J Digit Imaging (2013) 26:287–301
DOI 10.1007/s10278-012-9519-x

Wireless Capsule Endoscopy Video Reduction Based on Camera Motion Estimation

Hong Liu · Ning Pan · Heng Lu · Enmin Song · Qian Wang · Chih-Cheng Hung

Published online: 7 August 2012

Abstract Wireless capsule endoscopy (WCE) is a novel technology aiming for investigating the diseases and abnormalities in small intestine. The major drawback of WCE examination is that it takes a long time to examine the whole WCE video. In this paper, we present a new reduction scheme for WCE video to reduce the examination time. To achieve this task, a WCE video motion model is proposed. Under this motion model, the WCE imaging motion is estimated in two stages (the coarse level and the fine level). In the coarse level, the WCE camera motion is estimated with a combination of Bee Algorithm and Mutual Information. In the fine level, the local gastrointestinal tract motion is estimated with SIFT flow. Based on the result of WCE imaging motion estimation, the reduction scheme preserves key images in WCE video with scene changes. From experimental results, we notice that the proposed motion model is suitable for the motion estimation in successive WCE images. Through the comparison with APRS and FCM-NMF scheme, our scheme can produce an acceptable reduction sequence for browsing and examination.

Keywords Wireless capsule endoscopy · Bee algorithm · SIFT flow · Motion estimation

H. Liu (✉) · N. Pan · E. Song
Center for Biomedical Imaging and Bioinformatics,
School of Computer Science and Technology,
Huazhong University of Science and Technology,
Wuhan, Hubei 430074, China
e-mail: hl.cbib@gmail.com

N. Pan
e-mail: pannsp@gmail.com

H. Liu · N. Pan · E. Song
Key Laboratory of Education Ministry for Image Processing
and Intelligence Control,
Wuhan, Hubei 430074, China

H. Lu
Department of Gastroenterology, Nanjing Central Hospital
of Nanjing Military Command of Chinese PLA,
Nanjing, Jiangsu 210002, China

Q. Wang
School of Information and Safety Engineering,
Zhongnan University of Economics and Law,
Wuhan, Hubei 430073, China

C.-C. Hung
School of Computing and Software Engineering,
Southern Polytechnic State University,
Marietta, GA 30060, USA
e-mail: chung@spsu.edu

Introduction

Wireless capsule endoscopy (WCE) is a microelectromechanical system (MEMS), which consists of a miniature camera, white light-emitting diodes (LEDs), a battery, and a radio frequency emitter. The WCE, which is swallowed by the patient and propelled by gastrointestinal peristalsis [1], is mainly used for investigating the whole small intestine with a noninvasive manner. In gastrointestinal tract, WCE captures images of inner tract and wirelessly transmits these images to a receiver worn by the patient. WCE images can then be downloaded to a workstation for visualization, which helps clinicians to diagnose whether there are diseases and abnormalities or not in the small intestine.

In general, WCE captures images at a typical rate of two frames per second while moving forward along the digestive tract. For a complete examination, WCE takes about 6–8 h to traverse through the entire digestive tract and captures approximately 50,000 images of gastrointestinal wall [2, 3]. Currently, the major weakness of WCE examination is that a WCE clinician usually takes about 2–3 h to examine the whole WCE video. Even for an experienced physician, the work still needs about 1–2 h. Therefore, it is essential to reduce the examination time, and this study was motivated by attempts to devise a method for solving the problem.

In fact, WCE video has at least 10,000 frame images that carry useless clinical information. Moreover, clinical observations have revealed that only 20,000–30,000 frames about small intestine in a WCE video are concerned for a clinician. Although WCE captures two images per-second, a lot of images in WCE video have similar scenes, as shown in Fig. 1. By analyzing the redundant information, the number of images in a WCE video can be significantly reduced.

Although some methods have been proposed for WCE video processing, these methods mainly focus on computer aided diagnosis, such as the detection of obscure bleeding, polyp, ulcers, and tumor [4–8]. All of these are still in the experimental stage and have not been used in the clinic yet. Recently, some researchers begin to pay attention to the problem of summarizing the WCE video. The goal is to develop a useful approach that can reduce the number of images in a WCE video.

