CHAPTER 2: THEORETICAL BACKGROUND AND LITERATURE REVIEW

This chapter presents the theoretical background of concepts used throughout the research and a review of quality research papers published in refereed journals with search criteria as capsule endoscopy, engineering, and journal papers. A rigorous review of the available literature is conducted with the aim to figure out the following aspects:

- Identification of research component.
- Study of performance assessment measures involved.
- Current progress and challenges.
- The possible scope of improvement and future work.

Based on the above search criteria 62 papers are studied and categorized as:

- Video summarization and redundant image elimination.
- Image enhancement and interpretation.
- Segmentation and region identification.
- Computer-aided abnormality detection.
- Image and video compression.

Before we can embark on a detailed literature review, few technological pre-requisites are required to be discussed. Most of them are utilized in subsequent chapters also. The following section discusses the pre-requisites and subsequent section presents the need, methodology, strengths and limitations of all five categories of work as mentioned above.

2.1 Theoretical Background

2.1.1 Features

Features are the key components to identify an image. The features must be optimal in size and should be able to distinguish between normal and abnormal images. Since different abnormalities exhibit different properties, a wide range of features is used as observed in this study. Figure 2.1 shows the image of a few abnormalities [9].

(a) (b) (c)

Figure 2.1: Images of Various Abnormalities (a) Bleeding (b) Ulcer (c) Angioectasia

Textural, color, geometric, spatial, edge, temporal, and generic features are amongst the most used features. Novel hybrid features like curvlet based LBP, distance map and texton number is also experimented in a few publications. The significant features are extracted to create a feature vector which is then fed to a classifier. Lack of distinguishable features is one of the causes of failure of automatic abnormality detection in CE [10]. The CE images of GI tract diseases exhibit a wide range of color and texture. To address all the diseases under this study, the proposed feature set comprises features from the wavelet domain along with color, texture, and shape features from the spatial domain. Figure 2.2 presents a hierarchical idea of the feature set.

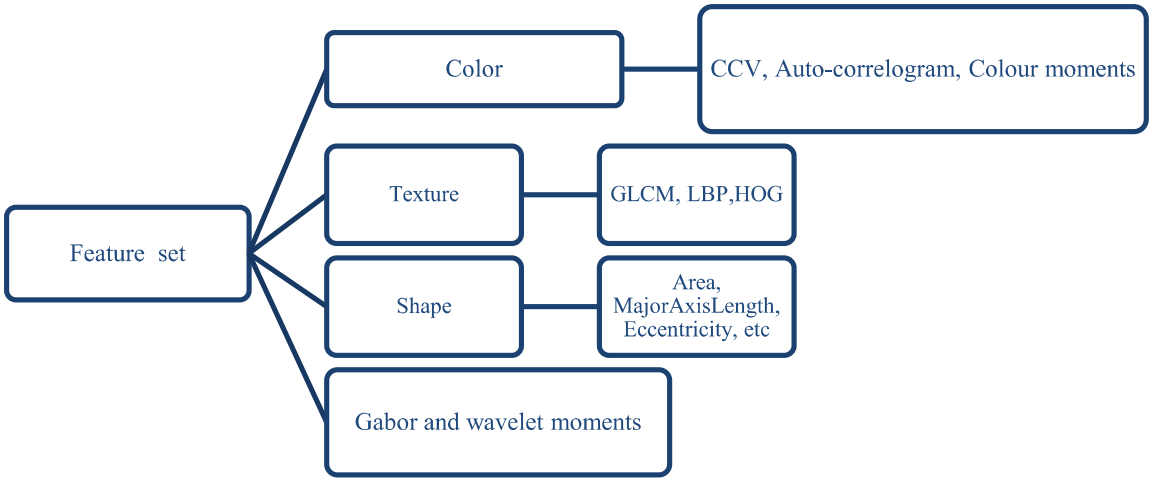

Figure 2.2: Hierarchical representation of the feature set

Before we discuss these hand crafted features in detail, it is worth noting that features can also be extracted from a pre-trained CNN model such as ResNet, GoogLeNet etc. by simply supplying the images at the input layer. The output of the last layer just before classification gives us the features. This is referred as the feature map. This feature map then acts as an input for classification. Thus either hand-crafted features or a feature map obtained from a pre-trained CNN can be fed to a classifier. This study makes use of both these approaches in the subsequent chapters.

2.1.1.1 Texture Features

The texture is a very interesting image feature that has been used for the classification of images. The texture is the repetition of a pattern or patterns over a region in an image. Brief details of some dominating texture features are discussed here.

2.1.1.1.1 Histogram of Oriented Gradients (HOG) and Local Binary Pattern (LBP)

CE images exhibit very discriminative texture and color properties. HOG as a feature descriptor deals with ambiguities related to texture and color [11]. Distribution (histogram) of intensity gradients can better describe the appearance of the object and shape within the image.

HOG captures the edge or gradient structure that is very characteristic of local shape [12]. After dividing the image into cells, a histogram of gradient directions is computed for each pixel in the cell. All the cells within a block are normalized, and concatenation of all these histograms is the feature descriptor. Figure 2.3 shows a sample CE image and visualization of the HOG descriptor. Uniform local binary patterns are fundamental properties of the texture of an image, and their occurrence histogram is a very powerful textural feature [13]. CE images suffer from illumination variations due to limited illumination capacity, limited range of vision inside GI tract and motion of the camera. It is learned that LBP performs robustly to illumination variations. A 3x3 neighborhood would produce up to $2^8 = 256$ local texture patterns. The texture feature descriptor is the LBP histogram of 256 bin occurrences calculated over the region. For pixel (x_c, y_c), (2.1) defines the LBP number of p members on the circle of radius r:

$$LBP_{p,r} = \sum_{p=0}^{p-1} s(g_p - g_c)2^p \tag{2.1}$$

Where s(x) is the sign function. The obtained LBP number as per (2.2) contain transitions from 0 to 1. Based on these, uniform patterns a novel rotation invariant LBP is proposed in [13] as:

$$LBP_{p,r}^{riu} = \begin{cases} \sum_{p=0}^{p-1} s(g_p - g_c) & \text{if } U(LBP_{p,r}) \leq 2 \\ p+1 & otherwise \end{cases} \tag{2.2}$$

$$where, U(LBP_{p,r}) = \left|s(g_{p-1} - g_c) - s(g_0 - g_c)\right| + \sum_{p=1}^{p-1} \left|s(g_p - g_c) - s(g_{p-1} - g_c)\right|.$$

Patterns are uniform if they contain at most two transitions on a circular ring from 0 to 1 or 1 to 0. Examples of uniform patterns are 11111111 (nil transitions), 01000000 (2 transitions). Thus, in a circularly symmetric neighbor set of p pixels, a total of p+1 uniform binary patterns are found.

