IEICE TRANS. INF. & SYST., VOL.E92–D, NO.3 MARCH 2009

PAPER

Controlling the Display of Capsule Endoscopy Video for Diagnostic Assistance

Hai VU[†a)], ***Nonmember***, **Tomio ECHIGO**[††], **Ryusuke SAGAWA**[†], ***Members***, **Keiko YAGI**[†††], **Masatsugu SHIBA**[††††], **Kazuhide HIGUCHI**[††††], **Tetsuo ARAKAWA**[††††], ***Nonmembers***, ***and*** **Yasushi YAGI**[†], ***Member***

SUMMARY Interpretations by physicians of capsule endoscopy image sequences captured over periods of 7–8 hours usually require 45 to 120 minutes of extreme concentration. This paper describes a novel method to reduce diagnostic time by automatically controlling the display frame rate. Unlike existing techniques, this method displays original images with no skipping of frames. The sequence can be played at a high frame rate in stable regions to save time. Then, in regions with rough changes, the speed is decreased to more conveniently ascertain suspicious findings. To realize such a system, cue information about the disparity of consecutive frames, including color similarity and motion displacements is extracted. A decision tree utilizes these features to classify the states of the image acquisitions. For each classified state, the delay time between frames is calculated by parametric functions. A scheme selecting the optimal parameters set determined from assessments by physicians is deployed. Experiments involved clinical evaluations to investigate the effectiveness of this method compared to a standard-view using an existing system. Results from logged action based analysis show that compared with an existing system the proposed method reduced diagnostic time to around 32.5 ± 7 minutes per full sequence while the number of abnormalities found was similar. As well, physicians needed less effort because of the systems efficient operability. The results of the evaluations should convince physicians that they can safely use this method and obtain reduced diagnostic times.

key words: *capsule endoscopy, color similarity, motion displacements, video display rate control.*

1. Introduction

Capsule Endoscopy (CE) involves a swallowable endoscopic device that is propelled by peristalsis through the GastroIntestinal (GI) tract. Through its image capturing ability, CE enables non-invasive examinations in the GI tract that are difficult to carry out by conventional endoscopic techniques. CE has been reported [1]–[3] to be particularly successful in finding causes of gastrointestinal bleeding of obscure origin, Crohn's disease, and suspected tumors of the small bowel. The clinical products, M2A and PillCam capsule [4], developed by Given Imaging Ltd, Israel, have become widely used with over 500,000 patients examined worldwide [5]. In a typical examination, CE takes approximately 7–8 hours to go through the GI tract for acquisition of images at a rate of 2 fps. The sequence obtained thus has around 57,000 images that can be used for reviewing and interpretation. With such a large number of images, an examination is time consuming and constitutes a heavy workload for physicians.

To reduce diagnostic time, different viewing modes for displaying images are provided in the RAPID Reader application [6], a CE annotation software developed by the capsule manufacturer. *Dual-view* mode reduces analysis time by concurrently displaying two consecutive frames. *Quad-view* reshapes four consecutive images into one. *Automatic-view* combines successive similar images to display representative frames. *Quick-view* mode allows a fast preview by showing only highlight images. The combination of *dual-view* and *automatic-view*, called a standard-view, is a common viewing mode for physicians. Following medical reports [2], [3], [7], [8] and a specific report by [9] that included the examination of 50 sequences, the average time taken to examine a sequence in a standard-view is reported to be approximately 76 ± 30 minutes. In *quad-view* mode, the average diagnostic time can be reduced to around 37 ± 13.4 minutes/sequence [10]. *Quick-view* allows preview sequences of around five minutes; however, in the application it is recommended that additional evaluations are required to confirm that there has been no loss of abnormal regions.

Although convenient methods that reduce diagnostic time for physicians are useful, they have the constraint that images must be displayed in the original/natural shape without any skipping of frames. This is because there are various challenges in the examination of CE videos that require careful attention, even by experienced physicians. For example, because of movements of the CE device caused by natural peristalsis, images are captured from different viewing directions that can make even normal anatomy look strange. The distorted images in *quad-view* mode can be difficult to interpret. An abnormality that may only be seen in a single frame, or in a few frames [3], is not easily identifiable in *quick-view* mode and may not appear if that image were to be skipped.

Many video analysis technologies to reduce the attention time for video editors, as well as to achieve reductions in the storage and transmission of video sequences, have been proposed. A survey [11] has categorized two kinds of

Manuscript received May 20, 2008.
Manuscript revised October 17, 2008.

[†]The authors are with the Institute of Scientific and Industrial Research, Osaka University, Ibaraki-shi, 567–0047 Japan.
[††]The author is with the Dept. of Engineering Informatics, Osaka Electro-Communication University, Neyagawa-shi, 572–8530 Japan.
[†††]The author is with the Kobe Pharmaceutical University, Kobe-shi, 658–8558 Japan.
[††††]The authors are with the Graduate School of Medicine, Osaka City University, Osaka-shi, 545–8585 Japan.
a) E-mail: vhai@am.sanken.osaka-u.ac.jp

DOI: 10.1587/transinf.E92.D.512

video summarization: the first technique is to select a small number of still images as key frames generated from scene change detection algorithms. Some tools [12]–[15] aim to segment data semi-automatically into domain objects that are meaningful to users for tasks related to video searching, browsing or retrieving. The second technique uses moving images to skim a video sequence. Some multimedia applications such as VAbstract [16] and MoCa Abstracting [17] have been developed to provide users with an impression of the complete sequence or highlights containing the most relevant parts of the original video. [18] proposed a method called "video fast forward", which aims to browse desired clips more quickly as opposed to using key frame-based summaries. [19], [20] describes a concept called "constant pace", that provides for varying the display speed by motion activity and semantic features such face or skin color appearance, speech, and music detection. However, considering the requirements for the display of CE images, intuitively, these techniques appear unsatisfactory.

In this scenario, the conditions of image acquisition are the important cues. As CE involves a passive device, its states during the capture of images depend on the motility patterns in the GI tract. The video sequence can be played at high speed in a stable state to save time, and the speed then decreased during rough changing states to more conveniently help identify suspicious regions. This fits with an opinion discussed in [3] that "it is probably unwise to read all the images at the fastest of the three available speeds" and the fact that physicians usually stop and inspect sequences frame-by-frame to recognize suspicious regions. Although physicians can adjust the display speed from 5 to 25 frames per second, changing this speed manually can break an examiners concentration for finding abnormal presentations. Therefore, in this paper, we propose a method to automatically control image display that is built upon the idea that durations for displaying frames (herein called the delay time) are adapted according to the different conditions governing image acquisition. It is notice that the proposed system is designed to reduce diagnostic time without the loss of any abnormal region under the same conditions as the existing system, assistant functions for automatically recognizing abnormal regions are not included. A comparison between images displayed according to the proposed method and those displayed at a fix frame rate is shown in Fig. 1. The contributions made by this article are as follows:

- The study proposes a method that effectively assists physicians by reducing the time for CE videos diagnoses. The main advantages are that entire sequences are displayed in the original shape without skipping any frames; thereby enabling the inspection of all data. Experimental results confirm that the diagnostic time is reduced to around 32.5 ± 7 minutes per full sequence. Compared with a standard-view using the existing system, Rapid Reader Version 4, the proposed method is 10 minutes less while the number of abnormalities found are similar under both systems. As well, the proposed system requires less effort because of its efficient operability.

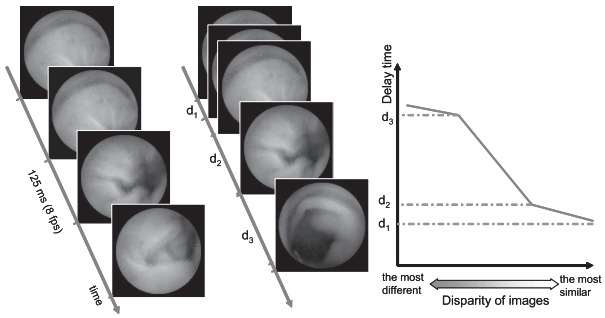


Fig. 1 Image display under the proposed method using adaptive display rate (middle), controlling by the disparity of images (right side) versus the conventional method with its fix frame rate (left side).

- To address issues of subjectivity in reducing diagnostic times, a series of clinical evaluations are conducted. We utilize a logged action based analysis to validate the proposed techniques. These results should convince physicians that they can safely use this approach in routine clinical work and still obtain reduced diagnostic times. Furthermore, to the best of our knowledge, it is the first time the detail actions of physicians are analyzed within the complete diagnostics procedure in respect to the application of CE. Logged actions provide comprehensive data for a better understanding of the behavior of physicians. This can be deployed for further research into areas such as appropriate education systems or assistance in recognizing the presence of abnormalities.

The rest of the paper is organized as follows: Section 2 introduces CE image properties and the techniques of feature selection and extraction. Section 3 describes a method for defining and classifying the states of image acquisition. Section 4 explains the functions used to compute the delay time and the techniques used to precisely control image display. In Sect. 5, we investigate the effectiveness of the proposed method through clinical evaluations. Finally, in Sect. 6, we conclude as well as discussing and suggesting areas for future research.

2. Capsule Endoscopy Image Properties and Feature Extractions

2.1 CE Image Properties and Feature Selections

The CE device is 11 mm by 26 mm, includes a CMOS sensor, a short focal length lens, four LED illumination sources and an antenna/transmitter (Referring to [2] and [7] for technical specifications). With its pill shape and small size, the CE device is easily ingested and then passed through the GI tract. Image data are transferred by radio transmission to a recording unit before being uploaded to a workstation for diagnosis. Image features include a 140^o field of view, 1 to 30 mm depth of view, 256×256 pixels and 24 bit color in RGB space.

CE images usually present homogeneous regions inside the GI tube. Similar to images acquired by conventional

endoscopy devices, CE images of the digestive organs show differences in texture, shape or color features. Because of the reflecting activities of the digestive system, the regions of the inner GI wall not only span continuous frames (in stationary phase) but also appear separately in an individual frame (in contraction situations). However, the visibility in CE images is sometimes limited because of the presence of intra-luminal gases, water bubbles, food or bile. To characterize the meaning of the various CE images, previous works have selected different image features depending on their goal. For example, to segment digestive organs, color features are favored in the works of Mackiewicz et al. ([21], [22]) and Coimbra et al. ([23], [24]). For calculating different plurality images, Glukhovsky et al. [25] use the average intensity of pixels or pixel clusters. To detect contractions in the small bowel, the gray-level intensity and edge features of intestinal folds are used in the works of Vilarino et al. ([26], [27]) and Spyridonos et al. ([28], [29]). To produce a map that represents the GI's surface in smooth-continuous frames, P. Szczypinski et al. ([30], [31]) estimate the motion displacement between frames. For the purpose of our investigation, to precisely control the image display, image features are selected so that the perceptual disparity of consecutive frames is as precise as possible. From observations in experiments, the changing of color features is useful for extracting global differences between images, whereas motion displacements are distributed unevenly in a small area or imply just local information. Therefore, we mainly focus on combinations of these features because changes in consecutive frames can adequately discriminate both global and local information.