An image mining approach that applies a computationally intensive clustering scheme (FCM-NMF) to summarizing the WCE video was proposed to reduce the time for the visual inspection [9, 10]. The method is time consuming and quite complicated. This clustering scheme can lead some images, which are not time-related grouped into the same cluster and produce a wrong representative frame. In [11], Li et al. proposed a WCE video reduction scheme, which reduces a WCE video sequence using motion features. In that paper, they use two typical motion analysis methods: adaptive rood pattern search block matching (ARPS) [11] and Bayesian multiscale differential optical flow (BMSD). In fact, ARPS assumes that the camera movements are in horizontal and vertical directions and needs the predicted motion vector (MV) for initializing the size of ARPS. If there is not a predicted MV, the ARPS cannot estimate motion accurately. Although BMSD takes a coarse-to-fine estimation scheme to estimate optical flow vector, it is difficult to estimate a larger displacement of WCE camera motion. Therefore, the motion models of ARPS and BMSD are not suitable for WCE imaging motion estimation. Olympus Ltd has announced a new feature of the improved Endo Capsule Software that can condense the entire examination into a maximum of 2,000 still images [12]. However, the method is similar to the ARPS, which compares images with explicit changes to pervious image. In other words, the method just detects the sudden changes in WCE image sequence. Since some small scale of diseases cannot make a significant scene changes, this method may ignore important information in WCE image sequence and lead to a wrong-decision result for medical diagnosis. A registration methodology was presented to reduce the amount of frames in [13, 14]. This method utilizes a segmentation scheme and a graph method to register similar regions in two adjacent WCE video frames. However, it difficult to produce an effective segmentation to register similar regions because the similar scene in intestine region may produce various WCE images due to illumination and local nonrigid deformation.

In this paper, we present a new reduction scheme consisting of three stages for WCE video. In the first stage, we estimate motion features in WCE video. In the second stage, we use motion features to measure the scene changes between two WCE images. In the third stage, we preserve those images, which have obvious scene changes. In order to estimate the WCE motion between two successive WCE images more accurately, we also propose a novel motion model for WCE video. In this model, the motion estimation of WCE video is divided into two levels, the coarse level and the fine level. In the coarse level, the WCE camera motion as the global rigid motion is estimated and viewed as the large displacement motion estimation. In the fine level, the gastrointestinal tract deformation as the local nonrigid motion is estimated based on the motion estimation of WCE camera, and the nonrigid motion is viewed as local motion estimation.

The rest of this paper is organized as follows. In "The Proposed Method," we will present the reduction scheme of

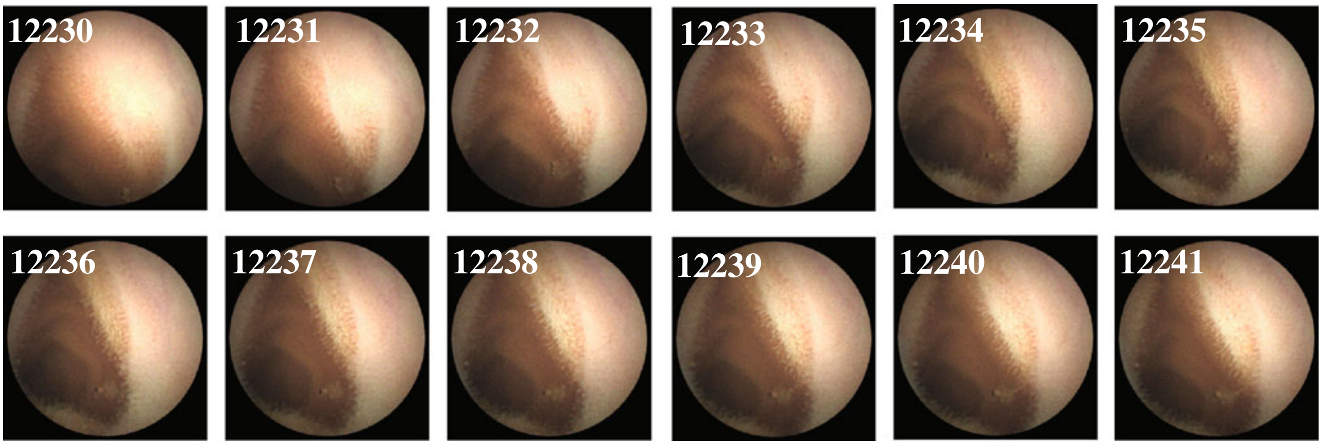


Fig. 1 Many images in the WCE video have similar scenes

WCE video in details. In "Results and Discussions," experimental results are comprehensively illustrated and evaluated. The last section summarizes the conclusions of our study.