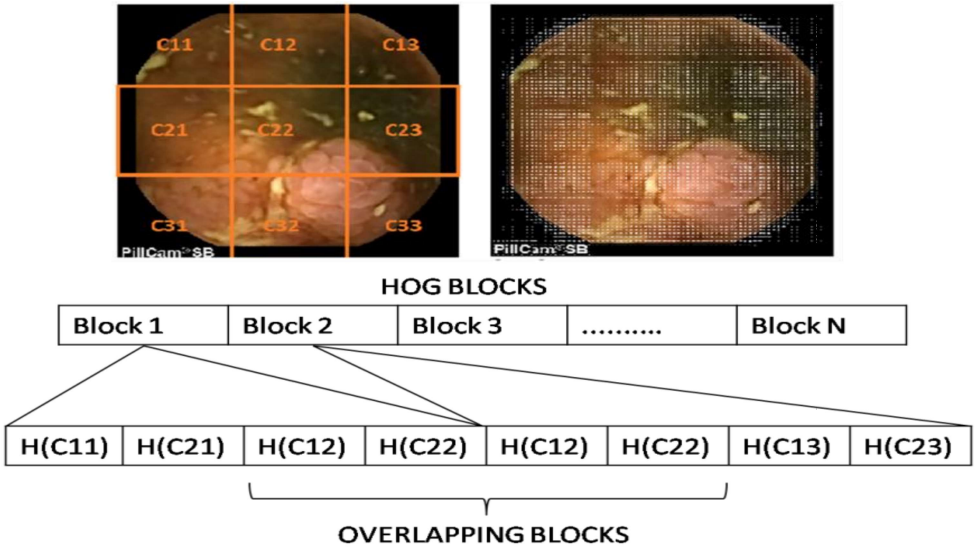

Figure 2.3: Sample CE image and HOG descriptor visualization

2.1.1.1.2 Gray level Co-occurrence Matrix (GLCM)

GLCM is a statistical approach for computing the co-occurrence probability of different combinations of grey levels in an image. The matrix element *G(p, q |Δx, Δy)* is the relative frequency, where two pixels are separated by a pixel distance (Δx, Δy) within a given neighborhood, one with intensity *p* and the other with intensity *q*. Let *I*(*x*, *y*) is an image with size *M*×*N and* gray levels *g* ranging from 0 to *g-1*. Then GLCM matrix for an image *I*, parameterized by an offset $(\Delta x, \Delta y)$ is given as:

$$G_{\Delta x,\Delta y}(p,q) = \sum_{x=1}^{M}\sum_{y=1}^{N}\begin{cases}1, \text{if } I(x,y) = p \text{ and } I(x+\Delta x, y+\Delta y) = q \\ 0, \qquad\qquad \text{otherwise}\end{cases} \tag{2.3}$$

2.1.1.2 Wavelet Transform based Features

Wavelet transform is a signal processing technique extensively used in texture analysis and extraction of visual texture features based on the multi-resolution decomposition of the images and representing textures in different scales. Wavelet transform transforms the images into a multi-scale representation with lower

computational cost. When we apply discrete wavelet transform (DWT) to the input images, it decomposes the images into four parts (LL, LH, HL and HH). Further, low-low sub-bands are decomposed and repeat for LL sub-band as desired number of decomposition as per requirement. Statistical features like mean, standard deviation, skewness, kurtosis etc. of the transform coefficients are used as a feature vector.

2.1.1.3 Gabor Filter based Features

Gabor filter is an example of linear wavelet filters, capturing energy at a specific frequency and a specific direction. It is frequently used in many image processing applications such as; synthesis of images, segmentation, edge detection, pattern recognition and many more. In all such applications, it is necessary to analyse the spatial frequency parts of an image in a localized manner using a Gaussian envelope. Frequency and orientation representations of Gabor filters are similar to those of the human visual system, and they have justified being appropriate for extracting useful texture features from an image. A two dimensional Gabor function $g(x, y)$ is defined as:

$$g(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} exp\left[-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right) + jw(xcos\theta + ysin\theta)\right] \qquad (2.4)$$

where space constants σ_x and σ_y define the Gaussian envelope along the *X and* Y-axes, w is modulation frequency $and\ \theta$ is orientation.

After applying Gabor filters on the images with a different orientation at different scale, we obtain an array of transformed coefficients. Mean square energy, Mean amplitude of these Gabor coefficients are used to represent the homogenous texture feature of the region.

2.1.1.4 Shape Features

The *shape* is another image feature applied in this study. The shape feature tries to quantify the pattern in such a way that it resembles the most with the human intuition. Shape features should be invariant to *translation, rotation,* and *scale,* for an effective system, when the arrangement of the objects in the image is not known in advance. To use *shape* as an image feature, it is essential to segment the image to detect object or region boundaries; and this is a challenge. Techniques for shape characterization can be divided into two categories. The first category is *boundary-based,* using the outer contour of the shape of an object. The second category is *region-based,* using the whole shape region of the object. The most prominent representatives of these two categories are Fourier descriptors, chain code, polygon approximation, moment invariants, curvature scale space descriptor, angular radial transform, image moments, and geometric features.

2.1.1.5 Color Features

Color is one of the most widely used visual features. While we can perceive only a limited number of gray levels, our eyes are able to distinguish thousands of colors and a computer can represent even millions of distinguishable colos in practice. Colour has been successfully applied to medical imaging because it has very strong correlations with the abnormalities. Moreover, colour feature is robust to background complications, scaling, orientation, perspective, and size of an image. There are variants of color features are used in literature, some introduction of few are as follow:

2.1.1.5.1 Colour Histogram

Colour histogram is one of the most important descriptors used in medical image processing; which shows, how many pixels in an image are of a particular color. The color histogram is represented as bar chart, where each bar (bin) represents a particular

color of the color space being used. For an M×N image I, the colors in that image are quantized to Q_1, Q_2…Q_{32}. The color histogram H (I) = [H_1, H_2, …, H_{32}], where H_i represents the number of pixels in color Q_i. The color histogram also represents the possibility of any pixel, in image I, that in color Q_i.

$$Probality(Prob \in Q_i) = \frac{H_i}{M \times N} \quad (2.5)$$

2.1.1.5.2 Colour Coherence Vector

One problem with the color histogram-based similarity measure approach is that the global color distribution doesn't reflect the spatial distribution of the color pixels locally in the image. This cannot distinguish whether a particular color is sparsely scattered all over the image or it appears in a single large region in the image. The color coherence vector-based approach was designed to accommodate the information of spatial color into the color histogram. Here we can classify each pixel in an image, based on whether it belongs to a large uniform region.

2.1.1.5.3 Color Moments

The color moment is a compact representation of color features to discriminate a color image. It has been shown that most of the color distribution information is captured by the three low-order moments. The average color value in the image and Standard deviation shows the color deviation from the mean, and skewness shows the degree of asymmetry in distribution.

2.1.1.5.4 Colour Correlogram

The weak point of the histogram method is the lack of space information in color. color correlogram is a technique proposed to integrate spatial information with color histograms. For each pixel in the image, the correlogram approach needs to go through

all the neighbors of that pixel. So the color correlogram shows how the spatial autocorrelation of color changes with distance.

2.1.1.5.5 Chromaticity Moments

This feature has does not require long histogram, only a small number of features, called chromaticity moments, are required to capture the spectral content (chrominance) of an image. Chromaticity moments are characterized by their two-dimensional shape and two-dimensional distribution.

2.1.2 Mapping of Relevant Features with Abnormalities

Table 2.1 presents a general observation of mappings between the various CE images and features. The extracted features are fed to a classifier for CAD of CE images.

