

UNIVERSITÀ DEGLI STUDI DI CATANIA
FACOLTÀ DI SCIENZE MATEMATICHE FISICHE E NATURALI
DOTTORATO DI RICERCA IN INFORMATICA

AUTOMATIC CLASSIFICATION OF FRAMES FROM
WIRELESS CAPSULE ENDOSCOPY

ELIANA GRANATA

A dissertation submitted to the Department of Mathematics and Computer Science and the committee on graduate studies of University of Catania, in fulfillment of the requirements for the degree of doctorate in computer science.

ADVISOR

Prof. Giovanni Gallo

COORDINATOR

Prof. Domenico Cantone

XXIII CICLO

Contents

1	INTRODUCTION	1
2	WIRELESS CAPSULE ENDOSCOPY OVERVIEW	4
2.1	Device description	6
2.2	Advantages and disadvantages of Wireless Capsule Endoscopy .	9
2.3	Common examples of capsule video images of the gut	10
2.4	Manual annotation	11
2.5	Intestinal content	13
3	LITERATURE REVIEW	15
3.1	Topographic segmentation	16
3.1.1	Event boundary detection	18
3.2	Event detection	20
3.2.1	Detection of intestinal contractions and juices	21
3.2.2	Abnormal patterns detection	23
3.2.3	Bleeding detection	27
3.3	Adaptive viewing speed adjustment	30
3.4	Image quality enhancement	31
4	FEATURE EXTRACTION	32
4.1	Energy and high frequency content	33
4.2	Gabor filters	35
4.3	Local Binary Pattern	37
4.3.1	The original LBP	39
4.3.2	Derivation	39

4.4	Derivative of Gaussian filter	43
4.5	Co-occurrences	46
5	TEXTONS	53
5.1	Background	54
5.2	The basic algorithm	55
6	INFORMATION THEORETIC METHOD	59
6.1	Introduction	59
6.2	Entropy	60
6.3	Kolmogorov complexity	63
6.4	Normalized Compression Distance	66
7	LINEAR DISCRIMINANT ANALYSIS OF WCE DATA	69
7.1	Different approaches to LDA	70
7.2	Mathematical operations	70
7.3	Fisher Analysis applied to WCE frames	75
8	DETECTION ALGORITHMS	80
8.1	Sudden changes detection in a WCE video	80
8.1.1	Pre-processing and feature extraction	81
8.1.2	Classification method	83
8.1.3	Finding sudden changes	84
8.1.4	Visual exploration of textons variability	85
8.1.5	Experimental results	86
8.2	Information Theory based WCE video summarization	91
8.2.1	The proposed method	91
8.2.2	Experimental results	93
8.2.3	Conclusion	95
8.3	LBP based detection of intestinal motility in WCE images	97
8.3.1	Contractions features description	98
8.3.2	Texton-based classification	99
8.3.3	Experiments	99
8.3.4	Conclusion and future work	102

<i>CONTENTS</i>	iii
8.4 Detection of intestinal motility using block-based classification	105
8.4.1 Preliminary classification	105
8.4.2 Extracting spatial local features	107
8.4.3 Results	108
9 CONCLUSION AND FUTURE WORK	111

Chapter 1

INTRODUCTION

Conventional endoscopic techniques for examining the small intestine are limited by its length (3.5-7 m) and by its looped configurations. There are a vast number of different techniques ranging from colonoscopy and push enteroscopy to full intraoperative endoscopy. Either the limitations or the intrusive nature of these techniques have made the small intestine the most uncharted section of the gastrointestinal tract, mostly due to its anatomical characteristics and difficult access.

Wireless Capsule Endoscopy (WCE) is a recent technological breakthrough to examine the entire small intestine without any surgery. With capsule endoscopy, a pill with a micro-camera attached to it is swallowed by the patient. During several hours, the capsule emits a radio signal which is recorded into an external device, storing a video movie of the trip of the capsule throughout the gut.

The application of this technique allows the specialist to overcome most of the difficulties associated to classical clinical procedures. However, capsule endoscopy carries a main drawback: the visualization analysis of the video frames is a tedious and difficult task, which deserves specifically trained staff,

and which may last for more than one hour for each study. Thus, although the information provided by capsule endoscopy is unique and there is no other current technique which improves the reach and quality of the capsule images, the consequent procedure of analysis of the video material makes this clinical routine not feasible. Besides, abnormalities in GI tract may be present in only one or two frames of video, so sometimes they may be missed by physicians due to oversight. Moreover, there may be some abnormalities that cannot be detected by naked eyes due to their size and distribution. In addition, different clinicians may have different findings when they review the same image. All these problems motivate researchers to develop reliable and uniform assisting methods to reduce the great burden of physicians.

In this work a machine learning system to automatically detect the meaningful events in video capsule endoscopy is proposed, driving a very useful but not feasible clinical routine into a feasible clinical procedure. Our proposal is divided into two different parts. The first part tackles the problem of the automatic detection of sudden changes in a video sequence to provide an automatic tissue discriminator.

It is reported that a medical clinician spends one or two hours to examine the output video and this limits the number of examinations and leads to considerable amount of costs. The problem is to label the video frames to automatically discriminate digestive organs such as esophagus, stomach, small intestine (duodenum, jejunum, ileum) and colon. So it is possible individuate event boundaries that indicate either entrance in the next organ or unusual event in the same organ, such as bleedings, intestinal juices or obstructions, etc. To this aim the construction of an indicator function that takes high value, whenever there is a sudden change, is proposed. Several features are extracted from images to build a robust classifier. The con-

struction of the function uses the statistical texton approach. Experimental results show that it is possible to take in account only the 30% of the video, that represents the percentage of relevant frames.

The second part of the current work tackles the problem of the automatic detection of specific events such as intestinal contractions, that may be related to certain gastrointestinal disorders. The proposal is based on the analysis of the wrinkle patterns, presenting a comparative study of several features and classification methods, and providing a set of appropriate descriptors for their characterization. Experiments have been conducted on over 2000 frames extracted from WCE videos. A recall of 99% and a precision of 95% have been reached. The effects of various parameters on our classification algorithm is discussed. The achieved high detection accuracy of the proposed system has provided thus an indication that such intelligent schemes could be used as a supplementary diagnostic tool in endoscopy.

In this dissertation a detailed analysis of the performance, achieved following several approaches both in a qualitative and a quantitative way, is provided.

Chapter 2

WIRELESS CAPSULE ENDOSCOPY OVERVIEW

Medical diagnosis is based on information obtained from various sources, such as results of clinical examinations and histological findings, patients history and other data that physician considers in order to reach a final diagnostic decision [1]. Wireless Capsule Endoscopy (WCE) has been proposed in 2000 ([2],[3]) and it integrates wireless transmission with image and video technology. It has been used to examine the small intestine non invasively. The WCE procedure consists of the ingestion of a small capsule whose front end has an optical dome where a white light illuminates the luminal surface of the intestine. This video produced by a micro colour camera, is emitted by radio frequency and recorded into an external device carried by the patient. Images have 256×256 pixel resolution with three 8-bit colour planes. Frame rate is approximately two per second, and an average exam has around 50.000 images where 1000 are from the gastrointestinal tract entrance (exterior, teeth, esophagus, etc.), 4.000 from the stomach, 30.000 from the small intestine (duodenum, ileum, etc.) and 3.000 from the large intestine (cecum,

colon, etc.) Once the study is finished, the final record can be easily downloaded into a PC with the appropriate software for its posterior analysis by the physicians.

Recently, several works have tested the performance of capsule endoscopy in multiple clinical studies. Some of these clinical scenarios include intestinal polyposis and the diagnosis of small bowel tumors, obscure digestive tract bleeding, Crohn's disease [4], Celiac disease [5] and occult bleeding [6]. In this direction, comparative studies have been published showing the main advantages and drawbacks of WCE in comparison with push enteroscopy, sonde enteroscopy and intraoperative endoscopy in this kind of pathologies. A more exhaustive review and summary about the current literature regarding wireless capsule endoscopy can be found in the following bibliographic references [7, 8, 9].

Imaging techniques have been extensively used, in the last decades, as a valuable tool in the hands of an expert for a more accurate judgment of patients condition. Medical specialists look for significative events in the WCE video by direct visual inspection manually labelling, in tiring and up to one hour long sessions, clinical relevant frames. This is a bottleneck of WCE usage, since it limits its general applicability. The solution might lie in Computer Vision, by creating automatic annotation tools that preselect all the important images. This can both reduce annotation times and automatically label data for clinical research. To automatically discriminate digestive organs such as esophagus, stomach, small intestine (duodenum, jejunum, ileum) and colon is hence of great advantage. Mostly relevant is to find event boundaries that indicate either entrance to the next organ or to find unusual events within the same organ, such as bleedings, intestinal juices or obstructions, etc. All of these events are characterized by a sudden

change in the video.

In this scenario, the need of an alternative procedure for the obtention of the meaningful events in video capsule endoscopy is mandatory. This urgent need boosted the collaboration with a group of gastroenterologists from Ospedale M. Raimondi in San Cataldo, so as to evaluate the possibility of starting a new research line in this fieldwork. From that moment on, our efforts were focused on the development of a machine learning based system for the automatic detection of specific events in video capsule endoscopy.

2.1 Device description

The American Food and Drug Administration (FDA) approved the endoscopic capsule in 2001, for the purpose of visualization of the small bowel mucosa as a tool in the detection of abnormalities of the small bowel. The capsule was developed by a team of Israeli and British scientists, and was marketed by Given Imaging Ltd., Israel [3]. Since its acceptance, and according to its distributor, over 400 thousand capsule exams have been performed worldwide. Another capsule distributor, Olympus [10], started marketing its own endoscopic capsule in 2006. This technology is performed by means of three main components: the capsule, the registration device and the proprietary data analysis software. The capsule is an ingestible device equipped with all the suitable technology for image acquisition, including illumination lamps and radio frequency emission.

Figure 2.1 shows a graphical scheme of the capsule together with the distribution of its components in scale. The exterior shell is a disposable plastic capsule weighting 3.7 g and measuring $11mm \times 26mm$. (1), which contains a lens holder (2) with one lens (3), four illuminating leds (4), a

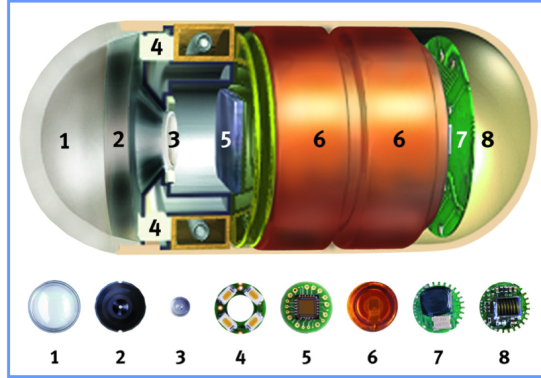


Figure 2.1: The components of the pillcam.

complementary metal oxide silicon (CMOS) image sensor (5), a battery (6), an application-specific integrated circuit (ASIC) transmitter (7), and a micro-antenna (8). The field of view of the lens spans 140-degree, very similar to that of standard endoscopy. The illuminating lamps are low consume white-light emitting diodes (LED). The video images are transmitted using UHF-band radio-telemetry to aeriels taped to the body which allow image capture, and the signal strength is used to calculate the position of the capsule in the body. Synchronous switching of the LEDs, the CMOS sensor and the ASIC transmitter minimize power consumption, which lets the emission of high-quality images at a frame ratio of 2 frames per second during 8 hours. The capsule is completely disposable and does not need to be recovered after use, being expelled by the body 10 to 72 hours after ingestion and it is passively propelled by peristalsis.

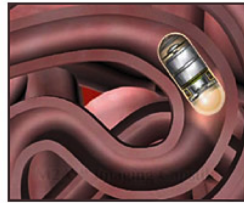


Figure 2.2: The M2A camera through the intestine.

The registration device consists of a set of aerial sensors for the RF signal reception, connected to a CPU with a hard disk for data storage. The registration device is carried by the patient fastened into a belt, altogether with a battery for power supply. The aerial sensors are taped to the body of the patient, forming an antenna array which collects the signal transmitted by the capsule and sends it to the receiver. The received data is subsequently processed and stored in the data storage by the CPU. Figure 2.3 shows a picture of (a) the external device, and (b) the belt worn by the patient.

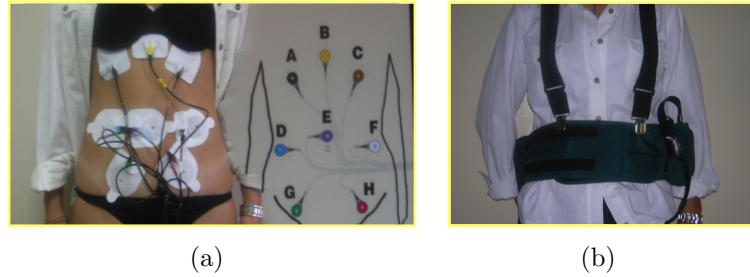


Figure 2.3: The external device and the belt, which are to be taped to the patients body.

The proprietary software is installed into a PC workstation. It allows the physicians to retrieve the data from the recorder and to transfer it to the workstation for additional processing and visualization on the display. The performed study can be stored independently on the workstation hard disk or be registered into a CD, a DVD or any other storage device, being ready for visualization and annotation on any computer in which the displaying software has been previously installed.

2.2 Advantages and disadvantages of Wireless Capsule Endoscopy

Capsule endoscopy overcomes most of the drawbacks related to manometry and other intestinal motility assessment techniques. It is much less invasive, since the patient simply has to swallow the capsule, which will be secreted in the normal cycle through the defecations. Depending on the type of study, during the whole process of the capsule endoscopy video recording the patient may lead an ordinary life. No hospitalization is needed nor special staff, since the video recording is performed without the need of any type of interaction. One of the main technological breakthroughs that this technology allows is the direct study and visualization of the entire small intestine, something that was not possible with the previous techniques so far [11]. In terms of cost, a US economic analysis in 2003, which was funded by Given Imaging Ltd., concluded that CEs per unit cost as a diagnostic tool for small intestine bleeding was comparable to that of other current endoscopic procedures.

Nowadays, however, capsule endoscopy cannot replace any of the other procedures in a general and exclusive way, and the investigation of the small intestine should include capsule endoscopy together with the rest of technologies. Since the capsule has no therapeutic capabilities, any lesion discovered by capsule endoscopy must be further investigated using other standard techniques. In addition, the capsule use is contraindicated in patients with cardiac pacemakers, defibrillators or implanted electromechanical devices (due to the risk of radio-interference with the UHF signal), and in those patients with known or suspected obstruction or pseudo-obstruction (due to the risk of causing bowel obstruction) [12]. Nowadays, the number of capsule studies performed worldwide is still small and it is too soon to appreciate the sensitiv-

ity and specificity of this technique. The study of a capsule endoscopy video takes between 1 and 2 hours, which means a heavy load for the physicians. In this sense, the research on computer-aided and intelligent systems, such as the one presented in this work, results highly interesting for the development of this technology.

2.3 Common examples of capsule video images of the gut

Capsule endoscopy video images are quite similar to those acquired by classical endoscopic techniques. Each video frame consists of a 256×256 pixel image, rendering a circular field of view of 240 pixels of diameter, which spans 140-degrees, in which the gut wall and lumen are visualized (Figure 2.4).



Figure 2.4: Appearance of a frame in capsule video endoscopy. The intestinal lumen and walls are rendered in a circular field of view. The black area has no information.

Typically, the aspect shown by each part of the gastrointestinal tube presents differences in texture, shape and colour, is patient-dependant and presents variability with several pathologies. For instance, gastric images appear with the common folded shape of the stomach wall and a pink tonality. This folded pattern is replaced in the jejunum by a plain pattern, describing

star-wise distributed wrinkles when a contractile event occurs and showing a colour tonality closer to orange. A hybrid pattern is shown in the duodenal zone, and almost no visibility is achieved in the cecal area. Figure 2.5 shows a set of sample images such as those described previously.

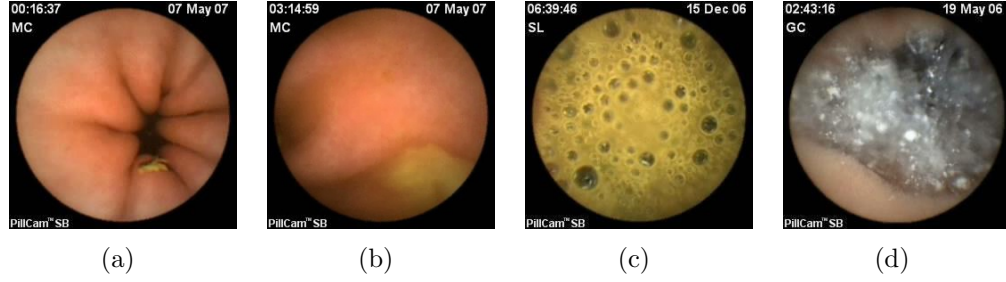


Figure 2.5: Different examples of capsule endoscopy video images. (a) Contraction. (b) Normal mucosa. (c) Residuals. (d) Intestinal juices.

The multiple patterns and appearances of the different images, and sequences of images will be the object of deep analysis in the next chapter.

2.4 Manual annotation

Video annotation is an essential characteristic of the capsule endoscopy technology. The general protocol is as follows. Once the study is downloaded to the workstation, the physician visualizes the whole video, selecting those frames where the object of interest is present (bleeding, polyp, wound, etc). Once the whole video is annotated, the expert can analyze, if necessary, each labelled frame in order to obtain information for clinical purposes, diagnosis, etc. The manufacturer itself provides a software tool to detect bleeding region; however, sensitivity and specificity of this system was reported to be 21.5% and 41.8%, respectively [13].

The expert visualizes the zone of interest where events are searched, and labels those frames where an event is detected. Figure 2.6 shows a snapshot

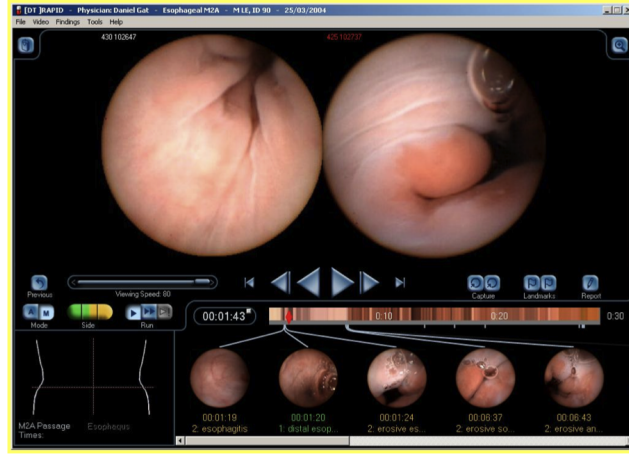


Figure 2.6: Annotation tool. The specialist saves into an external file the time of the frames of interest.

from the visualization tool provided by Given Imaging: Rapid Viewer. On the right, a time line indicates the real time (time in the real life experiment) and relative position of the frame in the video. Together with the time line, a set of findings labelled by the expert are shown in their relative position. The main screen renders the video sequence showing the gut wall and lumen. In addition, the relative position of the camera in the human body and a graph of illuminance variation can be visualized, if desired, in the lower part. The annotation process of intestinal events is, thus, not straightforward, time consuming and stressful. A typical study may contain up to 50,000 images obtained at a frame ratio of 2 images per second (6-7 hours of capsule recording). The visualization time can be adjusted from 5 to 25 frames per second. At a typical visualization rate of 15 frames per second, the specialist needs at least one hour only for visualization purposes, without taking into account the time consumed in labelling the findings.

2.5 Intestinal content

The gut can be seen as a long tube (about 4-7 meters) with its beginning in the mouth and its end in the anus. The constitution, function and visual appearance of each different part of the gut is multiple, and highly depends on the physiological task to which each part is devoted. In a general way, four main divisions can be enumerated (proximal -closer to the mouth- to distal -closer to the anus-): esophagus, stomach, small intestine and colon. The small intestine presents three different zones: duodenum, jejunum and ileum. The colon can be studied, in its turn, as ascending colon, transversal colon, descending colon and rectum. The typical visual aspect in capsule endoscopy of each one of these different parts is pictured in Figure 2.7.

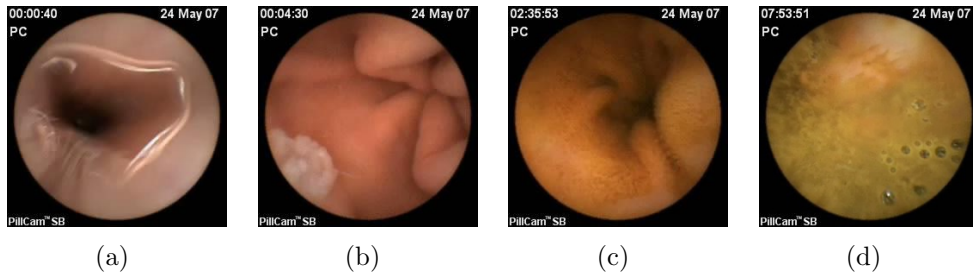


Figure 2.7: Typical visual aspect of (a) esophagus, (b) stomach, (c) duodenum, and (d) colon.

The transit of the pill through the esophagus is very fast, typically two or three seconds, with no useful information present in this area. The stomach is the first zone from which the specialists can obtain clinical information. The stomach has the shape of a sac with folded walls; these folded walls increase the overall surface of the stomach, allowing a higher performance in the physiological processes involved. It usually presents a pale colour close to pink. The typical aspect of the stomach walls in capsule endoscopy is shown in Figure 2.7 (b). The duodenum is the first part of the small intestine.

Outside the stomach, the intestinal lumen is visualized as a tunnel delimited by the intestinal walls. The intestinal walls are still folded in the duodenum, but with softer folds, presenting a colour ranging from pink to orange. A frame showing the common appearance of the duodenal walls can be observed in Figure 2.7 (c). The jejunum and ileum present a similar appearance: the intestinal walls are plain in the relaxation state, but they contract creating folds during the contractile activity. The colour appearance in these areas of the small intestine usually ranges from orange to red. Finally, the colon is the last part of the intestinal tube, see Figure 2.7 (d). The processes of assimilation of nutrients which take place in the colon are slow, in comparison with the previous stages. Moreover, since all the fecal content is released in the colon, the visual aspect is dark and the visualization quality is poor.

Chapter 3

LITERATURE REVIEW

In the previous section we have shown the importance of capsule endoscopy as a vital diagnostic procedure for a number of clinical conditions. Over 400 thousand capsule exams have been performed worldwide since 2001. It is quite clear that Computer Vision has to address the problem to overcome the most important problem of capsule endoscopy, the long exam annotation times. Fritscher Ravens and Swain [14] comment that with the predictable cost reductions of individual capsules, the time that a doctor needs to analyze the exam may become the most costly part of the procedure.

