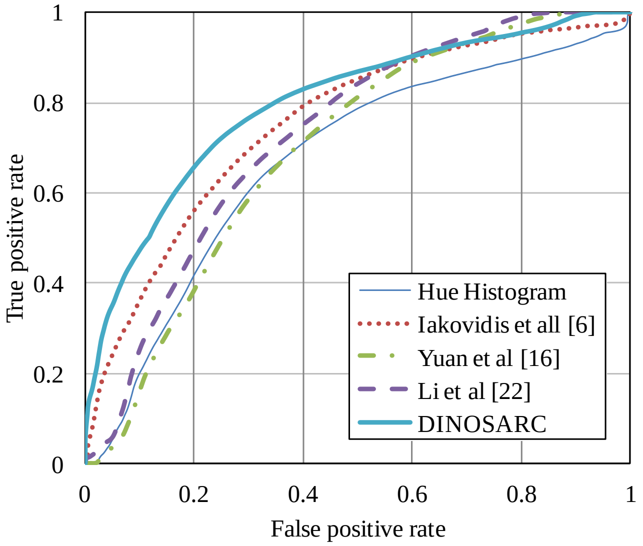


Figure 5.21: The ROCs corresponding to the AUCs reported in Table 5.8 for the classification of WCE images using global image descriptors.

5.7 Conclusions

In this chapter different approaches not only for the detection of video frames with possible existence of lesion, but also approaches for the localization of lesion in video frames were proposed and experimentally evaluated. Firstly, supervised classification scheme using weakly labeled images for automated lesion detection in WCE video frames was proposed following the BoW methodology and creating a visual dictionary encoding all extracted image features into visual words, and created BoW image descriptions, which were used to train SVM classifiers. This supervised approach was primarily investigated for inflammatory lesions (Vasilakakis et al. 2016). In the next step, the research for the contribution of supervised methodology using weakly labeled images was further investigated in the larger and more diverse datasets of MICCAI and KID (D. K. Iakovidis et al. 2018). The conclusions of this approach based on the results obtained can be summarized as follows:

- The results indicate that standard SURF features do not provide a reliable descriptor in the given problem
- The CIE-*Lab* color space is able to boost the performace of lesion detection and provide valuable results within the proposed supervised scheme using weakly labeled images
- The performance of supervised methodology using weakly labeled images was comparable to other state-of-the-art methodologies.

After the use of supervised methodology using weakly labeled images that could be used as an alternative to the demanding in terms of manual annotation effort, fully-supervised, methods, the motivation for the next step was the evaluation of supervised methods using weakly labeled images for the purpose of semantic interpretation of the whole content of GI tract. Multi-label classification methods for the purpose of semantic interpretation were investigated and the results validate that the effect of using multiple labels can enhance abnormality detection. This includes removal of the uninformative video frames, but also to avoid the removal of frames video frames with the intestinal content there is a chance to also miss frames with abnormalities that are present with the intestinal content. The results obtained show that by expressing the problem of abnormality detection as a multi-label classification problem can be beneficial. The research contributions in this direction include:

- DoM has been proposed as an alternative to the conventional salient point detection algorithms
- The results obtained by both of these methods are better than those obtained by state-of-the-art methods.

The localization of GI anomalies has also been addressed with unsupervised image segmentation methods. Such methods provide information about both the location and the area covered by an abnormality; therefore, they are also suitable for size measurement of GI anomalies. However, they are applicable only on images for which anomalies are present; otherwise by default they result in FP regions.

Considering the practical significance of this application the field has rapidly grown, with BoW and CNN-based approaches to play a protagonistic role. ICU that refines the result of DSD by inferring the most suspicious of the salient points, using solely image-level information. This algorithm is based on clustering; however, unlike conventional approaches, it does not use clustering for image segmentation, and it does not exploit any pixel-level annotation. The output of this algorithm is a very small set of points that can attract the attention of a GIE video reviewer, so as to thoroughly examine the respective image locations. Important outcomes that can be derived from this study about the proposed methodology include:

- The automatically extracted features from a CNN can provide significant better information about the characteristics of lesion than color features
- The ICU algorithm is able to localize in both an effective and efficient way the GI anomalies.
- The proposed methodology was challenged to detect and localize anomalies in MICCAI and KID dataset

In the last part of this chapter, DINOSARC(Vasilakakis, Iakovidis, et al. 2018), a color feature extraction methodology for WCE image analysis was presented. The proposed methodology aimed to the discrimination of various abnormal tissues from normal image contents. Major contributions of this study include:

- A novel salient point detection method, which considers saliency with respect to color differences observed in abnormality regions.
- A novel definition of regional saliency based on superpixel segmentation that extends the approach we previously proposed for bleeding detection (Iakovidis et al. 2015) . The extension relies on the fact that region-level saliency is defined based on DINOSARC salient points, and that point-level saliency is preserved to enable the localization of smaller abnormalities.
- A novel descriptor, which extends the descriptor we proposed in (Iakovidis & Koulaouzidis 2014b) by applying the calculations on an arbitrarily-shaped local region defined by a salient superpixel.
- The proposed methodology was applied for both supervised detection of abnormalities in a rich publicly available dataset. The supervised approach was based on the proposed local descriptors, and the supervised approach using weakly labeled images was based on global image descriptors derived from the local ones by application of the BoWV model.

The results showed that the proposed methodology can be more efficient and more effective than relevant state-of-the-art methods for the detection of abnormal images. More, specifically:

- The proposed salient point detection approach results in a smaller number of salient points, which are more likely to fall within regions of abnormality than other current approaches.
- The proposed local image descriptors result in better discrimination of the abnormalities from the normal image contents.
- The global image descriptors enable more accurate detection of the abnormal images in the WCE dataset.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 Summary of individual chapters

This doctoral research thesis investigated and developed computational signal and image analysis methods. The proposed methods were applied in biomedical applications. As the research of this thesis was within the scope of the project “Klearchos Koulaouzidis”, a lot of effort was invested in the investigation of methodologies for endoscopic image analysis, especially for images captured by Wireless Capsule Endoscopes. This concluding chapter summarizes the most important findings of this dissertation.

Chapter 2 provided the theoretical knowledge that was the basic background for the understanding of the methods that were described in the rest of this thesis’ chapters.

Chapter 3 introduced a novel constructive fuzzy representation model for signal classification. The proposed model called Fuzzy Phrases was inspired by the bag-of-words (BoW) feature extraction, which followed an intuitive approach of describing data, using histograms of data granules, referred to as words. Several experiments were conducted and presented. The performance of Fuzzy Phrases was comparable or better from other state-of-the-art fuzzy classification systems.

Chapter 4 provided a detailed review about the technological advantages in the challenging domain of Gastrointestinal Endoscopy. More specifically, this chapter was focused on Wireless Capsule Endoscopes and divided the literature of this domain in four research fields: *Capsule endoscopes*, *Capsule endoscopes localization*, *Image Enhancement and Abnormality detection software.*

Capsule endoscopes: This field was further divided in the current commercial capsule endoscopes and the research capsule endoscopes that were still developing. At the time of this thesis, the technological advancements of commercial capsule were only in the examination of the patients, without further capabilities for drug release or biopsy. A list with the available commercial capsule endoscopes and their specification were presented in **Table 4.1**. In order to extend the capabilities of commercial capsules and overcome their drawbacks, a lot of research had been made resulting in new capsule designs that were presented in **Table 4.2.**

Capsule endoscopes localization is a very important field in Capsule endoscopy. The detection of accurate location of the capsule identifying the location of the exact position of possible abnormalities detected, and can therefore guide further management such as surgery or local drug delivery. **Table 4.3** summarized the state-of-the-art methods for capsule endoscope localization. These methods were dived in three fields

- Computer Vision algorithms can assist the localization process of the capsule endoscope by exploiting the visual content of the raw CE video frames
- Magnetic localization methods exploit a permanent magnet inside the capsule
- Radiofrequency based localization techniques include Time-Of-Arrival (TOA), Time-Difference-Of-Arrival (TDOA), direction-of-arrival (DOA), and Received Signal Strength (RSS)

Image Enhancement the visualization of CE video streams was aiming for the increasing of diagnostic yield. **Table 4.4** provided of the later image enhancement used techniques.