Table 2.1 General observation of mappings between the various CE images and features

CE image type	Sample Image	Relevant features
Bleeding		Statistical measures, color and textural features.
Ulcer		Chromatic moments, textural, geometric, a combination of color and texture features
Angioectasia		Color and textural features
Normal		Textural, combination of statistical measures with DWT transform bands and Dif Lac analysis

General prediction about the presence or absence of abnormality uses features such as textural, combination of statistical measures with DWT transform bands and Dif Lac analysis. *Ulcer* detection uses features such as chromatic moments, textural, geometric, a combination of statistical measures over different color planes and the combination of color and texture features. *Motility analysis* uses features such as cluster-based, spatial, edge, temporal, a combination of color and textural and generic ones. *Bleeding* detection uses features such as a combination of statistical measures over different color space, cluster-based, color and textural features.

2.1.3 Performance Metrics

The common performance metrics are discussed here so that the literature review portion can be understood with a hassle. While most of the metrics are explained in this section a few topic specific performance criteria is discussed in the respective chapter of the thesis.

2.1.3.1 Coverage: The percentage of findings of interest represented by at least one frame in the video is called coverage. It can be estimated as follows:

$$Coverage = \frac{\sum_{i=1}^{n} C_i}{n} \tag{2.6}$$

Where $C_i = 1$ if at least one frame from the i^{th} finding exists; 0 otherwise; n= total number of abnormal findings.

2.1.3.2 Confusion Matrix: A *confusion matrix* is a useful tool for analyzing how well your classifier can recognize images of different classes:

Table 2.2 Confusion matrix to measure the performance of the classifiers

Predicted Class			
	Positive	Negative	Total
Positive	TP	FN	P
Negative	FP	TN	N

The performance of the classifiers is measured by the quantity of True positive (*TP*), True Negative (*TN*), False Positive (*FP*), False Negative (*FN*). Where *TP* (True Positive) is the number of positive instances that are classified as positive, *FP* (False Positive) is the number of negative instances that are classified as positive, *TN* (True Negative) is the number of negative instances that are classified as negative and *FN* (False Negative) is the number of positive instances that are classified as negative. By using these quantities accuracy, sensitivity, specificity, precision, *MCC* and *ROC* area performance measures are defined as:

2.1.3.3 Precision: Precision is defined as the proportion of instances classified as positive that are really positive.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2.7}$$

2.1.3.4 Recall: Recall is defined as the proportion of positive instances that are correctly classified as positive.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.8}$$

2.1.3.5 Accuracy: Accuracy is defined as the proportion of instances that are correctly classified.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{2.9}$$

2.1.3.6 *FP* **Rate:** It is the probability that a classifier produces erroneous results as positive results for negative instances.

$$FP\ Rate = \frac{FP}{FP+TN} \tag{2.10}$$

2.1.3.7 F-Measure: This measure is approximately the average of the two when they are close. The two measures are sometimes used together in the F1-**Score** (or **f-measure**) to provide a single measurement for a system.

$$F - Measure = \frac{2\ (Precision \times Recall)}{(Precision+Recall)} \tag{2.11}$$

2.1.3.8 Matthew's correlation coefficient (MCC): The *MCC* is a balanced measure that considers both true and false positives and negatives. The *MCC* can be obtained as

$$MCC = \frac{(TP)(TN)-(FP)(FN)}{\sqrt{[TP+FP][TP+FN][TN+FP][TN+FN]}} \tag{2.12}$$

2.1.3.9 Receiver operating characteristics (ROC): The *ROC* is a graph that shows the performance of a classifier by plotting the *TP* rate versus FP rate at various threshold settings. The area under the *ROC* curve (*AUC*) of a classifier is the probability that the classifier ranks a randomly chosen positive instance higher than a randomly chosen negative instance.

2.1.3.10 Jaccard Index: This statistic is used for comparing the similarity [14] and the diversity of sample sets. It is required in clustering techniques where we want to determine a measure of similarity between two observations. For sets A and B, Jaccard index is given as:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{2.13}$$

When A and B both are empty sets then J(A, B) = 1.

Receiver Operating Characteristic area under the curve (ROC AUC): In case of the imbalanced class say out 100 bleeding images and 50000 non-bleeding we want to

classify unknown instances then accuracy is a wrong metric to use. In such cases, ROC AUC is the correct metric to use. A ROC curve is plotting True Positive Rate (TP Rate) or sensitivity against False Positive Rate (FP Rate) or specificity. For a perfect classifier, there should be no false positives. It means that the curve should tend to the upper left corner. ROC AUC is just the area under the ROC curve. Higher the area, better the model.

2.1.3.11 Compression ratio (CR): It is used to quantify the reduction in data-representation size produced by a data compression algorithm.

$$CR = \frac{\text{Uncompressed Size}}{\text{Compressed Size}} \tag{2.14}$$

2.1.3.12 Peak signal to noise ratio (PSNR): It is the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation. Due to a wide dynamic range of signals, PSNR is usually expressed in the logarithmic decibel scale. PSNR is most commonly used to measure the quality of reconstruction of lossy compression for image compression. PSNR is an approximation to human perception of reconstruction quality. Although a higher PSNR indicates that the reconstruction is of higher quality, it may not always be true. It is only conclusively valid when it is used to compare results from the same codec and same content. For MxN image PSNR is calculated as:

$$PSNR = 20\ log_{10}\ \frac{255 * 255}{\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}[f(x,y) - f'(x,y)]}\ dB \tag{2.15}$$

where, $f(x,y)$ is the original image and $f'(x,y)$ is the new reconstructed image.

2.2 Literature Review

This section presents a detailed analysis of all the fields of study involved in medical image analysis for capsule endoscopy with a special focus on CAD systems for CE.

2.2.1 Video Summarization and Redundant Image Elimination

Out of 6 to 8-hour long capsule endoscopy videos, only 1 % of the video segment is of interest to the experts [15]. Depending on the experience of the examiner, the average inspection time ranges from 45 min to 120 min [16]. Constraints like energy, communication capability and limited resources of memory pose a huge challenge for analyzing and sharing these videos [17]. The time-consuming and tedious process inherently constrains the growth and spread of capsule endoscopy system. Thus there is an acute need for reducing the length of videos and video summarization, and elimination of redundant images is a solution. Figure 2.4 shows the idea of video summarization.

