Colour Image Analysis of Wireless Capsule Endoscopy Video: A Review

Mark Fisher and Michal Mackiewicz

Abstract Wireless capsule endoscopy (CE) has been available since 2001 and is now established as one of the principal approaches used to examine the small bowel, with a range of devices available from four manufacturers. But although its use is widespread the reading of CE videos remains an arduous and time consuming exercise for gastroenterologists because relevant frames which are of interest to the physician constitute only about 1 % of the video. CE exam viewing times vary from 40–90 minutes, depending on the clinician's experience, the complexity of the case and the small bowel transit time. Colour image analysis has been applied by manufacturers to speed up this process, for example, Given Imaging's *Rapid Reader* Software includes a suspected blood indicator (SBI) designed to detect bleeding in the video. However, some evaluations of this tool reported concerns with regard to its specificity and sensitivity and so it does not free the specialist from reviewing the entire footage and can only be used as a fast screening aid. Over the past decade a number of papers have proposed a range of colour image processing and computer vision applications to assist the gastroenterologist in the analysis of CE video. These techniques can be divided into three categories, the first considers the topographic segmentation of CE video into meaningful parts such as mouth, oesophagus, stomach, small intestine, and colon. The second involves the detection of clinically significant video events (both abnormal and normal) and the third attempts to adaptively adjust the video viewing speed. This chapter reviews this research focusing particularly on the role of colour and texture descriptors and concludes by suggesting possible future directions for CE analysis.

M. Fisher (✉) · M. Mackiewicz
School of Computing Sciences, University of East Anglia, Norwich Research Park, Norwich, NR4 7TJ, UK
e-mail: Mark.Fisher@uea.ac.uk

M. Mackiewicz
e-mail: M.Mackiewicz@uea.ac.uk

M.E. Celebi, G. Schaefer (eds.), *Color Medical Image Analysis*,
Lecture Notes in Computational Vision and Biomechanics 6,
DOI 10.1007/978-94-007-5389-1_7,

1 Introduction

Wireless Capsule Endoscopy (CE) is a non-invasive clinical procedure allowing the entire Gastrointestinal (GI) tract to be examined using a small encapsulated CMOS camera. The development of this system was heralded in 2000 [1] and the first commercial system was available from Given Imaging Ltd. following FDA (American Food and Drug Administration) clearance in August 2001. The system, initially marketed as *M2A* but later rebranded *PillCam SB* (SB denoting Small Bowel), consists of a small (11 mm $\times$ 26 mm) capsule, an associated data-recorder belt and application software. The disposable capsule is swallowed and propelled through the GI tract by peristalsis before being expelled naturally. A transparent optical dome at one end of the capsule contains an array of six white light emitting diodes which surround a camera designed to capture two (256 $\times$ 256) colour images a second. The images are compressed by JPEG and transmitted using radiotelemetry to the data recorder which is worn by the patient on a belt. Analysis of the RF signal received by an array of aerials fixed to the patient's body allows the position of the capsule to be determined and its trajectory to be logged. Two silver-oxide batteries located at the other end of the capsule enable the camera to operate for about 8 hours, after such time the belt is removed for analysis. A software application called *Rapid Reader* allows the stored data (approximately 50,000 images) to be downloaded to a PC workstation for analysis. The clinical procedure is simple. The patient is advised to fast overnight and in some cases a drug which prepares the bowel and reduces GI transit time is administered. On the following morning, antennas are attached to the patient and connected to the data recorder, which is worn on a belt. The physician removes the capsule from its holder and performs a visual check to confirm it is operational before it is ingested by the patient. Once the capsule has been swallowed the patient is free to undertake normal tasks (subject to certain limitations), returning to hospital after a period of 8 hours has elapsed.

In 2004 Given Imaging launched a second product called *PillCam ESO*, incorporating two CMOS cameras (one positioned at each end of the capsule) operating at a higher frame rate designed to target oesophageal disease. *PillCam COLON*, launched in 2006, represents another specialization of the concept, optimized for colon examinations. PillCam COLON also employs two cameras but after activation the capsule enters a sleep mode for two hours (allowing it to reach the colon) before resuming image transmission. In 2005 Olympus launched a system called *EndoCapsule* with similar functionality to *PillCam SB*. *EndoCapsule* uses a CCD camera system equipped with automatic brightness control (ABC) to provide automatic illumination adjustment designed to deliver higher resolution images of consistent quality. A unique feature of *EndoCapsule* is a real-time viewer which allows the clinician to observe images as they are captured in addition to reviewing the video using the more usual off-line analysis tool (*EndoView*). Given responded in 2007 by launching *PillCam SB 2* a second-generation product with a superior specification and additional features designed to improve workflow. Subsequent second generation versions denoted *PillCam ESO 2* and *PillCam COLON 2* followed. Since 2007, capsules called *MicroCam* developed by IntroMedic and a Chinese competitor called *OMOM* (jinshangroup.com) have become available. The stream of images

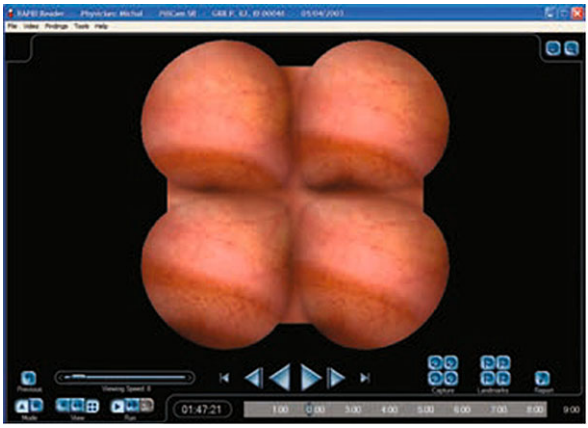

Fig. 1 *Rapid Reader* v4 (Given Imaging)

Fig. 2 *EndoView* (Olympus)

captured by the data recorder are presented as a video (typically 30 mins) and analyzed by a trained clinician using application software supplied by the manufacturer (e.g. Figs. 1 and 2).