Abnormality detection software was an important aspect for Capsule Endoscopy for the development of a computer-based system. In the field of abnormality detection in Capsule Endoscopy, based on the literature the abnormalities fall beneath four basic categories: Blood detection, Polyp and tumor detection, Inflammation detection and Hookworm detection

The most challenging part in the detection was the development of a software capable of detecting multiple abnormalities, as it was presented in the **Table 4.4**, where the specific abnormality detection software were outnumber the multiple detection software.

Chapter 5 investigated and presented novel algorithms for the detection of gastrointestinal abnormalities that could be included in a framework of a MDSS. In ***Chapter 5*** was presented for first time in the classification of Wireless Capsule Endoscopy images the concept of weakly annotated images. In supervised learning with weakly annotated images only the information of the existence or not of the abnormality was needed. The achieved AUC in the KID dataset was 81% and higher than other state-of-the-art methods proving that the weakly supervised learning was appropriate for the detection of gastrointestinal image abnormalities. A step further was the development of a weakly multi label learning classification scheme, in order to interpret the whole semantic content of the gastrointestinal tract. The experimental results presented further improvement as the AUC was 90% on KID dataset. Also, this chapter presented the DINOSARC algorithm, which was an unsupervised algorithm for the detection of interest point in Wireless Capsule Endoscopy images.

6.2 Future Work

The work presented is only in the beginning towards the challenges identified. Considering the continuous increase of signal data, the proposed methods can be further evolved and extended with respect to both their efficiency and effectiveness, e.g., for high throughput applications. Although the main focus of this thesis was on biomedical applications, extensions and generalizations are plausible.

The proposed Fuzzy Phrases framework was the first attempt to transform the input signal data exploiting fuzzy logic into multiple different words to describe and make decisions about real-world concepts. Future work and extensions of this framework include:

- Systematic evaluation of its robustness to noise and missing values
- The derivation of a rule extraction scheme from data
- The interpretability of black-box approaches, such as neural networks, by incorporating fuzzy phrases, e.g., into a CNN framework.

In the field of WCE the software demands are changing in dependence to the available capsule endoscopes. As the hardware technology of capsule endoscopes is advancing, more challenges for development of the appropriate software will arise. Thus, the proposed methodologies of this thesis for gastrointestinal abnormalities detection can be encountered as a basis for future application development. Future extensions of the proposed abnormality detection software of this thesis include:

- The experimental investigation in larger and more diverse datasets
- The evolution of the current salient point detector to be able to detect less salient points with higher accuracy in the lesion areas
- The localization of the abnormal region of the images and the identification of the size of the abnormal region.

APPENDIX

A List of publications

A.1: List of publications in Journals (5)

Vasilakakis, M., Koulaouzidis, A., Marlicz, W., & Iakovidis, D. (2019). The future of capsule endoscopy in clinical practice: from diagnostic to therapeutic experimental prototype capsules. *Gastroenterology Review*, *14*(1).

Vasilakakis, M., Koulaouzidis, A., Yung, D. E., Plevris, J. N., Toth, E., & Iakovidis, D. K. (2019). Follow-up on: Optimizing lesion detection in small-bowel capsule endoscopy and beyond: from present problems to future solutions. *Expert Review of Gastroenterology & Hepatology*.

Iakovidis, D. K., Georgakopoulos, S. V., Vasilakakis, M., Koulaouzidis, A., & Plagianakos, V. P. (2018). Detecting and Locating Gastrointestinal Anomalies Using Deep Learning and Iterative Cluster Unification. *IEEE Transactions on Medical Imaging*.

Vasilakakis, M. D., Iakovidis, D. K., Spyrou, E., & Koulaouzidis, A. (2018). DINOSARC: Color features based on selective aggregation of chromatic image components for wireless capsule endoscopy. *Computational and mathematical methods in medicine*, *2018*.

Vasilakakis, M. D., Diamantis, D., Spyrou, E., Koulaouzidis, A., & Iakovidis, D. K. (2018). Weakly supervised multilabel classification for semantic interpretation of endoscopy video frames. *Evolving Systems*, 1–13.

A.2: List of publications in International Conferences(5)

Vasilakakis, M., & Iakovidis, D. (2020). Constructive Fuzzy Representation Model for Data Classification. *2020 IEEE 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE(*Submitted*).

Vasilakakis, M., Iosifidou, V., Fragkaki, P., & Iakovidis, D. (2019). Bone Fracture Identification in X-Ray Images using Fuzzy Wavelet Features. *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)* (pp. 726–730). IEEE.

Vasilakakis, M. D., Iakovidis, D. K., Spyrou, E., Chatzis, D., & Koulaouzidis, A. (2017). *Beyond lesion detection: Towards semantic interpretation of endoscopy videos*. Communications in Computer and Information Science (Vol. 744, pp. 379–390).

Vasilakakis, M., Iakovidis, D. K., Spyrou, E., & Koulaouzidis, A. (2016). Weakly-supervised lesion detection in video capsule endoscopy based on a bag-of-colour features model. *Medical Image Computing and Computer Assisted Intervention*

(MICCAI) ; International workshop on computer-assisted and robotic endoscopy (pp. 96–103). Springer.

Georgakopoulos, S. V., Iakovidis, D. K., Vasilakakis, M., Plagianakos, V. P., & Koulaouzidis, A. (2016). Weakly-supervised Convolutional learning for detection of inflammatory gastrointestinal lesions. *IST 2016 - 2016 IEEE International Conference on Imaging Systems and Techniques, Proceedings* (pp. 510–514).

REFERENCES

Abadi, M. et al., 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

Achanta, R. et al., 2012. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11), pp.2274–2282.

Alcalá-Fdez, J. et al., 2011. Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic & Soft Computing*, 17.

Alemayehu, D. & Zou, K.H., 2012. Applications of ROC analysis in medical research: recent developments and future directions. *Academic radiology*, 19(12), pp.1457–1464.

Alizadeh, M. et al., 2017. Detection of small bowel tumor in wireless capsule endoscopy images using an adaptive neuro-fuzzy inference system. *Journal of Biomedical Research*, 31(5), pp.419–427.

Allen, R.L. & Mills, D., 2004. *Signal analysis: time, frequency, scale, and structure*, John Wiley & Sons.

Alotaiby, T.N. et al., 2014. EEG seizure detection and prediction algorithms: a survey. *EURASIP Journal on Advances in Signal Processing*, 2014(1), p.183.

Andrade, L.C., Oleskovicz, M. & Fernandes, R.A., 2016. Adaptive threshold based on wavelet transform applied to the segmentation of single and combined power quality disturbances. *Applied Soft Computing*, 38, pp.967–977.

Anthimopoulos, M. et al., 2016. Lung pattern classification for interstitial lung diseases using a deep convolutional neural network. *IEEE transactions on medical imaging*, 35(5), pp.1207–1216.

Athavale, Y. & Krishnan, S., 2017. Biosignal monitoring using wearables: Observations and opportunities. *Biomedical Signal Processing and Control*, 38, pp.22–33.

Al-Ayyoub, M., Hmeidi, I. & Rababah, H., 2013. Detecting Hand Bone Fractures in X-Ray Images. *JMPT*, 4(3), pp.155–168.

Bao, G., Pahlavan, K. & Mi, L., 2015. Hybrid Localization of Microrobotic Endoscopic Capsule Inside Small Intestine by Data Fusion of Vision and RF Sensors. *IEEE Sensors Journal*, 15(5), pp.2669–2678.