The Proposed Method

Overview of the Method

To achieve the reduction task, we assume that two successive frames have overlapping area (movement of WCE camera is continuous) and intrinsic parameters of WCE camera are constant in the entire examination procedure. We divide our scheme into three steps. First, the motion between successive WCE images is estimated based on a novel WCE video motion model. Then, the scene changes are measured with the motion estimation. Finally, images with obvious scene changes are preserved. The framework of our reduction scheme is presented in Fig. 2.

The Motion Estimation of WCE Video

In our scheme, the motion estimation is used to capture scene changes between successive WCE images. However, this is a difficult task because the WCE takes two images per second unlike a general video where it has more consecutive motion features. Therefore, we need a suitable motion model to describe the motion between successive WCE images. In this paper, we propose a novel WCE video motion model. In this model, the motion of WCE video includes two parts: One part is rigid motion, and the other is nonrigid motion.

The Motion Model of WCE Video

As shown in Fig. 3, when WCE works in the gastrointestinal tract, WCE is propelled by gastrointestinal tract peristalsis. If we just consider the WCE camera movement, then two successive WCE images in which the relationship can be described with homogeneous coordinates:

$$Y_{i-1} = M \times Y_i^T = \left[\begin{array}{c|c} s\cdot R & T \\ \hline p & 1 \end{array}\right] \times Y_i^T \tag{1}$$

where subscript i is an image index in the WCE video. Y_i and Y_{i-1} are 2D points in the source image i, and a neighborhood image $i-1$ of the source image. Operator × is a matrix multiplication. M is a 2D rigid transformation, which includes translation T, rotations R, scaling s, and perspective p parameters. These parameters describe the motion of WCE camera (global rigid motion) between two successive WCE images.

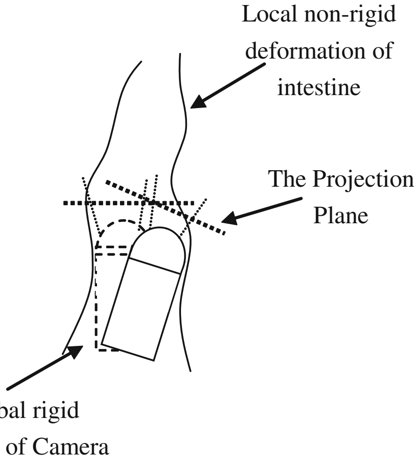

Local non-rigid deformation of intestine
The Projection Plane
Global rigid motion of Camera

Fig. 3 The graph demonstrates the motion of WCE in gastrointestinal tract

In fact, the motion between successive WCE images depends not only on the movement of WCE camera but also on nonrigid deformation (local nonrigid motion) of gastrointestinal tract due to its peristalsis. Therefore, both the movement of WCE camera and gastrointestinal tract need be considered simultaneously when we estimate the motion between successive WCE images. Then, Eq. (1) can be modified as follows:

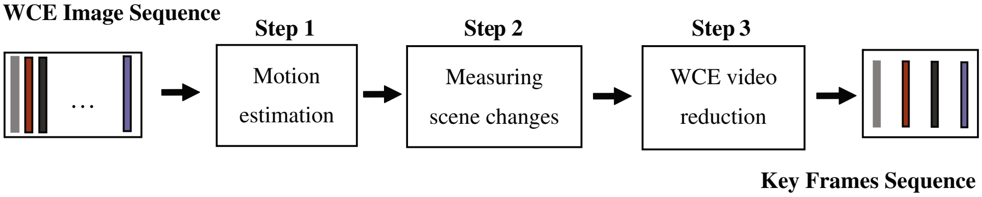
$$Y_{i-1} = M \times Y_i^T + \varepsilon_i \tag{2}$$

where ε_i is a local gastrointestinal tract displacement vector caused by local gastrointestinal tract nonrigid movement. We can integrate the rigid and nonrigid transformation as a universal form.