Although the use of image processing in WCE video analysis is still in its infancy, a significant number of papers have already been published. The applications of Computer Vision in capsule image analysis can be divided into four categories. The first category, which is clearly the most mature judging from the number and quality of papers published, considers the topographic segmentation of WCE video into meaningful parts such as mouth, esophagus, stomach, small intestine, and colon. The second category involves the detection of clinically significant video events (both abnormal and normal). Examples include bleeding, abnormality, intestinal fluids, intestinal contrac-

tions, and capsule retention. The third category considers video analysis with respect to changes in consecutive frames in an attempt to adaptively adjust the video viewing speed, and hence achieve a reduction in the viewing time. A final area of research has been identified, although not much work has been done on it. It focuses on using image processing techniques to enhance the viewing quality of raw images captured by the capsule. These approaches differ in how the features are extracted and images are classified.

What follows provides a general review of the literature related to classification of images extracted from WCE videos.

3.1 Topographic segmentation

A wireless capsule endoscope, being swallowed, is propelled by peristalsis through the entire gastrointestinal tract passing vital organs such as the mouth, esophagus, stomach, and small intestine; finally reaching the colon when its battery runs out. Different organs require different levels of attention from a clinical reviewer, so dividing the capsule video into meaningful gastrointestinal segments allows the expert to focus on particular areas of interest, thereby making the task of reporting easier. Moreover, gastric and intestinal transit times, which can be calculated from the topographic segmentation results, provide useful diagnostic cues for clinicians. It is also an important preprocessing step for more advanced automatic tools. As an example, users of Givens Rapid Reader software must interactively identify boundaries between the stomach and the intestine (pylorus); and intestine and colon (ileocaecal valve) before other functions (e.g., the suspected blood indicator (SBI) and capsule localization function) are enabled. Finding the pylorus in the video can be difficult and time-consuming, even for an ex-

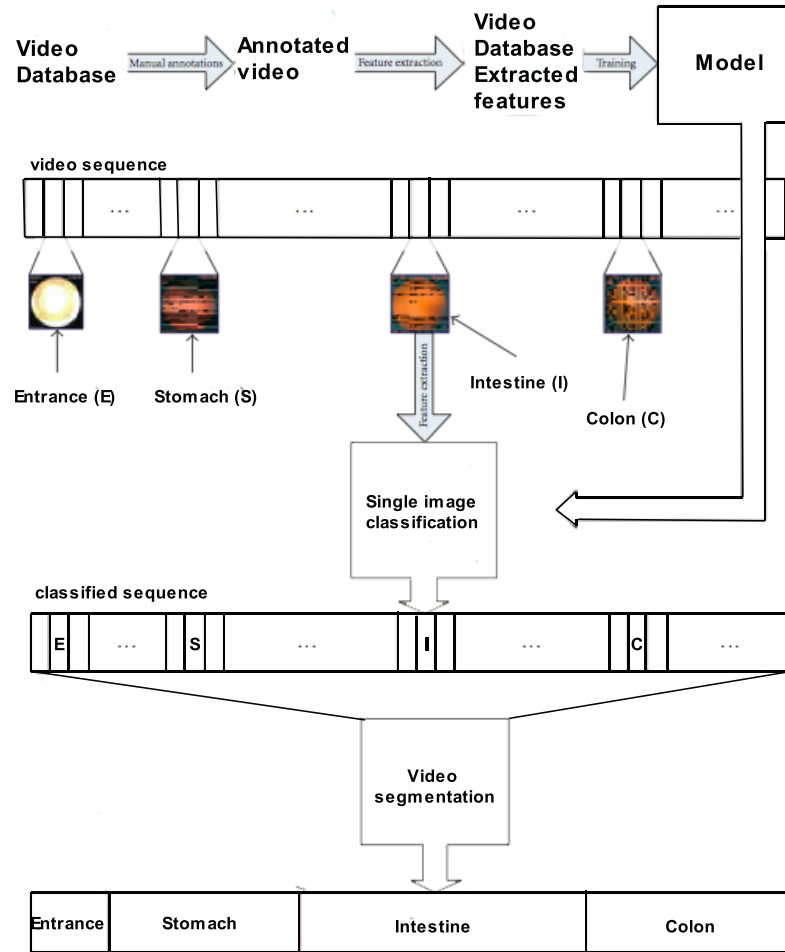


Figure 3.1: Typical structure of current topographic segmentation methods.

perienced viewer, as visually the stomach tissue in the pyloric region and the tissue at the beginning of the intestine appear very similar. A further difficulty presents itself at this point as tissues are often contaminated with faecal material that occludes the camera view. To summarize, accurate topographic segmentation is a difficult and time consuming task that is currently under-taken by clinical experts immediately before the WCE video can be reviewed. A number of Computer Vision algorithms have been developed addressing this problem, most of these algorithms can be divided into different

tasks as illustrated in the following subsections.

3.1.1 Event boundary detection

Lee et al. in [15] proposed a novel algorithm for event boundary detection in WCE based on energy function. The basic idea is that each organ has different pattern of intestinal contractions. So they first characterize the contractions using energy function in a frequency domain. Then, they segment WCE video into events by using a high frequency content (HFC) function. The detected boundaries indicate either entrance of the next organ or an anomaly in the organ, such as bleedings or intestinal juices, etc. The classification result is represented by a tree structure, which is called an event hierarchy of WCE. In order to characterize the contractions they extract energy-based feature in frequency domain from images and then they detect event boundaries by using a high frequency content function. The motility patterns are different since each digestive organ has different types of movements and functionalities. They mainly focus on the colour features because colours are the only feature values captured by the camera, so intestinal movements and contractions could change colour values. For the efficiency of processing first they converted the RGB colour space into HSI colour space for every frame in the video. Each part of the intestine has different pattern of colour sequence values. When the capsule enters the next digestive organ, the corresponding colour signal has a short-term change and the increase in energy. An event of WCE videos is defined as a sequence of continuous frames that include the same semantic contents. Events are for example esophagus, small intestine or anomalies. They choose a frequency domain method since is able to reveal the changes in overall energy and the energy concentration in frequency. In the frequency spectrum a sudden change appear as high frequency energy.

Event boundaries are detected by the energy based detection function. It is possible that a single event can be divided into several events because of the local maxima of the colour signal and the threshold value. To find the exact boundary they merged these events into a single event. So they built a tree using the energy function. They evaluate the performance of the proposed schemes by demonstrating that the proposed event boundary detection technique detect and classify accurately transitions of events in WCE videos and the proposed event hierarchy can provide the boundaries of digestive organs each of which has different types of intestinal contraction. The experimental results indicate that the recall and precision of the proposed event detection algorithm reach up to 76% and 51%, respectively.

In [16, 17, 18, 19, 20, 21, 22] authors present a method to discriminate automatically stomach and intestine tissue. They create a feature vector using colour and texture information; the colour features are derived from HSI histogram of the useful regions, compressed using hybrid transform, incorporating DCT and PCA. The texture features are derived by Singular Value Decomposition of the same tissue regions. After training the Support Vector Classifier, they apply a discriminator algorithm, which scans the video with an increasing step and builds up a classification result sequence. In [19] they observed that histograms built using the entire image will contain any visual contamination present in the image. In order to minimize the affect of visual contamination, they divide each WCE (256×256 pixel) image into a grid of 28 sub-images, 32×32 pixels each, covering most of the image area. Five parameters for each of the sub-images are derived: Mean Intensity, Saturation, Hue, and Standard Deviation of Intensity and Hue. The values for these parameters were set by experiment so that sub-images containing visual contamination (i.e. outside the expected colour range for the tissue type) are

rejected. In general the performance is performed by mean of the comparison of two classifiers: K-Nearest Neighbor and Support Vector Classifier (SVC). The best classification results were found using SVC. Although support vector machines are currently very popular in Computer Vision research due to their typically good performance, this highlights some characteristics of this specific scenario. The most important one is that using these features, Gaussian distributions provide a poor modeling of the system. A nonlinear classifier such as SVC yields better results but it provides very little in-sight about the behavior of the system itself. This is even more serious if they consider that different training data sets result in significant differences in the classifier performance.

Coimbra et al. [23] handled this task by studying the performance of the well known MPEG-7 visual descriptors for this specific scenario. Besides concluding that the two best visual descriptors have clearly scalable colour and homogenous texture, authors observed that better features are needed for more complex tasks such as event detection. Although good results were obtained in [24, 25], the authors have later shown that the key for successful automatic tools might lie in combining content with context features [26]. This means that not only features extracted directly from the images should be included, but other context features such as body spatial location and capsule displacement velocity.

3.2 Event detection

The hardest and most important challenge of endoscopic capsule Computer Vision research is, without doubt, automatic event detection. From a clinical perspective, this is exactly what doctors need: the removal of all unimportant

images and an automatic proposal for the annotation of all the relevant ones. It is hard to imagine a system that does not require human validation but if a doctor simply needs to validate/reject a small set of proposed annotations, then the time he needs to annotate a full exam is drastically reduced. It is thus obvious that all developed classifiers must have recall values very close to 100%. A single event could not be missed otherwise clinicians cannot trust this system and will view the whole video anyway.

3.2.1 Detection of intestinal contractions and juices

In [27] Spyridonos et al. propose a method based on anisotropic image filtering and efficient statistical classification of contractions and intestinal juices in the intestinal tract. Their proposal is based on a machine learning system which automatically learns and classifies contractions from a capsule video source, providing the expert with the portion of the video which is highly likely to contain the intestinal contractions. The prevalence of phasic contractions in video frames is low (about 1:50-70), which states an imbalanced problem. The omnipresent characteristic in these frames are the strong edges (wrinkles) of the folded intestinal wall, distributed in a radial way around the closed intestinal lumen. The procedure to encode in a quantitative way the wrinkle star pattern was accomplished in three steps. Firstly, the skeleton of the wrinkle pattern is extracted. Secondly, the center of the intestinal lumen is detected, as the point where the wrinkle edges converge using the image structure tensor. Finally, a set of descriptors were estimated taking into account the near radial organization of the wrinkle skeleton around the center of the intestinal lumen. Classification performance was tested by means of a SVM classifier with radial basis function kernel and employing the hold out cross validation method. The classification performance was estimated in

terms of sensitivity and specificity. On average the system detected correctly 90.84% of the positive examples, and 94.43% of the negative examples. In [28] the same authors detect contractions in video images. In imbalanced problems, even a small error rate results in an unacceptably large number of false positive classifications. They propose to use ROC curves to evaluate several classifier models, including classifier ensembles. The imbalanced recognition task of intestinal contractions was addressed by employing an efficient two-level video analysis system. At the first level, each video was processed resulting in a number of possible sequences of contractions. In the second level, the recognition of contractions was carried out by means of a SVM classifier. To encode patterns of intestinal motility, a panel of textural and morphological features of the intestine lumen were extracted. The system exhibited an overall sensitivity of 73.53% in detecting contractions. In order to automatically cluster the different types of intestinal contractions in WCE, the same authors have been developed a Computer Vision system which describes video sequences in terms of classical image descriptors [29]. The authors used Self-Organized Maps (SOM) to build a two-dimensional representation of the different types of contractions, which were clustered by the SOM in a non-supervised way.

In [30] Gabor filters for the characterization of the bubble-like shape of intestinal juices in WCE images have been applied. The authors present an algorithm which detects areas completely obscured by intestinal juices. Early detection of such regions is highly beneficial since they can be removed from the sequence presented to the clinician, resulting in a shortening of the reviewing time. Intestinal fluids appear as yellowish to brownish semi opaque turbid liquids often containing air bubbles as well as other artifacts. The authors point out that the most relevant feature of the intestinal fluids

is the presence of small bubbles of different sizes and circular shapes. The algorithm is based on texture analysis performed using Gabor filter banks.

3.2.2 Abnormal patterns detection

Based on the observation that endoscopic images possess rich texture information, Meng et al. in [31] search for regions affected by diseases, such as ulcers or coli. The main idea in this paper is the use of curvelet based on Local Binary Pattern (LBP) to extract textural features to distinguish ulcer regions from normal regions. The proposed new textural features can capture multi-directional features and show robustness to illumination changes. Curvelet transform has emerged as a new multi-resolution analysis tool recently. The basic idea of curvelet transform is to represent a curve as a superposition of functions of various lengths and widths obeying a specific scaling law. Regarding 2D images, it can be done first by decomposing an image into wavelet sub-bands, separating the object into a series of disjoint scales. Each sub-image of a given scale is then analyzed with a local ridgelet transform, another kind of new multiresolution analysis tool. Because images often suffer from illumination variations due to various imaging circumstances such as motion of camera and the rather limited range of illumination in digestive tract. Consequently, it is necessary to consider illumination variations effects on textures of endoscopic images because texture features are not constant to illumination variations. Uniform LBP shows rather robust performance to illumination variation. In addition, it has been demonstrated that uniform patterns can discern micro-structures such as bright spots and dark spots. Many diseases in endoscopic images show spot patterns including ulcer. Hence, they make use of uniform LBP to extract texture information after they apply curvelet transform to images. Using uniform LBP histogram,

they can obtain six statistical measurements of the histogram as features of texture in order to reduce the number of features. These features are standard deviation, skew, kurtosis, entropy, energy and mean of the histogram. To verify the performance of features, they deploy Multi Layer Perceptron (MLP) neural network and SVM to demonstrate their power in differentiating normal region and ulcer region in endoscopic images. The goal of using SVM and MLP simultaneously is to find which one is more suitable for the specific problem. Results on present ulcer data validate that this method is promising in recognizing ulcer regions since an impressive recognition rate of 92.37%, 91.46% and 93.28% in terms of accuracy, specificity and sensitivity, respectively, was obtained with MLP in YCbCr colour space.

In [32] Kodogiannis et al. present an integrated methodology for detecting abnormal patterns in WCE images. The implementation of an advanced neuro-fuzzy learning scheme has been adopted. In their research, an alternative approach of obtaining those quantitative parameters from the texture spectra is proposed both in the chromatic and achromatic domains of the image. In this research work they focused the attention on nine statistical measures (standard deviation, variance, skew, kurtosis, entropy, energy, inverse difference moment, contrast, and covariance). These statistical measures were estimated on histograms of the original image (1st-order statistics). The majority of the endoscopic research has focused on methods applied to grey-level images, where only the luminance of the input signal is utilized. Endoscopic images contain rich texture and colour information. All texture descriptors were estimated for all planes in both $RGB\{R(red), G(green), B(blue)\}$ and $HSV\{H(hue), S(saturation), V(intensity)\}$ spaces, creating a feature vector for each descriptor $D_i = (R_i, G_i, B_i, H_i, S_i, V_i)$. However, the histogram of the original image carries no information regard-

ing relative position of the pixels in the texture. An alternative scheme was proposed in this research study to extract texture features from the texture spectra in the chromatic and achromatic domains from each colour component histogram of the endoscopic images. The definition of texture spectrum employs the determination of the TU and TU number (NTU) values. The statistical features are then estimated on the histograms of the NTU transformations of the chromatic and achromatic planes of the image (R, G, B, H, S, V). An intelligent decision support system has been developed for endoscopic diagnosis based on a multiple-classifier scheme. Two intelligent classifier-schemes have been implemented in this research work. An adaptive neuro-fuzzy logic scheme that uses an alternative to the centroid defuzzification method, namely AOB (Area Of Balance) has been implemented while is then compared with an RBF network. This multiple-classifier approach using FI as a fusion method, provided encouraging results with a sensitivity, 97.18%, specificity, 98.55% and predictability 98.57%.

Four statistical measures, derived from the co-occurrence matrix in four different angles, namely angular second moment, correlation, inverse difference moment, and entropy, have been extracted by Karkanis [33]. These second-order statistical features were then calculated on the wavelet transformation of each image to discriminate among regions of normal or abnormal tissue. A software system, called CoLD, integrated the feature extraction and classification algorithms under a graphical user interface, which allowed both novice and expert users to utilise effectively all system functions. The detection accuracy of the proposed system has been estimated to be more than 95%.

Krishnan et al. [34] used endoscopic images to define features of the normal and the abnormal colon. New approaches for the characterization

of colon based on a set of quantitative parameters, extracted by the fuzzy processing of colon images, have been used for assisting the colonoscopist in the assessment of the status of patients and were used as inputs to a rule-based decision strategy to find out whether the colons lumen belongs to either an abnormal or normal category. The quantitative characteristics of the colon are: hue component, mean and standard deviation of RGB, perimeter, enclosed boundary area, form factor, and centre of mass [35]. The analysis of the extracted quantitative parameters was performed using three different neural networks selected for classification of the colon. The three networks include a two-layer perceptron trained with the delta rule, a multi-layer perceptron (MLP) with backpropagation (BP) learning and a self-organizing network. A comparative study of the three methods was also performed and it was observed that the self-organizing network is more appropriate for the classification of colon status [36]. A method of detecting the possible presence of abnormalities during the endoscopy of the lower GI system using curvature measures has been developed in [37]. In this paper, image contours corresponding to haustra creases in the colon are extracted and the curvature of each contour is computed after non-parametric smoothing. Zero-crossings of the curvature along the contour are then detected. The presence of abnormalities is identified when there is a contour segment between two zero-crossings having the opposite curvature signs to those of the two neighbouring contour segments. The proposed method can detect the possible presence of abnormalities such as polyps and tumours. Fuzzy rule-based approaches to the labelling of colonoscopic images to render assistance to the clinician have been proposed. The colour images are segmented using a scale-space filter. Several features are selected and fuzzified. The knowledge based fuzzy rule-based system labels the segmented regions as background,

lumen, and abnormalities (polyps, bleeding lesions) [38].

In [39] Lima et al. report a comparative study of Multilayer Perceptrons (MLP) and SVM in the classification of endoscopic images. Both have several advantages but their performance depends on the texture encoding process. Texture information is coded by second order statistics of colour image levels extracted from co-occurrence matrices. The co-occurrence matrices are computed from images rich in texture information. These images are obtained by processing the original images in the wavelet domain order to select the most important concerning texture description. The most relevant texture information often appears in the middle frequency channel. Hence are then modeled by using third and forth order moments in order to cope with non-Gaussianity, which appear especially in some pathological cases. They used several colour spaces such as RGB, HSI and YCbCr reaching the best results with HSI, which better separates light and colour information.

3.2.3 Bleeding detection

The manufacturers of the Given capsule system [3] provide only one automatic image analysis function in their Rapid Reader software: the suspected blood indicator (SBI), which is designed to report the location in the video of areas of active bleeding. However, this tool has been reported to have insufficient specificity and sensitivity [40].

In [41, 42], the authors report on the classification performance of the SBI for a multitude of patients, locations and different visual clarities of blood. Besides relating poor performance, they conclude that the SBI does not detect bleeding lesions in the stomach or altered blood anywhere in the GI tract, and does not reduce the time required for interpretation of the capsule endoscopy procedure.

Independent early work on this topic is described in [43] where authors propose an algorithm for detecting areas of bleeding in WCE videos, using expectation maximization (EM) clustering and a bayesian information criterion (BIC).

In [44] a study of various bleeding characteristics and a computer-based procedure to reduce the analysis time for detecting suspected bleeding diseases in patients is discussed. The proposed method for detecting gastrointestinal bleeding regions is divided into two main steps. The first step provides an efficient discrimination of the input videos that contain bleeding characteristics from those that do not correspond to bleeding. The second step proceeds with a further evaluation of the bleeding images and verifying if the initial classification really corresponds to active bleeding patterns. During this second phase the luminance-saturation relationship is explored to reduce the false positive detections. Another analysis tool is proposed, exploring the red colour component, to analyze the presence of food remains or bubbles, partly occluding the tissues. They differentiate four rules to classificate four levels of bleeding. *Level0* corresponds to a not-bleeding while *Level3* corresponds to an highly intensity bleeding. Furthermore a framework, called Capsule Endoscopy Supporting Software (CESS) has been developed. The system sensitivity reported is 88.3%.

Li and Meng [45, 46] propose to use local colour features based on chromaticity moments to discriminate normal regions and abnormal regions. They make full use of the Tchebichef polynomials and the illumination invariance of the HSI colour space. In the imaging process of WCE, the images suffer from illumination variation due to the specific imaging circumstances such as motion of the camera, the rather limited range of the illumination in the digestive tract. They consider illumination effects on colour because colour

features are very sensitive to illumination variation. The traditional multi-layer perceptron (MLP) neural network is employed to analyse the status of the WCE images because MLP has many advantages over other classifiers such as better generalization ability, robust performance, and less training data.

Penna et al. investigate bleedings in [47]. The core of the algorithm is represented by the Reed-Xiaoli (RX) detector, which is used to discriminate the bleeding regions from the surrounding normal tissues. In order to allow RX detector to target very specific blood areas, the data are pre-processed by means of a multi-stage filtering algorithm, and the final result is improved by means of morphological operations. The experimental results show that the proposed method achieves on average 92% and 88% of sensitivity and specificity respectively.

In [48] the authors compare the performance of the SVM classifier using features such as the raw data and colour histogram. In addition, for each feature, they compare the performance of different kernels, including the linear, polynomial (of degree 3), and radial basis functions (RBF). The accuracy for both sensitivity and specificity reached is over 99%.

Finally, Al-Rahayfeh et al. [49] use the purity of the red colour to detect the bleeding areas. They put range ratio colour for each of R, G, and B components. Therefore, they divide each image into multiple pixels and apply the range ratio colour condition for each pixel. Then the number of the pixels that achieved the condition is counted. If the number of pixels is greater than zero, then the frame is classified as a bleeding type. Otherwise, it is a non-bleeding. Using the range-ratio-colour feature overall accuracy of 98% is obtained.

3.3 Adaptive viewing speed adjustment

The main motivation for applying Computer Vision techniques to WCE video analysis is the potential improvement gained by reducing the overall time needed to review the data by alerting the expert to clinically significant video frames. This may be achieved not only by automatic detection of events or segmentation of the video into meaningful parts, but also by adjusting replay speed (number of frames displayed per second). The software supplied by both Given Imaging [3] and Olympus [10] includes such a control, although details of these algorithms are unknown. In [50], the authors propose a method for varying the frame rate in a capsule image sequence, which plays the video at high speed in stable regions and at a slower speed where significant changes between frames occur. The authors divide each frame into blocks and measure the similarity of colours between respective blocks in consecutive frames. In addition, the algorithm estimates motion displacement by extracting features using the KLT algorithm [51], tracking them using Newton-Raphson iterations. The authors conclude that using their method the viewing time may be reduced from 2 hours to around 30 minutes without loss of information. The most obvious remark to this type of methods is that their practical usefulness is highly subjective. There are several possibilities for measuring image disparity and then modeling how this should interact with video playback speed. How do we measure which one is best for a clinician, the faster one, the one that leads to smaller manual annotation errors. All research on this topic must handle this issue in a very convincing way, surely involving deployment in real clinical conditions for proper evaluation.