Bashar, M.K. et al., 2010. Automatic detection of informative frames from wireless capsule endoscopy images. *Medical Image Analysis*, 14(3), pp.449–470.

Bay, H. et al., 2008. Speeded-up robust features (SURF). *Computer vision and image understanding*, 110(3), pp.346–359.

Beccani, M. et al., 2015. A magnetic drug delivery capsule based on a coil actuation mechanism. *Procedia engineering*, 120, pp.53–56.

Berkaya, S.K. et al., 2018. A survey on ECG analysis. *Biomedical Signal Processing and Control*, 43, pp.216–235.

Bernal, J. et al., 2017. Comparative validation of polyp detection methods in video colonoscopy: results from the MICCAI 2015 endoscopic vision challenge. *IEEE transactions on medical imaging*, 36(6), pp.1231–1249.

Bernal, J. et al., 2015. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics*, 43, pp.99–111.

Blake, C.L. & Merz, C.J., 1998. UCI repository of machine learning databases, 1998.

Blaschko, M., Vedaldi, A. & Zisserman, A., 2010. Simultaneous object detection and ranking with weak supervision. In *Advances in neural information processing systems*. pp. 235–243.

Boostani, R., Karimzadeh, F. & Nami, M., 2017. A comparative review on sleep stage classification methods in patients and healthy individuals. *Computer methods and programs in biomedicine*, 140, pp.77–91.

Cai, Y. et al., 2019. Axiomatic fuzzy set theory-based fuzzy oblique decision tree with dynamic mining fuzzy rules. *Neural Computing and Applications*, pp.1–16.

Carneiro, G. et al., 2017. Automatic quantification of tumour hypoxia from multi-modal microscopy images using weakly-supervised learning methods. *IEEE transactions on medical imaging*, 36(7), pp.1405–1417.

Chan, T.F. & Vese, L.A., 2001. Active contours without edges. *IEEE Transactions on image processing*, 10(2), pp.266–277.

Chandran, S. et al., 2013. Risk stratification of upper GI bleeding with an esophageal capsule. *Gastrointestinal endoscopy*, 77(6), pp.891–898.

Chang, C.-C. & Lin, C.-J., 2011. LIBSVM: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3), p.27.

Chen, J., Wang, Y. & Zou, Y.X., 2015. An adaptive redundant image elimination for Wireless Capsule Endoscopy review based on temporal correlation and color-texture feature similarity. In *International Conference on Digital Signal Processing, DSP*. pp. 735–739.

Chen, J., Zou, Y. & Wang, Y., 2017. Wireless capsule endoscopy video summarization: A learning approach based on Siamese neural network and support vector machine. In *Proceedings - International Conference on Pattern Recognition*. pp. 1303–1308.

Chen, T. et al., 2016. Fuzzy rule weight modification with particle swarm optimisation. *Soft Computing*, 20(8), pp.2923–2937.

Chen, W.-W. et al., 2014. Design of micro biopsy device for wireless autonomous endoscope. *International journal of precision engineering and manufacturing*, 15(11), pp.2317–2325.

Chollet, F. & others, 2015. Keras.

Chou, C.-H., Hsieh, S.-C. & Qiu, C.-J., 2017. Hybrid genetic algorithm and fuzzy clustering for bankruptcy prediction. *Applied Soft Computing*, 56, pp.298–316.

clinicaltrials.gov, 2019. Available at: https://clinicaltrials.gov/ct2/results?cond=Capsule+endoscopy&term=Capsule+endoscopy&cntry=&state=&city=&dist=.

Cortes, C. & Vapnik, V., 1995. Support-vector networks. *Machine learning*, 20(3), pp.273–297.

Csurka, G. et al., 2004. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*. Prague, pp. 1–2.

Deeba, F. et al., 2018. Performance assessment of a bleeding detection algorithm for endoscopic video based on classifier fusion method and exhaustive feature selection. *Biomedical Signal Processing and Control*, 40, pp.415–424.

Demosthenous, P., Pitris, C. & Georgiou, J., 2016. Infrared Fluorescence-Based Cancer Screening Capsule for the Small Intestine. *IEEE Transactions on Biomedical Circuits and Systems*, 10(2), pp.467–476.

Deng, J. et al., 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, pp. 248–255.

Dimas, G., Iakovidis, D.K., Karargyris, A., et al., 2017. An artificial neural network architecture for non-parametric visual odometry in wireless capsule endoscopy. *Measurement Science and Technology*, 28(9), p.094005.

Dimas, G., Spyrou, E., et al., 2017. Intelligent visual localization of wireless capsule endoscopes enhanced by color information. *Computers in biology and medicine*, 89, pp.429–440.

Dimas, G., Iakovidis, D.K., Ciuti, G., et al., 2017. Visual Localization of Wireless Capsule Endoscopes Aided by Artificial Neural Networks. In *Computer-Based*

Medical Systems (CBMS), 2017 IEEE 30th International Symposium on. IEEE, pp. 734–738.

Domingo, M. et al., 2012. Evaluation of a telemedicine system for heart failure patients: feasibility, acceptance rate, satisfaction and changes in patient behavior: results from the CARME (CAtalan Remote Management Evaluation) study. *European Journal of Cardiovascular Nursing*, 11(4), pp.410–418.

Donnelley, M. & Knowles, G., 2005. Computer aided long bone fracture detection. In *Proceedings of the Eighth International Symposium on Signal Processing and Its Applications, 2005*. IEEE, pp. 175–178.

Drake, J. & Hamerly, G., 2012. Accelerated k-means with adaptive distance bounds. In *5th NIPS workshop on optimization for machine learning*.

Duan, L. et al., 2018. Time-series clustering based on linear fuzzy information granules. *Applied Soft Computing*, 73, pp.1053–1067.

Dubois, D.J., 1980. *Fuzzy sets and systems: theory and applications*, Academic press.

Dunn, S. et al., 2014. PTU-053 Is It Worth Repeating Previous Unremarkable Sb2 Capsules With The New Sb3? *Gut*, 63(Suppl 1), pp.A61–A62.

El Gamal, A. & Eltoukhy, H., 2005. CMOS image sensors. *IEEE Circuits and Devices Magazine*, 21(3), pp.6–20.

Elisseeff, A. & Weston, J., 2002. A kernel method for multi-labelled classification. In *Advances in neural information processing systems*. pp. 681–687.

Enns, R.A. et al., 2017. Clinical practice guidelines for the use of video capsule endoscopy. *Gastroenterology*, 152(3), pp.497–514.

Fawcett, T., 2006. An introduction to ROC analysis. *Pattern recognition letters*, 27(8), pp.861–874.

Fisher, L.R. & Hasler, W.L., 2012. New vision in video capsule endoscopy: current status and future directions. *Nature Reviews Gastroenterology and Hepatology*, 9(7), pp.392–405.

Fontana, R. et al., 2017. An Innovative Wireless Endoscopic Capsule with Spherical Shape. *IEEE Transactions on Biomedical Circuits and Systems*, 11(1), pp.143–152.

Fu, C. et al., 2019. Fuzzy granular classification based on the principle of justifiable granularity. *Knowledge-Based Systems*, 170, pp.89–101.

Fu, Q., Guo, S. & Guo, J., 2017. Conceptual design of a novel magnetically actuated hybrid microrobot. In *Mechatronics and Automation (ICMA), 2017 IEEE International Conference on*. IEEE, pp. 1001–1005.

Fu, Y. et al., 2014. Computer-aided bleeding detection in WCE video. *IEEE Journal of Biomedical and Health Informatics*, 18(2), pp.636–642.

Funt, B.V. & Finlayson, G.D., 1995. Color Constant Color Indexing. *IEEE Trans. Pattern Anal. Mach. Intell.*, 17, pp.522–529.