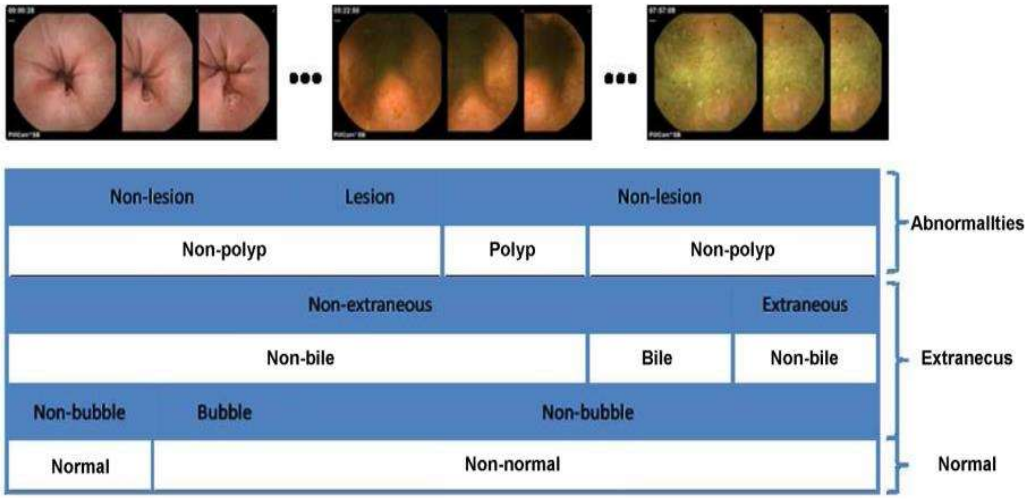

Figure 2.4: CE video summarised into multiple classes [18]

The natural peristaltic movement of the GI tract leads to the forward and backward motion of a capsule. At a frame rate of 2 fps, the capsule keeps on transmitting images to reviver irrespective of movement in forward or backward direction. This will lead to the generation of redundant images. Removal of redundant and non-informative frames will reduce examination time. Figure 2.5 depicts successive redundant image generation. The frames from l_1 to l_2 and l_2 to l_3 are redundantly produced due to backward and again the forward movement of the capsule.

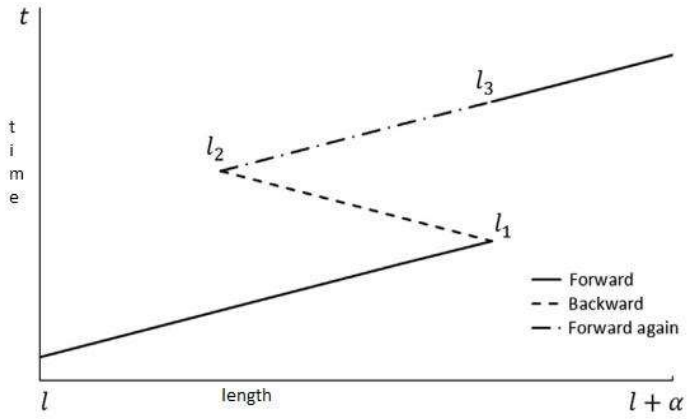

Figure 2.5: Successive redundant image generation [19]

Table 2.3 presents a comparative analysis of works published on video summarization and redundant image elimination. This comparison is based on approach, limitations, strengths, and dataset.

Table 2.3: Comparison of various works on video summarization and redundant image elimination

Work Reference	**Approach**	**Limitations**	**Strengths**	**Dataset**
[15]	Based on basic image descriptors 34 features are extracted; Eight individual classifiers and two classifier ensemble methods are applied to identify	Number of contraction images is very small leading to the imbalanced or cost-sensitive problem	ROC curve is used to optimize the trade-off between false positive and false-negative rates	20000 images

	contraction and non-contraction images			
[16]	Unsupervised data reduction algorithm is applied; an automatic tuning mechanism for the parameter controlling the number of frames to be extracted; clustering is applied to find useful images	High computational complexity	Reduces the time needed to inspect videos to 85%	24 videos; each of 5 minutes duration
[20]	Based on color and texture features, support vector machine (SVM) is used to segregate non-informative frames; Based on texture features, images with bubbles are segregated	Intra-video training and testing for supervised classification.	Higher detection accuracy	6 videos
[21]	Novel feature Local Colour-wise Binary Pattern (LCBP) is designed; elimination of un-useful image using statistical control charts	Small data set	Novel feature LCBP overcomes limitations of Local Binary Pattern (LBP) and Colour histogram (CH)	3200 images
[19]	Duplicate neighbor removal; backward motion removed and repeat forward removed using the Speed-up and robust feature (SURF)	Less accuracy	The higher frame rate of capsules may lead to better performance	10000 images
[17]	The Hu's image moments, curvature measures and multi-scale contrast are normalized in the range [0 1] and are fused to get a final saliency map. Based on saliency values and threshold, keyframes are selected	Computationally intense	Reduces storage cost	-
[22]	Relational motion histogram (RMH) based feature extraction; constraint formulation; Semi-supervised clustering and local scale learning (SS-LSL)	Works purely on global features, may miss local features present in CE	Ability to discover clusters of different sizes and densities.	150000 images

From the comparison shown in Table 2.3, it is observed that the size of data, computational cost and, features are the key design concerns for video summarization. Ismail *et al.* [22] uses the highest number of images amongst all to use machine learning-based algorithms and performs video summarization. For adopting machine learning, size of the dataset is of the at most importance. Better data leads to better machine learning. SS-LSL also minimizes the multi-term objective function leading to a sub-optimal solution. With a balanced trade-off between computational time and accuracy, a balanced system can prove important to gastroenterologists.

2.2.2 Image Enhancement and Interpretation

For the experts, it is very tiresome and stressful to interpret the images or videos [23]. It takes full concentration for up to 2 hours to examine the entire video. Also in view of computer-aided diagnosis, better image or video quality will lead to better performance of the system. Bad imaging conditions such as low complexity of GI tract, bad illumination, limitation of battery power and, short focal length leads to poor quality of images [24]. Thus, it is necessary to explore enhancement and interpretation techniques for improving the quality of captured CE image.

Essentially we need to process an image so that the result is more suitable than the original image for a CE analysis. CE images contain illumination related issues such as dark areas and we need to improve the visibility of dark areas while suppressing the noise.

For improved visualization, a mathematical model based on the motion of the capsule can be designed. There dimensional (3-D) reconstruction and surface map are an attempt to improve interpretability of capsule endoscopy videos. The core concept here is to provide a visual-friendly representation of GI tract. Figure 2.6 shows a surface map and 3-D reconstruction.

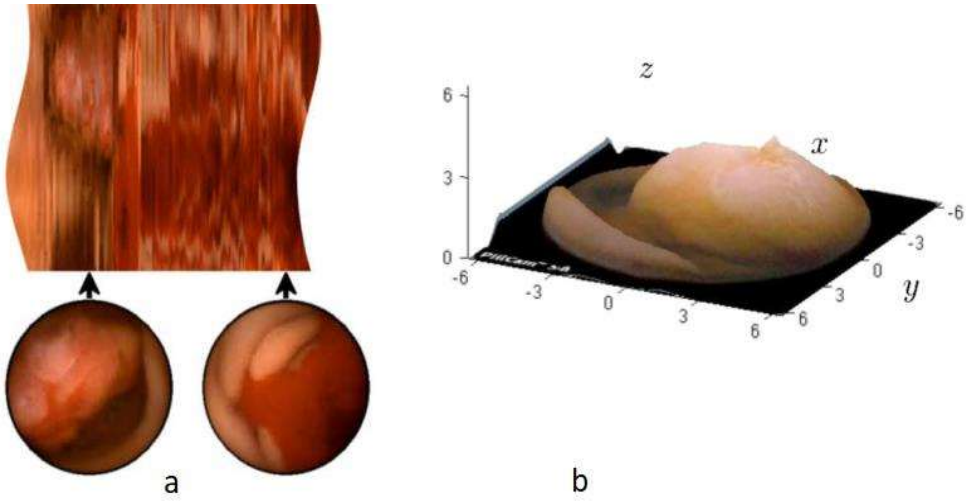

Figure 2.6: Surface map (a) [23] and 3-D reconstruction of protrusion (b) [25]

Szczypiński *et al.*[23] proposed a Model of deformable rings (MDR) which involves tracking movements of the GI tract by elastic matching of consecutive frames in a video. This model estimates the intensity of image motion outward and toward from the center of the frame in the video. By applying MDR, a map is generated giving a rough surface representation of lumen surface. This technique reduces overall time taken for interpretation also helps the investigator to focus on essential parts of the video.