In [25], Glukhovsky et al. introduced a framework for controlling the in vivo camera capture and display rate. After evaluating differences of the multiplicity of frames, they suggested an empirical database or a look-up table so that the display rate is varied accordingly. However, they leave unresolved the method needed to develop this type of database, look-up table, or a specific mathematical function. In our work, if the delay time between two consecutive frames is denoted by D_t, we express the correlating function between D_t and the disparity of images by:

$$D_t = \Theta(f(.), \xi_{skill}, \xi_{system}) \tag{1}$$

where $f(.)$ is a function to estimate perceptual differences between frames by color similarity and motion displacement. In preliminary versions of this study ([32]), we described the methods for extracting these image features. The sections below express in some depth the techniques used for the implementation of the proposed method.

2.2 Features Extractions

2.2.1 Color Similarity Extractions

Several methods to extract color features have been proposed for content-based image retrieval (CBIR). These include color histograms [33], color moments and color coherence vectors [34] and color correlograms [35]. Benchmarks and capacity color histograms have been reported in [36] and [37]. These reports show that color histograms are robust through a trade-off between performance and computation time. Therefore, the use of color histograms is a promising way of quickly indexing a large number of frames, such are found in a CE sequence.

Color descriptions can utilize different color spaces such as RGB or HSV. HSV color space is good for detecting abnormal regions because it offers improved perceptual uniformity. However, it is not so good for detecting time-varying color changes because the color space is not stable in a dark scene. Furthermore, there are no reasons to use different color spaces against the color spaces of the original input and display, respectively. From preliminary experiments using a bright scene, RGB and HSV color spaces showed a high correlation for the two similarity waveforms. Therefore, so that it is unnecessary to transform to another color space, the original color space of RGB is used for color histogram indexing.

In our implementations, CE images are divided into small blocks and a histogram is computed for each block. Block size value was decided heuristically through experiments with various block size values. With a small block size, image differences show sensitivity to the changes, whereas a too large block size can lose the changes in important regions. For a reasonable selection, the image is divided into $N_{blk} = 64$ blocks with a predetermined 32×32 pixels block size. The color histogram method [33] is applied to each block by dividing R, G, B components into a number of bins $N_{bins} = 16$. The distance of the local histograms is computed from the L1 distance:

$$D_{blk}(i) = \sum_{k=1}^{N_{bins}} (|H^n_{R,k} - H^{n+1}_{R,k}| + |H^n_{G,k} - H^{n+1}_{G,k}| + |H^n_{B,k} - H^{n+1}_{B,k}|) \tag{2}$$

where H is the histogram of each color component for block i and between frames $< n, n+1 >$.

Block matching between frames $< n, n+1 >$ is decided using a selected threshold value. The accumulation of matching blocks reveals overall similarity between two frames:

$$Sim(n) = \frac{1}{N_{blk}} \sum_{i=1}^{N_{blk}} sim_{blk}(i)$$
$$\text{With} \begin{cases} sim_{blk}(i) = 1 \text{ if } D_{blk}(i) \le Thresh_{blk} \\ sim_{blk}(i) = 0 \text{ otherwise} \end{cases} \tag{3}$$

Using (3), color similarity ($Sim(n)$) is normalized from zero to one, with the maximum value indicating the best match and the minimum value indicating that with the most difference. Additionally, the maximum distance between blocks $D_{maxblock}(n) = max_i\{D_{blk}(i)\}$ is also noted; this distance is particularly robust in situations when two images

have some common regions. In contrary situations, the minimum block $D_{minblock}(n) = min_i\{D_{blk}(i)\}$ is used to ascertain if the regions of the images are mostly different. We discuss utilizing these values in Sect. 3.

2.2.2 Motion Displacement Estimations

Using color similarity, the disparity between consecutive frames was evaluated in terms of global information. Motion features are cue information for representing the local displacement of adjacent frames. Motion is usually represented by set trajectories of the matching points of local features. In this study, the Kanade-Lucas-Tomasi (KLT) algorithm was utilized to estimate the displacement because it showed reliable results that emphasized [38] the accuracy and density of measurements for real image sequences. As well it has been reported to be successfully applied to conventional endoscopic images [39], [40]. This algorithm is a feature-tracking procedure developed for video by Tomasi and Kanade [41]. It is based on earlier work by Lucas and Kanade [42]. Extensions of the KLT algorithm [43] include support for a framework of a multi-resolution scheme [44] and constraints of affine transformation [45].

First, images are smoothed using a 2D Gaussian filter with standard deviation $\sigma = 1.5$. Applying a pre-filter before detecting good feature points is effective for improving the signal-to-noise ratio and reducing the non-linear components of any image that might tend to degrade the accuracy of subsequent gradient estimations. Smoothing also helps attenuate temporal aliasing and quantization effects in the input images.

For each pair of consecutive frames, the KLT algorithm automatically selects good features from the first image. A good feature is one that can be tracked throughout the following frames. The selection of good features is based on the requirement that the spatial gradient matrix computed on the corresponding frame location is above the noise level and is well conditioned. As defined [38], the gradient matrix G is computed by:

$$G = \int_{\omega} g(g^T)\omega dx = \sum_{i,j} \begin{bmatrix} grad_x(i,j) * grad_x(i,j) & grad_x(i,j) * grad_y(i,j) \\ grad_x(i,j) * grad_y(i,j) & grad_y(i,j) * grad_y(i,j) \end{bmatrix} \tag{4}$$

To compute the gradient in the x and y direction of the images, a Gaussian 2D kernel, with $\sigma = 1.0$ is applied. The gradient matrix is built from a patch window $\omega = 29 \times 29$ pixels size. The noise requirement implies that both the eigenvalues of matrix G must be sufficiently large, while the conditioning requirement means that the eigenvalues cannot differ by several orders of magnitude. To ensure that the noise requirement is satisfied and well conditioned, the patch window ω is accepted as a good feature if the two eigenvalues (λ_1, λ_2) of matrix G satisfy the condition: $min(\lambda_1, \lambda_2) > \lambda$, where λ is a predetermined threshold. A lower bound of λ is given by the distribution of elements of the gradient matrix with constant intensity, while the upper bound obtained is an area with variable intensity. In practice, to determine a good feature point we use $\lambda = 800$, chosen as the halfway point between the two bounds.

The process of selecting a good feature point finishes when the condition is reached by a certain number of points ($N_{points} = 80$) or distances between two good features are no smaller than a predefined value (7 pixels). With many homogeneous regions in endoscopic images, the trade-off against the computational cost of the number of good feature points required is not large and the minimum distance between them is not particularly small.

A computation framework for the measurement of visual motion also showed robust results when deployed by a multi-resolution scheme [44] in a coarse-to-fine manner. In our implementation, estimations are first produced at coarse scales by reducing the original size four fold; where the noise is assumed to be less severe, with velocities of less than 1 pixel/frame. These estimates are then used as the initial guesses for a finer scale (by restoring the original size of the images) to compensate for larger displacements.

The good features are then tracked in a second image at each scale using Newton-Raphson iterations to minimize the differences between the windows in successive frames. The tracking process stops when either the number of iterations (predefined value = 10) or the minimum distance is above a selected threshold value or the residue of patch windows is too large. Figure 2 shows the motion fields for some frames in a sequence that includes 16 continuous frames (upper panel). The results of frames 1 to 6 and 8 to 14 show that motion estimations are clear and realizable (as shown in Fig. 2 a and Fig. 2 c. At position (b) (frames 6 and 7) and (d) (frames 14 and 15) the results of the motion fields are a mess (as shown in Fig. 2 b and Fig. 2 d). These problems are resolved by the combination of color similarity, as described in Sect. 3.

Some methods for evaluating displacement from the motion field have been proposed. For example, [19] used the average magnitude of motion vectors. However, as shown in Fig. 2, the dominant movement is in an unrecognizable direction for endoscopic images so using an averaged value here is not feasible. Therefore, to evaluate motion signal strength between two adjacent frames $< n, n+1 >$, in this study the maximum magnitude of motion vectors notation $Motion_{orig}(n)$, is used.

To combine the color similarity feature, motion displacement is normalized in the range of $[0, 1]$. To avoid any bias resulting from non-realizable cases, the Z-Score normalization (Gaussian normalization) method [46] is used. From $Motion_{orig}$ data, the mean μ_k and standard deviation σ_k of a full sequence is calculated. The $Motion_{orig}(n)$ of a frame number n is normalized by:

$$Motion_{norm}(n) = \frac{Motion_{orig}(n) - \mu_k}{3\sigma_k} \tag{5}$$

The probability of normalization by (5) being in the

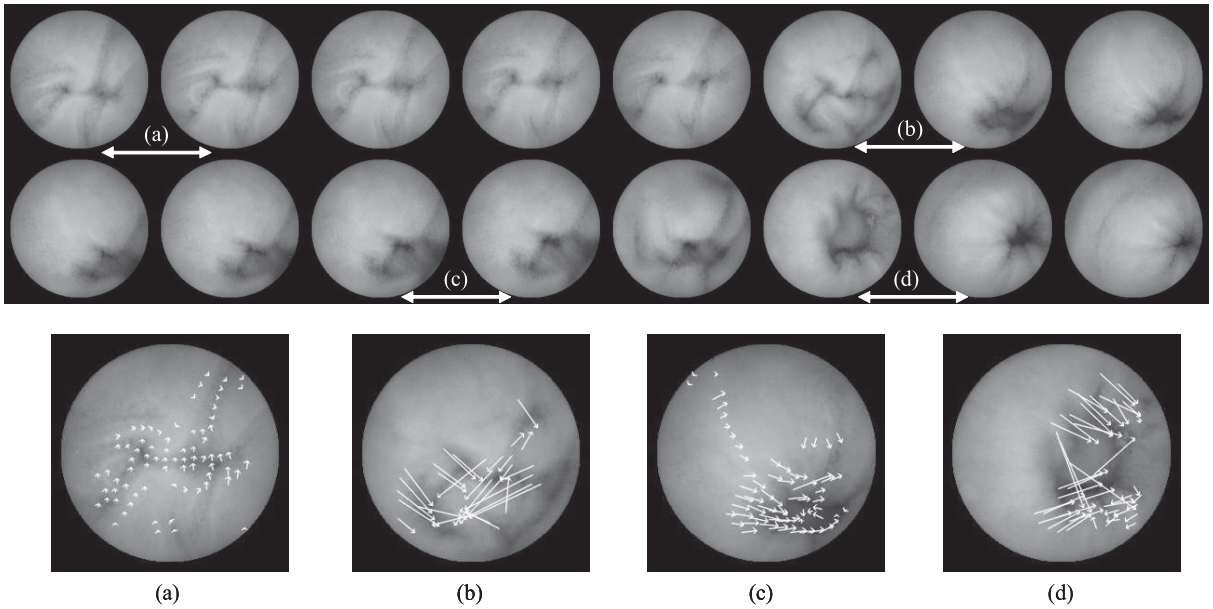


Fig. 2 A continuous image sequence of 16 frames (upper panel). Results of motion estimations at some positions are illustrated (bottom panel). At (a) and (c), the results of the motion fields are reliable, while at (b) and (d) the motion fields are not confident.