3.4 Image quality enhancement

To conclude, we cover an even more subjective and difficult to evaluate topic: image quality enhancement methods. Besides standard noise reduction methods, we can visually enhance the image somehow so that a clinician is faster and more accurate in detecting relevant events. The first commercial example is Olympus annotation software [10], which uses some sort of contrast and texture enhancement algorithms for displaying the captured images. Reactions by clinicians were mixed. The images do look more appealing but are they really improving our ability to correctly diagnose an exam? Or are they, in fact, creating misleading visual artifacts?

Chapter 4

FEATURE EXTRACTION

Feature is synonymous of input variable or attribute. In Computer Vision and Image Processing the concept of feature detection refers to methods that aim at computing abstractions of image information and making local decisions at every image point whether there is an image feature of a given type at that point or not. The resulting features will be subsets of the image domain, often in the form of isolated points, continuous curves or connected regions.

There is no universal or exact definition of what constitutes a feature, and the exact definition often depends on the problem or the type of application. Given that, a feature is defined as an "interesting" part of an image, and features are used as a starting point for many Computer Vision algorithms. Since features are used as the starting point and main primitives for subsequent algorithms, the overall algorithm will often only be as good as its feature detector. Consequently, the desirable property for a feature detector is repeatability: whether or not the same feature will be detected in two or more different images of the same scene.

Once features have been detected, a local image patch around the feature

can be extracted. This extraction may involve quite considerable amounts of image processing. The result is known as a feature descriptor or feature vector. No general theory exists to allow us to choose what features are relevant for a particular problem. Design of feature extractors is empirical and uses many ad hoc strategies. The effectiveness of any particular set can be demonstrated only by experiments.

4.1 Energy and high frequency content

Transitions, from an organ of the digestive tract to the next, are generally marked by frames that present a greater density of details like foldings, wrinkles, etc. The colour signal in WCE video is caused by only the contraction of digestive movements, and when the pillcam enters the next digestive organ, the corresponding colour signal has a short-term change that is the suddenness of the signal change and the increase in energy. This fact has been exploited in [15] to characterize transitions. For the efficiency of processing we choose the colour signal generated from intensity value of HSI colour domain.

In this work we define an “event” of WCE videos as a sequence of continuous frames that include the same semantic content. Hence, in our setting an event is a relevant anatomical locus (esophagus, pylorus, etc.), a pathological presence (bleedings, ulcerations, etc.) or a common non pathological disturbance (intestinal juices, bubbles, residuals, etc.). We design the event detection method by recognizing two signal properties associated with a short-term change, which are the suddenness of the signal change, and the increase in energy.

A frequency domain method is able to reveal non only changes in overall energy, but also the energy concentration in frequency. The frequency

location of energy is very important since the sudden changes in the signal cause phase discontinuities. In the frequency spectrum, this appears as high frequency energy. We define the energy function of colour signal, E as a sum of the squared magnitude of each frequency bin in a specified range. The energy function of the i^{th} frame, in a WCE video sequence, E_i is defined as:

$$E_i = \sum_{k=2}^{\frac{N}{2}+1} (|X_i(k)|^2) \quad (4.1)$$

where N is the FFT (Fast Fourier Transforms) array length, and $|X_i(k)|$ is the k^{th} bin of the FFT. In Equation (4.1), $\frac{N}{2} + 1$ indicates the frequency $\frac{F_S}{2}$ where F_S is the sample rate.

In particular we consider the weighted sum of the energy function of the i^{th} frame, linearly increased toward the high frequencies:

$$HFC_i = \sum_{k=2}^{\frac{N}{2}+1} (|X_i(k)|^2 * k) \quad (4.2)$$

where the range $0 \cdots N$ is the index number range of the FFT frequencies in the frame; $|X_i(k)|^2$ is the squared module of the k^{th} component of the FFT, and k is a weight to increase the relevance of higher frequencies.

In equation (4.1) and (4.2) we ignore the lowest two bins in order to avoid unwanted bias from low frequency components. Event boundaries of WCE video can be detected by the energy-based detection function.

The energy and HFC for each frame are plotted in Figure 4.1.

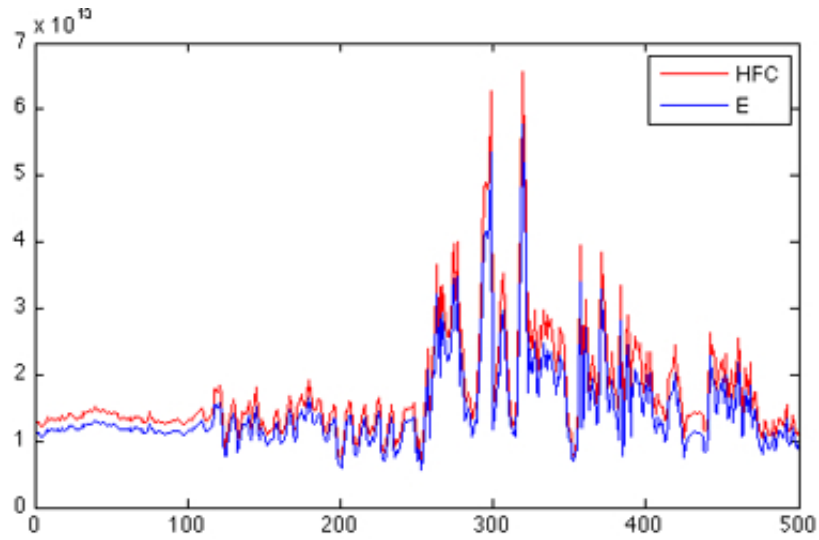


Figure 4.1: Energy and High frequency energy content in a WCE sequence

4.2 Gabor filters

Textures are powerful discriminators when one has to classify real world pictures. Indeed all the state of the art content based retrieval engines rely on texture analysis. It is hence natural to include texture descriptors among the features representing a WCE frame. Texture classification has a long history in Computer Vision, starting with Haralick proposed features [52] to the up today methods that use large sets of responses to family of linear filters. Since the sixties, texture analysis has been an area of intense research. Texture methods can also be used in medical image analysis, biometric identification, remote sensing, content-based image retrieval, document analysis, environment modeling, texture synthesis and model based image coding. Many methods have been proposed to extract texture features, e.g. the co-occurrence matrix [6], and the texture spectrum in the achromatic component of the image [9]. Signal processing based methods rely on texture filtering for extracting features in the spatial or frequency

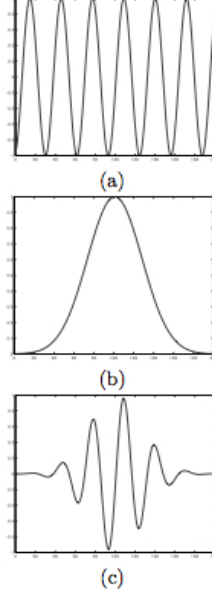


Figure 4.2: Gabor filter composition for 1D signals: (a) sinusoid, (b) a Gaussian kernel, (c) the corresponding Gabor filter.

domain. Spatial frequencies and their orientations are important characteristics of textures in images. Properly tuned Gabor filters [53] react strongly to specific textures and weakly to all others. The Gabor filters are band-pass filters with tuneable center frequency, orientation and bandwidth. A Gabor filter is obtained by modulating a sinusoid with a Gaussian. For the case of one dimensional (1D) signals, a 1D sinusoid is modulated with a Gaussian. This filter will therefore respond to some frequency, but only in a localized part of the signal. This is illustrated in Figure 4.2.

For 2D signals such as images, let us consider the sinusoid shown in Figure 4.3(a). By combining this with a Gaussian (Figure 4.3(b)), we obtain a Gabor filter - Figure 4.3(c).

For the sake of the present application a Gabor filter is defined as follows:

$$H(u, v) = \frac{1}{2\pi\sigma_u\sigma_v} e^{-\frac{1}{2} \left[\frac{(u-u_0)^2}{\sigma_u^2} + \frac{(v-v_0)^2}{\sigma_v^2} \right]} \quad (4.3)$$

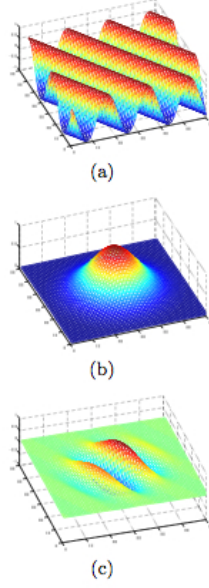


Figure 4.3: Gabor filter composition: (a) 2D sinusoid oriented at 30° with the x-axis, (b) a Gaussian kernel, (c) the corresponding Gabor filter. Notice how the sinusoid becomes spatially localized.

where $\sigma_x = \frac{1}{2\pi\sigma_u}$ and $\sigma_y = \frac{1}{2\pi\sigma_v}$ are the standard deviation of the Gaussian envelope along x and y directions. The set of parameters $(u_0, v_0, \sigma_x, \sigma_y)$ completely defines a Gabor filter.

In particular in our preliminary experiments we empirically found appropriate to choose as scale $\sigma_x = \sigma_y = 2, 4, 8$ and the following parameters set: *phase* : 0, 2, 4, 8, 16, 32 and four directions: $0^\circ, 45^\circ, 90^\circ, 135^\circ$. The rationale behind our choice has been to achieve a good compromise between recall and precision of the resulting classifier.

4.3 Local Binary Pattern

In most applications, image analysis must be performed with as few computational resources as possible. Especially in visual inspection, the speed

of feature extraction may play an enormous role. The size of the calculated descriptions must also be kept as small as possible to facilitate classification. Often, the Gabor filtering method is credited as being the current state-of-the-art in texture analysis. It has shown very good performance in a number of comparative studies. Although theoretically elegant, it tends to be computationally very demanding, especially with large mask sizes. It is also affected by varying illumination conditions. To meet the requirements of real-world applications, texture operators should be computationally cheap and robust against variations in the appearance of a texture. These variations may be caused by uneven illumination, different viewing positions, shadows etc. Depending on the application, texture operators should thus be invariant against illumination changes, rotation, scaling, viewpoint, or even affine transformations including perspective distortions. The invariance of an operator cannot however be increased to the exclusion of discrimination accuracy. It is easy to design an operator that is invariant against everything, but totally useless as a texture descriptor.

The local binary pattern (LBP) operator was developed as a gray-scale invariant pattern measure adding complementary information to the amount of texture in images. It was first mentioned by Harwood et al. (1993) and introduced to the public by Ojala et al. [54]. The approach brings together the separate statistical and structural approaches to texture analysis, opening a door for the analysis of both stochastic micro-textures and deterministic macro-textures simultaneously. The LBP operator can be made invariant against rotation, and it also supports multi-scale analysis.

4.3.1 The original LBP

The LBP operator was first introduced as a complementary measure for local image contrast [54]. The first incarnation of the operator worked with the eight-neighbors of a pixel, using the value of the center pixel as a threshold. An LBP code for a neighborhood was produced by multiplying the thresholded values with weights given to the corresponding pixels, and summing up the result (Figure 4.4). Since the LBP was, by definition, invariant to monotonic changes in gray scale, it was supplemented by an orthogonal measure of local contrast. Figure 4.4 shows how the contrast measure (C) was derived. The average of the gray levels below the center pixel is subtracted from that of the gray levels above (or equal to) the center pixel. Two-dimensional distributions of the LBP and local contrast measures were used as features.

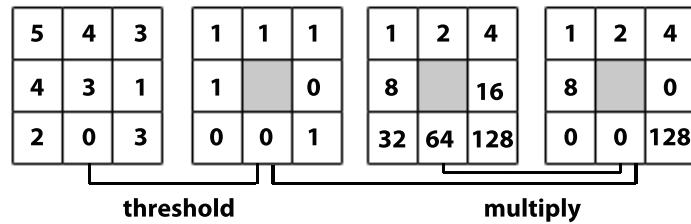


Figure 4.4: Calculating the original LBP code and a contrast measure. (a) $LBP = 1 + 2 + 4 + 8 + 128 = 143$. $C = (5+4+3+4+3)/5 - (1+0+2)/3 = 2.8$

4.3.2 Derivation

The basic version of the LBP operator considers only the eight neighbors of a pixel, but the definition has been extended to include all circular neighborhoods with any number of pixels. The derivation of the LBP was represented by Ojala et al. in 2002 [55]. Let us therefore define texture T in a local neighborhood of a gray-scale image as the joint distribution of the gray levels of

$P + 1(P > 0)$ image pixels:

$$T = t(g_c, g_0, \dots, g_{P-1}) \quad (4.4)$$

where g_c corresponds to the gray value of the center pixel of a local neighborhood. $g_P(p = 0, \dots, P - 1)$ corresponds to the gray values of P equally spaced pixels on a circle of radius $R(R > 0)$ that forms a circularly symmetric set of neighbors. This set of $P + 1$ pixels is later denoted by G_P . In a digital image domain, the coordinates of the neighbors g_p are given by $(x_c + R\cos(2\pi/P), y_c - R\sin(2\pi/P))$, where (x_c, y_c) are the coordinates of the center pixel. Figure 4.5 illustrates two circularly symmetric neighbor sets for

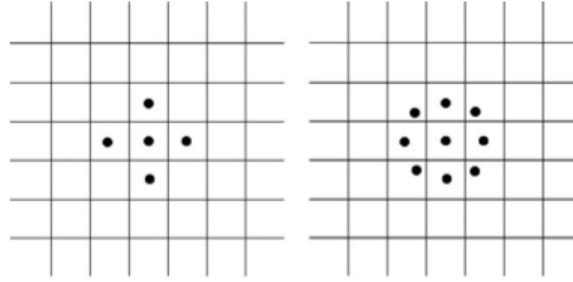


Figure 4.5: Circularly symmetric neighbours. The first represents $P = 4$ and Radius = 1; the second represents $P = 8$ and Radius = 1. The pixel values are bilinearly interpolated whenever the sampling point is not in the center of a pixel.

different values of P and R . The values of neighbors that do not fall exactly on pixels are estimated by bilinear interpolation. Since correlation between pixels decreases with distance, much of the textural information in an image can be obtained from local neighborhoods.

If the value of the center pixel is subtracted from the values of the neighbors, the local texture can be represented, without losing information, as a

joint distribution of the value of the center pixel and the differences:

$$T = t(g_c, g_0 - g_c, \dots, g_{P-1} - g_c) \quad (4.5)$$

Assuming that the differences are independent of g_c , the distribution can be factorized:

$$T \approx t(g_c)t(g_0 - g_c, \dots, g_{P-1} - g_c) \quad (4.6)$$

In practice, the independence assumption may not always hold. Due to the limited nature of the values in digital images, very high or very low values of g_c will obviously narrow down the range of possible differences. However, accepting the possible small loss of information allows one to achieve invariance with respect to shifts in the gray scale.

Since $t(g_c)$ describes the overall luminance of an image, which is unrelated to local image texture, it does not provide useful information for texture analysis. Therefore, much of the information about the textural characteristics in the original joint distribution (Eq. 4.5) is preserved in the joint difference distribution [56]:

$$T \approx t(g_0 - g_c, \dots, g_{P-1} - g_c) \quad (4.7)$$

The P dimensional difference distribution records the occurrences of different texture patterns in the neighborhood of each pixel. For constant or slowly varying regions, the differences cluster near zero. On a spot, all differences are relatively large. On an edge, differences in some directions are larger than the others. Although invariant against gray scale shifts, the differences are affected by scaling. To achieve invariance with respect to any monotonic transformation of the gray scale, only the signs of the differences are considered:

$$T \approx t(s(g_0 - g_c), \dots, g_s(P - 1 - g_c)) \quad (4.8)$$

where $s(x)$ is the sign function:

$$s(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0 \end{cases} \quad (4.9)$$

Now, a binomial weight 2^p is assigned to each sign $s(g_p - g_c)$, transforming the differences in a neighborhood into a unique LBP code. The code characterizes the local image texture around (x_c, y_c) :

$$LBP_{P,R}(x_c, y_c) = \sum_{p=0}^{P-1} s(g_p - g_c) 2^p \quad (4.10)$$

The name Local Binary Pattern reflects the functionality of the operator, i.e., a local neighborhood is thresholded at the gray value of the center pixel into a binary pattern.

In practice, Eq. (4.10) means that the signs of the differences in a neighborhood are interpreted as a P -bit binary number, resulting in 2^P distinct values for the LBP code. The local gray-scale distribution, i.e. texture, can thus be approximately described with a 2^P -bin discrete distribution of LBP codes:

$$T \approx t(LBP_{P,R}(x_c, y_c)) \quad (4.11)$$

Let us assume we are given an $N \times M$ image sample ($x_c \in \{0, \dots, N - 1\}$, $y_c \in \{0, \dots, M - 1\}$). In calculating the $LBP_{P,R}$ distribution (feature vector) for this image, the central part is only considered because a sufficiently large neighborhood cannot be used on the borders. The LBP code is calcu-

lated for each pixel in the cropped portion of the image, and the distribution of the codes is used as a feature vector, denoted by S :

$$S = t(LBP_{P,R}(x, y)), x \in \{\lceil R \rceil, \dots, N-1-\lceil R \rceil\}, y \in \{\lceil R \rceil, \dots, M-1-\lceil R \rceil\} \quad (4.12)$$

The original LBP, described in the previous section, is very similar to $LBP_{8,1}$, with two differences. First, the neighborhood in the general definition is indexed circularly, making it easier to derive rotation invariant texture descriptors. Second, the diagonal pixels in the 3×3 neighborhood are interpolated in $LBP_{8,1}$.

4.4 Derivative of Gaussian filter

Edges detection is one of the fundamental steps in image processing, image analysis, image pattern recognition, and Computer Vision techniques, particularly in the areas of feature detection and feature extraction, which aim at identifying points in a digital image at which the image brightness changes sharply or more formally has discontinuities. In the ideal case, the result of applying an edge detector to an image may lead to a set of connected curves that indicate the boundaries of objects, the boundaries of surface markings as well as curves that correspond to discontinuities in surface orientation. Thus, applying an edge detection algorithm to an image may significantly reduce the amount of data to be processed and may therefore filter out information that may be regarded as less relevant, while preserving the important structural properties of an image.

There are many methods for edge detection, but most of them can be

grouped into two categories, search-based and zero-crossing based. The search-based methods detect edges by first computing a measure of edge strength, usually a first-order derivative expression such as the gradient magnitude, and then searching for local directional maxima of the gradient magnitude using a computed estimate of the local orientation of the edge, usually the gradient direction. The zero-crossing based methods search for zero crossings in a second-order derivative expression computed from the image in order to find edges, usually the zero-crossings of the Laplacian or the zero-crossings of a non-linear differential expression. As a pre-processing step to edge detection, a smoothing stage, typically Gaussian smoothing, is almost always applied. The edge detection methods that have been published mainly differ in the types of smoothing filters that are applied and the way the measures of edge strength are computed. As many edge detection methods rely on the computation of image gradients, they also differ in the types of filters used for computing gradient estimates in the x and y directions. A filter that combines the gradient operator with a smoothing filtering is the Derivative of Gaussian (DroG). This operator corresponds to smoothing an image with Gaussian function and then computing the gradient. The Gaussian filter is a rotational symmetry function which equation in 2D is:

$$h(x, y) = \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right) = \exp\left(-\frac{r^2}{2\sigma^2}\right) \quad (4.13)$$

where σ is the standard deviation and r is the value for which $h(x, y)$ is reduced to $\frac{1}{\sqrt{e}}$ of his maximum.

A classical solution to have an useful mask is to sample the continuous function from the origin, that represents the point of application of the mask. σ sets the not null sample of the gaussian function. Hence, sets the width of the mask.

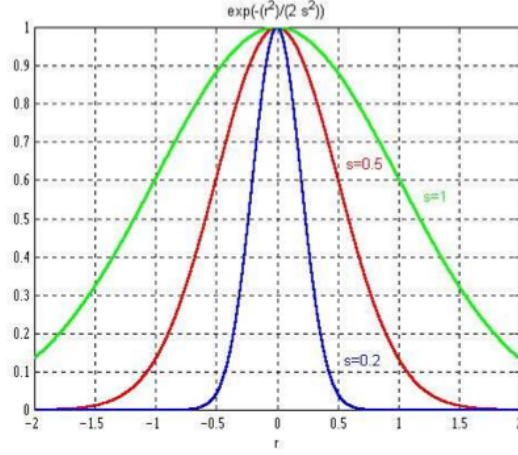


Figure 4.6: $h(x, y)$ represented for different values of σ .

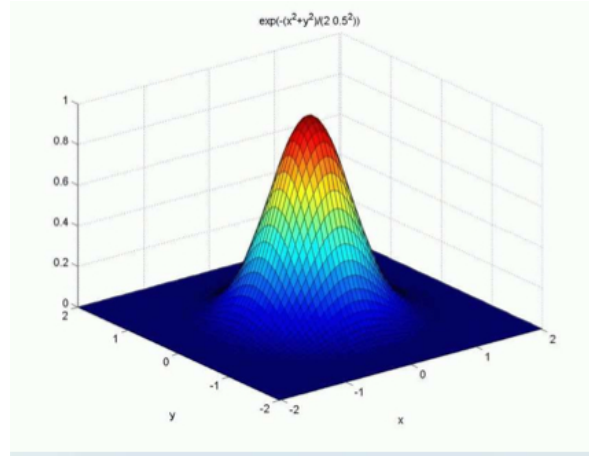
The responses to the function in two directions are obtained convolving the two gradient components with the gaussian function. Alternatively, because of the linearity of the derivate, the derivate of the gaussian in the two directions can be calculated:

$$\frac{dh}{dx} = -\frac{x}{\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right) \quad (4.14)$$

$$\frac{dh}{dy} = -\frac{y}{\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right) \quad (4.15)$$

Once a measure of edge strength is computed, the next stage is to apply a threshold, to decide whether edges are present or not at an image point. The lower threshold detects more edges, and the result will be increasingly susceptible to noise and detecting edges of irrelevant features in the image. Conversely a high threshold may miss subtle edges, or result in fragmented edges.

A commonly used approach to handle the problem of appropriate thresholds for thresholding is by using thresholding with hysteresis. This method

Figure 4.7: A 2D gaussian function with $\sigma = 0.5$

uses multiple thresholds to find edges. We begin by using the upper threshold to find the start of an edge. Once we have a start point, we then trace the path of the edge through the image pixel by pixel, marking an edge whenever we are above the lower threshold. We stop marking our edge only when the value falls below our lower threshold. This approach makes the assumption that edges are likely to be in continuous curves, and allows us to follow a faint section of an edge we have previously seen, without meaning that every noisy pixel in the image is marked down as an edge. Still, however, we have the problem of choosing appropriate thresholding parameters, and suitable thresholding values may vary over the image.