Karargyris *et al.* [25] Proposed a 3-D reconstruction of the digestive walls to recreate visually friendly and smooth interpolated images from consecutive frames while preserving the structure of the observed objects. The overall objective is to improve the quality of video and the viewing capability.

Li *et al.*[24] proposed a novel idea of adaptive contrast diffusion to enhance capsule endoscopy videos. This method performs image enhancement without noise amplification. It also facilitates computer-aided systems to automatically detect abnormalities as better input quality leads to better performance of the system.

As compared to the first-order derivative of gradient, the second-order derivative Hessian matrix is better for extracting the intensity variations. Ranging from unsupervised learning to the concept of 3-D reconstruction and surface representation is an attempt to improve the readability and interpretability of CE videos for both man and machine. The noise in CE is related to illumination issues. Modeling and removing it can certainly fulfill the objective.

2.2.3 Segmentation and Region Identification

Since the CE procedure generates thousands of images, it is impractical to check every image manually. Looking at the cost of clinical time, CE examination becomes a costly procedure [26]. By utilizing techniques of image processing, one can extract a region of interest and thereby reduce diagnosis time, enhance the quality of the image and improve the accuracy of diagnosis. In addition to edge detection, a region of interest in capsule endoscopy includes clear demarcation between esophagus, stomach, small intestine, large intestine, and colon. Based on symptoms and complaints of patients gastroenterologists often intend to focus on the particular region as compared to entire GI tract.

Segmentation aims to distinguish between region or object of interest [27] and helps to improve the detection process [28]. It can be performed manually, semi-automatically and automatically. The manual segmentation process is laborious and time-consuming [29]. The given image is de-noised, and after removing artifacts, a threshold-based technique is applied to differentiate between foreground and background. The threshold technique will fail in case of high intra-class similarity. In such case, one can use advanced techniques like k-mean clustering and, the adaptive threshold. Figure 2.7 shows the output of a typical segmentation process.

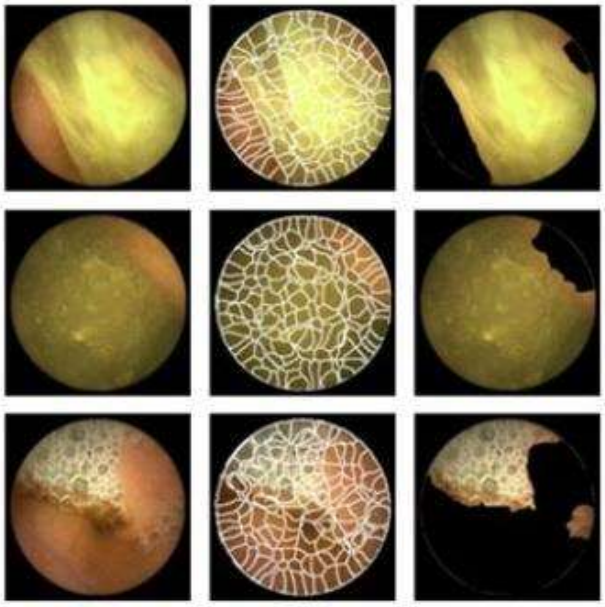

Figure 2.7: Images from left to right in each row shows the original image, intermediate segmentation stage and final segmented image [30]

Mackiewicz *et al.* [26] proposed a Hidden Markov model (HMM) based solution for segmentation which can also be extended for automatic abnormality detection. Shen *et al*. [31] proposed a scale-invariant feature transform (SIFT) algorithm for vocabulary building and probabilistic latent semantic analysis (pLSA) method for clustering to automatically classify the region of CE images. However, choosing appropriate codebook size is a challenge. Lan *et al.* [32] a total variation model combining edge and region information for segmentation. Arivazhagan *et al.* [30] proposed SURF and SVM based method for segmentation. Chen *et al.* [33] proposed octree-based convolutional neural networks (O-CNN) and HMM-based novel approaches to address this problem however, sensitivity needs to be improved.

Since the CE videos are lengthy, producing approximately 55000 frames; thus, segmentation and region identification plays a crucial role. Arivazhagan *et al.* [30] present a super pixel-based approach with acceptable accuracy to perform segmentation.

Automatic segmentation and region identification can help examiners to a great extent but, low contrast and non-guarantee of the closed contour is a challenge. For this we have seen a few attempts in literature review above but, still, a benchmark solution is yet to be designed. Also a generic segmentation technique capable of segmenting any CE image is not observed. Lack of availability of ground-truth annotated data in public domain is posing the toughest challenge.

2.2.4 Computer-Aided Abnormality Detection

In addition to a huge number of frames, GI tract appearance, intestinal dynamics, and need for constant concentration further complicate the diagnostic and analysis procedure. Using some automated features extraction and classification algorithms together formulated as a computer-aided diagnosis (CAD) system can be an excellent help for experts and physicians in detecting abnormalities [34]. A CAD system capable of analyzing and understanding the visual scene will certainly assist the doctor with a precise, fast and accurate diagnosis. After manual analysis of CE video, CAD can also provide a second opinion to a gastroenterologist [35]. In medical imaging, CAD is a prominent research area capable of providing precise diagnosis [36]. The ultimate goal of CAD is to reduce interpretation errors, reduce search errors and, reduce variation among observers [37]. In particular, a computer-aided medical diagnostic system for CE can consist of following units: (1) a data capturing and transmitting unit – the capsule (2) a data receiver and storage unit – the waist belt (3) a data processing unit for pre-processing and feature extraction (4) a machine learning-based classification unit or decision support system (5) a user interaction unit for final diagnostic report.

In general, a complete automated abnormality detection system includes pre-processing, segmentation, feature extraction, and classification of the abnormality. Figure 2.8 presents a diagrammatic representation of the entire process.

