The 2009 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 11-15, 2009 St. Louis, USA

In Situ Analysis of Capsule Endoscopy Images and Preliminary Results

Xiaona Wang, and Max Q.-H. Meng, *Fellow, IEEE*

***Abstract*— Capsule endoscopy has been proved efficient in examining the small intestine. A lot of work has been devoted to the study of capsule videos to assist the diagnosis. Different from previous approaches which worked off-line and couldn't be applied in active capsule controls, in this paper we proposed an in situ capsule video analysis method, which operates in real-time and provides the first event detection-based scheme for capsule techniques. Specifically, first we established a theoretical computing framework. The method addresses two key points: one is to merge the surgeon's expertise into the system, and the other is to identify unusual events. Then we evaluated the method by some empirical experiments. The preliminary results verified the usefulness of the method.**

I. Introduction

The capsule endoscopy is an emerging medical procedure in placement of the traditional endoscopes. It is designed to visualize the whole gastrointestinal tract for diagnosis of diseases [1], [2], [3]. The capsule encloses a camera for imaging and a wireless transceiver to send the captured images to an external data recorder. This feature not only exempts the patients from painful operations but also makes it possible to investigate the part of long and twisted small intestine. During the examination, the patient first swallows the capsule, and then the capsule is driven passively by the peristaltic movement and takes pictures at the same time. It is excreted naturally at the end of the examination. After the conclusion of the procedure, the videos are submitted to the doctors to make diagnosis and give treatments.

In current capsule endoscopy, doctors make diagnosis by reading the off-line videos. The reading time is dependent on the experience and expertise of the reader [4]. For a routine study in which more than $55,000$ images are obtained, up to 2 hours are required initially. The time may reduce to about 1 hour while the doctors are gathering experiences. What makes things worse is that even practiced readers may find that their accuracy deteriorates as they try to increase their inspection speed [5]. To solve the problem, several computing techniques have been proposed to help automate the checking process, filter out unnecessary video frames and only keep those that might be useful [6], [7], [8].

The next generation of capsule endoscopy [9], [10], [11], which will be equipped with an actuation mechanism and can be controlled wirelessly by the doctors, presents new challenges in computer vision and related areas. Because of the limited free space, the capsule cannot contain many sensors. It is advantageous that some feedback can be provided by the image method to the external control system [12]. A real-time analysis of the video frames is desired which may remind the doctors of the upcoming criticality.

Unfortunately, the existing off-line video analysis techniques in passive capsule endoscopy are not adequate any more, since these techniques generally assume that all the video frames have been obtained. While in active capsule endoscopy, we need to make predictions on the future images, based on the video frames acquired so far. In this paper, we start from the problems mentioned above and present a method that analyzes and predicts video images in an online manner, which has the potential to be used in active capsule endoscopies.

The paper is organized as follows. First, we review the developments of image-based analysis for current capsule endoscopy. Comparatively, we introduce our work which is designed towards both passive and active capsule endoscopies. Then we present in detail our solution. Next we report some promising experimental results. Finally we make conclusions and point out the future work.

II. Related Work

Our work is relevant to a few computing methods in computer vision. The techniques are proposed for passive capsule endoscopy, which focus on two lines.

- Feature Extraction: The first is to extract some features from the video frames (see [6], [7] and references therein). This involves the extraction of the features of blood, polyp, ulcer, the discriminating stomach, small intestine, colon, intestinal contractions, and so on. Different features may involve quite different extraction methods in practice.
- Classification: The second is to investigate the possibility of off-line classification of the images in capsule endoscopy (see [8] and references therein). Surely the feature extraction is a pre-requisite. Classification corresponds with a technique called supervised learning in artificial intelligence [13]. Using the technique, people wish to automate the process of identifying the patients' illness by constructing a classifier using labeled training data.

Manuscript received March 1, 2009. This research is supported by SHIAE project #8115021 of the Shun Hing Institute of Advanced Engineering of The Chinese University of Hong Kong, awarded to Max Meng.