range of $[-1, 1]$ is approximately 99%. In practice, all values can be considered within the range of $[-1, 1]$ by mapping the out-of-range value to 1 or -1. A shift operator will transform values to the range of $[0, 1]$ by:

$$Motion(n) = \frac{Motion_{norm}(n) + 1}{2} \tag{6}$$

The procedures for feature extractions were performed off-line on a Pentium IV 3.2 GHz, 2 GB RAM computer. Average computational cost for a full sequence was approximately 105 minutes, including 30 minutes for color similarity and 75 minutes for motion estimations.

3. Classification Scheme

Studies in the field of gastrointestinal motility show that motility patterns in the GI tract include two types of contractions. One is peristalsis where muscles contract in a synchronized way to move food in one direction. The other is segmentary contractions where muscles in adjacent parts squeeze to mix the contents but do not move the food [47]. Motility patterns are known to occur at infrequent intervals and vary depending on the phase of the contraction as well as the presence of various malfunctions. Recognizing motility patterns from CE image sequences is still a difficult task. However, the mechanisms reveal an idea for classifications into states of changes between two consecutive frames that correspond to the conditions of image acquisitions. Here, four states of image acquisitions can be defined. Descriptions of these states and a scheme for classifications based on the extracted image features are discussed below.

3.1 Descriptions of the States of Image Acquisition

For convenience, the four states corresponding to changes in contractions in the small bowel are presented in Fig. 3 a–d:

- *State 1*: Images are captured in a stationary condition. This state appears when the GI motility is in a stable phase. Thus, the position of capsule remains almost still. Figure 3 a shows 8 frames extracted from 195 successive frames, which are almost all the same. The adjacent images have high color similarity and motion displacements are small or nearly zero. This state impacts on the control of the display images by playing sequences at high speed to save time. When continuous frames are exactly the same, the display speed can reach a maximum value that can be set according to the limitations of the display system's hardware.

- *State 2*: The CE device captures images when it moves with just gradual transitions and there is no change in the viewing direction. This state corresponds with moments when the peristaltic contractions are strong enough to move the capsule by pushing it, but there is no effect from the segmentary contractions that mix or sweep the contents in the GI tract. Figure 3 b shows some frames at the beginning, middle, and the end parts of 52 continuous frames, being consecutive frames with small movements. There are not many differences in the changing colors and so the motions can be confidently estimated. In this state, the display of images is controlled at a medium speed so that observation is possible.

- *State 3*: Images are captured when the capsule undergoes larger movements. The strong contractions that sweep or mix the contents are considered to cause this state. As

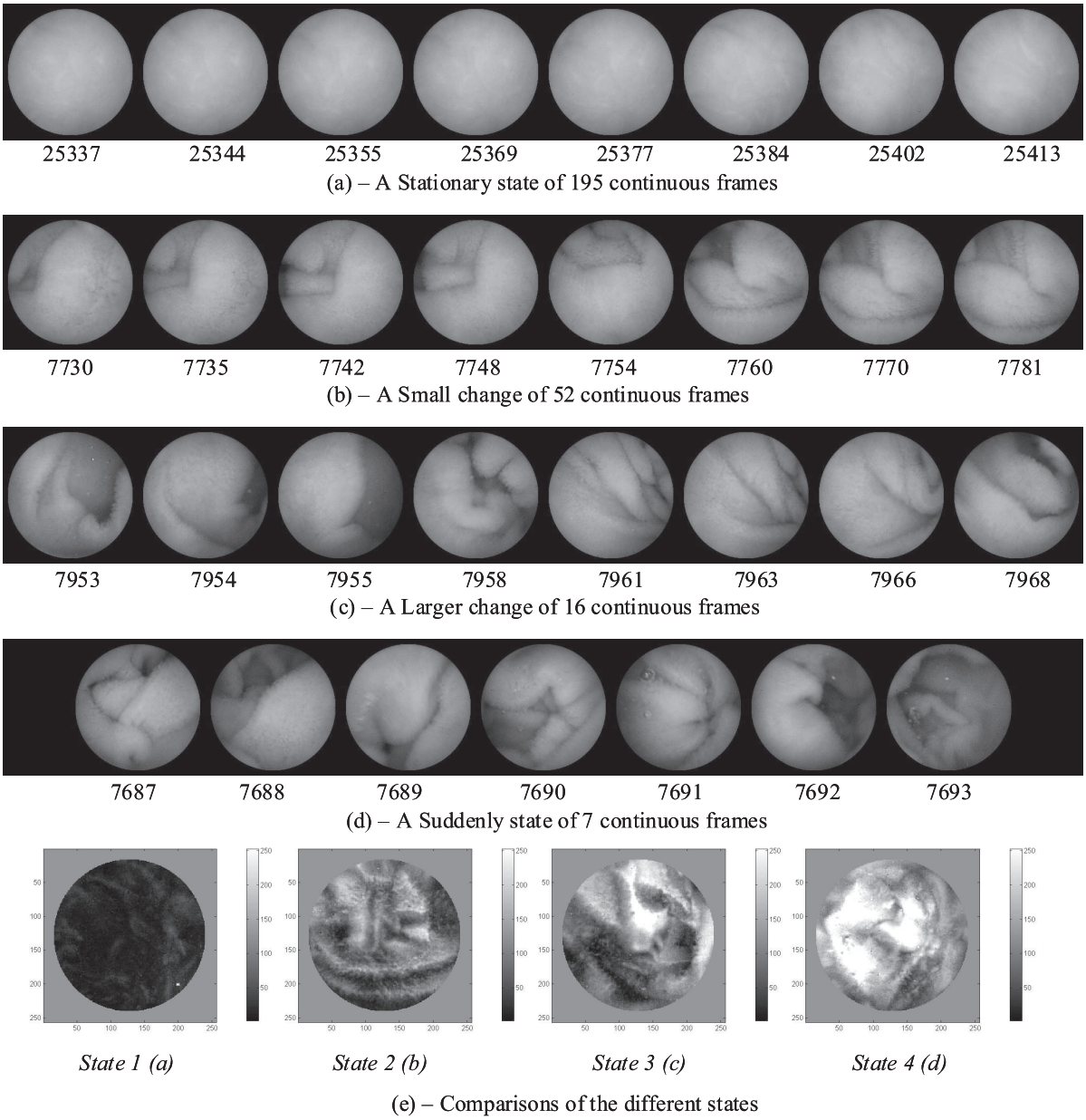


Fig. 3 (a–d) States of image acquisition. (e) A comparative differences between images for corresponding states. Pixel values at (i, j) in each sub-figure is plotted by maximum values from differencing of adjacent images $< t, k >$ shown in (a)–(d). The image differencing is calculated by $g_{i,j}^{t,k} = \sum_{R,G,B} |f_{i,j}^{t} - f_{i,j}^{k}|$. The gray scale bar presents image differencing with a brighter intensity showing a larger change.

shown in Fig. 3 c, in 8 of 16 continuous frames the movements in successive frames are larger and clearer than in the frames in Fig. 3 b. Images in this state would be displayed over a longer time so that physicians are able to clearly view them and focus better on the changing regions.

- *State 4*: This state occurs when there are brief bursts of contractions or giant migrating contractions. This type of contraction makes the capsule suddenly change direction and move. Figure 3 d shows images captured in this state with 7 continuous frames that are essentially different. Color similarity is minimal, and the motion vectors can not be confidently detected. The delay time is thus increased to the maximum to enable observations to be as easy as possible.

3.2 Decision Tree for Classifying States

With natural characteristics of GI motility, the states classification task is faced with the problem that a reasonable performance can only be achieved by using of a very large design set for proper training; probably much larger than the number of frames available. Such a difficulty can be overcome based on the above descriptions of the states in

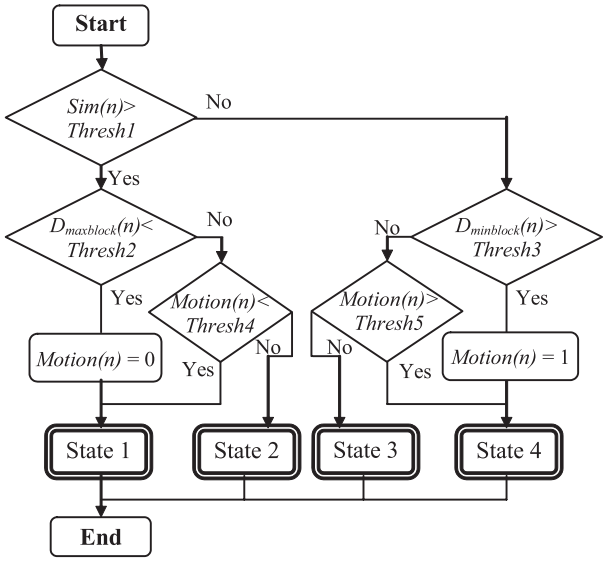


Fig. 4 A decision tree for classifying states.

which the color similarity is the most discriminating feature for separating global changes (e.g., stationary states (*State 1*) vs. abrupt changes (*State 4*)), while motion displacement is clearly used for discriminating small adjustments (e.g., stationary states (*State 1*) vs. gradually change (*State 2*)). With such discriminations of feature subsets, a "divide and conquer" principle, or a decision tree classifier, is usually applied. For classifying an unknown pattern into a class in successive stages, a decision function at a certain stage can perform rather well by using the discriminating feature [48]. Therefore, a decision tree as shown in Fig. 4 is proposed. In this, color similarity $Sim(n)$, maximum block $D_{maxblock}(n)$ and minimum block $D_{minblock}(n)$ are defined in Sect. 2.2.1; motion displacement $Motion(n)$ is calculated by (6).