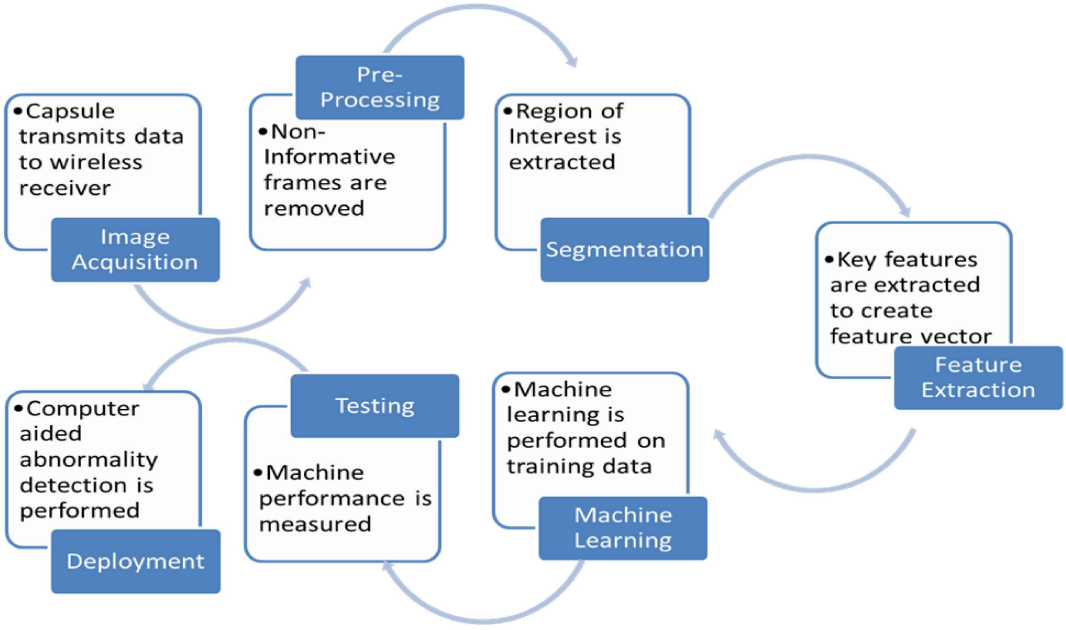

Figure 2.8: Diagrammatic representation of entire process

Details of each stage are as follows:

Step 1 - Image Acquisition:

Capsule captures and transmits data to the wireless receiver at a rate of 2 fps. This data is stored in a receiver belt tied on the waist of the patient. Later on these videos can be transferred from receiver to a computer for diagnosis.

Step 2 - Pre-Processing:

Non-Informative frames are removed in this stage. Such frames often contain bubbles, food remains, and so forth. Commonly observed difficulties in CE images are illumination changes, distortion and single-pixel randomness [38]. Such artifacts spoil the

learning of the system and hence, they are removed [39]. Also, methods such as contrast stretching are applied to improve the visibility of the images.

Step 3 – Segmentation:

The region of interest is extracted in this phase. Rather than processing the entire frame only the region where we are interested in processed to improve the efficiency of the system. For segmentation, threshold-based, region-based or edge-based methods are applied.

Step 4 - Feature Extraction:

Key features are extracted to create a feature vector. Lack of distinguishable features is one of the causes of failure of automatic abnormality detection in CE [10]. Colour, texture and statistical properties of the image are analyzed to figure out important features to distinguish the image. Wavelet-based, GLCM based and other statistical methods are used for feature extraction. Extremely large feature dimensions may not show consistent patterns [38]. Thus, out of all the extracted features only principally important features are selected based on techniques like chi-square, principal component analysis (PCA), and many others to reduce the dimensionality of the feature vector and thereby improve the efficiency of the system.

Step 5 - Machine Learning:

Machine learning is performed on training data. Based on the ground truth data the machine is trained to identify the abnormal and normal frames. SVM, convolution neural network (CNN), and, artificial neural network techniques are applied for machine learning.

Step 6 – Testing:

Machine performance is measured. Based on various performance measures such as recall, accuracy, precision, f-score, and so forth the performance of the system is evaluated. The confusion matrix is created, and various parameters such as true positive rate, false-positive rate, true negative rate, false-negative rate, and so forth are calculated to find out the quality of the learned system. Based on the results, the parameters and functioning of the machine learning technique are to be modified for achieving targeted accuracy and precession.

Step 7 – Deployment:

With a system expected to generate minimum false positives and acceptable accuracy, computer-aided abnormality detection is performed. The system is to be validated in real-time situations for automatic abnormality detection in capsule endoscopy. Figure 2.9 shows a clear idea encompassing the entire process of a computer-aided abnormality detection system.

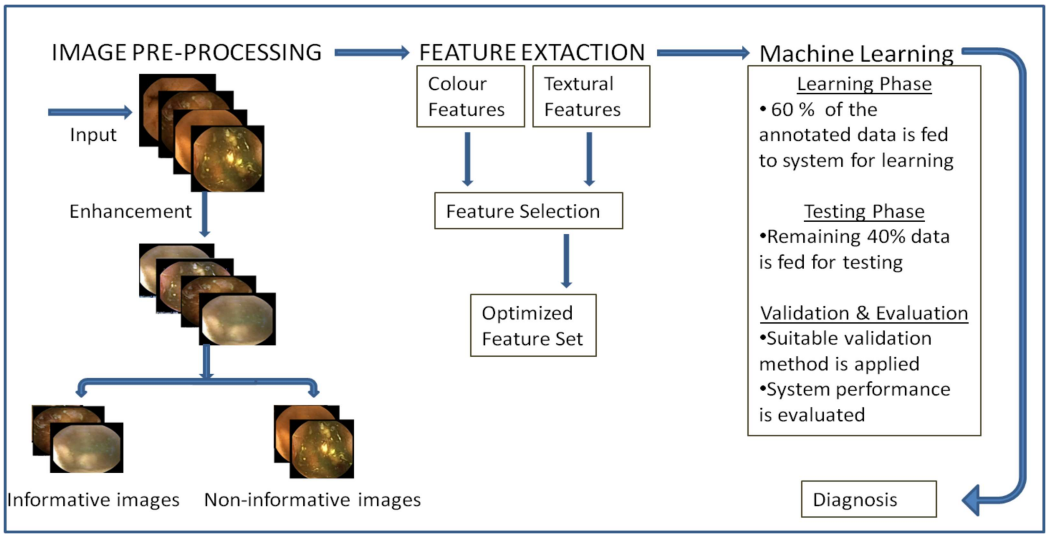

Figure 2.9: Overview of a computer-aided abnormality detection system

Table 2.4 presents a rigorous comparative analysis of works published in the field of computer-aided abnormality detection. This analysis is based on features, method/classifier, and type of abnormality, limitations, strengths, performance, and dataset. It is presented with the objective to provide current status and challenges in this field of study.

Table 2.4: Comparison of related work in the computer-aided abnormality detection

Reference	**Features**	**Technique used**	**Abnormality**	**Performance**	**Dataset**
[40]	Small intensity movements	-	Motility Analysis	-	-
[41]	Total features: 9*6 = 54. 9 measures for 6 color planes. { entropy, energy, inverse difference moment, standard deviation, variance, skew, kurtosis, contrast, covariance } *{R, G, B, H, S, V}	Multiple classifier Radial basis function (RBF) and Adaptive neuro-fuzzy logic system (AFLS)	Abnormal image	For RBF Predictability = 95.71% Sensitivity = 95.71% Specificity = 95.71% For AFLS Predictability = 98.57% Sensitivity = 97.18% Specificity = 98.55%	140 images
[42]	Total features: 9*6 = 54. 9 measures for 6 color planes. { entropy, energy, inverse difference moment, standard deviation, variance, skew, kurtosis, contrast, covariance } *{R, G, B, H, S, V}	Multiple classifier RBF+AFLS+ fuzzy inference neural network (FINN)	Abnormal image	For RBF Predictability = 95.71% Sensitivity = 95.71% Specificity = 95.71% Accuracy=91.43% For AFLS Predictability = 98.57% Sensitivity = 97.18% Specificity = 98.55% Accuracy=95.71 For FINN: Accuracy = 94.28%	140 images
[4]	Spatial ,edge and temporal features	Clustering	Motility Analysis	False alarm rate= 37% Detection rate=81%	6 videos