Xiaona Wang is with Department of Electronic Engineering, the Chinese University of Hong Kong, Hong Kong SAR. xnwang@ee.cuhk.edu.hk

Max Q.-H. Meng is with Department of Electronic Engineering, the Chinese University of Hong Kong, Hong Kong SAR. max@ee.cuhk.edu.hk

III. The Problem

A. Formulation

Different from previous work, we try to predict the upcoming of video clips that should arouse the doctors' attention. This technique potentially facilitates the next generation of active capsule endoscopy. To the best of our knowledge, little work has been done explicitly to attain this objective.

Mathematically, given a sequence of video frames $F_1, F_2, \cdots$, we are seeking a binary decision function $E\left(F_m, M\left(\{F_1, \cdots, F_{m-1}\}\right)\right)$ which outputs either 1 when the system captures an interesting event at frame F_m, or 0 otherwise. The function takes two arguments as input. The first is the current frame F_m. The second is $M\left(\{F_1, \cdots, F_{m-1}\}\right)$, which gives the information carried by the frames prior to F_m. For such frames that $E\left(F, M\right)$ outputs 1, we regard them as indications of unusual or interesting events.

B. Applications

We identify a number of potential applications for our proposed model, in both passive and active capsule endoscopies.

- Prediction of Interesting Video Clips: Out of the tens of thousands of video frames produced in a typical procedure of capsule endoscopy, quite often only a small fraction (maybe only tens or even less) of the frames contain the disease information a patient is suffering, while most other frames do not. To save the doctor's reading time, it will be useful if we provide a solution that helps identify and predict interesting video clips. Essentially this can be casted as a clustering problem in data processing or segmentation problem in video processing [14]. In passive capsule endoscopy, we may use conventional computing techniques to identify interesting video segments and reduce the doctor's reading time. In active capsule endoscopy, further applications can be sought along this line. For example, a signal may be produced to attract the doctor's attention (or to ask him back when he is not on-site) when the computer has identified a potential interesting spot and predicted that the upcoming frames may be critical.
- Active Control of Capsule Movement: Another related usage is to help control the capsule movement actively. Suppose a capsule has moved to a spot suspicious of diseases, it is desirable that the capsule moves slower for more observations. When a capsule has moved to a place that is impossible to have symptoms, it should move faster to speed up the examination. An online or sequential analysis of video frames will be of help in deciding when to exert a control signal and what the signal should be. Surely this control also heavily depends on the underlying actuation mechanisms that is being developed, which goes beyond our discussion.

It can be seen that one fundamental distinction distinguishes our work from the previous. As we have mentioned, previous video-processing methods in this area typically work in an off-line or batch mode. The recorder outside the human body first receives the video signals from the inside capsule. The whole video is processed after all the frames have been accepted. While our work hopes to provide a solution that operates both off-line and online, without requiring to have gathered all the frames.

C. Relationship with Clustering Technology

For further illustration of the difference, let us take the first application, prediction of interesting clips, as an example. This application is of help in both passive and active capsule endoscopies. But it has respective natures in the two areas. In passive capsule endoscopy, the whole video has been acquired, and this reduces to a video segmentation problem or a conventional clustering problem. We need to identify different video segments or group the video frames together according to some criteria after an appropriate representation of images is given.

For active capsule endoscopy, however, we need an in situ method. We cannot assume all the images have been acquired. Instead, we are coping with data streams and need to make predictions on the future data. This task becomes more difficult than the conventional one. In our case, the critical point is to make events detection by predicting the upcoming of interesting frames, based on the experiences from the frames accumulated so far.

IV. Methodology

In this paper, we propose a method that monitors the video frames sequentially and detects unusual events online. In prediction of video clips, we need to mark the beginning and the end of interesting clips. In active capsule endoscopy control, we need to decide when to slow down and when to speed up. The key assumption to this problem is that the events should be detected when the computer catches a sudden change of potential diseases or other interesting features from video frames. The two applications essentially comply with our problem formulation in section III.