Following this classifier, *State 1* is satisfied if $Sim(n)$ is larger than $Thresh1$ and $D_{maxblock}(n)$ is smaller than $Thresh2$. Because when $Thresh1$ is large and $Thresh2$ is small enough, the combination of these conditions means that all regions in the two images are stable. To evaluate gradual transitions of *State 2*, $Motion(n)$ is larger than $Thresh4$ (which separates the stationary *State 1*); most regions are similar ($Sim(n) > Thresh1$) and differences only appear in some regions by $D_{maxblock}(n)$ being larger than a predefined threshold ($Thresh2$). Similarly, *State 3* and *State 4* are defined as relying on $D_{minblock}(n)$ and $Motion(n)$ compared with $Thresh3$ and $Thresh5$. Particularly, if $D_{minblock}(n)$ is larger than $Thresh3$, it means that the abrupt changes that cause errors in the motion estimations result in the motion being assigned to *State 4* (the motion fields in the cases in Fig. 2 b and Fig. 2 d are avoided).

Note that following the classification scheme, motion features for *State 1* and *State 4* are always assigned 0 and 1, respectively. As such, no motion estimations are required in these cases. Thus, computational cost is reduced because motion extractions are only implemented for *State 2* and *State 3*.

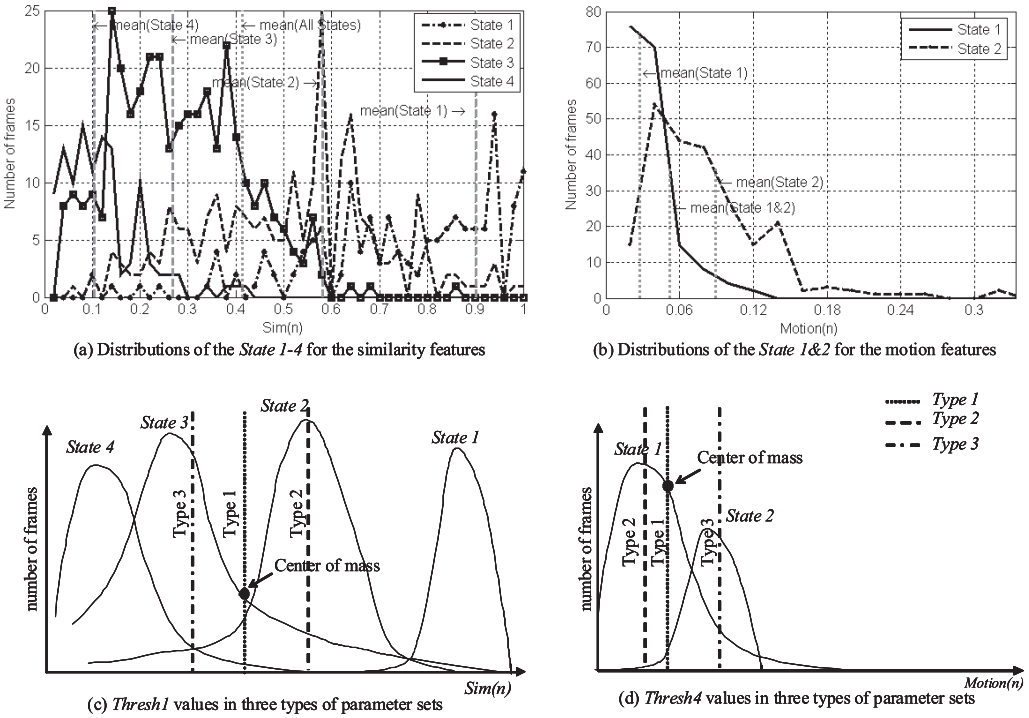
3.3 Selecting the Optimal Threshold Values

A combination of threshold values of the decision tree, named as a *parameter set*. The optimal *parameter set* was decided through an empirical study. The idea of this task is that we establish a series of *parameter sets* to enable an exhaustive search among the predetermined candidates to ascertain a reasonable decision tree. The steps taken in the empirical study were as follows.

First, a training data set that included one thousand frames was selected from small bowel regions. These regions were selected because they are usually the ones focused on by examining doctors. The training data set was build without any bias for the special positions along the small bowel. The image features of the training data were extracted and organized into histograms. For example, each curve in Fig. 5 a shows color similarity distributions for *States 1–4*, and Fig. 5 b shows motion displacement distributions for the *State 1* and *State 2*. Then the prototypes of these distributions are plotted in Fig. 5 c and Fig. 5 d, respectively. Because the center of mass of these distributions discriminate between the two groups of data shown, they suggest estimations of the threshold values. For example, $Thresh1$ is determined by the center of mass of the similarity curves of *State 1* and *State 2*. Similarly, the center of mass of the motion curves in (Fig. 5 d) suggest the $Thresh4$ value, that separates two groups *State 1* and *State 2*. The threshold values decided from training data set is named as a parameter set *Type 1*.

Based on the values for *Type 1*, the threshold values can be moved around the center of mass in the prototype figures. We defined two other parameter sets, *Type 2* and *Type 3*. The values for *Type 3* were determined so that a large number of frames belonged to *States 1* and *State 4* (approximately larger than 10% of data taken from the corresponding states in *Type 1*). Unlike *Type 3*, the *Type 2* values were decided so that the number of frames in *States 1* and *State 4* were small (less than 13% of corresponding states in *Type 1*). Examples of three values of *Thresh1* and *Thresh4* are marked by vertical lines in Fig. 5 c and Fig. 5 d. A series that includes three parameter sets are established in Table 1. We searched for the candidate utilizing the satisfaction evaluations of the examining doctors, as described in the second step below.

Thirty sequences of 90 minutes in length are selected and divided into 10 groups with each group including 3 sequences. Four physicians from the Graduate School of Medicine, Osaka City University, Japan, were asked to view all of the sequences in a certain group. The parameters sets were assigned randomly to evaluation sessions with a constraint that no type was selected twice in a group. For each evaluation, seven levels: *Poor, Quite Poor, Fair, Fairly Good, Good, Very Good, Excellent*, corresponding to scores from 1 to 7, were used to assess the physicians' satisfaction. Table 2 shows the total scores and experience examining CE image sequences of the examining doctors. From these data, *Type 3* was selected as the optimal parameter set because it



(a) Distributions of the *State 1-4* for the similarity features

(b) Distributions of the *State 1&2* for the motion features

(c) *Thresh1* values in three types of parameter sets

(d) *Thresh4* values in three types of parameter sets

Fig. 5 (a) Distributions of color similarity for *States 1–4* and (b) of motion displacement for *State 1* and *State 2* in the empirical study. The corresponding prototypes are plotted in (c) and (d). The center of the mass decides the threshold values. Vertical lines in (c) and (d) are different values of $Thresh1$ and $Thresh4$ in three types of parameters sets.

Table 1 Predetermined threshold values for each parameters set.

Para. set	$Thresh1$	$Thresh2$	$Thresh3$	$Thresh4$	$Thresh5$
Type 1	0.47	0.4	0.6	0.3	1
Type 2	0.7	0.2	0.7	0.2	0.7
Type 3	0.3	0.6	0.5	0.4	0.8

Table 2 Total scores of assessments by the examining physicians to select the optimal parameters set.

Para. set	*MD. A*	*MD. B*	*MD. C*	*MD. D*	*Avg.*
*Experiences**	115	87	121	15	
Type 1	49	51	51	53	51
Type 2	48	53	59	50	52.5
Type 3	53	51	62	51	54.25

*Experiences of the examining doctors by total CE sequences examined up to the evaluation time.

had the highest score (by consensus of the examining doctors). Moreover, in terms of diagnostic experience, *MD.A* and *MD.C*, who were more experienced than other doctors, also gave higher scores to the *Type 3* parameter set.

4. Calculating Delay Time and Controlling Image Display

4.1 Delay Time Functions

Delay time was defined in a general form in (1) ($D_t = \Theta(f(.), \xi_{skill}, \xi_{system})$). The sections below construct the detailed components of this definition.

The function $f(.)$ can be evaluated by adopting a method that queries the similarity/dissimilarity of images in a CBIR system. Given a query, the overall similarity/dissimilarity between the query and an image in a database is obtained from a combination of individual features $S(f_i)$ as below:

$$f(.) = \sum_i w_i S(f_i) \tag{7}$$

where the coefficients w_i are the weight of the features.

The coefficient ξ_{skill} indicates if a physician is accustomed to viewing such sequences, this is called the skill coefficient. This coefficient is treated differently for each state. In *State 1*, images are still so the skill level does not impact on the delay time or it is the same irrespective of the skill level. In *State 2* and *State 3*, the skill coefficients are linear coefficients corresponding to different images that gradually change. In *State 4*, with abrupt changes, the impact of the skill level on the delay time is an additional value. Thus, combinations of skill level and the disparity of the image for each state are defined by:

- For *State 1*, without ξ_{skill}:

$$\Im_t = \sum_i w_i S(f_i)$$

- For *State 2* and *State 3*, ξ_{skill} is a linear coefficient:

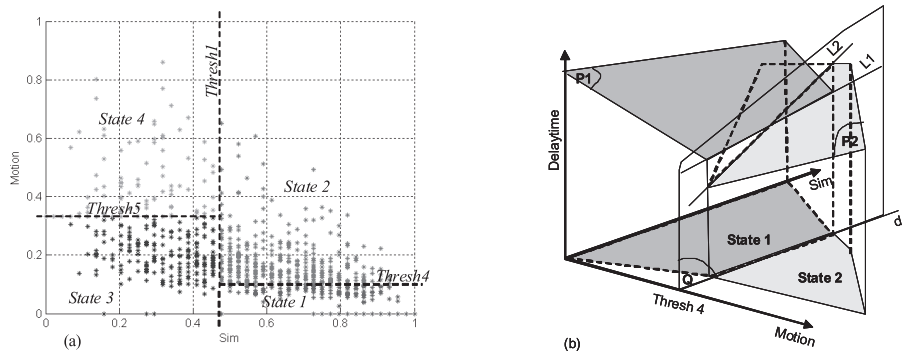


Fig. 6 (a) Results of the classifications of an example sequence. (b) "Jumping" exists between the delay time of *State 1* (P1) and *State 2* (P2) at the intersection of two planes.

$$\Im_t = \xi_{skill} * \sum_i w_i S(f_i) \tag{8}$$

- For *State 4*, ξ_{skill} is an additional value:

$$\Im_t = \sum_i w_i S(f_i) + \xi_{skill}$$

Assuming that the delay time is linearly proportional to $\Im_t$, the function Θ to calculate D_t can be determined by:

$$\Theta = r\Im_t + \xi_{system} \tag{9}$$

where r is a monotone of a non-increasing value for each state.