4.5 Co-occurrences

Gray Level Co-occurrence Matrix (GLCM) has proven to be a powerful basis for use in texture classification. Various textural parameters calculated from the gray level co-occurrence matrix help understand the details about the overall image content. The textural features based on gray-tone spatial de-

dependencies have a general applicability in image classification. Features generated using this technique are usually called Haralick features [57]. GLCM is also called as Gray Level Dependency Matrix. It is defined as a two dimensional histogram of gray levels for a pair of pixels, which are separated by a fixed spatial relationship. The basis for these features can reveal certain properties about the spatial distribution of the gray levels. This matrix is square with dimension N_g , where N_g is the number of gray levels in the image. Element (i, j) of the matrix is generated by counting the number of times a pixel with value i is adjacent to a pixel with value j and then dividing the entire matrix by the total number of such comparisons made. Each entry is therefore considered to be the probability that a pixel with value i will be found adjacent to a pixel of value j . GLCM of an image is computed using a displacement vector d , defined by its radius δ and orientation θ . A generalized GLCM for that image is shown in matrix G where $p(i, j)$ stands for number of times gray tones i and j have been neighbors satisfying the condition stated by displacement vector d .

$$G = \begin{bmatrix} p(1, 1) & p(1, 2) & \cdots & p(1, N_g) \\ p(1, 1) & p(1, 2) & \cdots & p(1, N_g) \\ \vdots & \vdots & \ddots & \vdots \\ p(N_g, 1) & p(N_g, 2) & \cdots & p(N_g, N_g) \end{bmatrix}$$

Since adjacency can be defined to occur in each of four directions in a 2D, square pixel image (horizontal, vertical, left and right diagonals, Figure 4.8)), four such matrices can be calculated.

Consider a 4×4 image represented by Figure 4.9 with four gray-tone values 0 through 3. The four GLCM for angles equal to 0° , 45° , 90° and 135° and radius equal to 1 are shown in Figure 4.10))



Figure 4.8: Four directions of adjacency as defined for calculation of the Haralick texture features. The Haralick statistics are calculated for co-occurrence matrices generated using each of these directions of adjacency.

0	0	1	1
0	0	1	1
0	2	2	2
2	2	3	3

Figure 4.9: Test image with four gray levels

Various research studies show δ values ranging from 1, 2 to 10. Applying large displacement value to a fine texture would yield a GLCM that does not capture detailed textural information. From the previous studies, it has been concluded that overall classification accuracies with $\delta = 1, 2, 4, 8$ are acceptable with the best results for $\delta = 1$ and 2. This conclusion is justified, as a pixel is more likely to be correlated to other closely located pixel than the one located far away. Also, displacement value equal to the size of the texture element improves classification.

Every pixel has eight neighboring pixels allowing eight choices for θ , which are $0^\circ, 45^\circ, 90^\circ, 135^\circ, 180^\circ, 225^\circ, 270^\circ, 315^\circ$. However, taking into consideration the definition of GLCM, the co-occurring pairs obtained by choosing θ , equal to 0 would be similar to those obtained by choosing θ , equal to 180° . This concept extends to $45^\circ, 90^\circ$, and 135° as well. Hence, one has four choices to select the value of θ . Sometimes, when the image is isotropic, or

4	2	1	0
2	4	0	0
1	0	6	1
0	0	1	2

(a)

6	0	2	0
0	4	2	0
2	2	2	2
0	0	2	0

(b)

4	1	0	0
1	2	2	0
0	2	4	1
0	0	1	0

(c)

2	1	3	0
1	2	1	0
3	1	0	2
0	0	2	0

(d)

Figure 4.10: (a) GLCM for $\delta = 1$ and $\theta = 0^\circ$; (b) GLCM for $\delta = 1$ and $\theta = 45^\circ$; (c) GLCM for $\delta = 1$ and $\theta = 90^\circ$; (d) GLCM for $\delta = 1$ and $\theta = 135^\circ$;

directional information is not required, one can obtain isotropic GLCM by integration over all angles.

The dimension of a GLCM is determined by the maximum gray value of the pixel. Number of gray levels is an important factor in GLCM computation. More levels would mean more accurate extracted textural information, with increased computational costs. The computational complexity of GLCM method is highly sensitive to the number of gray levels and is proportional to $O(G^2)$ [58]. Thus for a predetermined value of G , a GLCM is required for each unique pair of δ and θ . GLCM is a second-order texture measure. The GLCM's lower left triangular matrix is always a reflection of the upper right triangular matrix and the diagonal always contains even numbers. Various GLCM parameters are related to specific first-order statistical concepts. For instance, contrast would mean pixel pair repetition rate, variance would mean spatial frequency detection etc. Association of a textural meaning to each of these parameters is very critical. Traditionally, GLCM is dimensioned to the number of gray levels G and stores the co-occurrence probabilities $g(i, j)$. To determine the texture features, selected statistics are applied to each GLCM by iterating through the entire matrix.

The textural features are based on statistics which summarize the relative frequency distribution which describes how often one gray tone will appear in a specified spatial relationship to another gray tone on the image. Some statistics can be calculated from the co-occurrence matrix with the intent of describing the texture of the image:

$$Energy = \sum_i \sum_j g(i, j)^2 \quad (4.16)$$

This statistic is also called uniformity or angular second moment. It measures the textural uniformity that is pixel pair repetitions. It detects disorders in textures. Energy reaches a maximum value equal to one. High energy values occur when the gray level distribution has a constant or periodic form. Energy has a normalized range. The GLCM of less homogeneous image will have large number of small entries.

$$Entropy = \sum_i \sum_j g(i, j) \log_2 g(i, j) \quad (4.17)$$

This statistic measures the disorder or complexity of an image. The entropy is large when the image is not texturally uniform and many GLCM elements have very small values. Complex textures tend to have high entropy. Entropy is strongly, but inversely correlated to energy.

$$Contrast = \sum_i \sum_j (i - j)^2 g(i, j) \quad (4.18)$$

This statistic measures the spatial frequency of an image and is difference moment of GLCM. It is the difference between the highest and the lowest values of a contiguous set of pixels. It measures the amount of local variations present in the image. A low contrast image presents GLCM concentration

term around the principal diagonal and features low spatial frequencies.

$$Variance = \sum_i \sum_j (i - \mu)^2 g(i, j) \quad (4.19)$$

where μ is the mean of $g(i, j)$.

This statistic is a measure of heterogeneity and is strongly correlated to first order statistical variable such as standard deviation. Variance increases when the gray level values differ from their mean.

$$Homogeneity = \sum_i \sum_j \frac{1}{1 + (i - j)^2} g(i, j) \quad (4.20)$$

This statistic is also called as Inverse Difference Moment. It measures image homogeneity as it assumes larger values for smaller gray tone differences in pair elements. It is more sensitive to the presence of near diagonal elements in the GLCM. It has maximum value when all elements in the image are same. GLCM contrast and homogeneity are strongly, but inversely, correlated in terms of equivalent distribution in the pixel pairs population. It means homogeneity decreases if contrast increases while energy is kept constant.

$$Correlation = \frac{\sum_i \sum_j (ij) g(i, j) - \mu_x \mu_y}{\sigma_x \sigma_y} \quad (4.21)$$

where $\mu_x, \mu_y, \sigma_x, \sigma_y$ are the means and standard deviation of g_x and g_y . The correlation feature is a measure of gray tone linear dependencies in the image.

Of the textural features described above, the angular second moment, the entropy, the sum entropy, the difference entropy, the information measure of correlation and the maximal correlation features have the invariance property. Earlier studies [59] cite Energy and Contrast to be the most efficient parameters for discriminating different textural patterns. The general thumb

rules used in the selection of the textural features can be stated as follows:

- Energy is preferred to entropy as its values belong to normalized range.
- Contrast is associated with the average gray level difference between neighbor pixels. It is similar to variance however preferred due to reduced computational load and its effectiveness as a spatial frequency measure.
- Energy and contrast are the most significant parameters in terms of visual assessment and computational load to discriminate between different textural patterns.

Chapter 5

TEXTONS

“Textons” is a term used in different contests to name the fundamental micro-structures present in natural images (and videos). They have been first proposed as the atoms of pre-attentive human visual perception [60]. Unfortunately, the word “texton” has not been precisely defined and it remains a vague concept in the literature for lack of a universally accepted mathematical model. Texture analysis is important in many applications of computer image analysis and classification or segmentation of images based on local spatial variations of intensity or colour. A successful classification or segmentation requires an efficient description of image texture. Important applications include industrial and biomedical surface inspection, for example for defects and disease, ground classification and segmentation of satellite or aerial imagery, segmentation of textured regions in document analysis, and content-based access to image databases. Despite many potential areas of application for texture analysis in industry only a limited number of successful examples are reported in literature. A major problem is that textures in the real world are often not uniform, due to changes in orientation, scale or other visual appearance and invariance in texture analysis is still a hard

problem. In addition, the degree of computational complexity of many of the proposed texture measures is very high.

A wide variety of techniques for describing image texture have been proposed. Tuceryan and Jain [61] divided texture analysis methods into four categories: statistical, geometrical, model-based and signal processing. Due to the extensive research on texture analysis over the past 30 years, it is beyond the scope of this dissertation to propose an organized list of all published methods. The following sections provide a short introduction together with some key references. For surveys on texture analysis methods see Haralick [62], Van Gool et al. [63], Haralick and Shapiro [64], Reed and Du Buf [65], and Tuceryan and Jain [61]. Texture classification process involves two phases: the learning phase and the recognition phase. In the learning phase, the target is to build a model for the texture content of each texture class present in the training data while in the recognition phase the texture content of the unknown sample is first described with the same texture analysis method. The textural features of the sample are then compared with those of the training images using a classification algorithm, and the sample is assigned to the category with the best match. Texton models [66] have proven to be very discriminative for the recognition of grayvalue images taken from rough natural textures. Textons represent a statistical approach in which textures are modelled by the joint probability distribution of filter responses.

5.1 Background

Classifying textures from single images under general conditions is an hard challenge. The classification problem is, given an image of a textured material, to classify it into one of a set of pre-learnt classes.

The resulting image of a texture is primarily a function of the following variables: the texture surface, its albedo, the illumination, the camera and its viewing position. Most textures have large stochastic variations which make them difficult to model. Furthermore, often, two textures when photographed under very different imaging conditions can appear to be quite similar. The combination of both these factors makes the texture classification problem so hard.

Textures are modelled by the joint distribution of filter responses. This distribution is represented by texton (cluster centre) frequencies, and textons and texture models are learnt from training images. Classification of a novel image proceeds by mapping the image to a texton distribution and comparing this distribution to the learnt models. This approach is most closely related to those of Leung and Malik [67], Schmid [68] and Cula and Dana [69] and Varma and Zisserman [66]. Leung and Malik's method is not rotationally invariant and requires as input a set of registered images acquired under a (implicitly) known set of imaging conditions. Schmid's approach is rotationally invariant and texton clustering is in a higher dimensional space. Cula and Dana classify from single images, but the method is not rotationally invariant. The classifier developed by Varma and Zisserman [66] does not use colour information at all but rather normalises the images and filter responses so as to achieve partial invariance to changes in illuminant intensity.

5.2 The basic algorithm

The algorithm is based on a weak classifier and is divided in two phases: the learning stage and the classification stage. The first step of the learning stage is the construction of the texton dictionary (Figure 5.1). Multiple,

unregistered images from the training set of a particular texture class are convolved with a filter bank and features related to colour are extracted. The resulting ensemble is clustered into textons using the K-Means algorithm [70] to get a small set of recurrent and typical “visual words” and each cluster center was said to correspond to a texton.

Many criteria have been developed for determining cluster validity, all of which have a common goal to find the clustering which results in compact clusters which are well separated. The objective is to minimize this measure as we want to minimize the within-cluster scatter and maximize the between-cluster separation. Hence, the number of clusters is chosen to optimize the ratio of dispersion between cluster centers over the dispersion within clusters.

Textons from different texture classes are combined to form the texton dictionary. In this way we come to a high level representation of an image as a “bag of visual words”. This dictionary will subsequently be used to define the models based on texton frequencies learnt from training images.

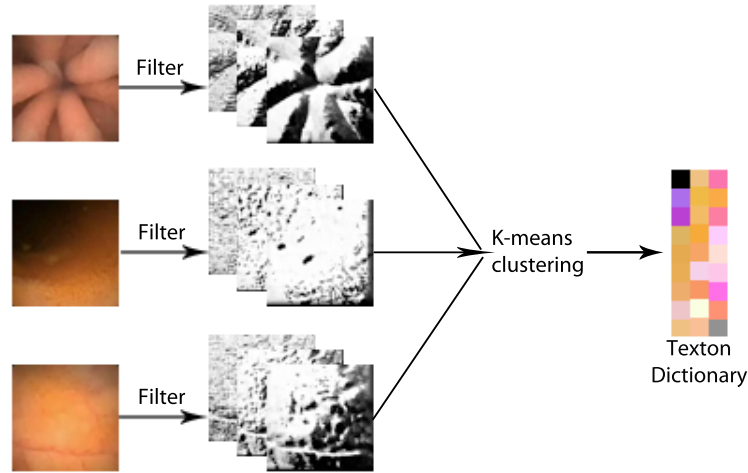


Figure 5.1: In the learning stage of the algorithm, every image of the training set is convolved with a filter bank. Filter responses are clustered using k-means algorithm to build a texton dictionary.

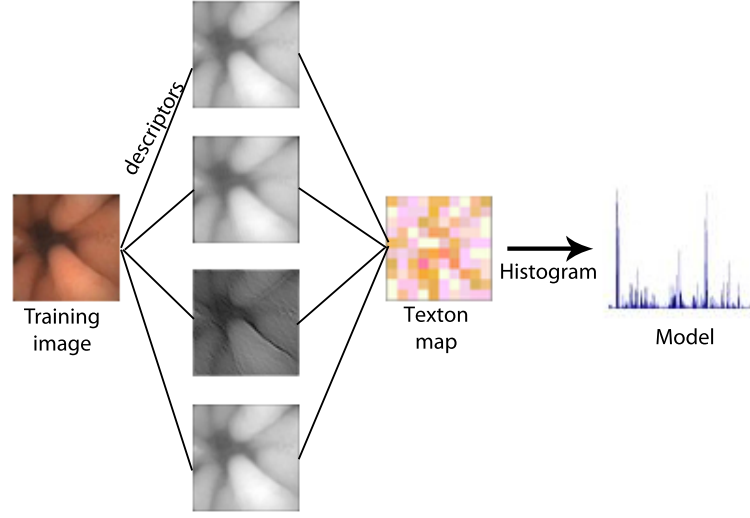


Figure 5.2: The second step is to learn models for each training image. Each filter response is labelled with the texton which lies closest to it in filter response space. The histogram represents the model corresponding to the training image.

In the classification stage, the same procedure is followed to build the histogram corresponding to the novel image. The histogram of textons, the frequency with which each texton occurs in the labelling, forms the model corresponding to the training image (Figure 5.2).

The problem of evaluating similarity between images is hence turned into the computation of histograms distance. The query image is declared as belonging to the texture class of the closest model using the k-nearest neighbor classifier [71] with a distance to measure the separability of classes (Figure 5.3). The Bhattacharyya [72] distance or the χ^2 [73] distance can be employed to compute the histograms distance to pick the closest model in the training set to the image to be classified.

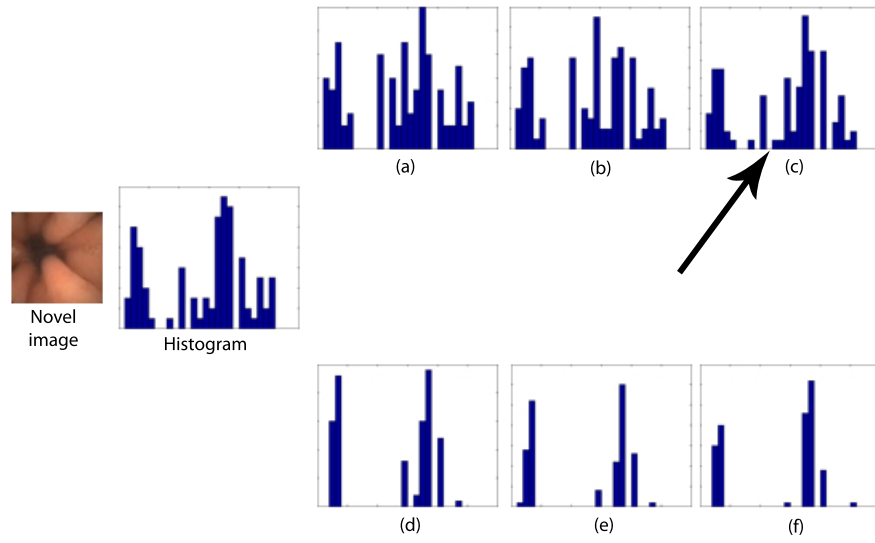


Figure 5.3: Classification stage of an image using K-NN algorithm. In this example the distance (Euclidean, Bhattacharyya, χ^2) is calculated between the histograms of the novel image and the images (a), (b), (c) and the novel image and (d), (e), (f). The minimum is found for the image (c).

Chapter 6

INFORMATION THEORETIC METHOD

6.1 Introduction

Information Theory is a branch of applied mathematics and electrical engineering involving the quantification of information. Information Theory was developed by Claude E. Shannon, in 1948, to find fundamental limits on signal processing operations such as compressing data and on reliably storing and communicating data. In “A mathematical theory of communication” [74] he defined measures such as entropy and mutual information, and introduced the fundamental laws of data compression and transmission.

Applications of fundamental topics of Information Theory include lossless data compression (e.g. ZIP files), lossy data compression (e.g. MP3s), and channel coding (e.g. for DSL lines). The field is at the intersection of mathematics, statistics, computer science, physics, neurobiology, and electrical engineering. Its impact has been crucial to the success of the Voyager missions to deep space, the invention of the compact disc, the feasibility of

mobile phones, the development of the Internet, the study of linguistics and of human perception, the understanding of black holes, and numerous other fields. Important sub-fields of information theory are source coding, channel coding, algorithmic complexity theory, algorithmic information theory, information-theoretic security, and measures of information.

An information source or source is a mathematical model for a physical entity that produces a succession of symbols called outputs in a random manner. The symbols produced may be real numbers such as voltage measurements from a transducer, binary numbers as in computer data, two dimensional intensity fields as in a sequence of images, continuous or discontinuous waveforms, and so on. The space containing all of the possible output symbols is called the alphabet of the source and a source is essentially an assignment of a probability measure to events consisting of sets of sequences of symbols from the alphabet.

In this chapter, some basic concepts of information theory and algorithmic information theory describing an absolute information-theoretic distance between bit strings, its practical approximation, and applications to real-world data are presented. A very good reference is the classic “Vitany trilogy” [75, 76, 77]. Other main reference used in this chapter is [78].

6.2 Entropy

Shannon asks himself: “Can we define a quantity which will measure, in some sense, how much information is produced by such a process, or better, at what rate information is produced?”

His answer is: “Suppose we have a set of possible events whose probabilities of occurrence are p_1, p_2, \dots, p_n . These probabilities are known but that is all

we know concerning which event will occur. Can we find a measure of how much choice is involved in the selection of the event or of how uncertain we are of the outcome?”

If there is such a measure, say $H(p_1, p_2, \dots, p_n)$, it is reasonable to require of it the following properties:

- H would be continuous in the p_i .
- If all the p_i are equal, $p_i = 1/n$, then H should be a monotonic increasing function of n . With equally likely events there is more choice, or uncertainty, when there are more possible events.
- If a choice is broken down into two successive choices, the original H should be the weighted sum of the individual values of H . The meaning of this is illustrated in Figure 6.1.

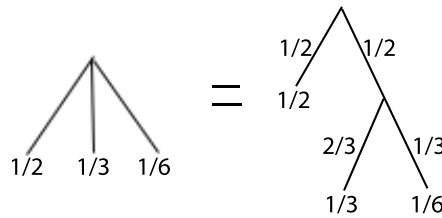


Figure 6.1: Grouping property of the entropy.

On the left, we have three possibilities $p_1 = \frac{1}{2}, p_2 = \frac{1}{3}, p_3 = \frac{1}{6}$. On the right, we first choose between two possibilities each with probability $1/2$, and if the second occurs, we make another choice with probabilities $\frac{2}{3}, \frac{1}{3}$. The final results have the same probabilities as before. We require, in this special case, that $H(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}) = H(\frac{1}{2}, \frac{1}{2}) + \frac{1}{2}H(\frac{2}{3}, \frac{1}{3})$. The coefficient $\frac{1}{2}$ is because this second choice only occurs half the time.

After these requirements, he introduces the following theorem: The only H satisfying the three above assumptions is of the form:

$$H = -k \sum_i^n p_i \log p_i \quad (6.1)$$

where k is a positive constant. When $k = 1$ and the logarithm is \log_2 , information is measured in bits. The Shannon entropy is the classical measure of information, where information is simply the outcome of a selection among a finite number of possibilities. Entropy also measures uncertainty or ignorance.

Thus, the Shannon entropy $H(X)$ of a discrete random variable X with values in the set $S = x_1, x_2, \dots, x_n$ is defined as

$$H = - \sum_i^n p_i \log p_i \quad (6.2)$$

where $n = |S|$, $p_i = \Pr[X = x_i]$ for $i \in 1, \dots, n$, the logarithms are taken in base 2 (entropy is expressed in bits), and we use the convention that $0 \log 0 = 0$, which is justified by continuity. We can use interchangeably the notation $H(X)$ or $H(p)$ for the entropy, where p is the probability distribution p_1, p_2, \dots, p_n , also represented by p_i . As $-\log p_i$ represents the information associated with the result x_i , the entropy gives us the average information or uncertainty of a random variable. Information and uncertainty are opposite. Uncertainty is considered before the event, information after. So, information reduces uncertainty. Note that the entropy depends only on the probabilities.

Some other relevant properties of the entropy are:

- $0 \leq H(X) \leq \log n$
- $H(X) = 0$ if and only if all the probabilities except one are zero,

this one having the unit value, i.e., when we are certain of the outcome.

– $H(X) = \log n$ when all the probabilities are equal. This is the most uncertain situation.

- If we equalize the probabilities, entropy increases.