[43]	Total features: 4*3*6 = 72. 4 measures {entropy, intensity, homogeneity, energy} For 3 Discrete wavelet transform (DWT) Components, For 6 colour bands {R,G,B,H,S,V}.	1D classifier with the leave one out method	Abnormal image	Accuracy = 94.7%	75 images
[44]	36 patches on 30 * 30 were identified and chromatic moment for each patch is a feature.	Neural Network (NN)	Ulcer, Bleeding	For Ulcer detection: Specificity=84.68± 1.80 Sensitivity=92.97± 1.05 For bleeding detection: Specificity=87.81± 1.36 Sensitivity=88.62± 0.44	100 images
[8]	6 measures {standard deviation, kurtosis, entropy, energy, skew, mean} of uniform LBP histogram.	Multi layer Perceptron (MLP), SVM	Ulcer	Accuracy=92.37%, Sensitivity=93.28% Specificity=91.46 %,	100 images
[45]	Total features: 256*3 = 768 The pixel intensity value of image block size 256 for each of color planes {R, G, B}.	SVM	Bleeding	Accuracy = 99%	640 images
[46]	6 measures {standard deviation, kurtosis, entropy, energy, skew, mean} of LBP histogram from I color space of HSI and Chromatic moments	MLP	Bleeding	Detection rate = 90%	200 images
[47]	GI myoelectrical activity	Fast Fourier transform (FFT)	Motility Analysis	-	2 videos
[48]	Blob, color and texture based 54 features	SVM	Motility Analysis	Sensitivity=70%	10 videos
[49]	Pixel intensity and distance map	Clustering	Bleeding	Sensitivity = 92%, Specificity = 95%	960 images
[50]	Geometric features {entropy, contrast, homogeneity, inverse moment}	SVM	Ulcer, Polyp	For ulcer detection: Sensitivity=75% Specificity=73.3% For polyp detection: Sensitivity=96.75%	50 images

				Specificity=72.45 %	
[51]	Total features: 3*3*6=54. 3 bands {LH,HL,HH} * 3 colour planes {R,G,B} * 6 statistical measures {standard deviation, skew, energy, mean, kurtosis, entropy }	Ensemble classifier	Tumor	Accuracy=90.50% Sensitivity=92.33% Specificity=88.67 %.	1200 images
[52]	colour features values of {R,G,B,H,S,I} Total features: 6	Probabilistic neural network (PNN)	Bleeding	Sensitivity=93.1% Specificity=85.5%.	14630 images
[53]	Dif lac analysis represents the feature vector.	Linear classifier, Mahalanobis, SVM-RBF, NN	Abnormal image	Mean accuracy > 95%	176 images
[54]	18 Uniform LBP features	SVM	Polyp	Accuracy=91.6% for RGB colour space	1200 images
[55]	6 measures (standard deviation, skew, kurtosis, entropy, energy, mean) of Texture spectrum histogram (TSH), RIULBP and Curvelet-based local binary pattern (CLBP) and Colour wavelet covariance (CWC) is taken as a feature vector	SVM	Tumor	For RIULBP color features Average accuracy = 83.50%	1200 images
[56]	Uniform LBP histogram {10}, 2 level DWT {7}, 3 bands for 3 colour space =10*7*3*3=630 features.	SVM	Tumor	Detection Accuracy = 92.4%	1200 images
[57]	-	Dynamic programming	Motility Analysis	Visual inspection is 4 times faster	videos
[58]	279 texture features {statistics of image brightness histogram, image gradient magnitude, grey-level co-occurrence matrix, run-length matrix, parameters of autoregressive model and energies of image signal within frequency bands	Vector supported convex hull method	Bleeding, Ulcer	For bleeding: Recall = 0.875 Precision =0.732 Jaccard index =0.487 For Focal Ulcer: Recall = 0.932 Precision =0.696 Jaccard index= 0.235 For Excessive Ulcer:	613 images

	obtained using Haar wavelet transform} And 21 color bands lead to 279*21=5859 features. Reduced to 2494 features			Recall=0.917 Precision =0.669 Jaccard index= 0.621	
[59]	5 measures { entropy, energy, mean, standard deviation, skew} from 3 colour space {R,G,B} Total features: 15	MLP	Bleeding	Specificity=90%, Sensitivity=96%, Accuracy=93%	100 images
[60]	130 texton histograms generated from Leung and Malik (LM) and LBP filter	k-nearest neighbour (KNN)	Bleeding, Erythema, Erosion, Ulcer, Polyp	Recall =92% and Specificity =91.8%	1750 images
[61]	Geometric feature	Binary classifier	Polyp	Sensitivity=81% Specificity =90%	18968 images
[18]	Total features: 24+80+14= 118 24 color, 80 edge, and 14 texture features.	SVM-RBF, HMM	Polyp	For Polyp: Accuracy=0.933 Recall =0.933	13 videos
[62]	Total Features: 6*3=18 Normalized GLCM and Haralick based 6 features{contrast, sum entropy, sum variance, difference variance, difference average, Entropy} for 3 colour bands {R,G,B}	SVM	Bleeding	Accuracy=99.19%, Sensitivity99.41%= and Specificity=98.95 %	2920 images
[63]	Color features	Binary classifier	Bleeding	Processing rate: 344 fps which is 256 times faster than sequential execution	
[64]	Total features: 85+24+12=121 MPEG-7 edge features, texture and color features.	SVM	Crohn's Disease	Precision and Recall > 90%	513 images
[65]	The saliency of color and texture	SVM	Ulcer	Accuracy=92.65% Sensitivity=94.12%	340 images
[66]	Motility bar	Mean shift clustering	Motility analysis	-	10 videos
[67]	Edge density estimation	The parallel executable algorithm	Inflammation	Basic approach: Accuracy=84% Advanced approach: Accuracy=90%	231 images

[7]	Total features: 310 6 texture descriptors will result in 126-second order statistics and 84 high order moments.	Genetic Algorithm (GA), SVM	Tumor	Accuracy=97.3%, Sensitivity=97.8%, Specificity=96.7%	1800 images
[68]	H,S,V colour space pixel intensity values	SVM	Bleeding	Sensitivity 94% , Specificity 91%, Accuracy 92%	8500 images
[69]	YCbCr and Joint diagonalization principal component analysis (JD-PCA) based features	ODR-BSMOTE-SVM (OB-SVM)	Cancer	Avg. Accuracy=91.94%, Avg. Standard deviations=0.0331 Avg. AUC=0.9593	1330 images
[70]	Colour features	SVM	Bleeding	Accuracy= 95.75%, Sensitivity= 92% , Specificity= 96.5%, AUC = 0.9771	2400 images
[71]	SIFT+CLBP based 384 features	SVM	Polyp	Accuracy=93.20%	2500 images
[72]	Histogram of average intensity (HAI) gives Uncurled tubular region (UTR)	Rusboost	Hookworm	Accuracy=78.2%, Sensitivity=77.2%, Specificity=77.9%	440000 images
[73]	Generic features	CNN	Motility Analysis	Mean accuracy=96%	120000 images
[38]	Cluster based features	k-mean clusturing, SVM	Bleeding	sensitivity =96.22%, specificity=98.54% , accuracy = 98.04%	2350 images
[74]	Total features: 5*3 planes =15. Mean, Standard deviation, Entropy, Skew and Energy	SVM	Bleeding	accuracy = 95%, sensitivity = 94% ,specificity = 95.3%	8872 images
[75]	Centroid, area, eccentricity, mean, standard deviation	Naïve Bayes	Bleeding	-	-
[2]	Total features: 54 (9 LBPV histograms of 9 DWT components * 6 statistical features mecomputed from LBPV	SVM, MLP	Abnormal images	Accuracy =97.0 %, Sensitivity =96.4 %, Specificity =98.5 %	1670 images
[3]	Statistical features	SVM	Ulcer	Accuracy =97.89%, Sensitivity =96.22%, Specificity =95.09%	48000 images
[10]	Kanade-Lucas-Tomasi (KLT) feature points	Affine transform	Red spot, Phlebectasi, Angiodysplasia,	-	120 images