$$Y_{i-1} = \left[\begin{array}{c|c} s\cdot R & [T+\varepsilon_i] \\ \hline p & 1 \end{array}\right] \times Y_i^T \tag{3}$$

According to this motion model, WCE image motion can be estimated in two stages. In the first stage of the coarse level estimation, the motion of WCE camera, which can be regarded as a large displacement of WCE image scene, is

Fig. 2 The framework of our WCE video reduction scheme, which is divided into three steps



estimated. In the second stage of the fine level estimation, the local gastrointestinal tract deformation based on the result of the first stage estimation is estimated. Actually, the first stage also can be thought of an approximate alignment between two successive WCE images, and the second stage is an image alignment of WCE images in the local detail. Once we have chosen a suitable motion model to describe the alignment between a pair of WCE images, we need to devise a method to estimate its motion parameters.

The Motion Estimation of WCE Camera

WCE is propelled by gastrointestinal peristalsis. In fact, the gastrointestinal wall is very close to WCE camera, which causes less projective deformation of WCE images and WCE camera motion can be described as 2D rigid deformation. Therefore, we can just focus on rigid motion and ignored projective p and local gastrointestinal tract nonrigid displacement ε parameters, when we estimate the motion of WCE camera.

A usual approach for this estimation is to extract distinctive features from each image and match features to establish a global correspondence, then to estimate transformation between the images (*feature-based methods*). In reality, this is very difficult to extract robust, stable and distinctive features between successive WCE images and establish a matching. A major reason is due to the low resolution, poor structural information of WCE image. In this paper, we use the Bee Algorithm (BA) [15] to search the best solution of the WCE camera motion parameters (BAME). BA is a new population-based search algorithm for many complex multiobjective optimization problems that cannot be solved exactly within the polynomial bounded computation times. As opposed to the feature-based methods, this method is often called direct (pixel-based) alignment methods. BAME can be described as seeking a minimal error:

$$T^* = \arg\min \ \text{error}_n(I_{i-1}, T_n(I_i)) : 1 \le n \le m, (i, m) \in \mathrm{Z} \tag{4}$$

where I_{i-1} and I_i are a neighborhood image and a source image in WCE video, T_n is nth transformation with deformation parameters being searched for in the solution space, and $\text{error}_n(I_{i-1}, T_n(I_i))$ is an error metric between the neighborhood image and source image. In our method, we use Mutual Information (MI) [16, 17] as error measure. MI is widely used in the medical image registration. In this paper, we use the gray-scale information to calculate MI because WCE images have similar intensity, color and hue. A pseudocode of the BAME is shown in Fig. 4.

The Motion Estimation of Local Gastrointestinal Tract

As mentioned earlier, gastrointestinal tract movement can be modeled as a nonrigid deformation, which is caused by gastrointestinal tract peristalsis. In this paper, we use the SIFT-flow [18] to predict the motion of the local gastrointestinal tract. The algorithm assumes that SIFT descriptors is able to establish the dense correspondences between a neighborhood image and a source image. SIFT descriptors have an excellent performance that is invariant in the local image illumination and encode local image structure. These properties make the matching more robust. SIFT flow can be formulated as an optimization problem on the correspondence search with the cost function:

$$E(\varepsilon) = \sum\nolimits_p \| s_{i-1}(p) - s_i(p + \varepsilon(p)) \|_1 + \frac{1}{\sigma^2} \sum\nolimits_p u_x^2(p) + u_y^2(p) + R(p, q) \tag{5}$$