The analysis software packages provided by different manufacturers comprise tools designed to improve the workflow and reduce the time spent on the analysis task (typically somewhere between 45–90 mins). In [2], the authors comment that with the expected reduction in capsule prices, the time needed by a clinician to analyse the exam may soon become the most expensive part of the procedure. Thus, a reduction of this time would be a major benefit, provided the quality of the diagnostic report was not reduced. The existing systems have user-friendly viewing interfaces, but with few exceptions lack automated tools that would highlight places of interest. Such tools could not only shorten the exam viewing time, but also improve the quality of patient's diagnosis by drawing attention to possible pathology, which could have been missed by the clinician among many thousands of normal frames. Incidentally, the manufacturers of the capsule try to reduce the video view-

ing time using additional viewing controls e.g. double and quad views in the *Rapid Reader* and *EndoView* software packages.

It is here where the computer vision can make a significant impact on the utility of CE. Ultimately, we would want the computer to take over from the clinician in stating the diagnosis allowing for a much cheaper screening technique. While this is still a very remote possibility, the development of computer vision methods for CE already allows or will soon allow for a significant aid in clinician's diagnosis. From segmenting CE video sequence into meaningful anatomical parts to detecting bleeding and other possible pathologies, computer vision methods have matured since the introduction of the first capsule a decade ago. The biggest challenge these algorithms face is to prove themselves that they can be trusted in practice i.e. perform the designated task at least as accurately as clinicians and hence allow for their wider adoption in clinical tools relieving clinicians from the burden of time consuming analysis. The ultimate bottom line measure here is the false negative ratio as for example for the pathological video event detection task, the exam evaluated as *normal* could skip manual inspection streamlining the population screening process.

The rest of this chapter reviews the computer vision research focusing particularly on the role of colour and texture descriptors and concludes by suggesting possible future directions for CE analysis. The main subjects of research are topographic video segmentation and filtering of non-informative frames, designed to provide a focus of attention, and classifiers for bleeding and abnormality detection. The following sections examine research in these areas, focusing in particular on the way that colour information is used in these tasks.

1.1 Feature Extraction

The distribution of colours in an image provides a useful cue for image indexing and object recognition. The colour distribution histogram is the most commonly used method of representing image colour information [3]. It is relatively invariant to image scale changes, translation and rotation about the viewing axis, and partial occlusion. Colour is an effective cue in CE image analysis and a salient feature of many proposed algorithms. Visually, the colour of the mouth is unsaturated, the stomach pinkish; the small intestine pinkish to yellowish; and the colon also pinkish to yellowish but often occluded by varying amounts of yellowish to greenish colour caused by faecal contamination. Moreover, different pathologies have their own distinct colour signatures. For example, ulcerations often contain yellowish and white colours surrounded by the overly reddish hues suggesting inflammation or bleeding.

CE video frames are stored as RGB triplets but very few researchers choose to analyse the data in this form. Fox [4], Bourbakis et al. [5] and Hwang et al. [6] are amongst a minority who extract colour features directly from RGB colour space. The authors claim their blood classifier appears to outperform the SBI tool provided by Given Imaging. Building on the work of Swain and Ballard [3], Berens [7] explored the scalability of colour indexing and extended their work by investigating

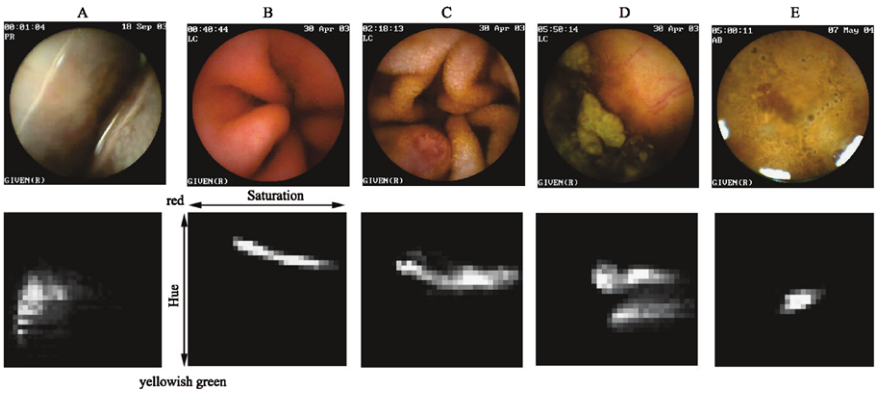

Fig. 3 CE images acquired from (**A**) Mouth, (**B**) Stomach, (**C**) Small Intestine, (**D**) partially occluded Colon and (**E**) completely occluded Colon; and below their respective equalized HS histograms. A visible shift in hue (vertical axis of the histogram) between the respective histograms is clearly visible. From [13]