			Lymphangiectasia, Erosion, Erythematous, Ulcer, and White-tipped villi		
[121]	Texture features obtained from Contourlet transform and Log Gabor filters on HSV images	SVM	Ulcer	Accuracy =94.16%, Sensitivity =96.92%, Specificity =91.67%	137 images
[132]	CNN based features	Mean Gaussian SVM(MGSVM)	Ulcer and bleeding	Accuracy =98.3%, Sensitivity =99%, Specificity =98.67%	10 videos
[133]	Pre-trained CNN	GoogLeNet, AlexNet	Ulcer	Accuracy =95.16%, Sensitivity =96.80%	1857 images
[134]	CNN based features	CNN	Aphthae, bleeding, ulcer,polyp, oedema, angiectasias, lymphangiectasias, cysts	AUC = 0.8	137 images
[135]	LBP features	SVM	Ulcer	Accuracy =95.61%, Sensitivity =97.68%, Specificity =94.4%	212 images
[136]	CNN based features	CNN	Bleeding	Recall=0.99 Precision=0.99 F1-score=0.99	10000 images
[137]	CNN based features	CNN , SVM	Celiac disease	Classification rate=97%	1661 images

Ross *et al.*[40] in 1964 discussed implications of radio telemetering capsule in motility analysis. Capsule endoscopy was introduced by Given Imaging Inc. in 2000, over 1,000,000 Pillcam small bowel (SB) capsules alone have already been swallowed in the past 10 years since the device was first approved by the U.S. food and drug administration (FDA) [18]. From the statistics of papers examined, one can observe that automatic disease detection has taken pace since 2009. In the above comparative analysis, it is observed that machine learning techniques in computer vision are tremendously used to solve problems related to this field. The texture is a very important visual feature [76] and it is widely used in abnormality detection. This analysis shows that abnormality detection

in CE is a multi-class classification problem. It needs to be noted that fewer data and unbalanced data leads to poor learning of machine. For a few cases, the huge dataset is used, but its unavailability in the public domain restricts the community from validating the results.

2.2.5 Image and Video Compression

Reducing power consumption will increase the lifespan of the capsule. To reduce power consumption, there is a need to develop a computationally in-expensive image compressor. For accurate pathological diagnostic, we need high-quality medical images. However, low-complexity compression algorithms maintain high image quality with a low compression rate [77]. Thus, it is challenging to design an effective image compressor that meets all these contradicting constraints.

Image compression is classified as lossy or lossless. The lossy method saves space and has better transmission speed at a higher compression ratio (CR) but at the cost of image quality [78]. Lossless compression is preferred for medical imaging. Lossy compression methods when used at low bit rates, produce compression artifacts. Lossy methods are suitable for natural images such as photographs in applications where the minor loss of fidelity is acceptable to achieve a substantial reduction in bit rate. Such techniques may be called visually lossless. Table 2.5 presents a detailed comparison of works published in image and video compression for CE images and videos. This comparison is performed on the basis of methodology, limitations, strengths, and performance of various works.

Table 2.5: Comparison of related work

Work Reference	Methodology	Limitations	Strengths	Performance
[79]	Based on the Joint photographic experts' group (JPEG) standard, but uses prediction as the central principle	Low compression ratio	Low power consumption	Compression Ratio=20
[80]	YCbCr color space is combined with efficient recoding of wavelet transformation	Other color space can be examined	Hardware efficient	-
[81]	A static prediction scheme combining Golomb-Rice and unary encoding is used	Exploring other color space can improve performance	Low power consumption	Compression ratio=73%
[82]	A hardware-based approach using field-programmable gate arrays (FPGA)-based image processing core is designed	Size of hardware can be a constraint	High processing efficiency	-
[83]	Differential pulse code modulation (DPCM) followed by Golomb–Rice coding is used	Performance for higher frame rate needs to be examined	Temporary storage like the buffer is not required	Compression ratio=80% and PSNR > 48dB
[84]	Novel color space-based low complexity approach is presented	overall results need to be discussed	Better color space is used	High PSNR and competitive CR
[85]	A novel video codec based on the principles of Wyner-Ziv coding	Validation may be required	Exhibits low encoding complexity	-
[86]	DPCM followed by Golomb–Rice coding is used	Another coding may be examined	Low complexity	Compression ratio=80% and PSNR > 45dB
[87]	A novel color-space and simple predictive coding based approach which supports both White-band imaging (WBI) and Narrow-band imaging (NBI) is proposed	Performance for higher frame rate may be examined	Dual-band image compression technique	For WBI band: Compression ratio = 80.4% For NBI band: Compression ratio = 79.2% PSNR > 43.7dB
[77]	A reversible color space transformation, quantization, sub-sampling, DPCM, and Golomb–Rice encoding is used	Validation needs to be done	Novel metric termed as colour space conversion gain is proposed	Avg. compression ratio =90.35% PSNR = 40.66dB

With the constant development in the hardware of capsule endoscopy especially in frame rate and image quality aspects, image compression also needs to be developed. Hardware integration is a major challenge.

2.3 Discussion

To speed up diagnosis time and reduce the cost of CE procedures, we need to develop a computer-aided diagnostic tool, which automatically analyses the CE images. Since one cannot steer the capsule, some points are likely to be missed in physical observation which a computer-aided system can detect. While abnormal findings are visible in fewer frames, the clinicians need to remain focused and undistracted for a long period. This condition increases the chances of error. Colour, size and texture complexity and versatility may lead to the wrong diagnosis. Looking at problems stated above it is automated abnormality detection system is required. Different abnormalities have different mathematical and physiological properties which pose a challenge to a computer-aided system for CE image analysis.

The automated tool for abnormality detection in CE is a multi-class classification problem. From the literature review, it is found that proposed systems are still undergoing testing and improvements leading to a wider scope of research. Especially the false alarms rate is required to be reduced.

2.4 Conclusion

In this chapter, the theoretical backgrounds related to capsule endoscopy image analysis as well as literature review are presented. A first brief overview of technological dependencies is given. It also provides the basis for computer vision applications as discussed in subsequent chapters of this thesis. Further, in this chapter, a literature survey

of all aspects of medical image analysis in CE is discussed with a special focus on the CAD system for CE.