The coefficient ξ_{system} is also added to (9) to ensure that the delay time function is adaptive to various display system platforms. In another expression, by combining (8) and (9), a delay time D_t between frames $< n, n+1 >$ can be computed by one of the parametric functions below:
- For *State 1*:

$$D_t = A_1(1 - Sim(n)) + A_2 Motion(n) + \xi_{system}$$

- For *State 2* and *State 3*:

$$D_t = [B(1 - Sim(n)) + (1 - B)Motion(n)]\xi_{skill} + \xi_{system} \tag{10}$$

- For *State 4*:

$$D_t = D_1(1 - Sim(n)) + D_2 Motion(n) + \xi_{skill} + \xi_{system}$$

where $Sim(n)$ and $Motion(n)$ are calculated by (3) and (6), respectively. The coefficients $< A_1, A_2, B, D_1, D_2 >$ are multiplied by monotone r and the weights of the selected features.

In term of variability in delay time values, (10) separately defines the functions for each state, while the classification scheme suggests that a principle of continuity exists between states. For example, Fig. 6 a shows the results of the classifications in which $Thresh4$ (motion feature) defines a border between *State 1* and *State 2*. The assumed results of the corresponding delay time are expressed in Fig. 6 b, there are "jumping" points at the intersection of two planes P_1 and P_2, that contain delay time values of *State 1* and *State 2*, respectively. Thus, a constraint that ensures "jumping" between states occurs smoothly must exist. This constraint intuitively creates ties between parameters A_1, A_2, B, D_1 and D_2 in (10). To find the relationships between them, analytic geometry is solved through the intersection of two straight lines L_1 and L_2, in which $L_1 = P_1 \cap Q$ and $L_2 = P_2 \cap Q$, with plane Q contains line d and is parallel with delay time axis. A solution for value of parameter B is expressed by:

$$B = \frac{A_1 Thresh1 + A_1\xi_{system}}{A_2 Thresh1 + A_1 Thresh1 + \xi_{system}} \tag{11}$$

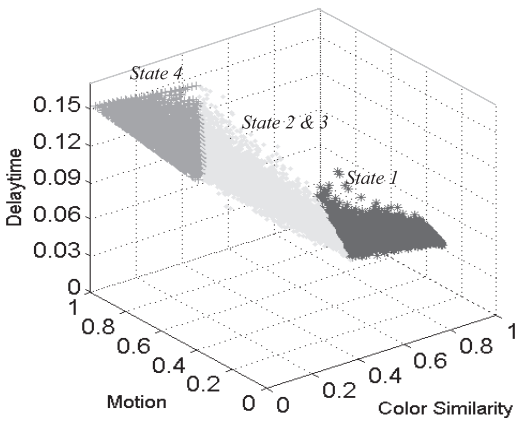


Fig. 7 Distribution of the delay time calculated from the motion displacement and similarity features of a sequence.

Figure 7 shows distributions of delay time D_t and selected features of a full sequence with smooth changing between states. Referring to the area of *State 1*, D_t is around 30 ms/frame. In *State 2* and *State 3*, the distributions are sloped and linearly proportional to the features of motion and color similarity. In *State 4* D_t is around 150 ms/frame. Thus, the delay time values spread in a range from 30 ms/frame to 150 ms/frame, corresponding to the disparity of images varying between stationary and suddenly changing. For comparing image display when the sequence is played at a fixed frame rate (i.e., 13 fps or a delay time with a constant value of 77 ms), the proposed method allows physicians to flexibility review the CE image sequence.

4.2 Post-Processing of Delay-Time Values

Post-processing improves the quality of the image display, but does not greatly impact on the values of the delay time. Two steps are carried out for delay times calculated using (10), and include data smoothing and solving artifact problems.

The delay time values include high frequency components that could cause negative effects on the observations of physicians because of uneven feelings when viewing a sequence. Data are thus smoothed by a Gaussian filter with full width at half maximum equal to 2 ($FWHM = 2$). In other words, the delay time of a frame is smoothed by the two nearest neighbor values to ensure that the transitions of consecutive frames are gradual. The function for smoothing data is as below:

$$D_t = \frac{1}{\delta\sqrt{2\pi}} e^{-(D_t)^2/2\delta^2} \text{ with } FWHM = 2\sqrt{2\ln 2\delta} \quad (12)$$

A solution to avoid the tearing artifact problem is also deployed. Such an artifact occurs when displaying of images losses synchronization with the frame rate of the screen. This is a problem caused by the possibility that the graphic adapter's display buffer updates at the wrong time with respect to the screen refresh rate. By adopting the solutions presented in [49], the tearing artifact problem is resolved by approximations of the delay time values to integer values of the refresh cycle of the screen. Thus, a new frame can only be updated to memory at the beginning of a refresh cycle.

To precisely display images, unlike other multimedia applications, our method emphasizes varying frame rates throughout the entire sequence. Thus, to precisely display corresponding values of delay time, we build a FIFO queue from the image stream and undertake the processing of the queue in a separate thread. A flip command (a function of Microsoft DirectX) is used to burn images from the buffer of the video graphic controller to the screen, with the flip timer being set by the delay time values.

5. Experimental Results

5.1 An Illustration of the Ability to Vary Display Rates

The ability of the proposed method is demonstrated through two cases below. To arrive at an expression that more conveniently describes varying display rates than Fig. 7, we count the total frames displayed in a second throughout entire sequence. Referring to an example sequence in Fig. 8 b, the degree of variability is from a minimum speed of 12 fps to a maximum one of nearly 60 fps.

For frames at position [A] in Fig. 8 b, images are displayed at around 20 fps. Their delay time and some representative frames are shown in detail in Fig. 8 c. Compared with playing the sequence at a constant frame rate (assumed as 13 fps), the images are displayed at twice the constant value (160 ms, compared with 77 ms). With a longer delay time, the frames in Fig. 8 c are clearer for physician interpretation. Contrarily, the display rates at position [B] in Fig. 8 b are increased. Some illustration frames at this position (Fig. 8 a) show obvious similarity. The lower panel in Fig. 8 a shows that the corresponding delay time is smaller than four times if the sequence is played at a fixed speed (around 20 ms, compared with 77 ms). Thus, the effectiveness of the method in this case is its significant reduction in diagnostic time.

The demonstrations above show that an adaptively controlling display rate is a promising way to reduce diagnostic time with less effort for the examining physician. However, this is not sufficient to confirm clinical issues such as that involving abnormal regions captured as well as system operability when reducing diagnostic time. To present more convincing evidence and for validating the subjectivity of reducing diagnostic time, the proposed method underwent clinical evaluations. These were conducted as below to compare the proposed technique against the standard-view mode used in the existing system.

5.2 Conducting Evaluations

To ensure that the conditions for the evaluation of both systems were as similar as possible, a GUI application (called *P system*) was developed for the proposed method so that normal diagnostic functions such as the capture of abnormal regions, the manual adjustment of viewing speeds and changes in viewing display, as well as functions for navigating and verifying suspicious regions were available. The delay time was calculated using the optimal parameters set from the results in Sect. 3.3. RAPID Reader Version 4 (the *G system*) is downloadable at [6]. Both systems were installed on a same PC with a Pentium IV 3.2 GHz, and 2 GB RAM.

We prepared six full sequences of patient data. The evaluations were implemented on both systems by the same four physicians from the Graduate School of Medicine, Osaka City University, who implemented the empirical study to select the optimal parameters set in Sect. 3.3. Thus, forty-eight evaluations were conducted. To facilitate unbiased evaluations, the order of the evaluations of a certain sequence were established so that the number of anterior/first evaluations on each system was equal. The physicians were asked to independently find and capture suspicious regions.

The main activities of the physicians as they used the two systems were recorded. These included: [*play* $\rightarrow$ *stop*], *browsing/scanning* frames to examine suspicious regions, *jumping* frames, *changing manually display speed* and *capturing* abnormal regions. The *P system* was programmed to record logs of the activities of the physicians in a database. To monitor their actions when using the *G system*, we developed a utility that captured the screen when the computer mouse was activated. Interpretation of these logs was implemented by manually reading the captured images. Figure 9 shows an example of the logged activities of