Fig. 4 The pseudocode of the BAME algorithm. We modified the standard Bee Algorithm to select the best optimal WCE camera motion parameters

```
Input: target image and source image $\{I_{i-1}, I_i\}$
• Initialize population of bees $m \in Z$
• Search solution randomly $TS=\{T_n=random(s, R, T): 1 \le n \le m, n \in Z\}$
• Evaluate the error of each solution $ES=\{error_n(I_{i-1}, T_n(I_i)): 1 \le n \le m\}$
While (stopping criterion not met)
    • Select the best solutions $eTS=\{T_n \in TS: error_n \le threshold, n \le p\}$
    for $i1=1$ to $p$
        • Recruiting new bees for optimal solution in each neighborhood. $eTS = update(eTS)$
        • $eES=\{error_n(eTS)\}$
    end
    • Assign remaining bees to search randomly $rTS=\{T_n=random(s, R, T): p < n \le m\}$
    • Evaluate the error $rES=\{error_n(I_{i-1}, T_n(I_i)): p \le n \le m\}$
    • $TS=eTS \cup rTS$, $ES=eES \cup rES$
• End
Output: $T^*=\{T_n \in TS: error_n=argmin(ES)\}$
```

$$R(p,q) = \sum\nolimits_{(p,q)\in N} \min \alpha\left|u_x(p) - u_x(q)\right|, d + \min \alpha\left|u_y(p) - u_y(q)\right|, d \tag{6}$$

where $\varepsilon(p)=(u_x(p),u_y(p))$ is a flow vector at pixel location $p=(x,y)$, $s_{i-1}(p)$ and $s_i(p)$ is the *SIFT* descriptor extracted at location p in the neighborhood image $i-1$ and source image i in a WCE video, and N is the spatial neighborhood of a pixel. Here, the source image is a transformed image using the result of the motion estimation of WCE camera. $R(p,q)$ constrains the flow vector which is consistent with adjacent pixels. The flow vector $\varepsilon(p)$ is regarded as local nonrigid motion component. Finally, motion estimation between two successive WCE images can be described as follows:

$$Z_{i-1} = M \times Y_i^T + \varepsilon_i^T \tag{7}$$

The right-hand side of the equation contains two terms. The first term is global rigid motion estimated by BAME and the matrix M is a transformation matrix of WCE camera. The second term is local nonrigid motion estimate by SIFT flow and ε is a local flow vector, the best local match can be found along with the flow vector between two successive WCE images. Z_{i-1} is an approximation of the neighborhood image points corresponding to Y_{i-1}.. The Eq. (7) can also be described as a universal form:

$$Z_{i-1} = \left[\begin{array}{c|c} s\cdot R & [T+\varepsilon_i^T] \\ \hline 0 & 1 \end{array}\right] \times Y_i^T \tag{8}$$

Measuring Scene Changes in WCE Images

In WCE video, a scene change means that two successive WCE images have more dissimilar context. For a robust measurement of scene changes between successive WCE images, we define a concept of the invalid region. An invalid region is a region in a WCE image (source image) that contains the scene of gastrointestinal wall that we cannot found it in neighborhood images (or key images). Here, key images are some images that should be preserved in the reduction process. As shown in Fig. 5, we take two steps to achieve the invalid region estimation. The first step is backward estimation. We estimate the scene deformation from image i to $i-1$ using Eq. (7), and seek for points of image i that not lie in image $i-1$ according to scene deformation. Those points are recorded as a set OB$=\{Y_i\text{: }\{M\times Y_i\} \notin Y_{i-1}\}$. The second step is forward estimation. We record points as a set OF$=\{Y_i\text{: }\{M\times Y_i\}\notin Y_{i+1}\}$ between image i and image $i+1$. Finally, an invalid region of current frame is a set:

$$IR = OF \cap OB \tag{9}$$

The area of invalid region can reflect the scene change between two successive WCE images. Figure 6 demonstrates an invalid region of image i. Even if the number of the points in the invalid region is large, it may still not indicate the large displacement. Because the invalid region of a WCE image is often distributed at the margin and is usually narrow. Actually, the maximum diameter of the invalid region can reflect a potential possibility of the size of the scene changes. Therefore, we simplify this measurement and use the diameter of the max-inscribed circle (DMC) of invalid region to evaluate whether there is a scene change between two images.