To make a computer identify and predict the critical frames like a medical expert, we need to do two things. The first is to encode the doctor's expertise into a computer system. The second is to let a computer analyze the video frames based on the expertise. Correspondingly, our solution has two steps. The first step is to construct a disease symptom space, which implicitly encodes the doctor's expertise, and represent each video frame as a point in the space. The second step is to analyze the importance of the data points based on a mathematically-consistent model.

A. Image Preprocessing

Preprocessing is a necessary procedure when analyzing images and videos. Similar to the previous work in feature extraction, we realized different processing methods for detection of different symptoms. In this paper, we restrict our discussion to two symptoms: bubble and bleeding. Surely the framework we developed here does not have this restriction. It is applicable to multiple symptoms, rather than two.

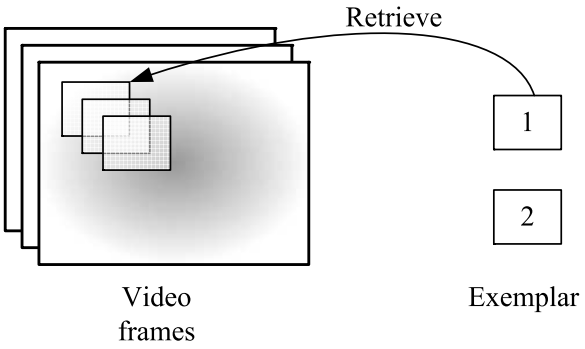

Fig. 1. An image retrieval model which maps each frame to a symptom space. We divide each frame into many regular grids, and use exemplar images to retrieve the grids. Each exemplar corresponds to a potential symptom in suspect. The highest similarity value between an exemplar and the grids is defined as the value of the frame to the corresponding symptom.

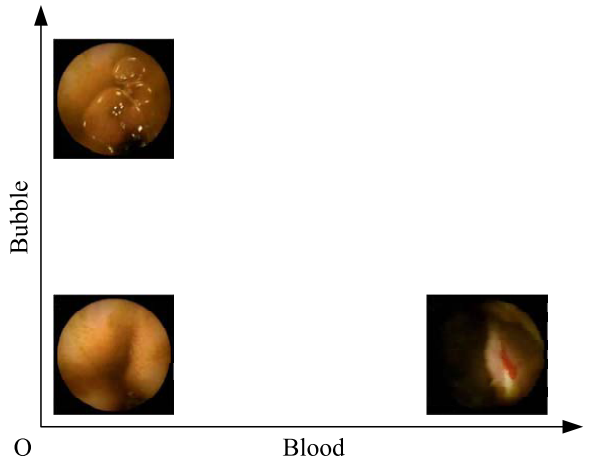

Fig. 2. A symptom space with two diseases. After retrieval, we associate each frame with a point in the symptom space. This example space has two symptoms: bubble and blood.