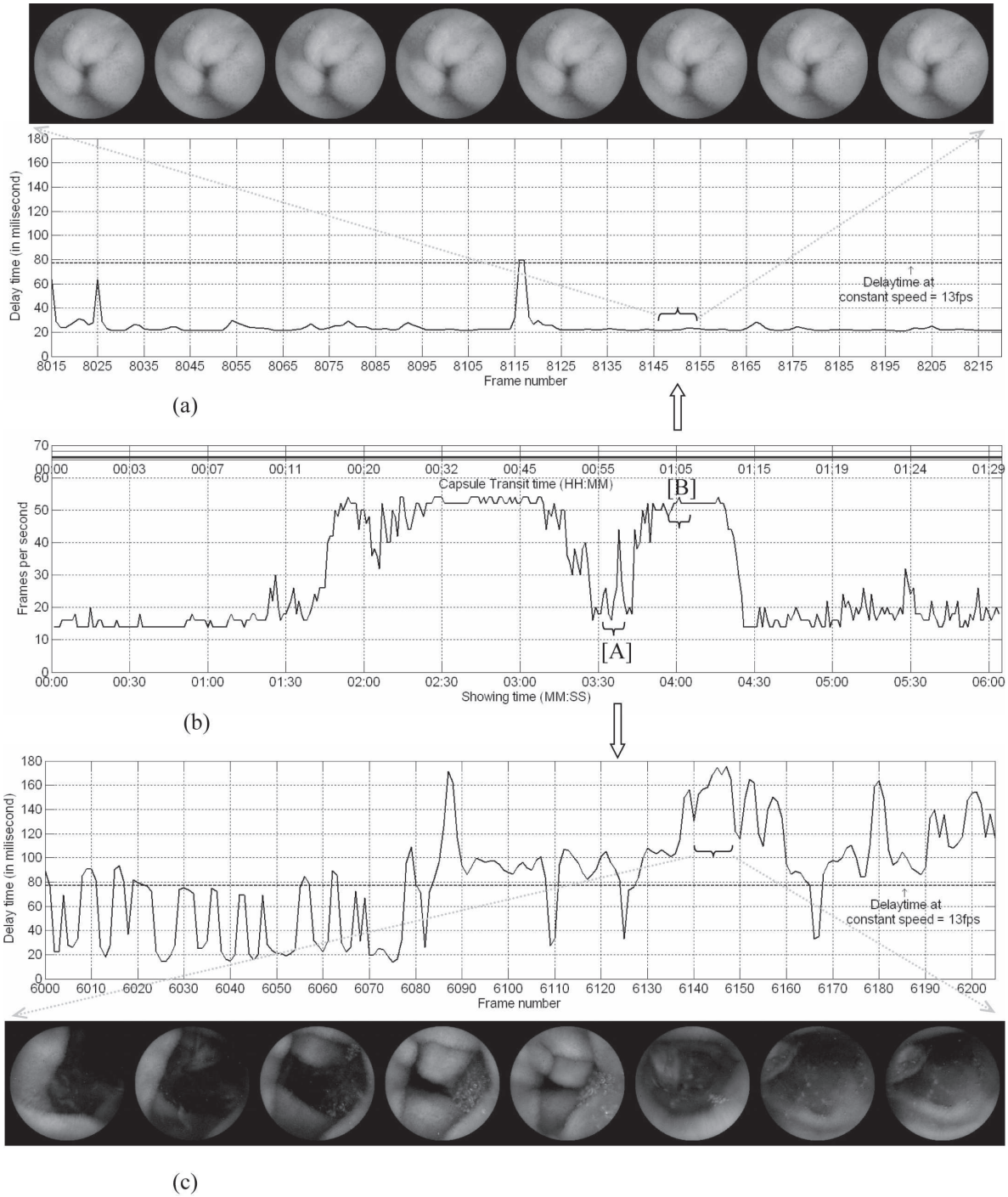


Fig. 8 (b) Varying frame rates of an example sequence. (a) Delay time at positions [B], the sequence is played at high speed (some continuous frames are displayed in the upper row). (c) Delay time at [A], the sequence is played at a slow speed (some continuous frames are suddenly changed, as displayed in the lower row).

physician (*MD. A*) for Seq. #3 under the two systems. From logs expressed in this figure, the logged action based analysis is described below to compare the performances of two systems through three criteria; diagnostic time, abnormal regions captured, and system operability.

5.3 Logged Action Based Analysis

5.3.1 Diagnostic Time

The physicians were asked to fill in evaluation forms when

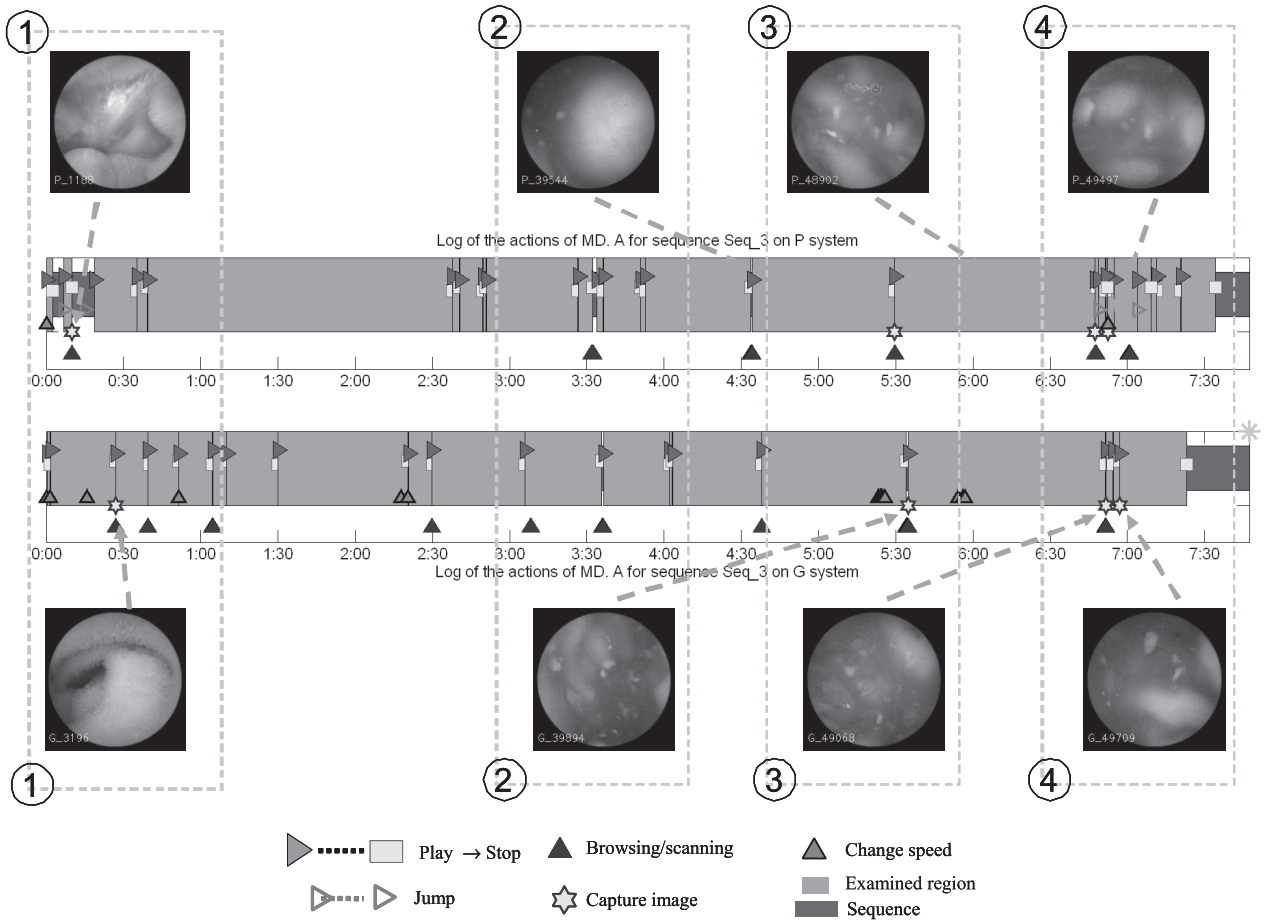


Fig. 9 Logged actions of *MD. A* for a CE image sequence (Seq. #3). The upper panel shows activities under the *P system*, the lower panel shows activities under *G system*. Same abnormal regions captured on both system are indicated by boxes.

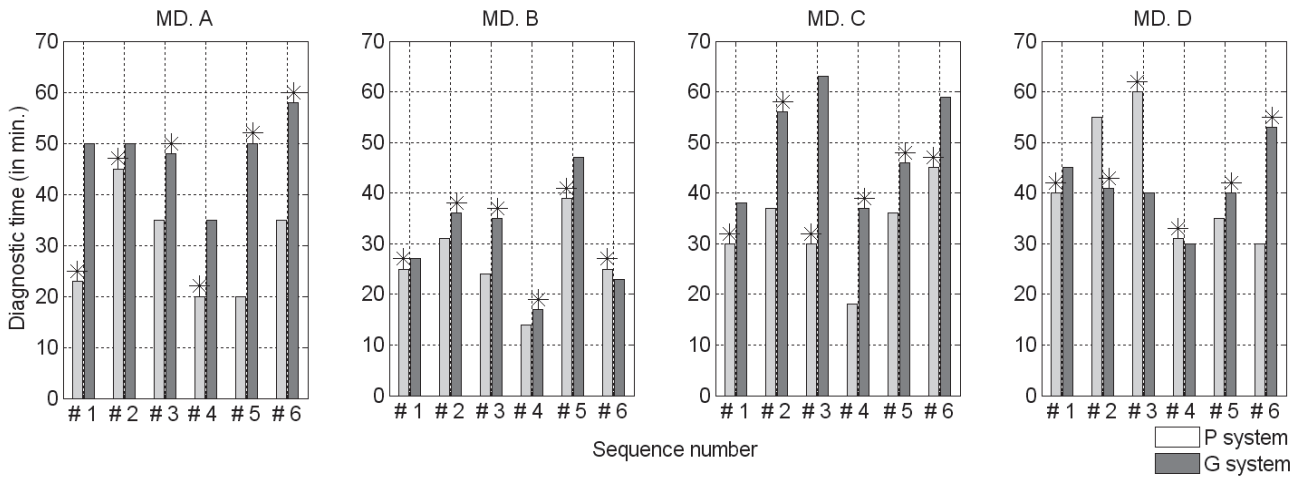


Fig. 10 Diagnostic times of physicians under the two systems. Asterisks mark the first evaluation of the corresponding sequence.

they started and finished a sequence evaluation. Diagnostic times were calculated from this data. The durations of activities such as continuously [*play* → *stop*], *browsing*/*scanning* frames, and *jumping* frames were summed by investigating the captured logs under both systems. These data were used to confirm the diagnostic times noted by the physicians.

Figure 10 compares the diagnostic times of the physi-

Table 3 The *MatchingRate* of evaluations on both systems (numerator is υ, denominator is χ of (13)).

Seq. No	*MD. A*		*MD. B*		*MD. C*		*MD. D*	
	P system	*G system*	*P system*	*G system*	*P system*	*G system*	*P system*	*G system*
# 1	2/3	3/3	3/3	2/3	2/2	2/2	2/4	3/4
# 2	3/3	3/3	4/5	5/5	4/5	5/5	5/6	5/6
# 3	4/4	4/4	3/4	3/4	5/6	6/6	7/7	7/7
# 4	2/3	2/3	1/2	2/2	3/3	3/3	5/5	4/5
# 5	5/5	4/5	3/4	3/4	5/5	4/5	5/6	6/6
# 6	5/5	5/5	6/6	6/6	8/8	8/8	2/2	2/2
Σ Reg. lost	1/23	2/23	4/24	3/24	2/29	1/29	4/30	3/30
Avg.	96%	91%	88%	92%	93%	96%	86%	90%
Average of *P system* = 91% and *G system* = 92%								

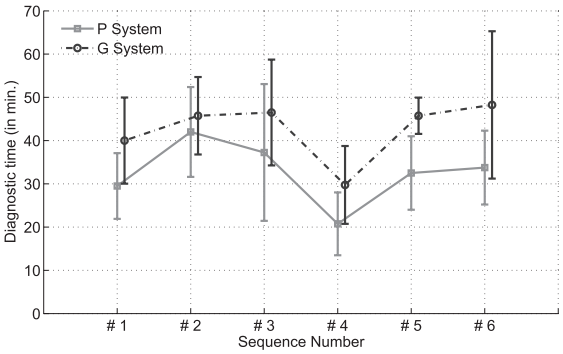


Fig. 11 Average diagnostic time by sequences.

cians for the sequences examined using the two systems. The first evaluation on the corresponding system for a certain sequence is also marked by asterisks in these figures. The diagnostic times using the proposed system were significantly reduced for most evaluations (approximately 16 min. for *MD. A*, 6 min. for *MD. B*, and 14 min. for *MD. C*). The diagnostic time of *MD. D* was equal in both systems.