Fürnkranz, J. et al., 2008. Multilabel classification via calibrated label ranking. *Machine learning*, 73(2), pp.133–153.

Ganz, M., Yang, X. & Slabaugh, G., 2012. Automatic segmentation of polyps in colonoscopic narrow-band imaging data. *IEEE Transactions on Biomedical Engineering*, 59(8), pp.2144–2151.

Gao, J. et al., 2016. Design and testing of a motor-based capsule robot powered by wireless power transmission. *IEEE/ASME Transactions on Mechatronics*, 21(2), pp.683–693.

Gao, T. et al., 2019. Conjugate gradient-based Takagi-Sugeno fuzzy neural network parameter identification and its convergence analysis. *Neurocomputing*, 364, pp.168–181.

Geng, Y. & Pahlavan, K., 2016. Design, Implementation, and Fundamental Limits of Image and RF Based Wireless Capsule Endoscopy Hybrid Localization. *IEEE Transactions on Mobile Computing*, 15(8), pp.1951–1964.

Georgakopoulos, S.V. et al., 2016. Weakly-supervised Convolutional learning for detection of inflammatory gastrointestinal lesions. In *IST 2016 - 2016 IEEE International Conference on Imaging Systems and Techniques, Proceedings*. pp. 510–514.

Ghoraani, B. & Krishnan, S., 2011. Time–frequency matrix feature extraction and classification of environmental audio signals. *IEEE transactions on audio, speech, and language processing*, 19(7), pp.2197–2209.

Ghosh, T., Fattah, S.A. & Wahid, K.A., 2018. Automatic Computer Aided Bleeding Detection Scheme for Wireless Capsule Endoscopy (WCE) Video Based on Higher and Lower Order Statistical Features in a Composite Color. *Journal of Medical and Biological Engineering*, 38(3), pp.482–496.

Gong, Y. et al., 2013. Deep convolutional ranking for multilabel image annotation. *arXiv preprint arXiv:1312.4894*.

Gonzalez, R.C. & Woods, R.E., 1992. Digital image processing addison-wesley. *Reading, Ma*, 2.

Gralnek, I. et al., 2013. Capsule endoscopy in acute upper gastrointestinal hemorrhage: a prospective cohort study. *Endoscopy*, 45(01), pp.12–19.

Gralnek, I. et al., 2008. Development of a capsule endoscopy scoring index for small bowel mucosal inflammatory change. *Alimentary pharmacology & therapeutics*, 27(2), pp.146–154.

Grauman, K. & Leibe, B., 2011. Visual object recognition. *Synthesis lectures on artificial intelligence and machine learning*, 5(2), pp.1–181.

Greenspan, H., Van Ginneken, B. & Summers, R.M., 2016. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging*, 35(5), pp.1153–1159.

Van Grinsven, M.J. et al., 2016. Fast convolutional neural network training using selective data sampling: Application to hemorrhage detection in color fundus images. *IEEE transactions on medical imaging*, 35(5), pp.1273–1284.

Gu, X., Chung, F.-L. & Wang, S., 2016. Bayesian Takagi–Sugeno–Kang fuzzy classifier. *IEEE Transactions on Fuzzy Systems*, 25(6), pp.1655–1671.

Gu, Y. et al., 2015. Design of endoscopic capsule with multiple cameras. *IEEE transactions on biomedical circuits and systems*, 9(4), pp.590–602.

Guillaumin, M. et al., 2009. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *2009 IEEE 12th international conference on computer vision*. IEEE, pp. 309–316.

Guo, J. et al., 2017. Development of a novel wireless spiral capsule robot with modular structure. In *Mechatronics and Automation (ICMA), 2017 IEEE International Conference on*. IEEE, pp. 439–444.

Guo, Z., Zhang, L. & Zhang, D., 2010. A completed modeling of local binary pattern operator for texture classification. *IEEE transactions on image processing*, 19(6), pp.1657–1663.

Gutkin, E. et al., 2013. Pillcam ESO® is more accurate than clinical scoring systems in risk stratifying emergency room patients with acute upper gastrointestinal bleeding. *Therapeutic advances in gastroenterology*, 6(3), pp.193–198.

Hale, M.F. et al., 2015. Magnetically steerable gastric capsule endoscopy is equivalent to flexible endoscopy in the detection of markers in an excised porcine stomach model: results of a randomized trial. *Endoscopy*, 47(07), pp.650–653.

Hany, U. & Akter, L., 2017a. Local parametric approach of wireless capsule endoscope localization using randomly scattered path loss based WCL. *Wireless Communications and Mobile Computing*, 2017.

Hany, U. & Akter, L., 2017b. Non-Parametric Approach of Video Capsule Endoscope Localization Using Suboptimal Method of Position Bounded CWCL. *IEEE Sensors Journal*, 17(20), pp.6806–6815.

Hany, U. & Akter, L., 2018a. Non-parametric Approach using ML Estimated Path Loss BoundedWCL for Video Capsule Endoscope Localization. *IEEE Sensors Journal*.

Hany, U. & Akter, L., 2018b. Non-parametric Method of Path Loss Estimation for Endoscopic Capsule Localization. *International Journal of Wireless Information Networks*, 25(1), pp.44–56.

Hany, U., Akter, L. & Hossain, F., 2017. Degree-based WCL for video endoscopic capsule localization. *IEEE Sensors Journal*, 17(9), pp.2904–2916.

Haralick, R.M., Shanmugam, K. & Dinstein, I.H., 1973. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6), pp.610–621.

He, J.-. et al., 2018. Hookworm Detection in Wireless Capsule Endoscopy Images with Deep Learning. *IEEE Transactions on Image Processing*, 27(5), pp.2379–2392.

He, X., Zheng, Z. & Hu, C., 2015. Magnetic localization and orientation of the capsule endoscope based on a random complex algorithm. *Medical Devices: Evidence and Research*, 8, pp.175–184.

Hinton, G., Srivastava, N. & Swersky, K., 2012. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on*, 14, p.8.

Hoai, M. et al., 2014. Learning discriminative localization from weakly labeled data. *Pattern Recognition*, 47(3), pp.1523–1534.

Hosoe, N. et al., 2016. Evaluation of performance of the Omni mode for detecting video capsule endoscopy images: A multicenter randomized controlled trial. *Endoscopy international open*, 4(8), p.E878.

Hsu, W. et al., 2019. Advancing Artificial Intelligence in Sensors, Signals, and Imaging Informatics. *Yearbook of medical informatics*, 28(01), pp.115–117.

http://english.jinshangroup.com/capsuleendoscopy.html, 2018. http://english.jinshangroup.com/capsuleendoscopy.html.

http://www.ankoninc.com.cn, 2018.

http://www.capsovision.com/physicians/product-specifications, 2018. *http://www.capsovision.com/physicians/product-specifications*,

http://www.intromedic.com/eng/main/, 2018. *http://www.intromedic.com/eng/main/*, http://www.shangxianinc.com/en/, 2018.

https://www.olympus-europa.com/medical/en/Products-and-Solutions/Products/Product/ENDOCAPSULE-10-System.html, 2018. https://www.olympus-europa.com/medical/en/Products-and-Solutions/Products/Product/ENDOCAPSULE-10-System.html.

Hu, E. et al., 2016. Bleeding and Tumor Detection for Capsule Endoscopy Images Using Improved Geometric Feature. *Journal of Medical and Biological Engineering*, 36(3), pp.344–356.

Hu, X. et al., 2019. Allocation of Information Granularity: A Multi-Objective Evolutionary Optimization Using Conflict Information. In *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, pp. 1–6.

Hu, X., Pedrycz, W. & Wang, X., 2018. Fuzzy classifiers with information granules in feature space and logic-based computing. *Pattern Recognition*, 80, pp.156–167.