the choice of colour space, coding of color histograms and techniques to provide invariance to illumination. Experiments undertaken by Berens [8, 9] showed that RGB colour space is not the best choice for image classification and that other perceptually relevant colour spaces such as HSI (Hue, Saturation, Intensity) produced better classification results. Consequently, Mackiewicz, working with Berens and Fisher, [10–13] also use HSI colour space, but due to the range of intensity variation in CE images, arising as the distance between the capsule and the intestine surface constantly varies, they ignore the intensity channel and favour HS histograms. The range of colour present in CE images is relatively small, mapping to a region covering just around 20 % of the possible HS colour space, so the histograms are equalized within this subset of red to yellowish-green colours. Figure 3 shows typical CE images acquired from the mouth, stomach, intestine and colon regions, and their respective HS histograms. It can be seen that the colour distribution of the stomach is slightly shifted towards red, compared to that within the intestine. It is also clear that the colour distribution of the colon tissue is highly similar to that of the small intestine, when it is free of faecal contamination. However, colon images are generally obscured by the presence of faecal contamination which has a distinct hue-saturation signature.

Texture features can play an important role, particularly in topographic video segmentation of CE video (Sect. 2). The most prominent texture pattern that distinguishes different organs are small finger-like projections called villi (responsible for food absorption), visible in Fig. 3C. These are present in the small intestine, but not in the neighbouring regions of stomach and colon. Mackiewicz et al. analyse texture by employing a 3D Local Binary Pattern (LBP) operator introduced by Connah and Finlayson [14] which extends the concept originally conceived by Mäenpää and Pietikäinen [15–17] who calculated 1D LBP histograms for the three colour channels independently. Because CE images are often obscured (to a varying degree) by

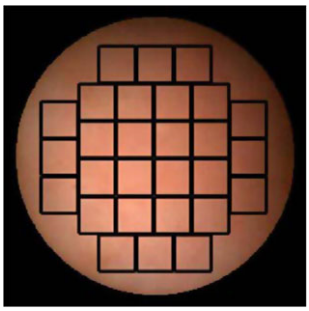

Fig. 4 Grid of 28 sub-image regions. From [13]

strong shadows, or by air bubbles and other artifacts such as mucus, bile, faeces, food etc. histograms built using the entire image will contain any visual contamination present in the image. To address this problem, some researchers extract only those parts of the image which contain only non-occluded tissue. In this respect, Mackiewicz divides each CE image frame into a grid of 28 sub-image regions arranged to cover most of the image area as shown in Fig. 4 and discards those regions which do not meet certain conditions.

The criteria are based on testing five parameters: Mean Intensity, Saturation, Hue, and Standard Deviation of Intensity and Hue against similar values derived from visually clear images of gastrointestinal tissue. The remaining sub-images form a so-called *sub-image region* (SubIR) that is used by the feature extraction process described previously. Figure 5 shows eight typical images acquired in the stomach and intestine showing only those sub-images selected by the procedure described above.

Another key result of Berens's work was that transform coding could be used to efficiently represent colour histograms without degrading their indexing performance [8, 18]. Mackiewicz applies this idea to colour histograms derived from CE images using both DCT and PCA transforms in a two stage algorithm to reduce the colour feature vector to just 8 values. Figure 6 shows the first three principal components calculated using the Hybrid Transform (DCT followed by PCA) from 1000 HS histograms extracted from one CE video. Each dot on the graph represents one histogram.

Jeongkyu Lee et al. [19] also address the problem of event boundary detection in CE arguing that there is compelling evidence to suggest HSI provides strong features that are highly correlated with topographic segments of the GI tract. Li et al. [20–22] also adopt the HSI colour space but in a similar approach to Mackiewicz they only use the HS components, summarizing this feature as a so called *chromaticity moment*. Coimbra et al. [23–27] favour colour and texture descriptors drawn from the MPEG-7 standard [28] and have evaluated these for detecting a variety of events in CE video. In [25] they conclude that the MPEG-7 *Scalable Colour* (SC) and *Homogeneous Texture* (HT) descriptors are the most adequate for the task of event detection. The SC descriptor is derived from the colour histogram defined in the HSI color space with fixed color space quantization of 16 bins. For compression, this information is encoded using the Haar transform, allowing scalable representa-

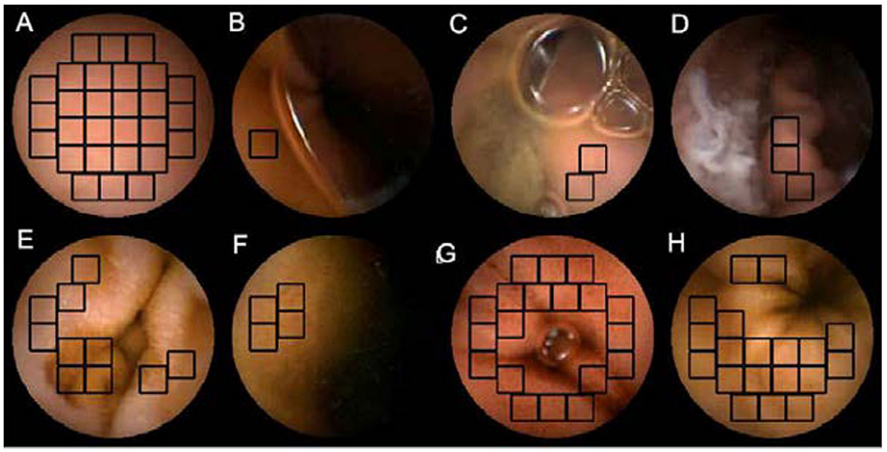

Fig. 5 CE images showing selected SubIRs. **A–D** Stomach; **E–H** Intestine. From [13]