Average diagnostic time by sequence is shown in Fig. 11. From this figure, the diagnostic time on the *P system* was seen as reduced for all six sequences. The average diagnostic time for the *P system* was 32.5 ± 7 minutes and it was 42.4 ± 9 minutes for the *G system*. Applying a T-test to measuring the significance of any difference of the average values, we found that the diagnostic time using the *P system* showed a significant difference from evaluations implemented using *G system* ($t = 3.1$, $df = 47$, $p < 0.05$).

5.3.2 Ability to Capture Abnormal Regions

The number of abnormalities present in evaluations differed according to the physician because it depended on factors such as personal judgment, skill level and concentration during the evaluation. Therefore, we took into account the abnormal regions captured by the same physician using the two systems. First, the abnormal regions χ of a sequence were considered by merging abnormal regions captured with both systems. For example, as shown in Fig. 9, abnormal regions captured by *MD. A* on both systems are matched. The matching rate was the ratio between abnormal regions υ captured in a particular system and the χ abnormal regions, as below:

$$MatchingRate = \frac{\upsilon}{\chi} 100(\%) \tag{13}$$

Table 3 shows the ratio of the evaluations by the physicians using both systems. The average value was 91% for the *P system*, approximating the matching rate on the *G system* (92%). The results implied there are no limitations in capturing abnormal regions when the display rates were controlled under the proposed technique.

Besides the above analysis for full sequences, we implemented evaluations to verify whether abnormal regions are lost in the stationary state because of the high speed display. The total diagnostic time when examining doctors examined frames in this state was calculated. As well, we compared the accuracy of abnormal regions captured in the stationary state. The results showed a reducing diagnostic time with no loss of abnormal regions when examining doctors implemented evaluations on the *P system*.

5.3.3 Operability of the Physicians

To evaluate operability in terms of a quantitative analysis, Fig. 9 can be used to illustrate the different activities implemented under the two systems. For qualitative indices, we used two criteria that can be impacted by the proposed technique:

- Comparing the number of *changing speed* actions on both systems. As shown in Fig. 12 (a), the number of evaluations with no *changing speed* actions was higher for the proposed system. The behavior of the physicians for this action clearly differs between the two systems.

- Another criterion is how the examining doctors perceived abnormal regions. Such an assessment might be achieved by counting events [*play* → *stop*] in the evaluations. Such events generally imply an action to verify or look for a suspicious region. As shown in Fig. 12 (b), these actions in the *P system* are less than in the *G system* for three of four examining doctors. In terms of the accuracy of capturing abnormal regions, Table 3 showed no significant different between the two systems. Therefore, automatic adjustments of the display rates using the proposed technique achieve substantial operability in the diagnostic procedures.

On the other hand, there is a skill level function sup-

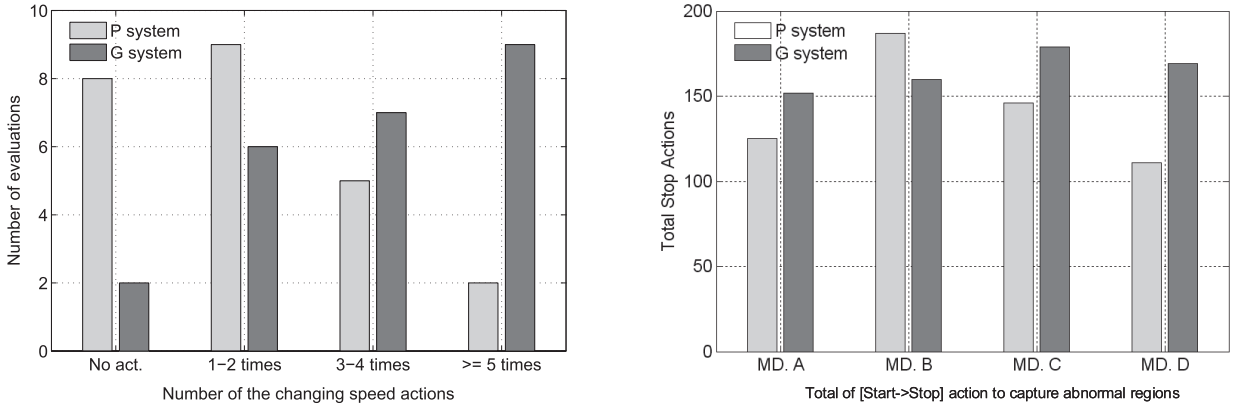


Fig. 12 Comparing the number of changing speed actions (a) and of [*play* → *stop*] actions to capture abnormal regions (b) in the evaluations between *P system* and *G system*.

ported in the *P system* that allows examining doctors to adjust the display speed to that suited to their expertise. From the GUI interface, examining doctors can manually select a level among seven skill levels. The results of analysis of the logged actions show that some levels were preferred. *MD. B* always used Level 7 in his examinations, *MD. A* and *MD.C* preferred level 5, 6, 7, while *MD. D* selected Level 5. Obviously, supporting different skill levels that takes into account the expertise of the examining doctor makes the system more flexible.

6. Discussions and Conclusion

6.1 Discussions

For feature extractions, generic color histogram indexing was selected to measure the similarity of consecutive frames. However, CE images usually present homogeneous regions of the GI tract wall, thus constructing a digestive color model is reasonable for the color histogram indexing method. Same as the observations of Mackeiwicz et al. in [21], the dominant color of the CE images is a pinkish color in the stomach and pinkish to yellowish in the small intestine. Intuitively, the range of color presented in CE images is relatively small (e.g. around 20% of the possible color space, in [21]). The result for measuring the similarity of consecutive frames, when indexing equalizations are focused within the dominant colors of this model, is thus more precise. For motion extractions in the current work, the accuracy of the measurements of the displacement of consecutive frames depended on the selection of predetermined parameter values. Research focused on motion estimations or the adoption of results of intestinal motility for the CE sequence could overcome current limitations. Furthermore, although the two selected features are a successful way to reflect peristaltic activity in the GI tract, more features can be evaluated and these could improve the results. For example, Combra et al. [24] analyzed a set of useful features for the discrimination of images extracted from CE image sequences. More computer vision research for CE images dedicated to intestinal motility should suggest the most relevant features.

Four states are classified in the scheme using a decision tree classifier that is learned from a one thousand training data set. To evaluate the size of the training data set used, a series of testing data of various sizes was established. The cross-correlation values of the feature distributions of the states between the testing data and the training data were examined. The results implied that the size of the training data set was sufficient for the state classification task. Indeed, with this scheme, the optimizing tree requires a lot of effort and exhaustive searching in the space of all possible structures, e.g., as in our solution in Sect. 3.3 for selection of the optimal parameters set. Moreover, an issue related to the tree classifier is that to obtain a higher classification performance, a very high performances is needed at each node of the tree. Therefore, non-hierarchical approaches such as HMM or a Support Vector Machine could be utilized in future research to overcome these limitations.

6.2 Conclusion

This paper presented a novel method to reduce the diagnostic time required to review and interpret CE videos through efficiently controlling the image display. The robustness of the method relies on the original images being displayed with no frame skipping. Major issues resolved included: 1) although images were captured at a low frame rate and in uncontrolled conditions, the differences between two consecutive frames were efficiently spread among the various conditions of image acquisition by combining the features of color and motion; 2) whereas recognizing GI motility patterns from CE videos still has limitations, an algorithm for classifying the states overcomes this problem; and 3) the functions to compute delay time are adaptable with the classified states and support the variable skill levels of physicians. The post-processing procedures enhanced and pre-

cisely controlled the display of images.

Clinical evaluations were conducted in experiments to investigate the effectiveness of the proposed system compared to the standard view using the Rapid Reader system. From these results, we concluded that the diagnostic time using our proposed system was 32.5 ± 7 minutes for each evaluation. This time was 10 minutes less than that for the same evaluations implemented using the Rapid Reader application. Moreover, the proposed method required less effort for the examining physicians while the number of abnormalities found with both system was similar. These results should convince physicians that the proposed technique can be safely used for routine clinical diagnoses.

Some limitations of the proposed method were discussed and areas suggested for future research. Effective indexing could be resolved by constructing a GI color space, whereas non-hierarchical approaches are suggested for research into recognizing patterns of GI motility. As well, using the action logs of physicians can be applied to clinical applications and for educational purposes. The expertise of physicians can be automatically evaluated and the system suitably adjusted for their skill level. These adjustments would allow for the effective and quick navigation of interesting parts of a sequence. The target of examinations can thus be more focused on suspicious regions rather than normal ones. This would have a major impact on diagnostic procedures.