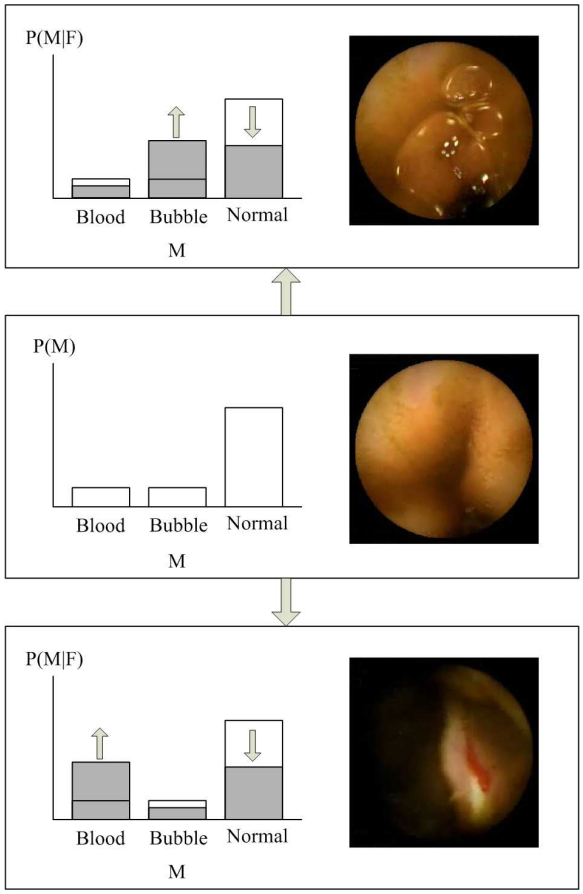

Fig. 3. An artificial example which depicts the change of scenes and the change of prior and posterior distributions, from the center to the upper or to the lower. Middle: Initially no symptoms are detected, and the possibility of normal is high. Upper: Something like a bubble suddenly appears, and the possibility of bubble increases significantly. Lower: Something like blood is suddenly detected, the possibility of blood increases significantly.

An insightful observation of bubble images may find a series of highlighted points in the image. These highlights are the reflections of the lights in the capsule. Correspondingly, our preprocessing mainly uses this feature as an identification and involves a blob extraction of highlights. For bleeding, it is relatively straightforward. We also use the blob extraction for the red regions.

B. Representation: Encoding the Medical Expertise

We realize a symptom world by a d-dimensional vector space, where d is the number of disease symptoms we are going to study. Each axis in the space is associated with a specific kind of symptom. The value in an axis indicates the degree of belief that a frame may contain the corresponding disease. We also construct for each symptom one or more exemplar images based on the medical knowledge. In doing so, we have implicitly encoded the medical expertise into the computer system.

Our current implementation adopts an enhanced histogram comparison method. Each frame in our experiments is an image having 640×480 pixels, with each pixel represented by an RGB color value. After preprocessing, a frame is further processed as follows. We divide it into 100×100 regular grids with overlaps, with each grid having 50×50 pixels. Then we use the exemplar images to retrieve all the grids in a frame (see fig. 1). To do this, we first make a color conversion from RGB representation to HSV representation [14]. Then we compute the corresponding intensity histogram by

$$g\left(h,s,v\right) = N \cdot \Pr\left(H=h, S=s, V=v\right)$$

where H, S, and V represent the three color channels, N is the number of pixels in the grid (in our case $N = 2,500$), and $\Pr$ denotes the probability of each specific color setting. Then we compare the histogram similarities between the exemplar images and the grids, based on the idea of intersection distance [15]

$$S\left(g,e\right) = \frac{\sum_h \sum_s \sum_v \min\left(g\left(h,s,v\right), e\left(h,s,v\right)\right)}{\min\left(\left|g\right|, \left|e\right|\right)},$$

where g denotes a grid's histogram, e denotes an exemplar image's histogram, and $|g|$ and $|e|$ give the magnitude of each histogram. Using this measure, colors not present in

the user's query images (exemplar images) do not contribute to the measure. This reduces the contribution of background colors. The summation is normalized within $[0, 1]$. A similarity value near 0 reflects that the grid does not contain much evidence of having the designated symptom; while a value near 1 reflects a high suspicion of having the symptom (see fig. 2).

After comparing the exemplar image with all the grids in a frame, the highest similarity value is used to be the value of the frame F in the corresponding axis of the disease symptom.

C. Model: Detecting Events When Capturing Sudden Changes

After pre-processing, each frame is represented as a d-dimensional vector, where d is the number of potential diseases we are interested in. The online analysis system (the observer) is initialized with some prior hypothesis on the possibility of having each disease. With the upcoming of each frame F, the hypothesis changes gradually.

This idea is formalized by Bayesian analysis approach. The observer captures the underlying frame information by the prior probability distribution $\{\Pr(M)\}_{M\in\mathcal{M}}$ over the hypotheses or models M in a model space $\mathcal{M}$. Given these prior beliefs, the fundamental effect of a new frame observation F on the observer is to change the prior distribution $\{\Pr(M)\}_{M\in\mathcal{M}}$ into the posterior distribution $\{\Pr(M|F)\}_{M\in\mathcal{M}}$ via Bayes theorem

$$\forall M \in \mathcal{M},\ \Pr(M|F) = \frac{\Pr(F|M)}{\Pr(F)} P(M).$$

Fig. 3 gives an example of belief updates with two symptoms.