By taking the consistencies out of a storage element, such as the redundancies, similar components, longest most used words, etc. compression increases the entropy of the file, until there is so much entropy in the file that it can't be further compressed. At the same time, the decompression algorithm adds more and more consistency and order to the entropic file, until it means something. One of the reasons that compression does not make a good form of encryption, is that there is information hidden in the entropy of the file, that indicates to the decompression program how to recover it. In fact in extreme compression the amount of data needed to recover the compressed text, is often a greater percentage of the file, than the actual remaining data that has yet to be compressed.

6.3 Kolmogorov complexity

In computer science, the concepts of algorithm and information are fundamental. In 1965 Andrey Nikolaevich Kolmogorov [79], a Russian mathematician, established the algorithmic theory of randomness via a measure of complexity, now referred as “Kolmogorov complexity”. According to Kolmogorov, the complexity of an object is the length of the shortest computer program that can reproduce the object. All algorithms can be expressed in a programming language based on Turing machine models with respect to

these models programs turn to be equally succinctly, up to a fixed additive constant term. The remarkable usefulness and inherent rightness of the theory of Kolmogorov complexity also called “descriptive complexity”, comes from this independence of the description method. The idea of Kolmogorov complexity first appeared in the 1960s in papers by Kolmogorov, Solomonoff and Chaitin. As specified by Schning and Randall [80], an algorithm can exhibit very different complexity behavior in the worst case and in the average case. The Kolmogorov complexity is defined as a probability distribution under which worst-case and average-case running time (for all algorithm simultaneously) are the same (up to constant factors). Quick sort algorithm has been widely taken as an example to show the applicability of Kolmogorov complexity since the algorithm takes $O(n \log n)$ time in average but $\omega(n^2)$ time at worst case. Later, the Kolmogorov complexity has been connected with Information Theory and proved to be closely related to Claude Shannon’s entropy rate of an information source. The basic theory of Kolmogorov complexity has also been extended to data compression and communication for the sake of true information measure.

Kolmogorov complexity of a string x , denoted $K(x)$, is the length $l(p)$ of the shortest binary program p that runs on a universal computing device (a Universal Turing Machine) and produces the string x as output, $\varphi(x) = p$. Mathematically, this is stated as follows [75]:

$$K(x) = \min_{\{p | \phi(p)=x\}} l(p) \quad (6.3)$$

Intuitively, the above equation describes a competitive selection of the shortest program (algorithmic description), denoted p^* , from an unbounded set of competing programs $\{p_0, p_1, \dots\}$, each capable of producing the desired output x . Experience has shown that every attempt to construct a theoretical

model of computation that is more powerful than the Turing machine has come up with something that is at the most just as strong as the Turing machine. This has been codified in 1936 by Alonzo Church as Church's Thesis [81]: the class of algorithmically computable numerical functions coincides with the class of partial recursive functions. Everything we can compute we can compute by a Turing machine and what we cannot compute by a Turing machine we cannot compute at all. Kolmogorov complexity can be used as a universal measure that will assign the same value to any sequence of bits regardless of the model of computation, within the bounds of an additive constant.

Kolmogorov complexity is not computable. It is nevertheless essential for proving existence and bounds for weaker notions of complexity. The fact that Kolmogorov complexity cannot be computed comes from the fact that we cannot compute the output of every program. More fundamentally, no algorithm is possible that can predict of every program if it will ever halt, as has been shown by Alan Turing in his famous work on the halting problem [82]. No computer program is possible that, when given any other computer program as input, will always output true if that program will eventually halt and false if it will not. Even if we have a short program that outputs our string and that seems to be a good candidate for being the shortest such program, there is always a number of shorter programs of which we do not know if they will ever halt and with what output. The uncomputability of Kolmogorov complexity has motivated several authors to seek useful approximations.

6.4 Normalized Compression Distance

In the trilogy of papers of Vitany et al. [75, 76, 77], authors describe practical approaches to approximate Kolmogorov complexity and the related notion of algorithmic information distance using common data compression algorithms such as “bzip”. In particular, the Normalized Compression Distance (NCD) measure has been shown to be a versatile and broadly applicable tool for pattern analysis. Specifically, the NCD measure approximates the idealized Normalized Information Distance (NID) using generic data compression algorithms [78]. Cilibrasi and Vitanyi [76] demonstrated the effectiveness of NCD across several applications in genomics, virology, languages, literature, music, handwritten digits, and astronomy. Following [78] the formulation of NCD is based on the relative Kolmogorov complexity between digitally represented objects (strings), rather than the Kolmogorov complexity of individual objects. Bennett et al. [75] define the absolute information distance between two strings x and y , denoted $E(x, y)$, as:

$$E(x, y) = \max \{K(x|y), K(y|x)\} \quad (6.4)$$

where $K(x|y)$ is the conditional Kolmogorov complexity of a string x relative to string y defined as the length of the shortest program to compute x if string y is provided to the universal computer as an auxiliary input. According to the Church’s thesis, $K(x)$ and $K(x|y)$ are machine independent up to an additive constant. $E(x, y)$ is the length of the shortest binary program that computes y from x , as well as x from y , while remaining unchanged itself, to within an additive logarithmic constant $O(\log \max\{K(y|x), K(x|y)\})$. Importantly, the lengths of the two strings need not be the same. Bennett et al. [75] show that $E(x, y)$ satisfies metric properties up to an additive fixed

constant. Image analysis problems of interest to us only require a relative or normalized distance metric, known as the NID [77] given as follows:

$$NID(x, y) = \frac{\max\{K(x|y), K(y|x)\}}{\max\{K(x), K(y)\}} \quad (6.5)$$

NID is an interesting theoretical concept with little practical value due to the noncomputability of its constituent Kolmogorov complexity terms. Cilibrasi and Vitany presented a method for approximating the *NID* using the *NCD*, a similarity metric between strings that is computed using off-the-shelf lossless compression programs such as zip, gzip, bzip2, etc. Compression algorithms excel at identifying and exploiting patterns in the data and the *NCD* measure exploits this ability. Intuitively, strings with similar patterns will compress better together versus when compressed separately. The *NCD* is computed as follows:

Let $C(x)$ denote the size of the compressed version of string x and $C(x, y)$ be the size of the compressed version of the concatenation of x and y .

$$NCD(x, y) = \frac{C(x, y) - \min(C(x), C(y))}{\max(C(x), C(y))} \quad (6.6)$$

There are no parameters needed to compute the *NCD*, except for the choice of compression algorithm and its settings. As shown by Vitany et al., the choice of compression algorithm has a negligible impact on the final analysis. *NCD* is a reasonable approximation to *NID* in that it is a nonnegative number in the range 0 and $1 + \epsilon$, where the ϵ arises from imperfections in real world compression algorithms and is typically less than 0.1. It is also approximately a metric and its deviations from metric properties depend upon the performance of the compression algorithm. *NCD* computation does not require any specific background knowledge about the data

and it can be computed meaningfully when x and y are of different lengths.

Practically speaking, the NCD can be computed for any set of image sequences in much the same manner, regardless of application details. Different compression algorithms may capture more or less of the structure in a string and exhibit different levels of compression performance. The normalization in (6.6) ensures that differences in NCD values resulting from the choice of different compression algorithms are modest.

Chapter 7

LINEAR DISCRIMINANT ANALYSIS OF WCE DATA

Linear discriminant analysis (LDA) and the related Fisher's linear discriminant [70] are methods used in statistics and machine learning to find a linear combination of features which characterizes or separates two or more classes of objects or events. The resulting combination may be used as a linear classifier, or, more commonly, for dimensionality reduction successive later classification. Computationally, discriminant function analysis is very similar to analysis of variance (ANOVA). Specifically, one can ask whether or not two or more groups are significantly different from each other with respect to the mean of a particular variable. If the means for a variable are significantly different in different groups, then we can say that this variable discriminates between the groups.

In the case of a single variable, the final significance test of whether or not a variable discriminates between groups is the F test. As described in Elementary Concepts and ANOVA /MANOVA, F is essentially computed as the ratio of the between-groups variance in the data over the pooled (average)

within-group variance. If the between-group variance is significantly larger then there must be significant differences between means. This method also helps to better understand the distribution of the feature data.

7.1 Different approaches to LDA

Data sets can be transformed and test vectors can be classified in the transformed space by two different approaches.

Class-dependent transformation: This type of approach involves maximizing the ratio of between class variance to within class variance. The main objective is to maximize this ratio so that adequate class separability is obtained. The class-specific type approach involves using two optimizing criteria for transforming the data sets independently.

Class-independent transformation: This approach involves maximizing the ratio of overall variance to within class variance. This approach uses only one optimizing criterion to transform the data sets and hence all data points irrespective of their class identity are transformed using this transform. In this type of LDA, each class is considered as a separate class against all other classes.

7.2 Mathematical operations

To explain discriminant analysis, let's consider a classification involving two target categories and two predictor variables. The following figure shows a plot of the two categories with the two predictors on orthogonal axes:

A visual inspection shows that category 1 objects (open circles) tend to have larger values of the predictor on the Y axis and smaller values on the X

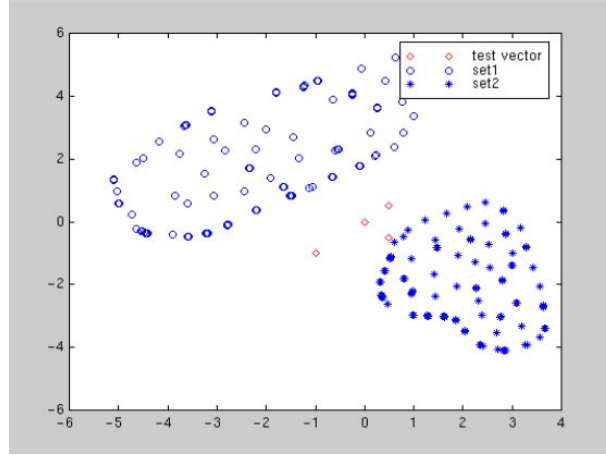


Figure 7.1: Data sets and test vectors in original

axis. However, there is overlap between the target categories on both axes, so we can't perform an accurate classification using only one of the predictors. For ease of understanding let us represent the data sets as a matrix consisting of features in the form given below:

$$set_1 = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{pmatrix} \quad set_2 = \begin{pmatrix} b_{1,1} & b_{1,2} & \cdots & b_{1,n} \\ b_{2,1} & b_{2,2} & \cdots & b_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m,1} & b_{m,2} & \cdots & b_{m,n} \end{pmatrix}$$

Compute the mean of each data set and mean of entire data set. Let μ_1 and μ_2 be the mean of set_1 and set_2 respectively and μ_3 be mean of entire data, which is obtained by merging set_1 and set_2 , is given by Equation (7.1).

$$\mu_3 = p_1 \times \mu_1 + p_2 \times \mu_2 \quad (7.1)$$

where p_1 and p_2 are the apriori probabilities of the classes. In the case of this simple two class problem, the probability factor is assumed to be 0.5.

In LDA, within-class and between-class scatter are used to formulate criteria for class separability. Within-class scatter is the expected covariance of each of the classes. The scatter measures are computed using Equation (7.2) and Equation (7.3).

$$S_w = \sum_j p_j \times (cov_j) \quad (7.2)$$

Therefore, for the two-class problem:

$$S_w = 0.5 \times cov_1 + 0.5 \times cov_2 \quad (7.3)$$

All the covariance matrices are symmetric. Let cov_1 and cov_2 be the covariance of set_1 and set_2 respectively. Covariance matrix is computed using the following equation.

$$cov_j = (x_j - \mu_j)(x_j - \mu_j)^T \quad (7.4)$$

The between-class scatter is computed using the following equation:

$$S_b = \sum_j (x_j - \mu_3)(x_j - \mu_3)^T \quad (7.5)$$

S_b can be thought of as the covariance of data set whose members are the mean vectors of each class. As defined earlier, the optimizing criterion in LDA is the ratio of between-class scatter to the within-class scatter. The solution obtained by maximizing this criterion defines the axes of the transformed space. However for the class-dependent transform the optimizing criterion is computed using equations Equation (7.4) and Equation (7.5). It should be noted that if the LDA is a class dependent type, for L-class L separate optimizing criterion are required for each class. The optimizing factors in

case of class dependent type are computed as:

$$J_j = inv(cov_j) \times S_b \quad (7.6)$$

For the class independent transform, the optimizing criterion is computed as:

$$J = inv(S_w) \times S_b \quad (7.7)$$

By definition, an eigen vector of a transformation represents a 1-D invariant subspace of the vector space in which the transformation is applied. A set of these eigen vectors whose corresponding eigen values are non-zero are all linearly independent and are invariant under the transformation. Thus any vector space can be represented in terms of linear combinations of the eigen vectors. A linear dependency between features is indicated by a zero eigen value. To obtain a non-redundant set of features all eigen vectors corresponding to non-zero eigen values only are considered and the ones corresponding to zero eigen values are neglected. In the case of LDA, the transformations are found as the eigen vector matrix of the different criteria defined in Equation (7.6) and Equation (7.7).

For any L-class problem we would always have L-1 non-zero eigen values. This is attributed to the constraints on the mean vectors of the classes in Equation (7.1). The eigen vectors corresponding to non-zero eigen values for the definition of the transformation.

Having obtained the transformation matrices, we transform the data sets using the single LDA transform or the class specific transforms which ever the case may be.

For the class dependent LDA,

$$transformedset_j = transformed_j^T \times set_j \quad (7.8)$$

For the class independent LDA,

$$transformedset = transformed_{spec}^T \times dataset^T \quad (7.9)$$

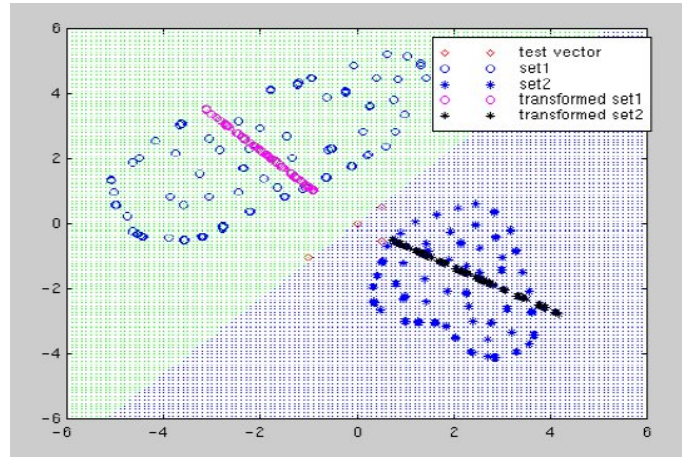


Figure 7.2: Data sets in original space and transformed space along with the transformation axis for class dependent LDA of a 2-class problem

Once the transformations are completed using the LDA transforms, Euclidean distance or RMS distance is used to classify data points. Euclidean distance is computed using Equation (7.10) where μ_{ntrans} is the mean of the transformed data set, n is the class index and x is the test vector. Thus for n classes, n euclidean distances are obtained for each test point.

$$dist_n = (transform_{nspec})^T \times x - \mu_{ntrans} \quad (7.10)$$

The smallest Euclidean distance among the n distances classifies the test vector as belonging to class n . The choice of the type of LDA depends on the data set and the goals of the classification problem. If generalization is

of importance, the class independent transformation is preferred. However, if good discrimination is what is aimed for, the class dependent type should be the first choice.

7.3 Fisher Analysis applied to WCE frames

To analyze WCE data we considered ten images in which we want to discriminate frames with contractions from frame without contractions. Images are partitioned into 25 blocks of 32×32 pixel and are labelled and each block is labelled following the protocol:

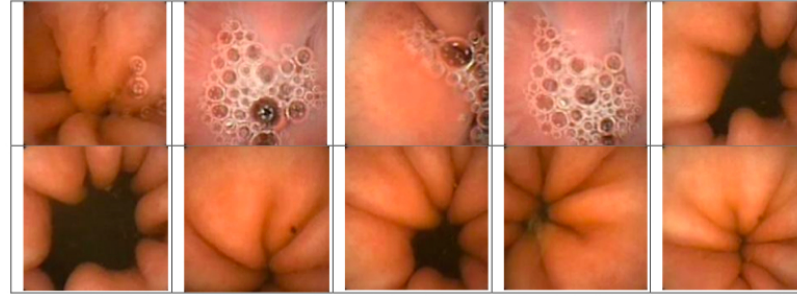
0 is the label associated to normal mucosa; 1 represents a wrinkle; 2 represents bubbles; 3 is a black part of the image that is associated to the intestinal lumen. In Figure 7.3 (a) the original dataset is reported. In Figure 7.3 (b) the manual labelling of blocks is indicated.

In Table 7.1 frames statistics over 10 images representing contractions/not-contractions extracted from WCE frames are reported. For each image we have the percentage of the four classes.

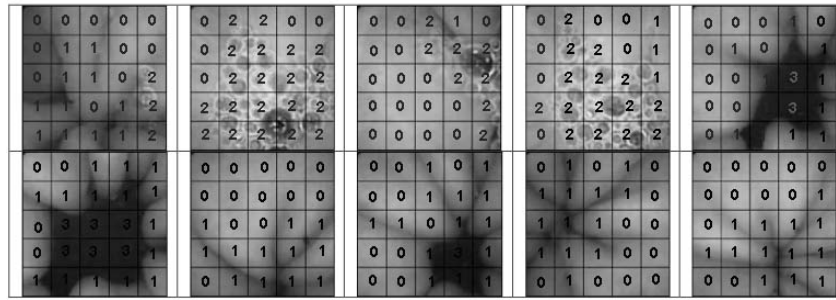
Table 7.1: Percentage of classes manually labelled in every image.

	0 (mucosa)	1 (wrinkle)	2 (bubble)	3 (lumen)	Total
<i>Img</i> ₁	44%	44%	12%	0%	100%
<i>Img</i> ₂	20%	0%	80%	0%	100%
<i>Img</i> ₃	32%	0%	68%	0%	100%
<i>Img</i> ₄	64%	4%	32%	0%	100%
<i>Img</i> ₅	44%	48%	0%	8%	100%
<i>Img</i> ₆	16%	60%	0%	24%	100%
<i>Img</i> ₇	52%	48%	0%	0%	100%
<i>Img</i> ₈	40%	56%	0%	4%	100%
<i>Img</i> ₉	48%	52%	0%	0%	100%
<i>Img</i> ₁₀	48%	52%	0%	0%	100%

Successively each block the following features are extracted: H, S, I, hfc,



(a)



(b)

Figure 7.3: (a) ten original images extracted from WCE video. (b) The same frames subdivided in 25 blocks of 32 x 32 pixel.

Drog, $LBP_{(8,1)}$. Each block is classified by mean of the use of K nearest neighbor following the method LOO (Leave On Out).

Labelling obtained with the application of K-NN is reported in Table 7.2. In the first four columns are reported the number of the classes of an image. The last column is the manual labelling in which "1" is associated to images that represent an intestinal contraction. Instead "0" is associated to images that are not a contraction.

Now we apply the linear discriminant analysis to separate the two classes, in particular we find a score function the results positive if the image represent an intestinal contraction and negative if the image is not a contraction.

The first step is the computation of the mean of all data and the subtraction to the same data. This corresponds to a change of the reference system

Table 7.2: Number of blocks labelled by K-NN. The last column represent the manual labelling of contraction/not-contraction (1/0)

	0 (mucosa)	1 (wrinkle)	2 (bubble)	3 (lumen)	Label (c/nc)
<i>Img₁</i>	10	14	1	0	1
<i>Img₂</i>	8	0	17	0	0
<i>Img₃</i>	12	0	13	0	0
<i>Img₄</i>	16	3	6	0	0
<i>Img₅</i>	7	15	1	2	1
<i>Img₆</i>	3	16	0	6	1
<i>Img₇</i>	14	10	1	0	1
<i>Img₈</i>	10	14	0	1	1
<i>Img₉</i>	9	16	0	0	1
<i>Img₁₀</i>	14	10	1	0	1

in the geometrical space so that the center of the "cloud" of the points corresponds to the origin of the axes. The total mean is given by the vector: (10.3, 9.8, 4.0, 0.9).

Subtracting the vector to the data we obtain 2 arrays that represent respectively contractions images and not-contractions images:

$$contr = \begin{bmatrix} -0,3 & 4,2 & -3 & -0,9 \\ -3,3 & 5,2 & -3 & 1,1 \\ -7,3 & 6,2 & -4 & 5,1 \\ 3,7 & 0,2 & -3 & -0,9 \\ -0,3 & 4,2 & -4 & 0,1 \\ -1,3 & 6,2 & -4 & -0,9 \\ 3,7 & 0,2 & -3 & -0,9 \end{bmatrix} \quad not - contr = \begin{bmatrix} -2,3 & -9,8 & 13 & -0,9 \\ 1,7 & -9,8 & 9 & -0,9 \\ 5,7 & -6,8 & 2 & -0,9 \end{bmatrix}$$

The second step is the computation of the covariance matrix and the mean of the single variances: In the following are reported the covariances for the two classes:

$$class1 = \begin{bmatrix} 14,95 & -8,71 & 1,12 & -7,36 \\ -8,71 & 6,62 & -0,88 & 2,98 \\ 1,12 & -0,88 & 0,29 & -0,52 \\ -7,36 & 2,98 & -0,52 & 4,90 \end{bmatrix} \quad class0 = \begin{bmatrix} 16 & 6 & -22 & 0 \\ 6 & 3 & -9 & 0 \\ -22 & -9 & 31 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Mean of the covariances of the two classes:

$$mean = \begin{bmatrix} 15,48 & -1,36 & -10,44 & -3,68 \\ -1,36 & 4,81 & -4,94 & 1,49 \\ -10,44 & -4,94 & 15,64 & -0,26 \\ -3,68 & 1,49 & -0,26 & 2,45 \end{bmatrix}$$

The score function is obtained in the following way:

$$S = inv(C)(m_1 - m_2)' \quad (7.11)$$

where m_1 and m_2 are the mean of the two classes.

The linear scores for the two classes are calculated as: $A * S$ and $B * S$ where A and B are the two classes of images.

$$A * S = \begin{bmatrix} 14.7264 \\ 15.1149 \\ 12.1579 \\ 1.0778 \\ 12.8928 \\ 21.2290 \\ 1.0778 \end{bmatrix} \quad B * S = \begin{bmatrix} -27.8956 \\ -29.1826 \\ -21.1984 \end{bmatrix}$$

Finally, considering that the elements of the contractions class are positives and the elements of the not-contractions class are negatives, the discriminant function classifies successfully all the images.

LDA has been applied to WCE data to understand how the two sets of contractions/not-contractions are separable. The analysis has been useful to apply the automatic classification to discriminate intestinal motility in WCE images.

Chapter 8

DETECTION ALGORITHMS

In this final chapter of the dissertation we called all the results obtained applying the methods described previously. In particular 8.1 ensemble experiments published in [83, 84, 85], 8.2 refers to [86], 8.3 refers to [87] and finally 8.4 refers to a work in progress in which we are improving the results.