The Scheme of the WCE Video Reduction

In our reduction scheme, WCE video is represented as the set $F=\{f_1,\ldots, f_i, f_{i+1},\ldots f_n\}$. We divide WCE video F into some segments named shots. A shot is defined as a finite subset of F: $SF=\{f_i, f_{i+1},\ldots,f_m\text{: } i, m\leq n\}$. Then, we define an *extract* function, which describes an extracting images procedure in a shot KF$_j$=extract(SF$_j$). KF represents a set

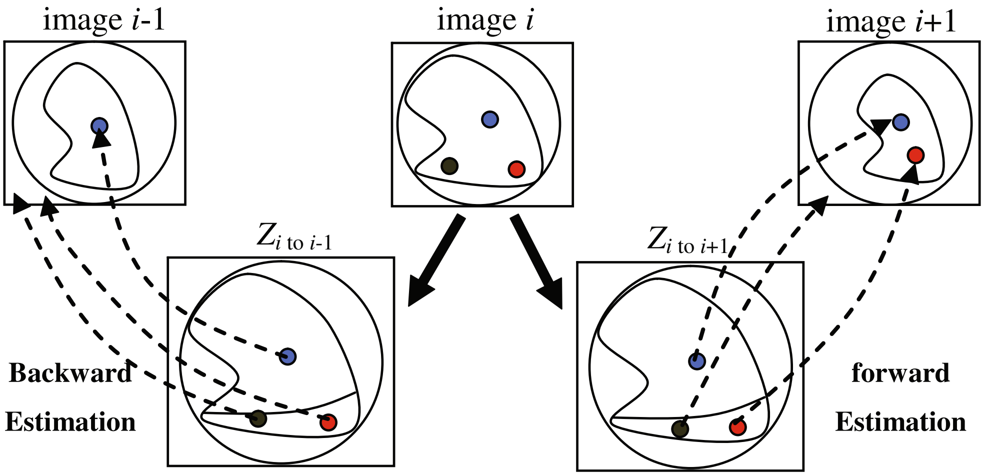


Fig. 5 An invalid region is measured in the forward and backward manner. For example, the image i has three points (*red*, *blue*, and *black*), after transformation, the black point cannot be found in the images $i-1$ and $i+1$

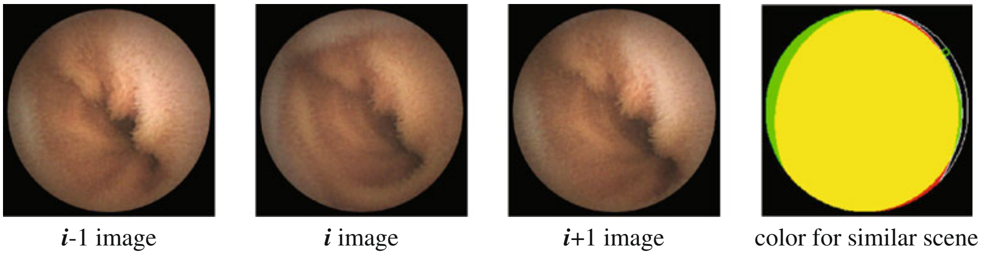


Fig. 6 Color for similar scenes: *green* and *red regions* that are the scenes of image i can only be found in $i-1$ and $i+1$ images respectively, *yellow region* (the scenes of image i) can be found both in $i-1$ and $i+1$ images. *Green circle* is a max-inscribed circle of the invalid region

consisting of some key images (key frames) should be preserved in a shot (j is a jth shot). If a WCE video is divided into the number of k shots, then our reduction mascheme can be described as an equation:

$$\mathrm{KFS} = \bigcup_{j=1}^{k} \mathrm{extract}(\mathrm{SF}_j) = \bigcup_{j=1}^{k} \mathrm{KF}_j \tag{10}$$

In our method, we set the first image in each shot as a key frame (KF), then we estimate the scene motion between source image i and key frame as backward estimation, and estimate the scene motion between source image i and neighborhood image $i+1$ as forward estimation. The invalid region is calculated after scene motion estimation. Finally, we determine whether current image should be preserved as a key frame or not according to the threshold τ of DMC. If a source image i is preserved as key frame, the following image will estimate scene motion with this new key frame in the backward estimation. This extraction procedure in a shot can be described as follows:

$$\mathrm{extract}(\mathrm{SF}_j) = \ \exists \mathrm{KF} \in \mathrm{SF}_j : \mathrm{DMC}(\mathrm{KF}) < \tau \tag{11}$$