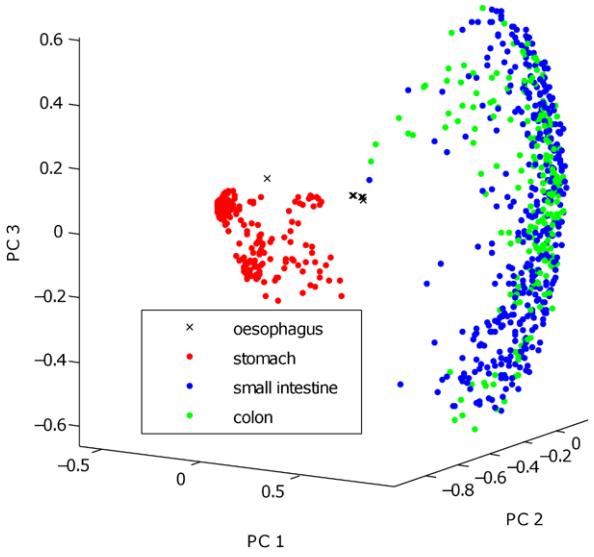

Fig. 6 Three first principal components representing compressed histograms extracted from four different video regions. From [13]

tion of the description and complexity scalability of feature extraction and matching procedures [29]. The HT descriptor encodes a precise statistical distribution of the image texture as a vector of 62 integers coming from the Gabor filter response of 30 frequency channels quantized in 30° radial segments in 5 octave bands [30]. Duda et al. [31, 32] also test MPEG-7 descriptors for CE image discrimination and conclude that the HT descriptor is the most reliable and the colour descriptors all performed similarly. They also selected HT and SC descriptors as features. In their work, Vilarino et al. [33–35] also surveyed a range of image descriptors and concluded that intensity, color and texture are the most relevant visual cues when processing en-

doscopic videos. However, since their focus is intestinal contractions they pursue a sequence-based rather than a frame-based approach, focusing on variations in image intensity.

All researchers use colour features as inputs, sometimes combined with other cues derived from motion, to classify single images and image sequences drawn from CE video. Applications fall broadly into three areas. Topographic video segmentation, the detection of clinically significant abnormalities, and attempts to control the speed at which frames are delivered to the viewer. These are reviewed in Sects. 2, 3 and 4.

2 Topographic Video Detection

The GI tract comprises mouth, oesophagus, stomach and duodenum (upper GI tract), the jejunum, ileum, colon and rectum. Typically the capsule takes a few seconds to pass through the oesophagus before reaching the esogastric junction and entering the stomach. The capsule remains in the stomach typically 15 minutes but this period might be extended to several hours before it passes through the pylorus (a valve between the stomach and the small intestine). The capsule takes about four hours to transit the small intestine before entering the colon. Three key landmarks are the esogastric junction (between oesophagus and stomach), pylorus (between stomach and small intestine), and ileocaecal valve (between small intestine and colon). Annotating the esogastric junction is quite easy as the features inside the mouth, oesophagus and stomach are quite different. Locating the pylorus in the video can be difficult and time consuming, even for those experienced in this task, as the stomach tissue near the pyloric valve and that of the small intestine are visually similar. The ileocaecal valve which marks the entry to the colon is even more difficult to locate as the tissue is often obscured by faecal material.

Topographic video segmentation considers the problem of segmenting the capsule video into meaningful parts such as mouth, oesophagus, small intestine and colon. Researchers have observed that the choice of the right features is probably the most important issue in this segmentation task and most support the view that image texture is an important cue. Mackiewicz and Coimbra classify single images based on information recovered from colour and texture descriptors (combined with a motion descriptor) and use these results to classify image features into the previously mentioned anatomical classes. Work by both Coimbra and Mackiewicz concludes that using a Support Vector Classifier rather than a Bayesian approach improves the results, which can then be used within a navigation tool and to estimate the capsule Gastric and Intestinal Transit Times, which are important factors in diagnosing certain medical conditions.

Mackiewicz investigates a number of recognition algorithms including various linear and non-linear classifiers: Multivariate Gaussian, kNN and Support Vector Classifier (SVC) to perform the actual video segmentation, i.e. label the transition points between anatomical regions. He performs a number of experiments to test his topographic segmentation approach using a data set comprising 76 annotated

CE videos provided by clinical collaborators at the Norfolk and Norwich University Hospital. The videos were annotated by an experienced clinician and segmented into meaningful parts: Entrance, Stomach, Intestine and Colon. The input feature set comprised both colour features derived from both whole images and subIRs. Single images are classified as Entrance/Stomach, Stomach/Intestine and Intestine/Colon. In these experiments the HS histograms were quantized into $32 \times 32 = 1024$ bins and LBP histograms were built using 8 sampling points to provide 7 unique patterns, $21(3 \times 7)$ bins for the independent 1D histogram and $343(7^3)$ for the joint 3D histogram.