Hwang, S., 2011. Bag-of-visual-words approach based on SURF features to polyp detection in wireless capsule endoscopy videos. In *Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV)*. The Steering Committee of The World Congress in Computer Science, Computer …, p. 1.

Iakovidis, D. et al., 2008. Unsupervised summarisation of capsule endoscopy video. In *Intelligent Systems, 2008. IS'08. 4th International IEEE Conference*. IEEE, pp. 3–15. Available at: http://dx.doi.org/10.1109/IS.2008.4670414.

Iakovidis, D.K. et al., 2015. Blood detection in wireless capsule endoscope images based on salient superpixels. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, pp. 731–734.

Iakovidis, D.K. et al., 2018. Deep Endoscopic Visual Measurements. *IEEE Journal of Biomedical and Health Informatics*.

Iakovidis, D.K. et al., 2018. Detecting and Locating Gastrointestinal Anomalies Using Deep Learning and Iterative Cluster Unification. *IEEE Transactions on Medical Imaging*.

Iakovidis, D.K. et al., 2016. Robotic validation of visual odometry for wireless capsule endoscopy. In *Imaging Systems and Techniques (IST), 2016 IEEE International Conference on*. IEEE, pp. 83–87.

Iakovidis, D.K. & Koulaouzidis, A., 2014a. Automatic lesion detection in capsule endoscopy based on color saliency: closer to an essential adjunct for reviewing software. *Gastrointestinal endoscopy*, 80(5), pp.877–883.

Iakovidis, D.K. & Koulaouzidis, A., 2014b. Automatic lesion detection in wireless capsule endoscopy—a simple solution for a complex problem. In *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE, pp. 2236–2240.

Iakovidis, D.K. & Koulaouzidis, A., 2015. Software for enhanced video capsule endoscopy: challenges for essential progress. *Nature Reviews Gastroenterology and Hepatology*, 12(3), p.172.

Iddan, G. et al., 2000. Wireless capsule endoscopy. *Nature*, 405(6785), pp.417–418.

Islam, M.N. & Fleming, A.J., 2014. A novel and compatible sensing coil for a capsule in wireless capsule endoscopy for real time localization. In *Proceedings of IEEE Sensors*. pp. 1607–1610.

Ben Ismail, M.M. & Bchir, O., 2016. Endoscopy video summarisation using novel relational motion histogram descriptor and semi-supervised clustering. *Journal of Experimental and Theoretical Artificial Intelligence*, 28(4), pp.629–653.

Ito, T., Anzai, D. & Wang, J., 2016. Hybrid toa/rssi-basedwireless capsule endoscope localization with relative permittivity estimation. *IEICE Transactions on Communications*, E99B(11), pp.2442–2449.

Jang, J. et al., 2018. 4-Camera VGA-resolution capsule endoscope with 80Mb/s body-channel communication transceiver and Sub-cm range capsule localization. In *Solid-State Circuits Conference-(ISSCC), 2018 IEEE International*. IEEE, pp. 282–284.

Jia, X. & Meng, M.Q.-., 2017. Gastrointestinal bleeding detection in wireless capsule endoscopy images using handcrafted and CNN features. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. pp. 3154–3157.

Jia, X. & Meng, M.Q.-H., 2016. A deep convolutional neural network for bleeding detection in wireless capsule endoscopy images. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, pp. 639–642.

Jia, X. & Meng, M.Q.-H., 2017. A study on automated segmentation of blood regions in wireless capsule endoscopy images using fully convolutional networks. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. IEEE, pp. 179–182.

Jia, Y. et al., 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, pp. 675–678.

Jia, Y., 2015. Polyps auto-detection in wireless capsule endoscopy images using improved method based on image segmentation. In *2015 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE, pp. 1631–1636.

Kalantarian, H., Sideris, C. & Sarrafzadeh, M., 2016. A hierarchical classification and segmentation scheme for processing sensor data. *IEEE journal of biomedical and health informatics*, 21(3), pp.672–681.

Karargyris, A. & Bourbakis, N., 2011. Detection of small bowel polyps and ulcers in wireless capsule endoscopy videos. *IEEE transactions on BioMedical Engineering*, 58(10), pp.2777–2786.

Karkanis, S.A. et al., 2003. Computer-aided tumor detection in endoscopic video using color wavelet features. *IEEE transactions on information technology in biomedicine*, 7(3), pp.141–152.

Kim, S.H., Yang, D.-H. & Kim, J.S., 2018. Current Status of Interpretation of Small Bowel Capsule Endoscopy. *Clinical endoscopy*, 51(4), p.329.

Kopylov, U. et al., 2017. Diagnostic yield of capsule endoscopy versus magnetic resonance enterography and small bowel contrast ultrasound in the evaluation of small bowel Crohn’s disease: Systematic review and meta-analysis. *Digestive and Liver Disease*, 49(8), pp.854–863.

Kopylov, U. & Seidman, E.G., 2014. Role of capsule endoscopy in inflammatory bowel disease. *World Journal of Gastroenterology: WJG*, 20(5), p.1155.

Koulaouzidis, A. et al., 2012. Diagnostic yield of small-bowel capsule endoscopy in patients with iron-deficiency anemia: a systematic review. *Gastrointestinal endoscopy*, 76(5), pp.983–992.

Koulaouzidis, A., Giannakou, A., et al., 2013. Do prokinetics influence the completion rate in small-bowel capsule endoscopy? A systematic review and meta-analysis. *Current medical research and opinion*, 29(9), pp.1171–1185.

Koulaouzidis, A. et al., 2017. KID Project: an internet-based digital video atlas of capsule endoscopy for research purposes. *Endoscopy international open*, 5(06), pp.E477–E483.

Koulaouzidis, A. et al., 2015. Optimizing lesion detection in small-bowel capsule endoscopy: from present problems to future solutions. *Expert review of gastroenterology & hepatology*, 9(2), pp.217–235.

Koulaouzidis, A., Rondonotti, E. & Karargyris, A., 2013. Small-bowel capsule endoscopy: a ten-point contemporary review. *World journal of gastroenterology: WJG*, 19(24), p.3726.

Koulaouzidis, G., Iakovidis, D. & Clark, A., 2016. Telemonitoring predicts in advance heart failure admissions. *International journal of cardiology*, 216, pp.78–84.

Krizhevsky, A., Hinton, G. & others, 2009. *Learning multiple layers of features from tiny images*, Citeseer.

Krizhevsky, A., Sutskever, I. & Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. pp. 1097–1105.

Le, V.H. et al., 2016. A soft-magnet-based drug-delivery module for active locomotive intestinal capsule endoscopy using an electromagnetic actuation system. *Sensors and Actuators A: Physical*, 243, pp.81–89.

LeCun, Y. et al., 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), pp.2278–2324.

LeCun, Y. et al., 1990. Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems*. pp. 396–404.

Lee, C. et al., 2015. Active locomotive intestinal capsule endoscope (ALICE) system: A prospective feasibility study. *IEEE/ASME Transactions on Mechatronics*, 20(5), pp.2067–2074.

Lee, S.-H., 2015. Feature selection based on the center of gravity of BSWFMs using NEWFM. *Engineering Applications of Artificial Intelligence*, 45, pp.482–487.

Leung, B.H.K. et al., 2017. A Therapeutic Wireless Capsule for Treatment of Gastrointestinal Haemorrhage by Balloon Tamponade Effect. *IEEE Transactions on Biomedical Engineering*, 64(5), pp.1106–1114.

Li, B. & Meng, M.Q.-H., 2012. Tumor recognition in wireless capsule endoscopy images using textural features and SVM-based feature selection. *IEEE Transactions on Information Technology in Biomedicine*, 16(3), pp.323–329.

Li, H. et al., 2016. Multi-level feature representations for video semantic concept detection. *Neurocomputing*, 172, pp.64–70.