During the analysis, a new frame F does not carry much information, or is not important, if it leaves the observer's beliefs unaffected, that is, if the posterior distribution is identical to the prior. Conversely, F carries much information, or should be important, if the posterior distribution resulting from observing F significantly differs from the prior. Therefore we formally measure the importance elicited by a frame F as a distance measure between the posterior and prior distributions. This is done using Kullback-Leibler (KL) divergence [16]. Thus, the importance of a frame F is defined by the average of the log-odd ratio

$$\begin{aligned} & I(F, M) \\ = \ & KL(Pr(M|F), Pr(M)) \\ = \ & \int_{\mathcal{M}} Pr(M|F) \log \frac{Pr(M|F)}{Pr(M)} dM \end{aligned}$$

taken with respect to the posterior distribution over the model space $\mathcal{M}$.

With the importance value for each frame, an interesting event can be identified by a boolean function

$$E(F, M) = \begin{cases} 1 & I(F, M) > \varepsilon \\ 0 & otherwise \end{cases}$$

where ε is a user-specified positive threshold. Now the observer is able to detect interesting events. An event occurs when the capsule returns a frame with a high importance value.

D. Justification of the Model

Although other methods (for example, a naïve thresholding method on the change of frame pixels) could be potentially useful in detecting unusual events in the frames, we are more interested in a solution that is mathematically consistent. Theoretically, our proposed method is based on a recent mathematical framework of *surprise* in computational neuroscience [17]. The notion is derived from principles and then formalized across general data types and data sources. Two elements are assumed when defining the concept. First, the surprise exists only in an uncertainty environment. Second, it is only defined in a relative, subjective manner and is related to the expectations of the observer. The same data may carry different amount of surprise for different observers, or even for the same observer taken at different times.

In probability theory, under quite mild assumptions, it can be shown that the only consistent and optimal way for reasoning about uncertainty is through the Bayesian inference [18]. So the work is done within the Bayesian theory of probability. In Bayesianism, probabilities correspond to subjective degrees of beliefs in hypotheses or models which are updated, as data is acquired. Bayes' theorem is used as the fundamental tool for transforming prior belief distributions into posterior belief distributions. In doing so, the work gives a consistent definition for inference.

Practically, the study has verified the relationship between Bayesian surprise and human attention. Experiments have revealed that human attention and gaze are often attracted by a scene with a high surprise score.

In our problem, we adopt this idea. We design a computer program to simulate a doctor's observation process. We make a hypothesis that a doctor's attention should be reminded by an event when the capsule observes a sudden change in the video frames. Here the "change" is problem dependent. It does not refer to a naïve change of pixels in scenes, but actually means a suddenly increased possibility of video clips having potential diseases. It is noteworthy that this is also the difference between our work and the mathematical theory. In practice, we need to encode the medical knowledge into the expert system, rather than simply applying a general mathematical theory.

E. Extension to Other Applications

So far, our discussion has focused on the first application, the prediction of interesting video frames. The idea can also be applied in the second application naturally, the active control of capsule directions. When a sudden explosion to diseases is observed, the computer may decide to adjust the capsule's direction and speed, so as to have a closer examination. Due to the delay between video signals and actual capsule movement, it is not appropriate that we make adjustment on the capsule's movement when it has returned a clear evidence of diseases. Instead, we need to do so

TABLE I

COMPARISON BETWEEN HUMAN OBSERVATIONS AND COMPUTER FIND-OUTS. EACH COLUMN INDICATES AN EXPERIMENT.

	EXP I	EXP II	EXP III
Test:	bubble	blood	bubble/blood
#(frames):	100	200	1000
#(human obs.):	4	4	10
#(computer obs.):	6	8	30
mark coverage:	100%	100%	80%
#(human frame):	17	24	65
#(computer frame):	23	33	127
frame coverage:	100%	100%	86.2%

beforehand when a sudden explosion of diseases just begins to appear.