It is worth noting that the choice of anatomical regions to be segmented varied between researchers. The most popular set was mouth/entrance; oesophagus; stomach; small intestine and colon. However, Duda et al. attempted to classify the CE images from only the upper part of the GI tract into a larger number of distinctive regions. They chose six anatomical regions: (A) oesophagus, (B) cardia, (C) fundus, (D) corpus of the stomach, (E) pylorus and (F) duodenal cap. They used Neural Networks as the image feature classifiers. The authors reported only the classification results and did not attempt to segment the actual videos. Lee et al. chose yet another set of anatomical regions namely: oesophagus; stomach; duodenum and jejunum; ileum; and colon. The idea for their algorithm is based on the fact that each digestive organ has different patterns of intestinal contractions. The analysis of the frequency functions associated to these patterns leads to the event boundaries which indicate either entrance to the consecutive organ or unusual events in the same organ, such as intestinal juices, bleeding, ulceration, and unusual capsule movements. These events can then be classified and if necessary merged into higher level events that represent digestive organs leading to a tree-like representation of the capsule endoscopy topography. The authors report that the performance on ileum and colon is worse than on the upper digestive organs which confirms the earlier findings regarding difficulties with locating the entrance to the colon reported by Mackiewicz and Coimbra.

Some researchers have produced clinical demonstrator systems by combining their classifiers within a search framework that allows the user to search and navigate within and between topographic regions. Both Coimbra and Mackiewicz have found a Hidden Markov Model (HMM) to be the best strategy for this purpose.

3 Detection of Clinically Significant Events

Another important research area involves the detection of clinically significant video events (both abnormal and normal). Examples include physical abnormality (e.g. ulceration, polyp, cancer), intestinal fluids, intestinal contractions and capsule retention. This category also includes bleeding, an area which has received considerable focus in the literature and one that has been addressed by the manufacturers in their proprietary software packages.

Blood detection is a focus for much of CE research, perhaps motivated by early reports that questioned the performance of the SBI shipped with *PillCam SB*.

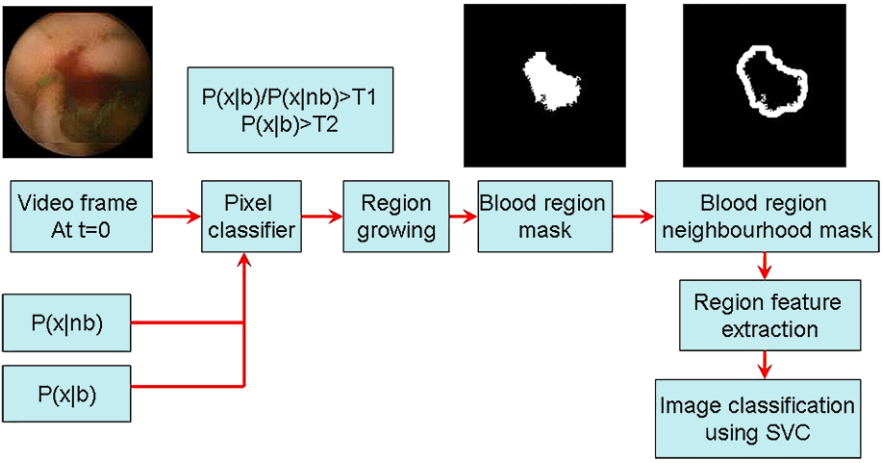

Fig. 7 Bleeding detection system flow chart. From [13]

In [4, 6], the authors propose a new algorithm that they claim can detect bleeding areas in the capsule videos. The algorithm uses Expectation Maximization (EM) clustering and Bayesian Information Criterion (BIC). The authors manually segmented around 200 images into blood and non-blood regions. Then, they selected 16,000 bleeding and 45,000 non-bleeding pixels and modelled the colour distribution of these regions using Gaussian mixtures in RGB colour space. A Bayesian decision rule was used. The algorithm chooses those pixels x to be bleeding candidates for which conditional probability $p(x—bleeding)$ of a pixel x given by bleeding pixels is significantly larger than conditional probability $p(x—non\text{-}bleeding)$ of a pixel x given by non-bleeding pixels; and also it is larger than a certain predefined threshold. In the final step of the algorithm, the areas of bleeding regions are calculated and all segmented regions containing less than 1,000 pixels are rejected. To test the results of bleeding detection, the authors selected 15,222 capsule images of which 1,731 contained blood from three different videos. On this test set, the reported specificity and sensitivity were 98,10 % and 92,55 % respectively.

Contrary to [4, 6], who use parametric bleeding colour distribution models, Mackiewicz chooses a different method using the similar feature set as described in the previous section [36]. A simplified flowchart of the bleeding detection system is shown in Fig. 7. First, each pixel is classified as bleeding or non-bleeding using a HSI model. Then, a region growing operation merges candidate pixels into regions of at least 250 pixels. If a blood region is detected, associated colour and texture features are extracted. These features are also extracted from the region surrounding the suspicious region. Then, after searching for specular highlights in order to check if the frame contains air bubbles, these features are used to identify the frame as containing suspicious regions. The images are classified using a Support Vector Classifier into three classes: Bleeding, Lesion/Abnormality or Normal, reporting figures of 97 %, 92 % and 92 % respectively using on ten-fold cross validation with a database comprising 84 full-length CE videos.