8.1 Sudden changes detection in a WCE video

We consider the problem of classifying frames of a WCE video without imposing any constraints on the viewing or illumination conditions under which these images were specially obtained.

The basic idea is that each digestive organ has a different visual pattern. Each pattern may be characterized by specific values of a set of observed features. Several candidate features may be considered: the texton method [66] allows to statistically combine all of them to produce a classifier. Our proposal integrates texture-based features, obtained as response to a bank of Gabor filters, with features like high frequency content of energy, colour and luminance, that are customarily of primary relevance to the clinician.

The general goal of automatic WCE segmentation is to split a video into shorter sequences each with the same semantic content. In the following with the term “event” we indicate a very short frame sequence (6 consecutive frames) that testifies an abrupt and significative change in the video. We made precise this term in agreement with the medical expert that have manually labelled the sequences. Our definition of “event” includes boundary transitions from an organ to another one, intestinal juices, bile, bubbles, pathologies, etc.

The features, that we have selected to build a classifier, are: luminance and colour, high-frequency energy content and the responses to a bank of Gabor filters. These features are computed separately on frame sub-blocks. We apply C-means clustering to the set of feature vectors to build a texton dictionary. Frames are hence represented by mean of the histograms over the resulting dictionary. The computation of a function to compare histograms provides a way to assign a distance between frames. High values correspond to an abrupt change in the frames sequences.

8.1.1 Pre-processing and feature extraction

The frames coming from the WCE videos have been pre-processed as follows. The original frames have a dimension of 576×576 pixels in which there is a large black background and textual annotations. We restrict the Region Of Interest within the circular area of the video, hence for each frame only a sub-image is considered. More precisely only the maximum square inscribed in the circular image is considered. The information loss is not relevant, since the left over “lunettes” are typically out of focus because of the dome structure of the camera-pill (Figure 8.1)

Each extracted ROI has been initially transformed into the HSI colour

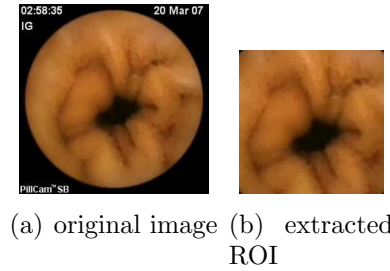


Figure 8.1: Example of an image extracted from a WCE video.

space. This colour space has been chosen for the well known robustness in image processing [88]. Frames, moreover, are partitioned into squares each of 16×16 pixels. For each one of these square sub-images we extract the features used for automatic classification. Direct visual inspection by clinicians is largely based on the consideration of the chrominance value of the frames. For this reason we choose to include the average values of the hue, saturation and intensity of each of the blocks of a frame among the representative features of the frame. These features, although informative, are not sufficient to effectively classify the frames and have to be integrated with more features as follows.

Transitions from an organ of the digestive tract to the next are generally marked by frames that present a greater density of details. This fact has been exploited in [15] to characterize transitions. For this reason we include the high frequency energy content of blocks among the features used by the classifier. When the capsule enters the next organ the corresponding colour signal has short-term change, that is the suddenness of the signal change, and an increase in energy. A frequency domain method it is able to reveal non only changes in overall energy, but also the energy concentration in frequency.

Following [15] we consider the weighted sum of the energy function, as described in 4.1, linearly increased toward the high frequencies, and we ignore the lowest two bins in order to avoid unwanted bias from low frequency

components. We include the HFC of each 16x16 sub-image among the representative features of a frame.

Textures are powerful discriminators when one has to classify real world pictures. Indeed all the state of the art content based retrieval engines rely on texture analysis. It is natural to include texture descriptors among the features representing a WCE frame. We choose a Gabor filter bank [53] for texture representation. In particular in our preliminary experiments we empirically found appropriate to choose as scale $\sigma_x = \sigma_y = 2, 4, 8$ and the following parameters set: *phase* : 0, 2, 4, 8, 16, 32 and four directions: $0^\circ, 45^\circ, 90^\circ, 135^\circ$. The rationale behind our choice has been to achieve a good compromise between recall and precision of the resulting classifier. In our proposal a frame comes to be eventually represented as a vector of 28×484 components. The 28 features includes information about average colour and luminance (3 elements), HFC (1 element) and Gabor filter responses (24 elements) for each block.

8.1.2 Classification method

In order to achieve a more abstract representation we pool together the vector of all of the 16x16 blocks of the frames in the video. In our experiments each of the videos is made of 500 frames. This leads to an ensemble of 242000 vectors. The ensemble has been clusterized to get a small set of recurrent and typical “visual words”. Clusterization is performed with a standard K-clustering. The number of clusters is chosen to optimize the ratio of dispersion between cluster centers over the dispersion within clusters. We empirically found that a suitable value for the number of clusters in our experiments is 100. In this phase the dictionary obtained provides the buckets to compute, for each frame, the relative frequencies of “visual word”. In this

way we come to a high level representation of a frame as a “bag of visual words”.

8.1.3 Finding sudden changes

The problem of evaluating similarity between frames is hence turned into the computation of histograms distance. Among the several available choices for histogram distance estimation we choose the Bhattacharya distance [72], that is normally used to measure the separability of classes. In this way we define $d(f_i, f_j)$, as the distance between frame f_i and f_j , the Bhattacharya distance between the corresponding histograms. Direct computation of the Bhattacharya distance of a pair of consecutive frames is generally a weak indicator of changes in the video. This happens because occasionally a frame can be quite different from the previous one just because of casual disturbances and trasmission noise. To have a more robust indicator of sudden changes in the video we consider for each frame f_i the function $C(i)$ defined as follows:

$$C(i) = \frac{1}{9} \sum_{k=i}^{i+2} \sum_{j=3}^5 d(f_k, f_{i+j}) \quad (8.1)$$

$C(i)$ averages the distances between frames in a short sequence and it provides high values when a sudden change is happening or low values in more homogeneous segments (Figure 8.2). Thresholding the function $C(i)$ will naturally lead to select frames that are very likely to be loci of sudden changes (Figure 8.3).

Thresholding the function $C(i)$ will naturally lead to select frames that are very likely to be loci of sudden changes (Figure 8.4).

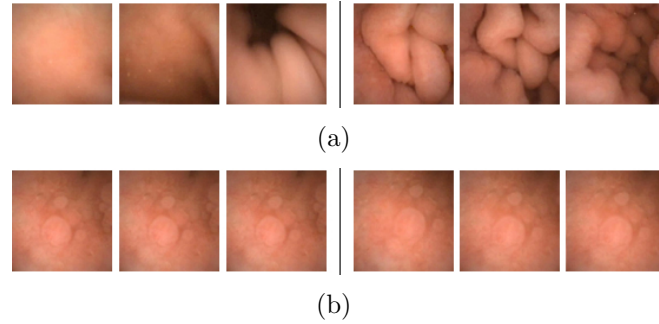


Figure 8.2: The two rows represent two sequences of consecutive frames. The row (a) is relative to the pylorus: function $C(i)$ for it takes the value 0.41. The row (b) is relative to a smooth portion of stomach: the $C(i)$ takes the value 0.21

8.1.4 Visual exploration of textons variability

A smart visualization of the frequency variations of the visual words across the video may also supply the medical specialists with a powerful and effective way to explore WCE data. Because our visual dictionary is made of 100 words, it is possible to represent the frequency of each word with a column of coloured pixels. Collecting all of these columns in a strip, one gets a direct visual representation of a whole video from the textons point of view. In this strip each pixel column is the histogram of a frame, and each row is the temporal sequence of values assumed by a texton. The “jet” colourmap has been chosen to provide sufficient contrast. In Figure 8.5(a) the textons are in the same arbitrary order produced by C-means clustering; Figure 8.5(b) shows the same histogram sequence after sorting the textons by row entropy, to bring together rows carrying more information. This improves the visual impact of the representation and helps to highlight both singular events and slow changes in the video; notice that the light stripe near the upper left corner of Figure 8.5(b) was completely invisible in Figure 8.5(a). In an interactive environment, a clinician could click on an interesting zone and

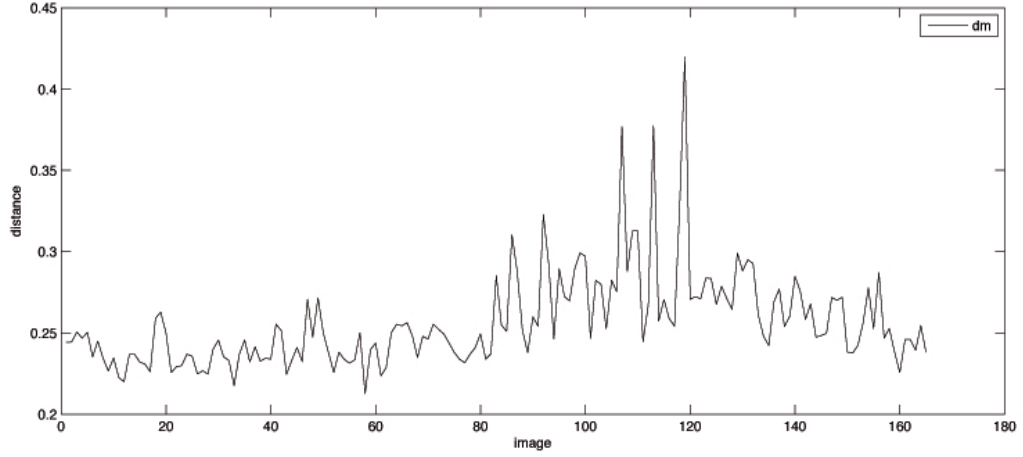


Figure 8.3: Plot of the function $C(i)$ for a WCE video sequence; peaks indicate loci of sudden change. The sequence in the upper row of Figure 8.2 corresponds to the maximum (interval of frames: 355-360) while the sequence in the lower row of Figure 8.2 corresponds to the minimum (interval of frames: 172-177)

immediately see the corresponding video frames for further investigations.

8.1.5 Experimental results

To assess the validity of the use of the indicator function $C(i)$ we have analyzed the performance of this index over 10 manually labelled sequences from patients of the Hospital "Maddalena Raimondi" in the period between 2005 and 2008. The manual labelling protocol has been the following. Let $(f_1 \dots f_N)$ be the sequence of frames in a video. We have formed the sequence of intervals $(I_1 \dots I_{\frac{N-3}{3}})$ where interval I_i is made of the six frames $(f_{3i-2} \dots f_{3i+3})$. For each interval the expert has judged if there is a significative change between the first 3 frames with respect to the last 3 frames. If this is the case the interval has been labelled as "event". Hence, in our setting an event is a relevant anatomical locus (esophagus, pylorus, etc.), a pathological presence (bleedings, ulcerations, etc.) or a common non pathological disturbance

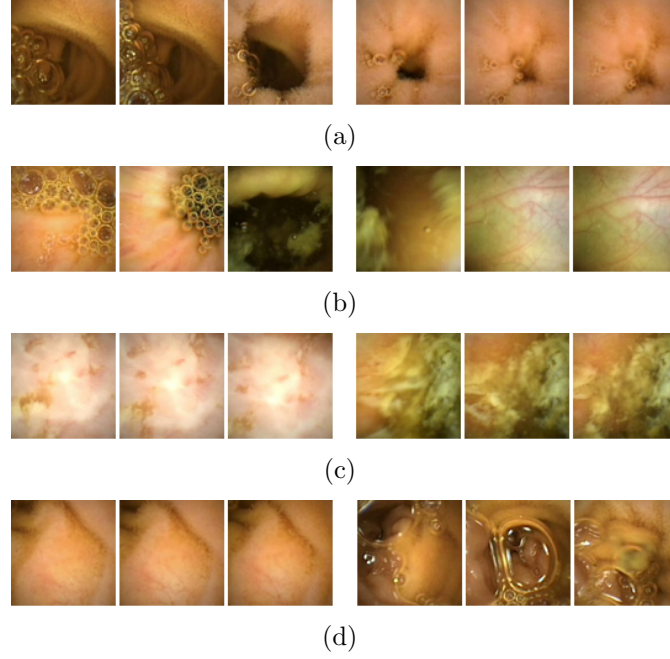


Figure 8.4: Examples of events among the WCE video frames; the row (a) corresponds to a pylorus; (b) is relative to the ileo-caecal valve; (c) represents frames with faecal residuals; (d) show the presence of bubbles.

(intestinal juices, bubbles, etc.) (Figure 8.4). Thresholding the normalized $C(i)$ function is a simple and direct way to eliminate a large percentage of the video frames from the need of medical direct visual inspection.

The performance of the index $C(i)$ has been tested as follows: intervals I_i of each video have been sorted according to the decreasing value of $C(i)$. We have hence partitioned the sorted I_i 's into ten groups of the same size. The first group contains the intervals with the top 10% $C(i)$ score and so on until the last group contains the interval with the lowest 10% $C(i)$ score. For each group we have counted the number of intervals that the expert has labelled as event (true events) (Table 8.1, Table 8.2).

We repeated the classification tests using three different settings for the bank of Gabor filters in the features extraction phase. More precisely we tried

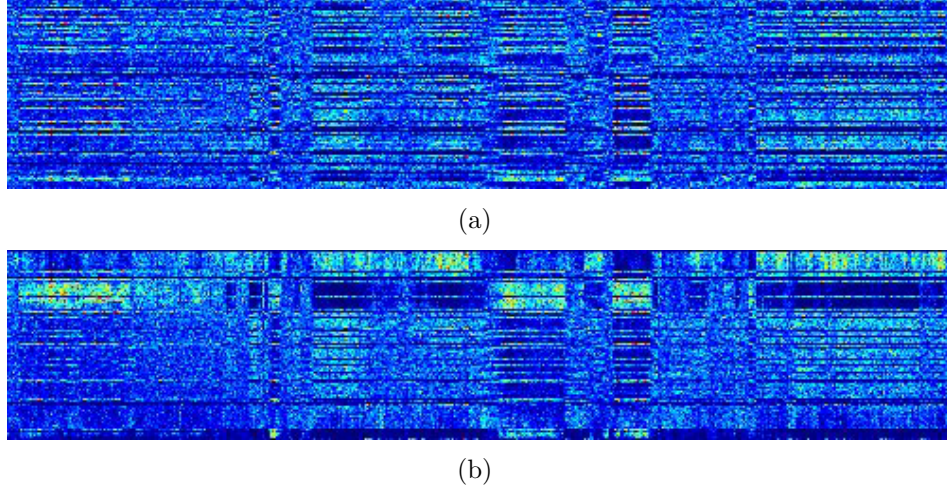


Figure 8.5: Compact representation of frequency variations of the visual words across a video. In (a) the rows are in arbitrary order, while in (b) they are sorted by row entropy.

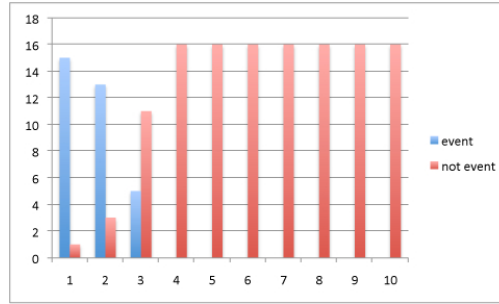


Figure 8.6: Percentage of events and not-events in a WCE video sequence partitioned into ten intervals representing the 10% of $C(i)$ score.

them separately with $\sigma_x = \sigma_y = 2, 4, 8$. We report also results considering the three scale together without appreciating important performances. As it is shown in Table 8.1 and Table 8.2 scale is not of great relevance; but in any case the best results have been obtained at the smallest scale ($\sigma_x = \sigma_y = 2$). It is evident from the experimental data that the proposed method may safely provide a filter to the clinician that could indeed concentrate the visual inspection on the intervals that score at the top 30% of the $C(i)$ index. Although this reduction is significative, the number of intervals that have to

Table 8.1: Percentage of true events in the ten groups made of all intervals sorted by $C(i)$ value. Each row represents the mean of the values for the ten videos examined

σ	10 th	9 th	8 th	7 th	6 th	5 th	4 th	3 rd	2 nd	1 st
2	95	76	14	1	1	0	1	0	1	0
4	93	69	23	2	1	0	1	1	0	0
8	94	66	25	3	1	0	1	0	1	0
2,4,8	93	71	17	7	1	1	0	1	0	0

Table 8.2: Presence in % of true events in the ten groups of intervals sorted by $C(i)$ value.

σ	10 th	9 th	8 th	7 th	6 th	5 th	4 th	3 rd	2 nd	1 st
2	51	40	7	1	0	0	0	0	0	0
4	50	37	11	1	0	0	0	0	0	0
8	51	34	13	1	1	0	1	0	1	0
2,4,8	50	38	8	3	0	0	0	0	0	0

be examined is still high for a human observer; however, it may allow the realistic application of computationally more expensive pattern recognition algorithms to this restricted set of intervals. The statistics relative to the test data are reported in Table 8.3.

Table 8.3: Values of Recall and Precision calculated for the ten video sequences at the top 20% score and top 30% score

Video	top 30%		top 20%	
	Precision	Recall	Precision	Recall
Video 1	100%	69%	85%	88%
Video 2	100%	50%	100%	75%
Video 3	93%	79%	78%	100%
Video 4	100%	52%	100%	78%
Video 5	94%	69%	80%	88%
Video 6	100%	67%	91%	91%
Video 7	100%	54%	100%	81%
Video 8	100%	64%	100%	94%
Video 9	97%	65%	88%	88%
Video 10	100%	52%	96%	75%
mean	98%	62%	83%	78%

8.2 Information Theory based WCE video summarization

In this experiment a method of automatically discriminating intestine tissue which can significantly speed-up the video analysis time is presented. The texton method is connected to the Normalized Compression Distance (NCD) [89] to create a robust binary classifier. The original contribution of this work is the use of an information theoretic approach to summarize meaningful changes in WCE image sequences inspired to [78].

8.2.1 The proposed method

As described in the previous section (8.1.1) images are pre-processed and features relative to colour and texture are extracted. In the early experiments NCD distance adopting several compression algorithms (dzip, gzip, etc.) has been tested. Although these algorithms are supposed to grant a good performance because of their ability to exploit the sequential redundancies in the data, their usage is costly. We found that, for the problem at hand, the gain obtained in this way is not relevant and for this reason a simplified (although rough) version of NCD based on Shannon's entropy is introduced:

$$NCD_{entropy}(x, y) = \frac{E(x, y) - \min((E(x), E(y)))}{\max(E(x), E(y))} \quad (8.2)$$

where $E(x)$ is the Shannon's entropy for the string x and $E(x, y)$ is the entropy of the concatenation of the string x and y . In the application considered here x is the string obtained concatenating the "symbols" made with the textons dictionary. In other words a frame from a WCE video is represented here as a sequence of visual words. Following a common practice in Com-

puter Vision we disregard the sequential order of the words and represent a frame as a "bag" of visual words. This observation justify the substitution of a compression algorithm with the much less expensive use of Shannon's entropy. The use of entropy in place of Kolmogorov complexity is not novel even in image domain, see for example [90, 91]. Observe that if sequentiality is disregarded, the entropy of the string of visual words obtained concatenating the representation of two frames is the entropy relative to the averaged histogram of the visual words frequencies in two frames.

For the WCE application, however, it makes sense to bias the difference between frames not only considering the visual differences but taking into account the proximity of the frames within the video. To this aim a new similarity distance SIM is introduced as follows:

$$SIM(x, y) = \alpha * NCD_{entropy}(x, y) + \beta * |i(x) - i(y)| \quad (8.3)$$

$i(x)$ and $i(y)$ represent the index of two frames in the video sequence and $\alpha + \beta = 1$. In this experiments the best results have been obtained with $\alpha = 0.8$ and $\beta = 0.2$.

In particular following the previous works, for each frame f_i , a new function $Score(i)$ is defined as follows:

$$Score(i) = \frac{1}{9} \sum_{k=i}^{i+2} \sum_{j=3}^5 SIM(f_k, f_{i+j}) \quad (8.4)$$

$Score(i)$ averages the distances between frames in a short sequence and it provides high values when there is an abrupt change or low values in segments with similar frames. Thresholding the function $Score(i)$ will lead to select interval of frames in which there is a sudden change in pattern.

8.2.2 Experimental results

In this section a number of experiments are undertaken in a real problem domain to demonstrate the efficacy of the proposed method. In our experiments we use ten video sequences provided by the “Maddalena Raimondi” Hospital. We use the labelling protocol explained in [83]. Let $(f_1 \dots f_N)$ be the sequence of frames in a video. We have formed the sequence of intervals $(I_1 \dots I_{\frac{N-3}{3}})$ where interval I_i is made of the six frames $(f_{3i-2} \dots f_{3i+3})$.

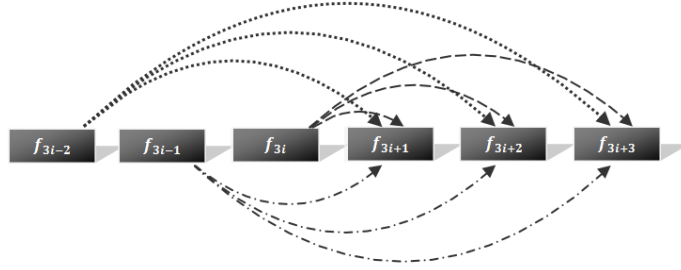


Figure 8.7: The computation of function $Score(i)$ (8.4)

In our setting an event includes every change in pattern in a short video sequence like a boundary transition, a pathology or a common disturbance like intestinal juices, residuals, bubbles, etc. For each interval the clinician has judged if there is a significative change between the first 3 frames with respect to the last 3 frames. If this is the case the interval has been labelled as an “event”. To grant greater robustness the labelling has been performed independently by two human experts. Only those intervals that both of them have labelled “event” are considered real event in the following experiments. The two independent labelling agree on 93% of the cases.

Intervals I_i of each video have been sorted according to the decreasing value of their $Score(i)$ indicator. We have hence partitioned the sorted I_i 's into ten groups of the same size. The first group contains the intervals with the top 10% of $Score(i)$, the last group contains the interval with the lowest

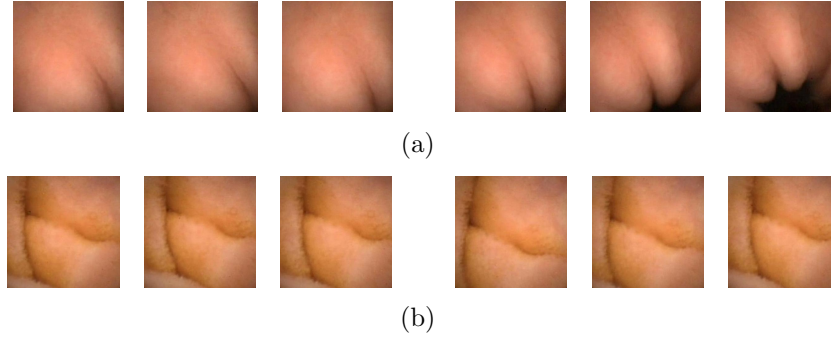


Figure 8.8: The two rows represent two sequences of consecutive frames. The row (a) represents an event. The row (b) is relative to an homogeneous tract.