Similarly, when the computer "feels" no possibility of symptoms at the capsule's current position, the capsule may be accelerated, and the direction is adjusted. The direction change instruction is made when a sudden deviation from the correct one is observed.

V. EVALUATIONS

To test the performance of the method, we carried out several experiments. We used three video samples: two were acquired from the medical school, the Chinese University of Hong Kong; one was downloaded from the web. The samples ranged from 100 to $1,000$ frames[1], roughly 1 to 10 minutes. We had three tests. The first tried to identify the bubbles during the examination; the second tried to identify bleeding; the third tried to identify the both symptoms.

A. Accuracy

In the first experiment, we first let a person with basic medical knowledge watch the clips. He made a mark when he think the video has reached a sudden appearance of bubbles, which indicated a beginning of bubble frames, or a sudden disappearing of bubbles, which indicated an end of bubble frames. The results are summarized in table I.

The experiment used 100 frames. The human participant marked 4 frames. Then we run the computer program and computed the importance value for all frames, and selected 6 marks with highest importance values. And we found all the 4 marks observed by the human were contained in the 6 frames found out by the computer program if we allow for $+1$ or -1 frame deviation[2]. Based on the marks selected by the human observer, there were totally 17 frames relevant with bubbles, which were 100% covered by the 23 frames based on the computer marks.

The second experiment used a video clip with 200 frames to test bleeding. The human observer selected 4 marks which might be of interest to check, which were covered by 8 marks selected by the computer. The frame coverage is also 100%.

[1]The largest one with $1,000$ frames is actually a combination of several video clips.

[2]Besides identical observations, we count the case that the person marks frame s while the computer selects frame $s+1$, or vice versa. This deviation is also allowed in frame coverage calculation.

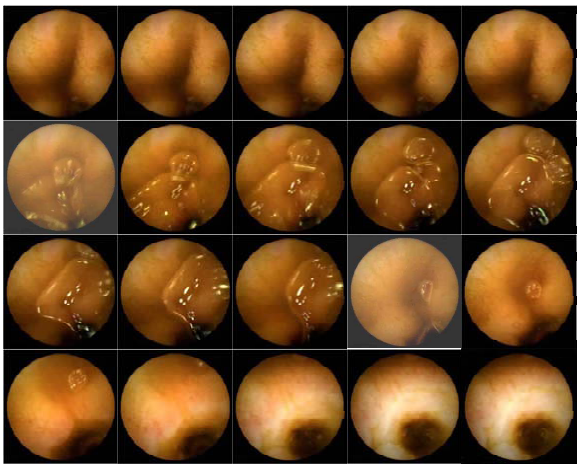

Fig. 4. For a sequence of video frames, the method detected two frames (with transparency) which marked the begin and the end of the bubble occurrences.

The third experiment used a video clip with 1000 frames to test both bubble and bleeding. The human observer selected 10 marks, most of which (8 out of 10) were covered by the 30 marks with highest importance values selected by the computer. The frame coverage (86.2%) is also acceptable.

As an example, figure (4) depicts the two marks found by the computer in a 10-second episode of the first experiment.

B. Speed

As for the computational speed, we implemented the method in MATLAB without special optimization and run the code on a computer with intel Pentium IV-2.53GHz CPU and 1G memory. On this PC, the method gave satisfactory performance. The total computation time on each frame, including histogram analysis and importance calculation, did not exceed 0.1 second. Since the method had left much room for real-time response, we did not carry out the theoretical analysis on the method's complexity any more.

VI. CONCLUSIONS

The capsule endoscope is widely used in examination of the small intestine diseases. The active control mechanism represents a direction for the next generation capsule endoscopy. So we consider to design computing techniques towards the upcoming active techniques.

In this paper, we have proposed a new method of analyzing the video frames in capsule endoscopy. Comparing with existing solutions, it provides an active way in analyzing the videos, both off-line and online. To the best of our knowledge, this provides the first event detection-based technique for the capsule endoscopy and can be potentially applied in active controls.

Two key points to our solution are to encode the medical knowledge into the system and to detect interesting or unusual events. We tackled the two points in a simple way. Although our current work mainly focuses on the theoretical part, the preliminary evaluations have reported promising results of the solution.