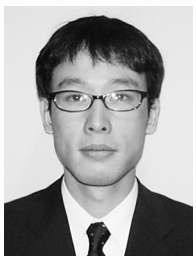
References

[1] D.G. Adler and C.J. Gostout, "Wireless capsule endoscopy - state of art," Hospital Physician, pp.14–22, May 2003.
[2] P. Swain, "Wireless capsule endoscopy," GUT, vol.52, pp.48–50, 2003.
[3] P. Swain and A. Fritscher-Ravens, "Role of video endoscopy in managing small bowel disease," GUT, vol.53, pp.1866–1875, 2004.
[4] G. Iddan, G. Meron, A. Glukovsky, and P. Swain, "Wireless capsule endoscope," Nature, vol.405, p.417, May 2000.
[5] Given Imaging's Homepage, "http://www.givenimaging.com/en-us/Patients/Pages/pagePatient.aspx," April 2008.
[6] Given Imaging, "http://www.givenimaging.com/en-us/HealthCareProfessionals/Products/Pages/Software.aspx," Oct. 2007.
[7] American Society for Gastrointestinal Endoscopy - ASGE, "Technology status evaluation report wireless capsule endoscopy," Gastrointestinal Endoscopy, vol.56, no.5, pp.1866–1875, Aug. 2002.
[8] Medical Advisory Secretariat, "Wireless capsule endoscopy, health technology literature review," Ontario Ministry of Health and Long-term Care, Canada, May 2003.
[9] M. Hadathi, G. Heine, A. Jacobs, A.V. Bodegrawn, and L. Mdlder, "A prospective study comparing video capsule endoscope followed by double balloon enteroscopy for suspected small bowel disease," Program & Abstract of the International Conference Capsule Endoscope, p.203, 2005.
[10] M. Keuchel, S. Al-Harthi, and F. Hagenmuller, "New automatic mode of rapid 4 software reduces reading time for small bowel pillcam studies," Program & Abstract of the International Conference Capsule Endoscope, p.93, 2006.
[11] Y. Li, T. Zhang, and D. Tretter, "An overview of video abstraction techniques," HP Laboratory Technical Report HPL-2001-191, July 2001.
[12] B. Shahraray, "Scene change detection and content-based sampling of video sequences," Proc. IS&T/SPIE, pp.2–13, 1995.
[13] D. Swanberg, C. Shu, and R. Jain, "Knowledge guided parsing in video database," Proc. IS&T/SPIE, pp.13–24, 1993.
[14] S. Smoliar and H. Zhang, "Content-based video indexing and retrieval," IEEE Multimedia, vol.1, no.2, pp.62–72, 1994.
[15] C. Toklu and S.P. Liou, "Automatic keyframe selection for content-based video indexing and access," Proc. IS&T/SPIE, pp.554–563, 2000.
[16] S. Pfeiffer, R. Lienhart, S. Fisher, and W. Effelsberg, "Abstracting digital movies automatically," J. Visual Communication and Image Representation, vol.7, no.4, pp.345–353, Dec. 1996.
[17] R. Lienhart, "Dynamic video summarization of home video," Proc. IS&T/SPIE, pp.378–389, 2000.
[18] N. Petrovic, N. Jojic, and T.S. Huang, "Adaptive video fast forward," Multimedia Tool and Applications, vol.26, no.3, pp.327–344, Aug. 2005.
[19] K.A. Peker, A. Divakaran, and H. Sun, "Constant pace skimming and temporal sub-sampling of video using motion activity," Proc. IEEE Conf. on ICIP, pp.414–417, 2001.
[20] K.A. Peker and A. Divakaran, "An extended framework for adaptive playback-based video summarization," Mitsubishi Electric Research Laboratory Technical Report TR-2003-115, Sept. 2003.
[21] M. Mackiewicz, J. Berens, M. Fisher, and G. Bell, "Colour and texture based gastrointestinal tissue discrimination," Proc. IEEE Int. Conf. Acoust. Speech Signal Process. ICASSP, pp.597–600, May 2006.
[22] M. Mackiewicz, J. Berens, and M. Fisher, "Wireless capsule endoscopy video segmentation using support vector classifiers and hidden markov models," Proc. International Conference on Medical Image Understanding and Analyses, June 2006.
[23] M. Coimbra, P. Campos, and J.P.S. Cunha, "Topographic segmentation and trasit time estimation for endoscopic capsule exam," Proc. IEEE Int. Conf. Acoust. Speech Signal Process., pp.1164–7, May 2006.
[24] M. Coimbra and J.P.S. Cunha, "Mpeg-7 visual descriptors - contributions for automated feature extraction in capsule endoscopy," IEEE Trans. Circuits Syst. Video Technol., vol.16, no.5, pp.628–637, May 2006.
[25] A. Glukhovsky, G. Meron, D. Adler, and O. Zinati, "System for controlling in vivo camera capture and display rate," Patent number PCT WO 01/87377 A2.
[26] F. Vilarino, L. Kuncheva, and P. Radeva, "Roc curves and video analysis optimization in intestinal capsule endoscopy," Pattern Recognit., vol.27, no.8, pp.875–881, June 2006.
[27] F. Vilarino, P. Spyridonos, J. Vitria, F. Azpiroz, and P. Radeva, "Linear radial patterns characterization for automatic detection of tonic intestinal contractions," Proc. CIARP, pp.178–187, 2006.
[28] P. Spyridonos, F. Vilarino, J. Vitria, F. Azpiroz, and P. Radeva, "Identification of intestinal motility events of capsule endoscopy video analysis," Proc. Advanced Concepts for Intelligent Vision Systems, pp.531–537, 2005.
[29] P. Spyridonos, F. Vilarino, J. Vitria, F. Azpiroz, and P. Radeva, "Anisotropic feature extraction from endoluminal images for detection of intestinal contractions," Proc. Medical Image Computing and Computer-Assisted Intervention, pp.161–168, 2006.
[30] P.M. Szczypinski, P.V.J. Sriram, R.D. Sriram, and D.N. Reddy, "Model of deformable rings for aiding the wireless capsule endoscopy video interpretation and reporting," Proc. International Conference on Computer Vision and Graphics 2004, pp.167–172, 2006.
[31] P.M. Szczypinski, "Selecting a motion estimation method for a model of deformable rings," Proc. International Conference on Signals and Electronic Systems, pp.297–300, Sept. 2006.
[32] V. Hai, T. Echigo, R. Sagawa, M. Shiba, K. Higuchi, T. Arakawa, and Y. Yagi, "Adaptive control of video display for diagnostic assistance by analysis of capsule endoscopic images," Proc. 18th ICPR, pp.980–983, Aug. 2006.
[33] M. Swain and D. Ballard, "Color indexing," Int. J. Comput. Vis.,

vol.7, no.1, pp.11–32, June 1991.
[34] G. Pass, R. Zabih, and J. Miller, "Comparing images using color coherence vectors," Proc. 4th ACM Conf. on Multimedia, Nov. 1996.
[35] J. Huang, S. Kumar, M. Mitra, W. Zhu, and R. Zabih, "Image indexing using color correlograms," Proc. IEEE Computer Society Conference on Vision and Pattern Recognition, pp.762–768, 1997.
[36] W.Y. Ma and H. Zhang, "Benchmarking of image features for content-based retrieval," Proc. 32nd Asilomar Conference on Signal, System and Computers, 1998.
[37] M. Stricker and M. Swain, "The capacity of color histogram indexing," Proc. Computer Vision and Pattern Recognition 1994, pp.704–708, 1994.
[38] J. Barron, D. Fleet, and S. Beauchemin, "Performance of optical flow techniques," Int. J. Comput. Vis., vol.12, no.1, pp.43–77, Feb. 1994.
[39] C.H. Wu, Y.C. Chen, C.Y. Liu, C.C. Chang, and Y.N. Sun, "Automatic extraction and visualization of human inner structures from endoscopic image sequences," Proc. IS&T/SPIE, pp.464–473, 2004.
[40] P. Suchit, R. Sagawa, T. Echigo, and Y. Yagi, "Deformable registration for generating dissection image of an intestine from annular image sequence," Proc. Computer Vision for Biomedical Image Applications, pp.271–280, 2005.
[41] C. Tomasi and T. Kanade, "Detection and tracking of point features," tech. rep., 1991.
[42] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," Proc. Intl. Joint Conf. on Artificial Intelligence, pp.674–679, 1981.
[43] S. Birchfield's Homepage, KLT: Kanade-Lucas-Tomasi Feature Tracker, http://www.ces.clemson.edu/~stb/klt/," April 2006.
[44] P. Anandan, "A computation framework and an algorithm for the measurement of visual motion," Int. J. Comput. Vis., vol.2, no.3, pp.283–310, Jan. 1989.
[45] J. Shi and C. Tomasi, "Good features to track," Proc. IEEE Computer Society Conference on Vision and Pattern Recognition, pp.593–600, 1994.
[46] A. Feinstein, Principles of Medical Statistics, CRC Press Company, Boca Raton, Florida, 2002.
[47] D. Grundy, GastroIntestinal Motility - The Integration of Physiological Mechanisms, MTP Press Limited, Hingham, MA, 1985.
[48] J.M. de Sá, Pattern Recognition - Concepts, Method and Applications, Springer, 2001.
[49] D.H. Nowlin, Intel Corporation, Video Frame Display Synchronization, "http://www.intel.com/cd/ids/developer/asmo-na/eng/dc/digitalmedia/optimization/239118.htm," April 2006.

Hai Vu received the B.Sc. degree in Electronic and Telecommunication in 1999, and M.Sc. degree in Information Processing and Communication in 2002, both from Hanoi University of Technology, Vietnam. He is currently a Ph.D. student in the Graduate School of Information Science and Technology, Osaka University, Japan. His research interests are in computer vision, medical imaging.

Tomio Echigo received the B.Sc. and M.Sc. degrees in electrical engineering from the University of Osaka Prefecture, Osaka, Japan in 1990 and 1982, respectively, and the Ph.D. degree in engineering science from Osaka University, Osaka, Japan in 2003. He joined IBM Japan, Ltd., in 1982, where he was an Advisory Researcher at the Tokyo Research Laboratory, IBM Research, Kanagawa, Japan. From 2003 to 2006, he has been a Visiting Professor at Osaka University, Osaka, Japan. Since 2006, he has been with Osaka Electro-Communication University, Osaka, Japan, where he is currently Professor of the Department of Engineering Informatics. His research interests include image and video processing, medical imaging, and video summarization.



Ryusuke Sagawa is an Assistant Professor at the Institute of Scientific and Industrial Research, Osaka University, Osaka, Japan. He received a B.E. in Information Science from Kyoto University, Kyoto, Japan, in 1998. He received a M.E. in Information Engineering in 2000 and Ph.D. in Information and Communication Engineering from the University of Tokyo, Tokyo, Japan in 2003. His primary research interests are computer vision, computer graphics and robotics (mainly geometrical modeling and visualization).

Keiko Yagi is a lecturer at the Department of Clinical Pharmacy, Kobe Pharmaceutical University, Japan. She received a Ph.D. from Osaka City University. Her research interests are in the field of Clinical Pharmacology.

Masatsugu Shiba received a Doctor degree from Osaka City University Graduate Medical School, Japan, in 2002. Since 1995, he has been with Osaka City University Medical School Hospital, where he is Assistant Professor of Gastroenterology in 2002. His research interests focus on Gastrointestinal Endoscopy and Medical Informatics.

Kazuhide Higuchi is a Professor at the Second Department of Internal Medicine, Osaka Medical College, Japan. He received M.D. degree 1982 and the Ph.D. degree in 1992, from Osaka City University. His interests are in the fields of gastroenterology, gastrointestinal endoscopy, capsule endoscopy, therapeutic endoscopy.

Tetsuo Arakawa is a Professor at the Department of Gastroenterology, Osaka City University Graduate School of Medicine, Japan. He received M.D. degree 1975 and the D.M.Sc. degree in 1981, both from Osaka City University Medical School. He is Director of Japanese Society of Senile Gastroenterology from 2001 and Japanese Gastroenterological Association from 2004.

Yasushi Yagi is a Professor at the Institute of Scientific and Industrial Research, Osaka University, Japan. He received B.E. and M.E. degrees in control engineering, 1983 and 1985, respectively and the Ph.D. degree in 1991, from Osaka University. His interests are in the fields of computer vision, image processing and medial imaging.