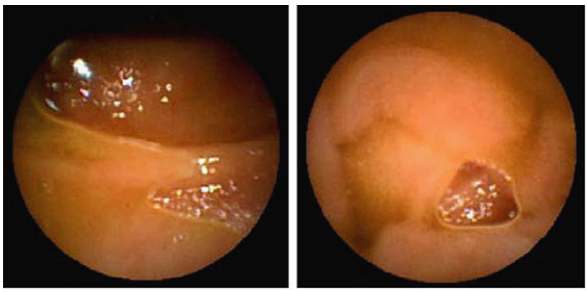

Fig. 8 Two images containing air bubbles with specular highlights. From [13]

Another idea for aiding capsule endoscopy video review involves removing non-informative frames from the video sequence. Early detection of such regions is highly beneficial since they can be removed from the sequence, before it is presented to the clinician, resulting in a shortening of the reviewing time. Intestinal fluids are one type of non-informative content. They appear as yellowish to brownish semi-opaque turbid liquids often containing air-bubbles as well as other artifacts (Fig. 8). Removal of such frames was first proposed by [34] who presented an algorithm which detects areas in the WCE video comprising images completely obscured by intestinal fluids. The authors observe that the most relevant feature of the intestinal fluids is the presence of small bubbles of different sizes and quasi-circular shapes. Their algorithm is based on texture analysis performed using Gabor filter banks. In order to construct a filter bank, the authors used four different directions oriented at 0°, 45°, 90°, 135° and consisting of four Gaussian scales (sigma values of 1, 2, 4 and 8 pixels), resulting in a bank of 16 filters. Mackiewicz also addresses the problem of air bubbles as these can cause problems when attempting to identify frames containing blood because the healthy tissue colour distribution seen through the air bubble is similar to the blood colour distribution, thus triggering false positives. He observed that air bubbles often contain specularities which can be detected using an approach due to Ortiz and Torres [37].

Vilarino et al. [33–35] as well as Igual et al. [38] studied detection of intestinal contractions and intestinal motility disfunction. Villarino claim a sensitivity of 70 % in respect of their approach which involves the analysis of textural, colour and blob features using a Support Vector Machine (SVM).

Recently, Li and Meng [20] proposed a method of bleeding and ulceration detection by means of chromaticity moments constructed from the Tchebichef polynomials. The authors divide the circular CE image into a grid of 36 non-overlapping blocks (30×30 pixels) (similar grid was also used for feature extraction in [12], see Fig. 4), from which they calculate six chromaticity moments. Next, they performed an experiment in which 5400 (1800 normal, 1800 bleeding and 1800 ulceration) image blocks were selected from 300 non-consecutive CE images extracted from 10 patient video sequences. The blocks were randomised and classified using an MLP Neural Network. Finally, the authors reported sensitivity and specificity figures obtained from the block classifications. This was a preliminary study that was not performed on the full length videos.

More recently, Li et al. [22] presented a study with an aim to develop a computer aided system to diagnose small bowel tumours. They proposed a textural feature that is built on wavelet and local binary pattern. They employed a classifier ensemble consisting of k-nearest-neighbor, multilayer perceptron neural network and support vector machine. Results obtained from the single image classification of 600 normal and 600 abnormal capsule images showed the promising performance for small bowel tumour detection.

4 Viewing Speed

Attempts to automatically adapt the viewer's focus of attention based on video content have focused on automatically adjusting the viewing speed and filtering of non-informative frames. Hai et al. [39] proposed video speed is adjusted by an algorithm which plays the video at high speed in stable regions and at slower speed where significant changes between frames occur, signifying the possibility of pathologies. The authors divide each frame into 64 blocks and measure the similarity of colours between respective blocks in consecutive frames. RGB histograms quantized to 163 bins are used to describe each image block. The distance between local histograms is computed using the L1 norm, formally:

$$D_{blk}(i) = \sum_{k=1}^{nbins} \left(\left\| H_{R,k}^{n} - H_{R,k}^{n+1} \right\| + \left\| H_{G,k}^{n} - H_{G,k}^{n+1} \right\| + \left\| H_{B,k}^{n} - H_{B,k}^{n+1} \right\| \right)$$

which is later used to calculate the similarity between two frames:

$$Sim(n) = \frac{1}{n_{blocks}} \sum_{i=1}^{N_{blocks}} sim\_block(i)$$

where

$$sim\_block(i) = \begin{cases} 1: & D_{blk}(i) > Thresh_{block} \\ 0: & otherwise \end{cases}$$

These features are used together with estimates of motion displacement to classify the frame in one of four states and these in turn adjust a delay which controls the speed at which frames are presented to the viewer. The authors conclude that using their method the viewing time may be reduced from 2h to around 30 minutes without 'loss of information'.

The software supplied by both Given Imaging (Rapid Reader) and Olympus (EndoView) also include play speed control. Unfortunately, the details of these algorithms remain unknown. Moreover, in the more recent versions of Given's Rapid Reader, the clinician is given an option of watching a video in either "Normal Mode" or in the "Quick View Mode". Although the "Quick View" mechanism is not precisely explained in the documentation, we noticed that it uses an approach similar to that described above to reduce the viewing time of the video. It must be added

though, that the "Quick View" mode skips some frames, displaying only the most suspicious (at least to the algorithm that is used by Given Imaging), which makes it different to the algorithms described above.

The obvious conclusion regarding these methods must be that they are highly subjective. All research on this topic has to include particularly extensive clinical evaluations in order to make sure that the increase in the viewing speed does not increase the number of false negatives [40].