Related to the future, more evaluations will be carried out to test the applicability of the method. On one hand, we have used a simple histogram comparison-based approach to identify bubble and bleeding. More sophisticated features, for example geometric features, will be used to identify other symptoms of diseases, such as Crohn's disease, gastric ulcers, and colon cancer. On the other hand, a close collaboration with the medical experts will be sought before the deployment of the real-time monitoring system in empirical applications.

ACKNOWLEDGMENT

The authors acknowledge the discussion with Dr. Wenye Li in University of Alberta.

REFERENCES

[1] G. Iddan, G.Meron, A. Glukhovsky, and P. Swain, "Wireless capsule endoscopy," *Nature*, vol. 405, p. 717, 2000.

[2] A. Fritscher-Ravens and C. P. Swain, "The wireless capsule: New light in the darkness," *Digestive Diseases*, vol. 20, pp. 127–133, 2002.

[3] W. A. Qureshi, "Current and future applications of the capsule camera," *Nature reviews drug discovery*, vol. 3, pp. 447–450, May 2004.

[4] E. Rondonotti, J. Herrerias, M. Pennazio, A. Caunedo, M. Mascarenhas-Saraiva, and R. de Franchis, "Complications, limitations, and failures of capsule endoscopy: a review of 733 cases," *Gastrointestinal Endoscopy*, vol. 62, no. 5, pp. 712–716, 2005.

[5] D. Rowbotham, "Ulcers,lies,and video speed:does clinical experience matter in wireless capsule endoscopy?" *Gastrointest Endosc.*, vol. 57, p. M1877, 2003.

[6] J. Berens, M. Mackiewicz, and D. Bell, "Stomach, intestine, and colon tissue discriminators for wireless capsule endoscopy images," in *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, vol. 5747, 2005, pp. 283–290.

[7] M. T. Coimbra and J. P. da Silva Cunha, "MPEG-7 visual descriptors - contributions for automated feature extraction in capsule endoscopy," *IEEE Trans. Circuits Syst. Video Techn*, vol. 16, no. 5, pp. 628–637, 2006.

[8] F. Vilariño, L. I. Kuncheva, and P. Radeva, "ROC curves and video analysis optimization in intestinal capsule endoscopy," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 875–881, 2006.

[9] A. Menciassi and P. Dario, "Bio-inspired solutions for locomotion in the gastrointestinal tract: Background and perspectives," *Phil. Trans. R. Soc. Lond. A*, pp. 03TA1109/1–03TA1109/11, 2003.

[10] M. Q. H. Meng, T. Mei, J. Pu, C. Hu, X. Wang, and Y. Chan, "Wireless robotic capsule endoscopy: State-of-the-art and challenges," in *Proceedings of 5th World Congress on Intelligent Control and Automation (WCICA04)*, Hangzhou, June 2004, pp. 5561–5565.

[11] A. Kornbluth, P. Legnani, and B. S. Lewis, "Video capsule endoscopy in inflammatory bowel disease: Past, present, and future," *Inflammatory Bowel Diseases*, vol. 10, no. 3, pp. 278–285, 2004.

[12] X. Wang and M. Q.-H. Meng, "A magnetic stereo-actuation mechanism for active capsule endoscope," in *the 29th International Conference of the IEEE Engineering in Medicine and Biology Society*, Lyon, France, August 2007.

[13] S. Russel and P. Norvig, *Artificial Intelligence - A Modern Approach*. Englewood Cliffs: Prentice-Hall, 1995.

[14] D. A. Forsyth and J. Ponce, *Computer Vision: A Modern Approach*. Prentice Hall, 2003.

[15] M. J. Swain and D. H. Ballard, "Color indexing," *International Journal of Computer Vision*, vol. 7, no. 1, pp. 11–32, 1991.

[16] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: John Wiley & Sons, 1991.

[17] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," in *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Schölkopf, and J. Platt, Eds. Cambridge, MA: MIT Press, 2006, pp. 547–554.

[18] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, 1988.