10% of $Score(i)$. For each group we have counted the number of intervals labelled as event. In this experimental session two experts have labelled the sequences and the resulting ensemble has been given by their intersection.

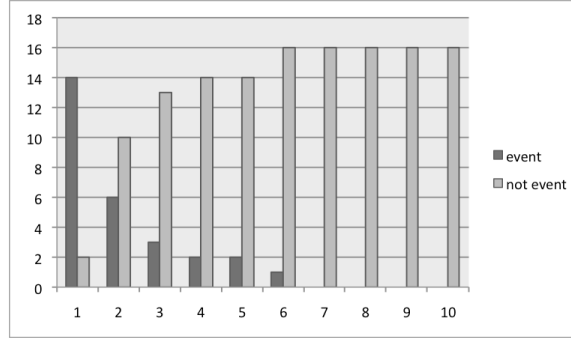


Figure 8.9: Percentage of events and not-events in a WCE video

The bar plot of Figure 8.9 shows the average percent of intervals that have been labelled as event vs the intervals that have been labelled not-event in the ten groups. The use of precision-recall analysis is investigated in Figure 8.10. As the ROC curve shows the discrimination obtained using the proposed method is comparable with the results in [83]. The slightly less robust discrimination shown by the novel method is justified by the proposed usage of $NCD_{entropy}$ instead of classical NCD . This loss in discrimination power is however justified by the greater efficiency that the usage of $NCD_{entropy}$

provides with respect to NCD .

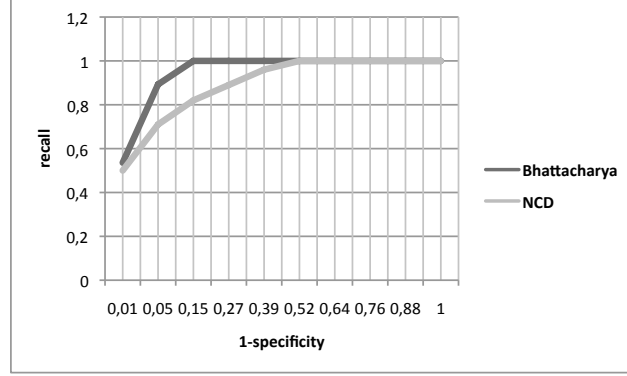


Figure 8.10: Two ROC curves compare the performance of tested methods

Table 8.4: Summary of experimental results

Intervals	$NCD_{entropy}$	$NCD_{\alpha\beta}$
top 20%	72%	71%
top 20%	68%	66%
top 30%	85%	86%
top 30%	53%	54%

In particular we compare the results obtained with the formula of $NCD_{entropy}$ (8.2), the modified version that uses the concept of entropy, and $NCD_{\alpha\beta}$ (8.3). Results are shown in Table 8.4.

Examples of images are shown in Figure 8.11. The first row corresponds to an event found in the first 10% until the last row corresponds to an event found in the sixth interval.

8.2.3 Conclusion

In this experiment we have presented an algorithmic information-theoretic method applied to find sudden changes in WCE video sequences. We used a modified formula NCD to compute the distance between the histograms





































f_{3i-2}	f_{3i-1}	f_{3i}	f_{3i+1}	f_{3i+2}	f_{3i+3}	$Score(i)$
						0.0291
						0.0203
						0.0158
						0.0133
						0.0131
						0.0112

Figure 8.11: Examples of events found with the proposed method

obtained with the textons approach explained in [83]. Experimental results have been shown that using the entropy, in combination with two parameters α and β , we reach a recall of 90% with a precision of 52% discarding the 30% of the video. Future works will extend the usage of NCD-like distance since the early stage of textons dictionary construction.

8.3 LBP based detection of intestinal motility in WCE images

Small intestine motility dysfunctions are shown to be related to certain gastrointestinal disorders which can be manifest in a varied symptomatology. Small intestine contractions are among the motility patterns which reveal many gastrointestinal disorders, such as functional dyspepsia, paralytic ileus, irritable bowel syndrome, bacterial overgrowth. Several techniques have been developed and tested in a wide range of modalities to analyze intestinal contractions. Nevertheless, all of these techniques suffer from important drawbacks: they are highly invasive, they usually generate patient discomfort, they need hospitalization and specialized staff, etc. In this section the use of Local Binary Pattern (LBP) combined with the powerful textons statistics, to find the frames of the video related to contractions, is proposed. Recognizing intestinal contractions from WCE image sequences provides a non-invasive method of measurement, and suggests a solution to the problems of traditional techniques for assessing intestinal motility [37]. Based on the characteristics of contractile patterns and information on their frequencies, the contractions can be investigated using essential image features extracted from WCE videos. The methodology proposed in this paper is made in two phases. The first implements the extraction of image features while in the second phase a textons-based [66] classifier is employed to perform the contraction detection task. In particular the definition and extraction of quantitative parameters from endoscopic images based on colour and texture information is at the core of the proposed technique. Several techniques for the texture analysis of images have been reported in the literature [52, 53]. In this section, the Local Binary Pattern (LBP) approach proposed by Ojala et

al. [54, 56], that provides highly discriminative texture information, is used. Among the advantages of LBP are its invariance to any monotonic change in gray level and its computational simplicity. Following [83] colour and high frequency energy content are included as features to build the model for each image. Feature evaluation in detecting positive examples of contractions has been performed by means of texton method, finally using the K-Nearest Neighbors (K-NN) with Bhattacharyya [72] distance to classify the images. Experiments have been conducted on over 6000 frames extracted from WCE videos. Results suggest that the technique is suitable for clinical applications. We also discuss the effects of various parameters on our classification algorithm such as the choice of filter bank, the size of the texton dictionary as well as the number of training images used. Furthermore we compare the performance of texton classifier with the Support Vector Machine (SVM) [70].

8.3.1 Contractions features description

Some pre-processing steps are applied before going ahead with any learning or classification. First, before convolving with any of the filter banks, a central region is cropped and retained from every image and the black background data and textual information discarded. All processing is done on these cropped regions and they are converted to HSI colourspace for the well known robustness in image analysis. Following [83] we divide each image into blocks of the same size and then we calculate HSI and the High Frequency Energy Content (HFC) for each block. In this paper we use LBP filter to characterize the texture of each block. In particular we use a filter bank with two parameters: radius 8 and neighbors 1, radius 16 and neighbors 4.

8.3.2 Texton-based classification

The goal of classification in general is to select the most appropriate category for an unknown object, given a set of known categories. To this purpose we build a training set containing positive and negative examples with different orientations and illuminations conditions. Hence the textons are learned using k-means clustering for the resulting feature vector. The histogram of textons forms the model corresponding to the training image. In the classification stage a novel image is classified by forming its histogram and then using a nearest neighbour classifier and the Bhattacharyya distance to pick the closest model.

Based on these experiments, it can be said that texton classification combined to multi-scale LBP features is a relatively effective classification method in wrinkled endoscopic images.

8.3.3 Experiments

This section investigates the performance of the textons method with the use of the LBP texture features to classify contractions images. The typical aspect of these frames is characterized by strong wrinkles of the folded intestinal wall, distributed in a radial way around the closed intestinal lumen. In the case of the visual pattern of the intestinal contractions we label the image as a positive example.

We perform our experiments to assess texture classification rates over 6000 images extracted from WCE videos. The original images have a dimension of 256×256 pixels. The cropped area is of 170×170 pixels and includes the square inscribed in the circular area. Each frame has been divided into 16×16 blocks and the vector feature has been extracted as we explained in

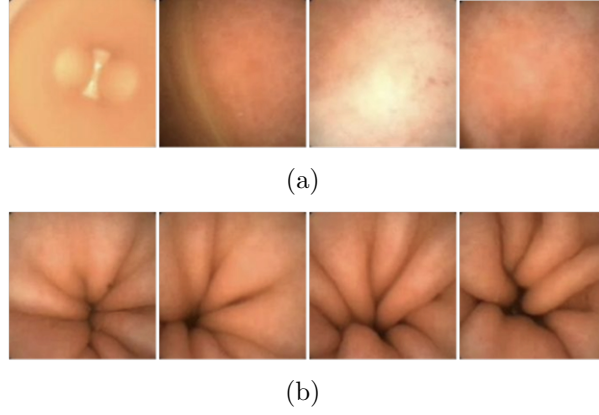


Figure 8.12: The row (a) is relative to sequence of not contractions. The row (b) is relative to frames that represent a contraction.

the previous section.

To avoid to have very similar visual words in the dictionary, we found that a suitable value for the number of textons is 25. This number is chosen to optimize the ratio of dispersion between cluster centers over the dispersion within clusters. In our experiments, the classification is performed by a K-NN classifier, with the value of $k = 1, 3, 5$. Bhattacharyya distance is used to measure the difference between histograms that represent, for each frame, the relative frequencies of each visual word in the dictionary.

In Figure 8.13 the visual texton representation of positives and negative examples is reported.

We compared results using ten different training datasets of 500 images extracted in random way from the whole set of 6000 frames. In Table 8.5 the percentage of positives and negatives, for each training set, are listed.

The effectiveness of our algorithm is measured by mean of the use of sensitivity and specificity. Results, for the ten extracted training set, are reported in Table 8.5. The features used are: H, S, I, HFC, $LBP_{8,1}$, $LBP_{16,4}$.

To provide a comparison, we repeated our experiment extracting texture

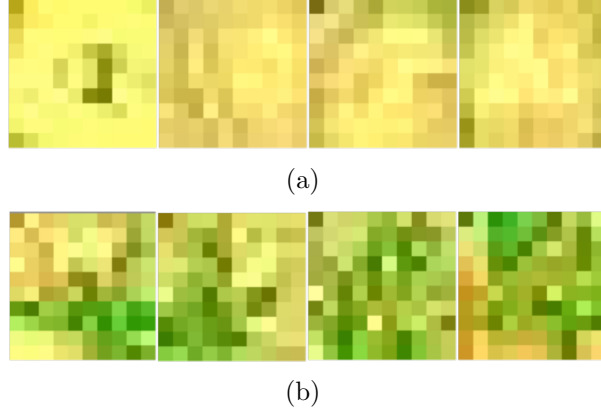


Figure 8.13: (a) textons in a sequence of not contractions. The row (b) is relative to textons present in contraction frames.

Table 8.5: Statistics over the ten training set

	Ts_1	Ts_2	Ts_3	Ts_4	Ts_5
positives	36%	35,4%	36,6%	39,2%	41,6%
negatives	64%	64,6%	63,4%	60,8%	58,4%
	Ts_6	Ts_7	Ts_8	Ts_9	Ts_{10}
positives	41,6%	39,4%	39,3%	39,4%	39,4%
negatives	58,4%	60,6%	60,6%	60,6%	60,6%

features using the Gabor filter bank with different parameters for σ , phase and orientation. In our settings we found that applying several gabor filters the feature vector present many repeated values. Hence we obtained that suitable parameters are: $\sigma_x = \sigma_y = 2$; phase= 0, 16; orientation = 0° , 45° . Results are reported in Table 8.7.

A comparison with the SVM classifier is written in Table 8.8. We used the Matlab implementation with the radial basis function.

We can conclude that results do not depend too much on the structure of the training set. Furthermore, the experiments show that using LBP texture features and Gabor features leads to classifier of comparable performance. LBP on the other hand is up to 10 times faster than Gabor filtering. Com-

Table 8.6: Sensitivity and specificity for different values of k for K-NN-classifier using LBP texture extraction

	Sensitivity			Specificity		
	$k = 1$	$k = 3$	$k = 5$	$k = 1$	$k = 3$	$k = 5$
Ts_1	60%	68%	72%	100%	100%	100%
Ts_2	66%	78%	76%	100%	100%	100%
Ts_3	62%	66%	66%	100%	100%	100%
Ts_4	68%	72%	78%	100%	100%	100%
Ts_5	74%	78%	76%	98%	100%	100%
Ts_6	74%	82%	80%	100%	100%	100%
Ts_7	74%	82%	78%	100%	100%	100%
Ts_8	74%	84%	92%	100%	100%	100%
Ts_9	64%	60%	68%	100%	100%	100%
Ts_{10}	58%	62%	76%	100%	100%	100%
mean	67,4%	73,2%	76,2%	99,8%	100%	100%
$\sqrt{\sigma^2}$	6	8	7	0	0	0

parison of texton based techniques with a standard SVM classifier shows moreover that this general methodology achieves greater sensitivity paying a high price in terms of number of false alarms. Our proposed solution has the advantage of reducing complexity and making the methodology more suitable for real-time application.

8.3.4 Conclusion and future work

In this paper a new method is developed for classification of intestinal contractions in a WCE video sequence. The good experimental results suggest its high potentiality to contribute to a totally intelligent auto-diagnosis endoscopy system. The main advantage of the proposed method is the computational simplicity of LBP operator, used for texture feature extraction, that permits applicable the technique in a real time modality. In our experiments we improve the results using ten different training sets and we can

Table 8.7: Sensitivity and specificity for different values of k for K-NN-classifier using Gabor filters

	Sensitivity			Specificity		
	k=1	k=3	k=5	k=1	k=3	k=5
Ts_1	60%	74%	74%	100%	100%	100%
Ts_2	60%	72%	82%	100%	100%	100%
Ts_3	56%	66%	68%	98,03%	100%	100%
Ts_4	84%	82%	90%	100%	100%	100%
Ts_5	78%	86%	90%	98%	100%	100%
Ts_6	60%	68%	76%	100%	98%	100%
Ts_7	66%	72%	80%	98,03%	100%	100%
Ts_8	78%	88%	86%	100%	100%	100%
Ts_9	56%	68%	74%	100%	100%	100%
Ts_{10}	72%	70%	76%	100%	100%	100%
mean	67%	74,6%	79,6%	99,6%	99,8%	100%
$\sqrt{\sigma^2}$	10	8	7	0	0	0

conclude that the algorithm does not depend on this selection. In this work we compare the proposed method with the usage of Gabor filters to extract texture features and SVM classifier to detect intestinal wrinkled frames. We demonstrated that the proposed method reach a sensitivity of about 80% and a specificity of about 99%. Future work will address in the extensions the results to recognize other kind of events (such as bleedings, cancer, polyps, etc.) in a video sequence. This will help the phycisian to reduce the time inspection and to make capsule endoscopy as a clinical routine.

Table 8.8: Sensitivity and specificity for the SVM classifier

	LBP		Gabor	
	Sensitivity	Specificity	Sensitivity	Specificity
Ts_1	76%	44%	94%	44%
Ts_2	96%	32%	90%	50%
Ts_3	96%	32%	84%	18%
Ts_4	88%	70%	94%	64%
Ts_5	98%	24%	82%	32%
Ts_6	96%	36%	98%	34%
Ts_7	96%	26%	74%	70%
Ts_8	100%	50%	100%	54%
Ts_9	96%	44%	100%	42%
Ts_{10}	100%	32%	89%	47%
mean	94,2%	42,2%	90,5%	45,5 %
$\sqrt{\sigma^2}$	7	16	8	15

8.4 Detection of intestinal motility using block-based classification

The goal of this experiments is the recognition of image contractions among a set of images, representing several scenarios, extracted from different videos.

In this set of experiments we present a novel method to discriminate contraction images in a video, classifying each block in the image into one of a set of fixed categories. We propose a 2-step approach to solve this problem, first estimating image classes through a preliminary block-based classification, which provides initial classification of the blocks as belonging to a fixed class, then performing recognition of contraction/not-contraction by classifying each image using the co-occurrence matrices derived from neighborhood relations. Extensive results are presented comparing several methods for blocks classification, image features extraction, etc. The performance of the method is evaluated across the number of hits reached, and finally is compared to the performance of the previous methods implemented. For these experiments the conclusion is that the method needs further investigations because it does not achieve the best results.

8.4.1 Preliminary classification

As previously described, images are pre-processed and for every frame only the central square is considered, discarding the black area and the textual information. The first step is the building of the training set, constituted by 40 images extracted from different WCE videos. Every image is divided into 25 blocks of 32×32 pixels and each block is labelled from the expert as belonging to a specific class, that are fixed as follows: 0 corresponds to a block representing normal mucosa; 1 corresponds to a region containing

a wrinkle; 2 is represented by an obscure region that can be assimilated to a lumen; 3 corresponds to an image with residuals and finally 4 presents intestinal bubbles.

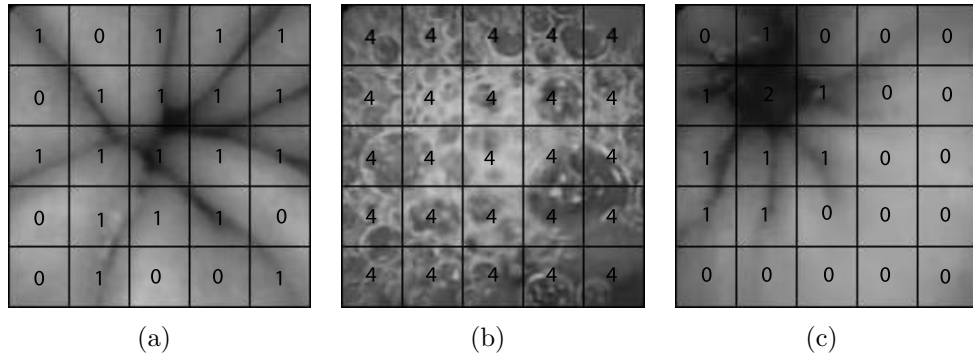


Figure 8.14: Samples of blocks labelling of the training set.

In Table 8.9 the statistics over the 40 images of the training set are reported.

Table 8.9: Percentage of classes in the training set

Img	class 0	class 1	class 2	class 3	class 4	Total
Total	38,1%	38,3%	1,3%	11,9%	10,4%	100%

Statistics show that most blocks represent normal mucosa or an arc. For every frame a further label (0,1) is assigned, it indicates if the image is a contraction or not. As control set a number of 5876 frames is considered. The classification step has been obtained extracting first the features for both training and control set and then applying a classification algorithm to reach the representation of the control set by mean of a set of visual words, according to the classes described above. The features considered for this set of experiments are chosen among: hue, saturation, intensity, high frequency energy content to extract information about colour. For texture extraction Gabor filters, LBP filters and Drog filters are alternatively applied.

Classification, achieved applying a leave-n-out cross-validation, is performed using essentially the standard K-NN and the Self Organizing Maps (SOM) obtaining a representation of every image by mean of the classification of his blocks.

8.4.2 Extracting spatial local features

Interpreting endoscopic images is still a significant challenge, especially since one single still image may not always contain enough information to make a robust diagnosis. To aid the physicians, some local feature-based retrieval methods are proposed to provide, given a query image, the recognition of contractions images. The central idea of this method is explained in [92] in which authors combine image retrieval and mosaicing for endomicroscopic images using a spatial criterion derived from the co-occurrence matrix of local features.

Bag of Visual Words (BVW) method has been successfully used in many applications of Computer Vision. For example, on a well-defined non medical application, by using this method on a large variety of images of natural or artificial textures.

The BVW method, as described in section 5, aims at extracting a local image description that is both efficient to use and invariant with respect to viewpoint changes, e.g., translations, rotations and scaling, and illumination changes, e.g., affine transformation of intensity. Its methodology consists in first finding and describing local features, then in quantizing them into clusters named visual words, and in representing the image by the histogram of these visual words. The BVW retrieval process can thus be decomposed into four steps: detection, description, clustering and similarity measuring, possibly followed by a classification step for image categorization. A problem

is that the spatial relationship between the local features is lost in the standard BVW representation of an image, whereas the spatial organization of blocks is highly discriminative in contraction images. To extract discriminative information over the entire image field, the proposed method measures a statistical representation of this spatial geometry. The spatial organization of the blocks can be included in the retrieval process because it could be substantial to differentiate contractions images. This is achieved by exploiting the co-occurrence matrix of the visual words labelling the local features in the image. Thus, we are able to store in a co-occurrence matrix M of size $K \times K$ the probability for each pair of visual words of being adjacent to each other. In Figure (8.15) an example of co-occurrence matrix building is illustrated.

In order to best differentiate the images of the contractions from other images, we looked at the most discriminative linear combination W of some elements m of M . Similarity is evaluated calculating the χ^2 and euclidean distance to find the model, in the training set, closest to the image query. Our results show that taking into account the spatial relationship between local features of images improves the retrieval accuracy.

8.4.3 Results

In this section we present our experimental results conducted over about 6000 images extracted randomly from several WCE videos. As previously described, training set is builded by 40 frames and the control set by 5876 frames, divided into 25 blocks. In assessing the quality of our experiments we first determined some parameters before running the preliminary classification. These parameters include the combination of features, the value of k for the K-NN for the initial classification of the blocks and the value of k for the K-NN for the final classification of frames. We tested that for the pre-

liminary classification the best value for k is 1 and for the final classification is 5. The dataset has been normalized before the preliminary classification.

Table 8.10: Percentage of hits reached using different features

Features	Hits
$H, S, I, hfc, Drog, LBP81, Gabor2200, Gabor221690, R, G, B$	76.41%
$H, S, I, hfc, Drog, Gabor2200, Gabor221690$	76.48%
$H, S, I, hfc, Gabor2200, Gabor221690, Gabor2232135$	71.80%
$R, G, B, hfc, Drog, Gabor_{2200}, Gabor221690$	72.39%

Our experimental results indicate that the techniques discussed here are promising but are not the best. Our future work in this area will include a better choice of features for the initial blocks classification. The main problem checked is that the labelling of a block of 32×32 pixel presents itself ambiguity. Hence the automatic classification of blocks reports many errors. Future work will explore the possibility to change the dimension of the blocks and to introduce the correlatons analysis to capture the relationship between the spatial correlation of all possible pairs of visual words as a function of distance in the image. Correlograms capture both local and global shape information, both short and long range spatial interactions. This makes correlograms suitable to represent the whole object or just a part of it.