5 Future Directions for CE Research

In a recent review of ten years of CE, Mackiewicz [41] considers a number of exciting opportunities for further research in the field CE video analysis. Firstly, there is the possibility of focusing on specific pathological events possibly addressed by some of the specialist capsules now being marketed (e.g. *PillCam ESO* and *PillCam COLON*). Adaptive control of the speed at which the video is reviewed is also a promising area as it draws from experience gained in other research in the field of video summarization and beyond this, there is the possibility of tools for automated reporting and annotation of CE video. The prospect of more advanced capsules which might be controlled by the physician are probably no more than a decade away. One of the main challenges for CE research is in providing sufficient quantities of annotated training data to enable classifiers to be built. Given that a typical CE exam may contain around 50,000 images, but only a few abnormal events, a reasonably conservative figure for a training set might be 100 exams (about 50 hours of video). The fact that few researchers have access to a database of this size probably explains the lack of significant progress in the field, even following the publication of hundreds of individual papers.

6 Discussion

All of the significant investigators of CE video analysis [12, 21, 27] have used colour and texture features. The preferred colour feature is the HSI colour histogram, encoded using the Haar (MPEG-7) or hybrid (DCT + PCA) transform. Many researchers chose MPEG-7 features, possibly due to the freely available reference software [42], the established track record of these techniques in other content based image retrieval applications, and the work due to Coimbra et al. [25]. All researchers agree that colour texture is a very important component. The groups using MPEG-7 favour the HT descriptor, based on the Gabor wavelet. However, the comparative success of other methods such as LBP adopted by Mackiewicz in an implementation developed by colleagues Connah and Finlayson [14], suggests that the MPEG-7 descriptors may not be the most suitable for this purpose. A number of classifiers have been tested using both feature sets and there is widespread agreement that the Support Vector Classifier yields marginally better results than other methods.

Evaluation of the approaches varies considerably and the lack of a large reference data set is a major drawback. The quality of evaluations undertaken by groups in Norwich, Porto, HongKong and Barcelona is largely due to the support by collaborators at local hospitals or manufacturers. Manufacturers Given Imaging, Olympus and OMOM (jinshangroup.com) all provide example CE video data but primarily motivated by a desire to promote sales, marketing, and training rather than support the development of algorithms by the wider scientific community. The availability of CE video has made a significant impact on the medical imaging community since its introduction in 2001 and there is no doubt that it will become increasingly important, as the number of CE examinations grows.

References

1. Iddan G, Meron G, Glukhovsky A, Swain P (2000) Wireless capsule endoscopy. Nature 405:725–729
2. Ravens AF, Swain P (2002) The wireless capsule: new light in the darkness. Dig Dis 20:127–133
3. Swain M, Ballard D (1991) Color indexing. Int J of Computer Vision 11–32
4. Cox J (2005) Finding blood in capsule endoscopy. Master's thesis, The University of Texas at Arlington
5. Bourbakis N, Makrogiannis S, Kavraki D (2005) A neural network-based detection of bleeding in sequences of WCE images. In: Proceedings of the 5th IEEE symposium on bioinformatics and bioengineering (BIBE'05)
6. Hwang S, Oh J, Tang SJ (2006) Expectation maximization based bleeding detection for wireless capsule endoscopy (WCE) images. In: Proceedings of SPIE, vol 6144, pp 577–587
7. Berens J (2002) Image indexing using compressed colour histograms. PhD thesis, School of Information Systems, University of East Anglia, Norwich, UK
8. Berens J, Finlayson GD, Qiu G (2000) Image indexing using compressed colour histograms. IEE Proc, Vis Image Signal Process 147(4):349–354
9. Berens J, Finlayson GD (2002) An efficient coding of three dimensional colour distributions for image retrieval. In: CIVR2002. Lecture notes in computer science, vol 2383, pp 245–252
10. Berens J, Mackiewicz M, Bell GD (2005) Stomach, intestine and colon tissue discriminators for wireless capsule endoscopy (WCE) images. In: Proceedings of SPIE, vol 5747, pp 283–290
11. Berens J, Mackiewicz M, Fisher M, Bell GD (2005) Using colour distributions to discriminate tissues in wireless capsule endoscopy images. In: Proceedings of medical image understanding and analyses 2005 conference, Bristol, UK, July 2005, pp 107–110
12. Mackiewicz M, Berens J, Fisher M (2008) Wireless capsule endoscopy colour video segmentation. IEEE Trans Med Imaging 27(12):1769–1781
13. Mackiewicz M (2007) Computer-assisted wireless capsule endoscopy video analysis. PhD thesis, School of Computing Sciences, University of East Anglia, Norwich, UK
14. Connah D, Finlayson GD (2006) Using local binary pattern operators for colour constant image indexing. In: CGIV 2006 conference proceedings
15. Ojala T, Pietikäinen M, Mäenpää T (2002) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans Pattern Anal Mach Intell 24(7):971–987
16. Mäenpää T, Pietikäinen M (2004) Classification with color and texture: jointly or separately. Pattern Recognit 37:1629–1640
17. Pietikäinen M, Hadid A, Zhao G, Ahonen T (2011) Computer vision using local binary patterns. Springer, Berlin