Li, Z. et al., 2016. Capsule Design for Blue Light Therapy against Helicobacter pylori. *PloS one*, 11(1), p.e0147531.

Liedlgruber, M. & Uhl, A., 2011. Computer-aided decision support systems for endoscopy in the gastrointestinal tract: a review. *IEEE reviews in biomedical engineering*, 4, pp.73–88.

Liu, G. et al., 2016. Detection of small bowel tumor based on multi-scale curvelet analysis and fractal technology in capsule endoscopy. *Computers in biology and medicine*, 70, pp.131–138.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2), pp.91–110.

Mahoney, A.W. & Abbott, J.J., 2016. Five-degree-of-freedom manipulation of an untethered magnetic device in fluid using a single permanent magnet with application in stomach capsule endoscopy. *International Journal of Robotics Research*, 35(1-3), pp.129–147.

Mallat, S.G., 1989. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (7), pp.674–693.

Mamonov, A.V. et al., 2014. Automated polyp detection in colon capsule endoscopy. *IEEE Transactions on Medical Imaging*, 33(7), pp.1488–1502.

Manivannan, S. & Trucco, E., 2015. Learning discriminative local features from image-level labelled data for colonoscopy image classification. In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*. IEEE, pp. 420–423.

Mateen, H. et al., 2017. Localization of Wireless Capsule Endoscope: A Systematic Review. *IEEE Sensors Journal*, 17(5), pp.1197–1206.

Mehmood, I., Sajjad, M. & Baik, S.W., 2014a. Mobile-cloud assisted video summarization framework for efficient management of remote sensing data generated by wireless capsule sensors. *Sensors (Switzerland)*, 14(9), pp.17112–17145.

Mehmood, I., Sajjad, M. & Baik, S.W., 2014b. Video summarization based tele-endoscopy: A service to efficiently manage visual data generated during wireless capsule endoscopy procedure. *Journal of medical systems*, 38(9).

Meltzer, A.C. et al., 2013. Video capsule endoscopy in the emergency department: a prospective study of acute upper gastrointestinal hemorrhage. *Annals of emergency medicine*, 61(4), pp.438–443.

Mencıa, E.L. & Furnkranz, J., 2008. Pairwise learning of multilabel classifications with perceptrons. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. IEEE, pp. 2899–2906.

Mitselos, I.V. & Christodoulou, D.K., 2018. What defines quality in small bowel capsule endoscopy. *Annals of Translational Medicine*.

Mohammed, A. et al., 2017. Sparse Coded Handcrafted and Deep Features for Colon Capsule Video Summarization. In *Proceedings - IEEE Symposium on Computer-Based Medical Systems*. pp. 728–733.

Mohri, M., Rostamizadeh, A. & Talwalkar, A., 2012. The Foundations of Machine Learning.

Mura, M. et al., 2016. Vision-based haptic feedback for capsule endoscopy navigation: a proof of concept. *Journal of Micro-Bio Robotics*, 11(1-4), pp.35–45.

Nafchi, A.R., Goh, S.T. & Zekavat, S.A.R., 2014. Circular arrays and inertial measurement unit for DOA/TOA/TDOA-based endoscopy capsule localization: Performance and complexity investigation. *IEEE Sensors Journal*, 14(11), pp.3791–3799.

Nawarathna, R. et al., 2014. Abnormal image detection in endoscopy videos using a filter bank and local binary patterns. *Neurocomputing*, 144, pp.70–91.

Nickolls, J., Buck, I. & Garland, M., 2008. Scalable parallel programming. In *2008 IEEE Hot Chips 20 Symposium (HCS)*. IEEE, pp. 40–53.

Nyquist, H., 1928. Certain topics in telegraph transmission theory. *Transactions of the American Institute of Electrical Engineers*, 47(2), pp.617–644.

Ojala, T., Pietikäinen, M. & Harwood, D., 1996. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1), pp.51–59.

Ojala, T., Pietikäinen, M. & Mäenpää, T., 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (7), pp.971–987.

Oliva, A. & Torralba, A., 2007. The role of context in object recognition. *Trends in cognitive sciences*, 11(12), pp.520–527.

Omori, T., Nakamura, S. & Shiratori, K., 2015. Localization of the patency capsule by abdominal tomosynthesis. *Digestion*, 91(4), pp.318–325.

Pahlavan, K. et al., 2015. A novel cyber physical system for 3-D imaging of the small intestine in vivo. *IEEE Access*, 3, pp.2730–2742.

Papandreou, G. et al., 2015. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of the IEEE international conference on computer vision*. pp. 1742–1750.

Park, J., Cho, Y.K. & Kim, J.H., 2018. Current and Future Use of Esophageal Capsule Endoscopy. *Clinical endoscopy*, 51(4), p.317.

Parker, C.E. et al., 2015. Capsule endoscopy–not just for the small bowel: a review. *Expert review of gastroenterology & hepatology*, 9(1), pp.79–89.

Partio, M. et al., 2002. Rock texture retrieval using gray level co-occurrence matrix. In *Proc. of 5th Nordic Signal Processing Symposium*. Citeseer.

Peng, H. et al., 2017. Fault diagnosis of power systems using intuitionistic fuzzy spiking neural P systems. *IEEE transactions on smart grid*, 9(5), pp.4777–4784.

Pham, D.M. & Aziz, S.M., 2014. A real-time localization system for an endoscopic capsule using magnetic sensors. *Sensors (Switzerland)*, 14(11), pp.20910–20929.

PillCam^{TM}, 2016. *PillCam^{TM} Capsule Endoscopy User Manual PillCam^{TM} Desktop Software Version 9.0 DOC-2928-02 November 2016,*

Pota, M., Esposito, M. & De Pietro, G., 2017. Designing rule-based fuzzy systems for classification in medicine. *Knowledge-Based Systems*, 124, pp.105–132.

Provost, F.J., Fawcett, T. & others, 1997. Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In *KDD*. pp. 43–48.

Qian, R. et al., 2016. Visual attribute classification using feature selection and convolutional neural network. In *2016 IEEE 13th International Conference on Signal Processing (ICSP)*. IEEE, pp. 649–653.

Rahman, I. et al., 2014. 219 Magnet Assisted Capsule Endoscopy (MACE) in the upper GI tract is feasible: first human series using the novel Mirocam-Navi System. *Gastrointestinal Endoscopy*, 79(5), p.AB122.

Read, J. et al., 2016. Meka: a multi-label/multi-target extension to weka. *The Journal of Machine Learning Research*, 17(1), pp.667–671.

Read, J., Pfahringer, B. & Holmes, G., 2008. Multi-label classification using ensembles of pruned sets. In *8th IEEE international conference on data mining*. IEEE, pp. 995–1000.

Ribeiro, E., Uhl, A. & Häfner, M., 2016. Colonic polyp classification with convolutional neural networks. In *2016 IEEE 29th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, pp. 253–258.

Ronneberger, O., Fischer, P. & Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, pp. 234–241.

Rui, T. et al., 2018. Convolutional neural network feature maps selection based on LDA. *Multimedia Tools and Applications*, 77(9), pp.10635–10649.

Sánchez, J. et al., 2013. Image classification with the fisher vector: Theory and practice. *International journal of computer vision*, 105(3), pp.222–245.

Schwiegerling, J. & others, 2004. Field guide to visual and ophthalmic optics. In Spie.

Sekuboyina, A.K., Devarakonda, S.T. & Seelamantula, C.S., 2017. A convolutional neural network approach for abnormality detection in Wireless Capsule Endoscopy. In *Proceedings - International Symposium on Biomedical Imaging*. pp. 1057–1060.

Shi, Y. et al., 2015. Micro-intestinal robot with wireless power transmission: Design, analysis and experiment. *Computers in biology and medicine*, 66, pp.343–351.