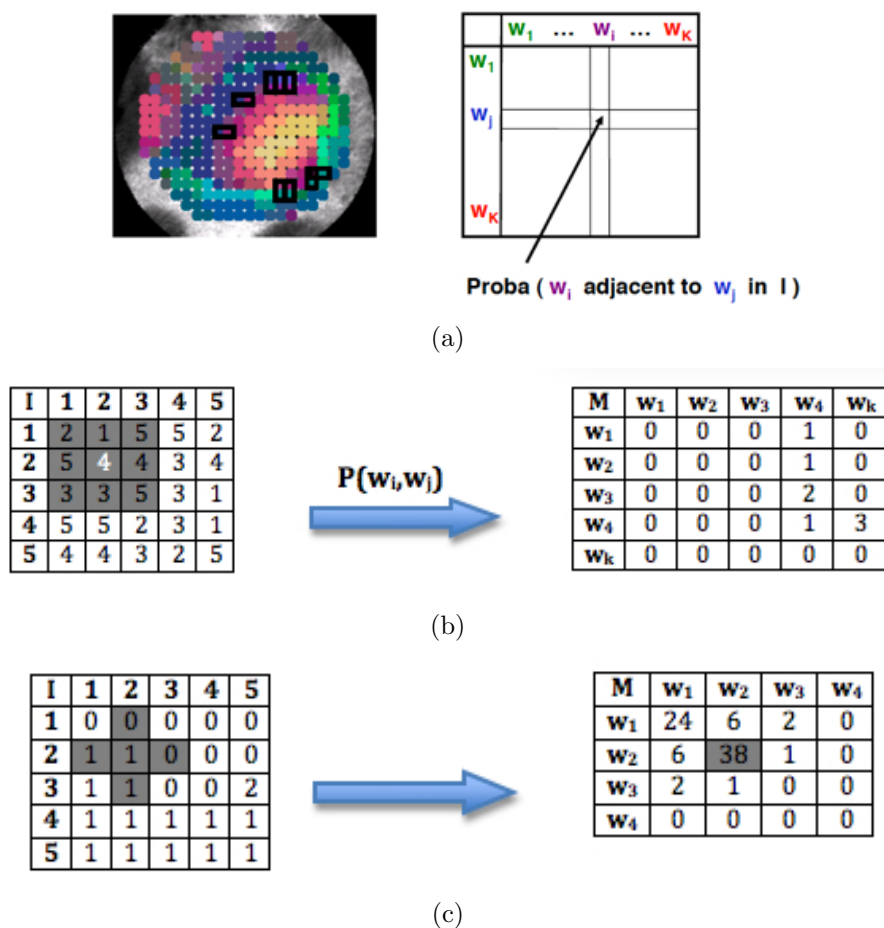


Figure 8.15: (a) The spatial relationship feature is given by the co-occurrence matrix of visual words. (b) Example of calculating the co-occurrence matrix in a 8-adjacency. (c) Example of calculating the co-occurrence matrix in a 4-adjacency following four directions ($0^\circ, 90^\circ, 180^\circ, 270^\circ$).

Chapter 9

CONCLUSION AND FUTURE WORK

In this thesis we have presented several distinctive methods to tackle the problem of automatic classification of image frames belonging to a WCE video. One of the limitation of the application of this diagnostic tool as a feasible routine is the long annotation time that each exam needs from a trained specialist. The duration of this assessment typically varies from one to two hours. Therefore, it is expected to substantially reduce the number of images to be manually analyzed to provide a diagnose proposal, allowing a more widespread use of WCE.

In this dissertation we focuses two areas: sudden changes discrimination and intestinal motility detection in a WCE video.

The general goal of automatic WCE segmentation is to split a video into shorter sequences each with the same semantic content. In particular we indicate with "event" an abrupt and significative change in the video. Our definition of event includes boundary transitions from an organ to another one, intestinal juices, bile, bubbles, pathologies, etc. Event detection is the

hardest and most important challenge from a clinical perspective because clinicians are interested to remove unimportant images and annotate only the relevant ones. To this aim we have constructed an indicator function that reveals a sudden change in a video. Several features are extracted and compared to build a robust classifier. The construction of the function uses the statistical texton approach. In this thesis we have presented an algorithmic information-theoretic method applied to find sudden changes in WCE video sequences. We also used a modified formula NCD to compute the distance between the histograms obtained with the textons approach. The best results can be achieved considering a combination of features related to colours, texture and energy information. The experiments have been demonstrated that the proposed method may eliminate up to 70% of the frames from further processing while retaining all the clinically relevant frames.

Intestinal motility is investigated in the second part of this work. Intestinal contractions may reveal the presence of different malfunctions. The definition and extraction of quantitative parameters from endoscopic images based on colour and texture information is at the core of the proposed technique. In our experiments we propose the use of the Local Binary Pattern (LBP) to analyze intestinal wrinkled patterns, presenting a comparative study of diverse features and classification methods. We also discuss the effects of various parameters on the classification algorithm such as the choice of filter bank, the size of the texton dictionary as well as the number of training images used. Furthermore we compare the performance of our texton classifier with a standard Support Vector Machine. We demonstrated that the proposed method reach a sensitivity of about 80% and a specificity of about 99%. The achieved high detection accuracy of the proposed system has provided thus an indication that such intelligent schemes could

be used as a supplementary diagnostic tool in endoscopy. Recognition of image contractions has been studied by classifying each image using the co-occurrence matrices derived from neighborhood relations. Extensive results are presented comparing several methods for block-based classification, image features extraction, etc. The performance of the method is evaluated across the number of hits reached, about the 71.80%, and finally is compared to the performance of the previous methods implemented. For these experiments the conclusions are that the method needs further investigations because it does not achieve the best results.

The presented tests showed promising results. Future work will address to achieve more sophisticated classification techniques. We think that the enrichment of features to characterize WCE images could help to achieve better results in classification.

Recognizing other kind of events (such as bleedings, cancer, polyps, etc.) in a video sequence will help the phycisian to reduce the time inspection and to make capsule endoscopy as a clinical routine. Finally, we are in a continuous feedback with the experts in order to improve the current methods, create optimal protocols and include faster and more efficient versions of our solutions for their use in a real clinical scenario in a close future.

Bibliography

- [1] Y. Haga, M. Esashi, Biomedical microsystems for minimally invasive diagnosis and treatment, *Proc. of IEEE* 92 (2004) 98–114.
- [2] G. Iddan, A. Glukhovsky, P. Swain, Wireless Capsule Endoscopy, *Nature* 405 (2000) 725–729.
- [3] G. Imaging, <http://www.givenimaging.com>, Ltd Israel.
- [4] G. Gay, M. Delvaux, J. Key, The role of video capsule endoscopy in the diagnosis of digestive diseases: A review of current possibilities, *Endoscopy* 36 (2004) 913–920.
- [5] P. Swain, *Gut* 54 (2005) 323–326.
- [6] A. Culliford, J. Daly, B. Diamond, M. Rubin, P. H. R. Green, The value of Wireless Capsule Endoscopy in patients with complicated celiac disease, *Gastrointestinal Endoscopy* 62 (1) (2005) 55–61.
- [7] Z. Fireman, A. e. a. Glukhovsky, Wireless Capsule Endoscopy, *Israel Medical Association Journal* 4 (2002) 717–719.
- [8] F. Rey, e. a. Gay, G., Guideline for video capsule endoscopy, *Endoscopy* 36 (2004) 656–658.

- [9] R. Eliakim, Wireless Capsule Video Endoscopy: Three years of experience, *World Journal of Gastroenterology* 10 (9) (2004) 1238–1239.
- [10] Olympus global <http://www.olympus-global.com/en/global/>.
- [11] S. Bar Meir, E. Bardan, Wireless Capsule Endoscopy pros and cons., *Israel Medical Association Journal* 4 (2002) 726.
- [12] D. G. Adler, C. Gostout, Wireless Capsule Endoscopy, *Hospital Physician* (2003) 14–22.
- [13] R. Franci, Sensitivity and specificity of the red blood identification (rbis) in video capsule endoscopy, *Proceedings of the 3rd International Conference on Capsule Endoscopy*, Miami, FL, USA,.
- [14] A. Fritscher-Ravens, C. P. Swain, The wireless capsule: new light in the darkness, *Digestive Diseases* 20 (2) (2002) 127–133.
- [15] J. Lee, J. Oh, S. K. Shah, X. Yuan, S. J. Tang, Automatic classification of digestive organs in Wireless Capsule Endoscopy videos, in: *SAC '07: Proceedings of the 2007 ACM symposium on Applied computing*, 2007.
- [16] M. Mackiewicz, J. Berens, M. Fisher, Wireless Capsule Endoscopy color video segmentation, *IEEE Transaction on Medical Imaging* 27 (12).
- [17] J. Berens, M. Mackiewicz, B. G.D., Stomach, intestine, and colon tissue discrimination for Wireless Capsule Endoscopy images, *SPIE Medical Imaging 2005: Image Processing* 5747 (2005) 283–290.
- [18] J. Berens, M. Mackiewicz, M. Fisher, B. G.D., Using colour distributions to dicriminate tissues in Wireless Capsule Endoscopy images, *Proceedings of Medical Image Understanding and Analyses Conference (MIUA '05)*.

- [19] M. Mackiewicz, J. Berens, M. Fisher, B. G.D., Colour and texture based gastrointestinal tissue discrimination, Proceedings of IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP '06) 2 (2006) 597–600.
- [20] M. Mackiewicz, J. Berens, M. Fisher, Wireless Capsule Endoscopy video segmentation using Support Vector Machine and Hidden Markov Models, Proceedings of Medical Image Understanding and Analyses Conference (MIUA '06).
- [21] J. Berens, M. Mackiewicz, B. G.D., C. Jamieson, Can we detect when a Wireless Capsule Endoscope leaves the stomach using computational colour techniques? a pilot study, Endoscopy 36 (1).
- [22] M. Mackiewicz, J. Berens, M. Fisher, B. G.D., C. Jamieson, Computational colour techniques can speed up the viewing of Wireless Capsule Endoscopy images as well as determine gastric and intestinal transit times (gtt and itt), Endoscopy 54 (2) (2005) A10.
- [23] M. Coimbra, S. C. J. P., MPEG-7 visual descriptors - contributions for automated features extraction in capsule endoscopy, IEEE Trans. on Circuits and System for Video Tech. 16 (5) (2006) 628–636.
- [24] M. Coimbra, P. Campos, S. C. J. P., Extracting clinical information from endoscopic capsule exams using mpeg-7 visual descriptors, Proc. of EWIMT '05 (2005) 105–110.
- [25] M. Coimbra, P. Campos, S. C. J. P., Topographic segmentation and transit time estimation for endoscopic capsule exams, ICASSP '06 2 (2006) 1164–1167.

- [26] M. Coimbra, J. Kustra, P. Campos, S. C. J. P., Combining color with spatial and temporal position of the endoscopic capsule for improved topographic classification and segmentation, SAMT '06.
- [27] P. Spyridonos, F. Vilarino, J. Vitria, F. Azpiroz, P. Radeva, Anisotropic features extraction from endoluminal images for detection of intestinal contractions, in: Proceedings of the 9th MICCAI, Copenhagen, Denmark, 2006.
- [28] F. Vilarino, L. I. Kuncheva, P. Radeva, ROC curves and video analysis optimization in intestinal capsule endoscopy, Pattern Recognition Letters 27 (8) (2006) 875–881.
- [29] F. Vilarino, P. Spyridonos, J. Vitria, P. Radeva, Self Organized Maps for intestinal contractions categorization with Wireless Capsule Video Endoscopy, Proceedings of the Third European Medical and Biological Engineering Conference, EMBEC'05.
- [30] F. Vilarino, P. Spyridonos, O. Pujol, J. Vitria, P. Radeva, Automatic detection of intestinal juices in Wireless Capsule Video Endoscopy, ICPR '06 4 (2006) 719–722.
- [31] B. Li, Q. Max, H. Meng, Texture analysis for ulcer detection in capsule endoscopy images, Image and Vision Computing 27 (2009) 1336–1342.
- [32] V. Kodogiannis, J. N. Lygouras, Neuro-fuzzy classification system for Wireless Capsule Endoscopic images, World Academy of Science, Engineering and Technology 45 (2008) 620–628.
- [33] S. Karkanis, G. Magoulas, D. Iakovidis, D. Maroulis, N. Theofanous, Tumor recognition in endoscopic video images, Proceedings of the 26th EUROMICRO Confernce, Netherlands (2000) 423–429.

- [34] S. Krishnan, P. Wang, C. Kugean, M. Tjoa, Classification of endoscopic images based on texture and neural network., Proceedings of the 23rd Annual IEEE International Conference in Engineering in Medicine and Biology 4 (2001) 3691–3695.
- [35] S. Krishnan, P. Goh, Quantitative parameterization of colonoscopic images by applying fuzzy technique, Proceedings of 19th International Conference IEEE/EMBS 3 (1997) 1121–1123.
- [36] S. Krishnan, C. Yap, K. Asari, P. Goh, Neural network based approaches for the classification of colonoscopic images, Proc. of 20th International Conference on IEEE Engineering in Medicine and Biology Society 2 (1998) 1678–1680.
- [37] S. Krishnan, X. Yang, K. Chan, S. Kumar, P. Goh, Intestinal abnormality detection from endoscopic images, International Conference of the IEEE on Engineering in Medicine and Biology Society 2 (1998) 895–898.
- [38] S. Krishnan, X. Yang, K. Chan, P. Goh, Region labeling of colonoscopic images using fuzzy logic, Proceedings of the First Joint BMES/EMBS Conference Serving Humanity, Advancing Technology 2 (1999) 1149.
- [39] C. Lima, J. Correia, J. Ramos, D. Barbosa, Texture classification of images from endoscopic capsule by using MLP and SVM. a comparative approach, World Congress on Medical Physics and Biomedical Engineering, 25 (5) (2009) 271–275.
- [40] S. C., F. Villa, E. Rondonotti, C. Abbiati, G. Beccari, R. De Franchis, Sensitivity and specificity of the suspected blood identification system in video capsule enteroscopy, Endoscopy 37 (12) (2005) 1170–1173.

- [41] S. A. Zanati, S. J. Tang, E. e. a. Dubcenco, Value of the suspected blood indicator in Wireless Capsule Endoscopy, *Gastrointestinal Endoscopy* 59 (5) (2004) 167.
- [42] P. N. D'halluin, M. Delvaux, L. M. G. et al., Does the suspected blood indicator improve the detection of bleeding lesions by capsule endoscopy?, *Gastrointestinal Endoscopy* 61 (2) (2005) 243–249.
- [43] S. Hwang, J. Oh, J. Cox, S. J. Tang, H. F. Tibbals, Blood detection in Wireless Capsule Endoscopy using expectation maximization clustering, *SPIE Medical Imaging 2005: Image Processing* 6144 (2006) 11.
- [44] P. Y. Lau, P. L. Correia, Detection of bleeding patterns in WCE video using multiple features, *Proc. of the 29th Annual International Conference of the IEEE EMBS* (2007) 5601–5604.
- [45] B. Li, H. Meng, Analysis of Wireless Capsule Endoscopy images using chromaticity moments, *Proc. of IEEE International Conference on Robotics and Biomimetics* (2007) 87–92.
- [46] B. Li, H. Meng, Computer-based detection of bleeding and ulcer in Wireless Capsule Endoscopy images by chromaticity moments, *Computers in Biology and Medicine* 39 (2009) 141–147.
- [47] B. Penna, T. Tillo, M. Marco Grangetto, E. Magli, G. Olmo, A technique for blood detection in Wireless Capsule Endoscopy images, *17th European Signal Processing Conference (EUSIPCO 2009)* (2009) 1864–1868.
- [48] J. Liu, X. Yuan, Obscure bleeding detection in endoscopy images using support vector machines, *Optim Eng* 10 (289-299).

- [49] A. A. Al Rahayfeh, A. A. Abuzneid, Detection of bleeding patterns in WCE images, *The International Journal of Multimedia and its Applications (IJMA)* 2 (2) (2010) 1–10.
- [50] V. Hai, T. Echigo, R. Sagawa, Adaptive control of video display for diagnostic assistance by analysis of capsule endoscopic images, *Proceedings of the 18th International Conference on Pattern Recognition (ICPR '06)* 3 (2006) 980–983.
- [51] Q. A. Qureshi, Current and future applications of the capsule camera, *Nature Reviews Drug Discovery* 20 (2) (2004) 447–450.
- [52] R. Haralick, L. Shapiro, Image segmentation techniques, *Computer Vision Graphics Image Processing* 29 (1985) 100–132.
- [53] C. T. Gabor, K. M. Sanjit, *Modern filter theory and design*, Wiley. John & Sons.
- [54] M. P. T. Ojala, M. Pietikainen, D. Harwood, A comparative study of texture measures with classification based on feature distributions,, *Pattern Recognition* 29 (1996) 51–59.
- [55] M. P. T. Ojala, T. Maenpaa, Multiresolution grey-scale and rotation invariant texture classification with local binary patterns, *IEEE Trans. Pattern Anal. Mach. Intell.* 24 (7) (2002) 971–987.
- [56] M. P. T. Ojala, K. Valkealahti, E. Oja, M. Pietikainen, Texture discrimination with multidimensional distributions of signed gray-level differences, *Pattern Recognition* 39 (3) (2001) 727–739.

- [57] R. M. Haralick, K. Shanmugam, I. Dinstein, Textural features for image classification, *IEEE Transactions on Systems, Man and Cybernetics* 3 (1973) 610–621.
- [58] D. A. Clausi, An analysis of co-occurrence texture statistics as a function of gray level quantization, *Can. J. Remote Sensing* 28 (1) (2002) 45–62.
- [59] Z. Wang, A. C. Bovik, L. Lu, Why is image quality assessment so difficult?, *Acoustics, Speech and Signal Processing, IEEE International Conference* 4.
- [60] B. Julesz, Textons, the elements of texture perception, and their interactions, *Nature* 290 (5802) (1981) 91–97.
- [61] M. Tuceryan, A. K. Tuceryan, *Handbook of Pattern Recognition and Computer Vision*, World Scientific Publishing Co., Inc., River Edge, NJ, 1993.
- [62] R. M. Haralick, Statistical and structural approaches to texture., *Proc. of IEEE* 65 (5) (1979) 786–804.
- [63] L. Van Gool, P. Dewaele, A. Oosterlinck, Texture analysis, *Comput. Vision Graphics Image Process.* 29 (1985) 336–357.
- [64] R. M. Haralick, L. Shapiro, *Computer and robot vision*, Addison Wesley.
- [65] T. R. Reed, J. M. H. Du Buf, A review of recent texture segmentation and feature extraction techniques, *CVGIP: Image Understanding* 57 (3) (1993) 359–372.
- [66] M. Varma, A. Zisserman, A statistical approach to texture classification from single images, *International Journal of Computer Vision* 62 (1–2) (2005) 61–81.

- [67] T. Leung, J. Malik, Representing and recognizing the visual appearance of materials using three-dimensional textons, *International Journal of Computer Vision* 43 (1) (2001) 29–44.
- [68] C. Schmid, Constructing models for content-based image retrieval, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2 (2001) 39–45.
- [69] O. G. Cula, K. J. Dana, 3d Texture recognition using bidirectional feature histograms, *International Journal of Computer Vision* 59 (1) (2004) 33–60.
- [70] R. Duda, P. Hart, S. D.G., *Pattern Recognition*, 2000.
- [71] B. Dasarathy, *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques.*, ISBN 0-8186-8930-7, 1991.
- [72] A. Bhattacharyya, On a measure of divergence between two statistical populations defined by probability distributions, *Bull. Calcutta Math. Soc.* 35 (1943) 99–109.
- [73] E. Wilson, M. Hilferty, The distribution of chi-square, *Proceedings of the National Academy of Sciences, Washington* 17 (1931) 684–688.
- [74] C. E. Shannon, Prediction and entropy of printed english, *The Bell System Technical Journal* (1951) 30–64.
- [75] C. Bennett, P. Gacs, P. Vitanyi, W. H. Zurek, Information distance, *IEEE Trans. Information Theory* 44 (1998) 1407–1423.
- [76] R. Cilibrasi, P. Vitanyi, Clustering by compression, *IEEE Trans. Information Theory* 51 (2005) 1523–1545.

- [77] M. Li, X. Chen, P. Vitanyi, The similarity metric, *IEEE Trans. Information Theory* (2004) 3250–3264.
- [78] A. Cohen, C. S. Bjornsson, S. Temple, G. Banker, B. Roysam, Automatic summarization of changes in biological image sequences using algorithmic information theory, *IEEE Trans. Pattern Analysis and Machine Intelligence* 31 (8) (2009) 1386–1403.
- [79] J. O’ Connor, E. F. Robertson, Andrey Nikolaevich Kolmogorov, School of Mathematics and Statistics, University of St. Andrews, Scotland.
- [80] U. Shoning, P. Randall, *Gems of theoretical computer science*, Springer Verlag.
- [81] A. Church, A set of postulates for the foundation of logic, *Annals of Mathematics*, second series 33 (1932) 346–366.
- [82] A. Turing, On computable numbers with an application to the entscheidungsproblem, *Proc. London Math. Soc.* 2 (42) (1936) 230–265.
- [83] G. Gallo, E. Granata, G. Scarpulla, Wireless Capsule Endoscopy video segmentation, *IEEE International Workshop on Medical Measurements and Applications* (2009) 236–340.
- [84] G. Gallo, E. Granata, G. Scarpulla, Sudden changes detection in WCE video, *International Conference on Image Analysis and Processing* 5716 (2009) 701–710.
- [85] G. Gallo, E. Granata, WCE video segmentation using textons, *SPIE Medical Imaging 2010 Image Processing* 7623 (2010) 76230X.
- [86] G. Gallo, E. Granata, A. Torrisi, Information theory based WCE video summarization, *ICPR 2010* (2010) 4198–4201.

- [87] G. Gallo, E. Granata, LBP based detection of intestinal motility in wce images, SPIE Medical Imaging 2011: Image Processing.
- [88] Y. Du, C. I. Chang, P. Thouin, An unsupervised approach to color video thresholding, International Conference on Multimedia and Expo 3 (2003) 337–340.
- [89] C. Bennett, P. Gacs, L. Ming, P. Vitanyi, W. Zurek, Information distance, IEEE Trans. Information Theory 44 (4) (1998) 1407–1423.
- [90] A. Bardera, M. Feixas, I. Boada, M. Sbert, Compression-based Image Registration, IEEE International Symposium on Information Theory (2006) 436–440.
- [91] A. Kaltchenko, Algorithms for estimating distance with application to bioinformatics and linguistics, <http://xxx.arxiv.cornell.edu/abs/cs.CC/0404039>.
- [92] B. Andr , T. Vercauteren, A. Perchant, M. B. Wallace, A. M. Buchner, N. Ayache, Introducing space and time in local feature-based endomicroscopic image retrieval, Proceedings of the MICCAI Workshop - Medical Content-based Retrieval for Clinical Decision 5853 (2009) 18–30.