18. Berens J, Finlayson GD (2002) An efficient coding of three dimensional colour distributions for image retrieval. In: Proc. CIVR2002. Lecture notes in computer science, vol 2383, pp 245–252
19. Lee J, Oh J, Shah SK, Yuan X, Tang SJ (2007) Automatic classification of digestive organs in wireless capsule endoscopy videos. In: Proceedings of the 2007 ACM symposium on applied computing, SAC'07. ACM, New York, pp 1041–1045
20. Li B, Meng MQ-H (2009) Computer-based detection of bleeding and ulcer in wireless capsule endoscopy images by chromaticity moments. Comput Biol Med 39:141–147
21. Li B, Meng MQ-H (2009) Computer-aided detection of bleeding regions for capsule endoscopy images. IEEE Trans Biomed Eng 56(4):1032–1039
22. Li B, Meng MQ-H, Lau JYW (2011) Computer-aided small bowel tumor detection for capsule endoscopy. Artif Intell Med 52(1):11–16
23. Coimbra M, Campos P, Silva Cunha JP (2005) Extracting clinical information from endoscopic capsules exams using MPEG-7 visual descriptors. In: 2nd European workshop on the integration of knowledge semantic and digital media technologies, November 2005. IEE, New York
24. Coimbra M, Campos P, Silva Cunha JP (2006) Topographic segmentation and transit times estimation for endoscopic capsule exams. In: Proceedings of the IEEE international conference on acoustics, speech, and signal processing, vol II, Toulouse, France, May 2006, pp 1164–1167
25. Coimbra M, Campos P, Silva Cunha JP (2006) MPEG-7 visual descriptors—contributions for automated feature extraction in capsule endoscopy. IEEE Trans Circuits Syst Video Technol 16:628–637
26. Coimbra M, Kustra J, Silva Cunha JP, Campos P (2006) Combining color with spatial and temporal position of the endoscopic capsule for improved topographic classification and segmentation. In: Proceedings of the 1st international conference on semantic and digital media technologies, Athens, Greece, December 2006
27. Silva Cunha JP, Coimbra M, Campos P, Soares JM (2008) Automated topographic segmentation and transit time estimation in endoscopic capsule exams. IEEE Trans Medical Imaging 27(1):19–27
28. Chang S-F, Sikora T, Puri A (2001) Overview of the MPEG-7 standard. IEEE Trans Circuits Syst Video Technol 11(6):688–695. Special issue on MPEG-7
29. Manjunath BS, Ohm J-R, Vasudevan VV, Yamada A (2001) Color and texture descriptors. IEEE Trans Circuits Syst Video Technol 11(6):703–715
30. Wu P, Ro YM, Won CS, Choi Y (2001) Texture descriptors in MPEG-7. In: Skarbek W (ed) Proceedings CAIP 2001. LNCS, vol 2124, pp 21–28
31. Duda K, Zielinski T, Fraczek R, Bulat J, Duplaga M (2007) Localization of endoscopic capsule in the GI tract based on MPEG-7 visual descriptors. In: IEEE international workshop on imaging systems and techniques, IST'07, May 2007, pp 1–4
32. Duda K, Zielinski T, Duplaga M, Grega M, Leszczuk M (2007) VQ classification based on MPEG-7 visual descriptors for video endoscopic capsule localisation in the gastrointestinal tract. In: 15th European signal processing conference (EUSIPCO 2007), Poznan, Poland, September 2007
33. Vilarino F, Kuncheva LI, Radeva P (2005) ROC curves and video analysis optimization in intestinal capsule endoscopy. Pattern Recognit Lett, Special Issue on ROC analysis
34. Vilarino F, Spyridonos P, Puyol O, Vitrià J, Radeva P (2006) Automatic detection of intestinal juices in wireless capsule video endoscopy. In: Proceedings of ICPR, pp 531–537
35. Vilarino F, Spyridonos P, DeIorio F, Vitrià J, Azpiroz F, Radeva P (2010) Intestinal motility assessment with video capsule endoscopy: automatic annotation of phasic intestinal contractions. IEEE Trans Med Imaging 29(2):246–259
36. Mackiewicz M, Fisher M, Jamieson C (2008) Bleeding detection in wireless capsule endoscopy using adaptive colour histogram model and support vector classification. In: Proceedings of SPIE, vol 6914

37. Ortiz F, Torres F (2006) Automatic detection and elimination of specular reflectance in color images by means of MS diagram and vector connected filters. IEEE Trans Syst, Man, Cybern, Part C 36(5):681–687
38. Igual L, Seguí S, Vitrià J, Azpiroz F, Radeva P (2007) Eigenmotion-based detection of intestinal contractions. Lect Notes Comput Sci 4673(2007):293–300
39. Hai V, Echigo T, Sagawa R, Yagi K, Schiba M, Higuchi K, Arakawa T, Yagi Y (2006) Adaptive control of video display for diagnostic assistance by analysis of capsule endoscopic images. In: Proceedings of ICPR, pp 531–537
40. Zheng Y, Hawkins L, Wolff J, Goloubeva O, Goldberg E (2012) Detection of lesions during capsule endoscopy: physician performance is disappointing. Am J Gastroenterol
41. Mackiewicz M (2011) Capsule endoscopy—state of the technology and computer vision tools after the first decade. In: Pascu O (ed) New techniques in gastrointestinal endoscopy. InTech, New York
42. Multimedia content description interfaces. Part 6: reference software. MPEG-7 ISO/IEC 15938-6, 2003