Simonyan, K. & Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Singh, A. et al., 2013. Timing of video capsule endoscopy relative to overt obscure GI bleeding: implications from a retrospective study. *Gastrointestinal endoscopy*, 77(5), pp.761–766.

Sivic, J. & Zisserman, A., 2008. Efficient visual search of videos cast as text retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 31(4), pp.591–606.

Sliker, L.J. & Ciuti, G., 2014. Flexible and capsule endoscopy for screening, diagnosis and treatment. *Expert review of medical devices*, 11(6), pp.649–666.

Smeulders, A.W. et al., 2000. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (12), pp.1349–1380.

Son, D., Dogan, M.D. & Sitti, M., 2017. Magnetically actuated soft capsule endoscope for fine-needle aspiration biopsy. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*. IEEE, pp. 1132–1139.

Song, S. et al., 2014. 6-D magnetic localization and orientation method for an annular magnet based on a closed-form analytical model. *IEEE Transactions on Magnetics*, 50(9), pp.1–11.

Song, S., Hu, C. & Meng, M.Q.-., 2016. Multiple Objects Positioning and Identification Method Based on Magnetic Localization System. *IEEE Transactions on Magnetics*, 52(10).

Sonnenberg, A., 2015. Modeling Lengthy Work-ups in Gastrointestinal Bleeding. *Clinical Gastroenterology and Hepatology*, 13(3), pp.433–439.

Spada, C. et al., 2015. Colon capsule endoscopy. *Techniques in Gastrointestinal Endoscopy*, 17(1), pp.19–23.

Spyrou, E. et al., 2015. Comparative assessment of feature extraction methods for visual odometry in wireless capsule endoscopy. *Computers in biology and medicine*, 65, pp.297–307.

Spyrou, E. & Iakovidis, D.K., 2014. Video-based measurements for wireless capsule endoscope tracking. *Measurement Science and Technology*, 25(1).

Stewart, F. et al., 2017. Development of a therapeutic capsule endoscope for treatment in the gastrointestinal tract: Bench testing to translational trial. In *Ultrasonics Symposium (IUS), 2017 IEEE International*. IEEE, pp. 1–4.

Suman, S. et al., 2017. Feature selection and classification of ulcerated lesions using statistical analysis for WCE images. *Applied Sciences (Switzerland)*, 7(10).

Sun, Z.-J. et al., 2014. Preliminary study of a legged capsule robot actuated wirelessly by magnetic torque. *IEEE Transactions on Magnetics*, 50(8), pp.1–6.

Sung, J.J. et al., 2016. Use of capsule endoscopy in the emergency department as a triage of patients with GI bleeding. *Gastrointestinal endoscopy*, 84(6), pp.907–913.

Swain, M.J. & Ballard, D.H., 1991. Color indexing. *International journal of computer vision*, 7(1), pp.11–32.

Swets, J.A., 1979. ROC analysis applied to the evaluation of medical imaging techniques. *Investigative radiology*, 14(2), pp.109–121.

Szegedy, C. et al., 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 1–9.

Tajbakhsh, N., Gurudu, S.R. & Liang, J., 2015. Automatic polyp detection in colonoscopy videos using an ensemble of convolutional neural networks. In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*. IEEE, pp. 79–83.

Tasoulis, S.K., Tasoulis, D.K. & Plagianakos, V.P., 2013. Random direction divisive clustering. *Pattern Recognition Letters*, 34(2), pp.131–139.

Than, T.D. et al., 2012. A review of localization systems for robotic endoscopic capsules. *IEEE Trans. Biomed. Engineering*, 59(9), pp.2387–2399.

Than, T.D. et al., 2017. Enhanced Localization of Robotic Capsule Endoscopes Using Positron Emission Markers and Rigid-Body Transformation. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*.

Theodoridis, S. & Koutroumbas, K., 2008. *Pattern Recognition*, Elsevier/Academic Press.

Theodoridis, S. & Koutroumbas, K., 2008. *Pattern Recognition, Fourth Edition* 4th ed., Orlando, FL, USA: Academic Press, Inc.

Tortora, G. et al., 2016. An ingestible capsule for the photodynamic therapy of helicobacter pylori infection. *IEEE/ASME Transactions on Mechatronics*, 21(4), pp.1935–1942.

Tsoumakas, G. & Katakis, I., 2007. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3), pp.1–13.

Turan, M. et al., 2017. A deep learning based fusion of RGB camera information and magnetic localization information for endoscopic capsule robots. *International journal of intelligent robotics and applications*, 1(4), pp.442–450.

Turan, M. et al., 2018. Deep endovo: A recurrent convolutional neural network (rcnn) based visual odometry approach for endoscopic capsule robots. *Neurocomputing*, 275, pp.1861–1870.

Turan, M. et al., 2018. Sparse-then-dense alignment-based 3D map reconstruction method for endoscopic capsule robots. *Machine Vision and Applications*, 29(2), pp.345–359.

Tuytelaars, T., 2010. Dense interest points. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 2281–2288.

Tuytelaars, T., Mikolajczyk, K. & others, 2008. Local invariant feature detectors: a survey. *Foundations and trends® in computer graphics and vision*, 3(3), pp.177–280.

Umadevi, N. & Geethalakshmi, S., 2012. Multiple classification system for fracture detection in human bone x-ray images. In *2012 Third International Conference on Computing, Communication and Networking Technologies (ICCCNT'12)*. IEEE, pp. 1–8.

Umay, I. & Fidan, B., 2016. Adaptive magnetic sensing based wireless capsule localization. In *International Symposium on Medical Information and Communication Technology, ISMICT*.

Usman, M.A. et al., 2016. Detection of small colon bleeding in wireless capsule endoscopy videos. *Computerized Medical Imaging and Graphics*, 54, pp.16–26.

Vasilakakis, M., Iosifidou, V., et al., 2019. Bone Fracture Identification in X-Ray Images using Fuzzy Wavelet Features. In *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE, pp. 726–730.

Vasilakakis, M., Koulaouzidis, A., Yung, D.E., et al., 2019. Follow-up on: Optimizing lesion detection in small-bowel capsule endoscopy and beyond: from present problems to future solutions. *Expert Review of Gastroenterology & Hepatology*.

Vasilakakis, M., Koulaouzidis, A., Marlicz, W., et al., 2019. The future of capsule endoscopy in clinical practice: from diagnostic to therapeutic experimental prototype capsules. *Gastroenterology Review*, 14(1).

Vasilakakis, M. et al., 2016. Weakly-supervised lesion detection in video capsule endoscopy based on a bag-of-colour features model. In *Medical Image Computing and Computer Assisted Intervention (MICCAI) ; International workshop on computer-assisted and robotic endoscopy*. Springer, pp. 96–103.

Vasilakakis, M. et al., 2017. *Weakly-supervised lesion detection in video capsule endoscopy based on a bag-of-colour features model*,

Vasilakakis, M. & Iakovidis, D., 2020. Constructive Fuzzy Representation Model for Data Classification. In *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, p. Submitted.

Vasilakakis, M.D. et al., 2017. *Beyond lesion detection: Towards semantic interpretation of endoscopy videos*,

Vasilakakis, M.D., Iakovidis, D.K., et al., 2018. DINOSARC: Color features based on selective aggregation of chromatic image components for wireless capsule endoscopy. *Computational and mathematical methods in medicine*, 2018.

Vasilakakis, M.D., Diamantis, D., et al., 2018. Weakly supervised multilabel classification for semantic interpretation of endoscopy video frames. *Evolving Systems*, pp.1–13.

Viola, P. & Jones, M., 2001. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*. IEEE, pp. I–I.

Wang, H., Xu, Z. & Pedrycz, W., 2017. An overview on the roles of fuzzy set techniques in big data processing: Trends, challenges and opportunities. *Knowledge-Based Systems*, 118, pp.15–30.

Wang, J. et al., 2016. Cnn-rnn: A unified framework for multi-label image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 2285–2294.

Wang, J. et al., 2010. Locality-constrained linear coding for image classification. In *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, pp. 3360–3367.

Wang, S. et al., 2016. Computer-aided endoscopic diagnosis without human-specific labeling. *IEEE Transactions on Biomedical Engineering*, 63(11), pp.2347–2358.

Wang, Z. et al., 2019. A faster convergence and concise interpretability TSK fuzzy classifier deep-wide-based integrated learning. *Applied Soft Computing*, 85, p.105825.

Wilson, R., Calway, A.D. & Pearson, E.R., 1992. A generalized wavelet transform for Fourier analysis: the multiresolution Fourier transform and its application to image and audio signal analysis. *IEEE Transactions on Information Theory*, 38(2), pp.674–690.

Wimmer, G., Vécsei, A. & Uhl, A., 2016. CNN transfer learning for the automated diagnosis of celiac disease. In *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*. IEEE, pp. 1–6.

Winstone, B. et al., 2017. Toward Bio-Inspired Tactile Sensing Capsule Endoscopy for Detection of Submucosal Tumors. *IEEE Sensors Journal*, 17(3), pp.848–857.

Witten, I.H. et al., 2016. *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann.

Woods, S.P. & Constandinou, T.G., 2016. A compact targeted drug delivery mechanism for a next generation wireless capsule endoscope. *Journal of micro-bio robotics*, 11(1-4), pp.19–34.

Woodward, Z., Williams, J.L. & Sonnenberg, A., 2016. Length of endoscopic workup in gastrointestinal bleeding. *European journal of gastroenterology & hepatology*, 28(10), p.1166.

Wu, X. et al., 2016. Automatic Hookworm Detection in Wireless Capsule Endoscopy Images. *IEEE Transactions on Medical Imaging*, 35(7), pp.1741–1752.

Wyszecki, G. & Stiles, W.S., 1982. *Color science*, Wiley New York.

Ye, Y. et al., 2014. Comparative performance evaluation of RF localization for wireless capsule endoscopy applications. *International Journal of Wireless Information Networks*, 21(3), pp.208–222.

Yim, S. et al., 2014. Biopsy using a magnetic capsule endoscope carrying, releasing, and retrieving untethered microgrippers. *IEEE Transactions on Biomedical Engineering*, 61(2), pp.513–521.

Yim, S. & Jeon, D., 2014. Magnetic mechanical capsule robot for multiple locomotion mechanisms. *International Journal of Control, Automation and Systems*, 12(2), pp.383–389.

Yu, L., Yuen, P.C. & Lai, J., 2012. Ulcer detection in wireless capsule endoscopy images. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. IEEE, pp. 45–48.

Yu, W. et al., 2015. A smart capsule with GI-tract-location-specific payload release. *IEEE Transactions on Biomedical Engineering*, 62(9), pp.2289–2295.

Yuan, Y. et al., 2015. Saliency Based Ulcer Detection for Wireless Capsule Endoscopy Diagnosis. *IEEE Transactions on Medical Imaging*, 34(10), pp.2046–2057.

Yuan, Y., Li, B. & Meng, M.Q.-., 2016a. Bleeding Frame and Region Detection in the Wireless Capsule Endoscopy Video. *IEEE Journal of Biomedical and Health Informatics*, 20(2), pp.624–630.

Yuan, Y., Li, B. & Meng, M.Q.-., 2016b. Improved Bag of Feature for Automatic Polyp Detection in Wireless Capsule Endoscopy Images. *IEEE Transactions on Automation Science and Engineering*, 13(2), pp.529–535.

Yuan, Y., Li, B. & Meng, M.Q.-., 2017. WCE abnormality detection based on saliency and adaptive locality-constrained linear coding. *IEEE Transactions on Automation Science and Engineering*, 14(1), pp.149–159.

Yuan, Y., Li, D. & Meng, M.Q.-H., 2017. Automatic polyp detection via a novel unified bottom-up and top-down saliency approach. *IEEE journal of biomedical and health informatics*, 22(4), pp.1250–1260.

Yung, D.E., Rondonotti, E., Giannakou, A., et al., 2017. Capsule endoscopy in young patients with iron deficiency anaemia and negative bidirectional gastrointestinal endoscopy. *United European gastroenterology journal*, 5(7), pp.974–981.

Yung, D.E., Koulaouzidis, A., et al., 2017. Clinical outcomes of negative small-bowel capsule endoscopy for small-bowel bleeding: a systematic review and meta-analysis. *Gastrointestinal endoscopy*, 85(2), pp.305–317.

Yung, D.E. et al., 2017. Clinical validity of flexible spectral imaging color enhancement (FICE) in small-bowel capsule endoscopy: A systematic review and meta-analysis. *Endoscopy*, 49(3), pp.258–269.

Yung, D.E., Rondonotti, E., Sykes, C., et al., 2017. Systematic review and meta-analysis: is bowel preparation still necessary in small bowel capsule endoscopy? *Expert review of gastroenterology & hepatology*, 11(10), pp.979–993.

Zadeh, L.A., 1988. Fuzzy logic. *Computer*, 21(4), pp.83–93.

Zadeh, L.A., 1999. Fuzzy logic= computing with words. In *Computing with Words in Information/Intelligent Systems 1*. Springer, pp. 3–23.

Zadeh, L.A., 1965. Fuzzy sets. *Information and control*, 8(3), pp.338–353.

Zadeh, L.A., 1996. Knowledge representation in fuzzy logic. In *Fuzzy Sets, Fuzzy Logic, And Fuzzy Systems: Selected Papers by Lotfi A Zadeh*. World Scientific, pp. 764–774.

Zadeh, L.A., 1975. The concept of a linguistic variable and its application to approximate reasoning—I. *Information sciences*, 8(3), pp.199–249.

Zadeh, L.A., 1975a. The concept of a linguistic variable and its application to approximate reasoning—II. *Information sciences*, 8(4), pp.301–357.

Zadeh, L.A., 1975b. The concept of a linguistic variable and its application to approximate reasoning-III. *Information sciences*, 9(1), pp.43–80.

Zhang, M.-L. & Zhou, Z.-H., 2013. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8), pp.1819–1837.

Zhang, M.-L. & Zhou, Z.-H., 2007. ML-KNN: A lazy learning approach to multi-label learning. *Pattern recognition*, 40(7), pp.2038–2048.

Zhang, M.-L. & Zhou, Z.-H., 2006. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE transactions on Knowledge and Data Engineering*, 18(10), pp.1338–1351.

Zhang, R. et al., 2016. Automatic detection and classification of colorectal polyps by transferring low-level CNN features from nonmedical domain. *IEEE journal of biomedical and health informatics*, 21(1), pp.41–47.

Zhang, Y., Ishibuchi, H. & Wang, S., 2017. Deep Takagi–Sugeno–Kang fuzzy classifier with shared linguistic fuzzy rules. *IEEE Transactions on Fuzzy Systems*, 26(3), pp.1535–1549.

Zheng, Y. et al., 2012. Detection of lesions during capsule endoscopy: physician performance is disappointing. *The American journal of gastroenterology*, 107(4), p.554.

Zhong, Y., Du, R. & Chiu, P.W., 2015. Tadpole endoscope: a wireless micro robot fish for examining the entire gastrointestinal (GI) tract. *HKIE Transactions*, 22(2), pp.117–122.

Zhu, R., Zhang, R. & Xue, D., 2015. Lesion detection of endoscopy images based on convolutional neural network features. In *2015 8th International Congress on Image and Signal Processing (CISP)*. IEEE, pp. 372–376.

Ziólko, B., Emms, D. & Ziólko, M., 2017. Fuzzy evaluations of image segmentations. *IEEE Transactions on Fuzzy Systems*, 26(4), pp.1789–1799.