In addition, we use the multiplication technique to deal with the case that has an interval between source image i and key frame, as shown in Fig. 7. The motion estimation between source image 2 (CF2) and key frame (KF) is given by:

$$Z_{\mathrm{KF}} = M_1 \times M_2 \times Y_{\mathrm{CF2}}^{T} \tag{12}$$

Here, M_1 and M_2 are transformation matrices in Eq. (8), Y_{CF2} points in CF2 and Z_{KF} points transformed from current points Y_{CF2} using the multiplication of transformation matrices M_2 and M_1.

Results and Discussions

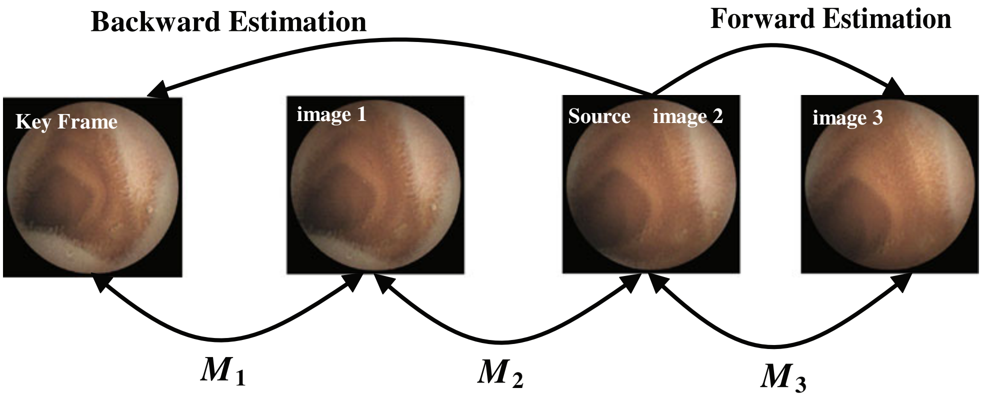
Material and Evaluation of the Proposed Method

We test our method with various WCE image sequences from different patients provided by Nanjing General Hospital of Nanjing Military Command. We divide our experiments into two parts. The first part verifies whether the proposed WCE video motion model is suitable for WCE image motion estimation and the second part tests whether our reduction scheme is significant on WCE video.

In the experiments of the motion estimation, we use both image registration and PSNR, MSE, SSIM, and MI to evaluate the motion estimation performance. In reduction experiments, we use recall and precision to measure the reduction performance, which are widely used in the field of information retrieval. We also compare our reduction scheme with ARPS and FCM-NMF. Recall and precision in the reduction of WCE video is given as follows:

$$\mathrm{Recall} = \frac{\text{Number of keyframes correctly detected}}{\text{Number of keyframes in the ground truth}} \tag{13}$$

Fig. 7 The reduction scheme in a WCE video shot. We use multiple multiplication technique to calculate scene changes between source image i and key frame



$$\text{Precision} = \frac{\text{Number of keyframes correctly detected}}{\text{Number of keyframes detected}} \tag{14}$$

Experiments of WCE Video Motion Estimation

Both WCE camera and local gastrointestinal tract motion estimation are tested in our experiments. The source and neighborhood images as a pair of images are chosen from WCE video including three types of gastrointestinal tract images: stomach, small intestine, and colon, and motion estimation performance is evaluated with image registration and PSNR, MSE, SSIM, and MI. In the experiments of WCE camera motion estimation as shown in Fig. 8, we notice that the image registration of BAME is clearer in structural information and less visual artifacts than the direct image alignment method. We also run BAME on some nongastrointestinal tract images and notice that BAME is also effective in the general image scene. The residual error between the neighborhood image and transformed image with BAME is presented in Fig. 11. BAME does not show noticeable geometrical difference, which can explain that BAME can estimate WCE camera motion precisely.

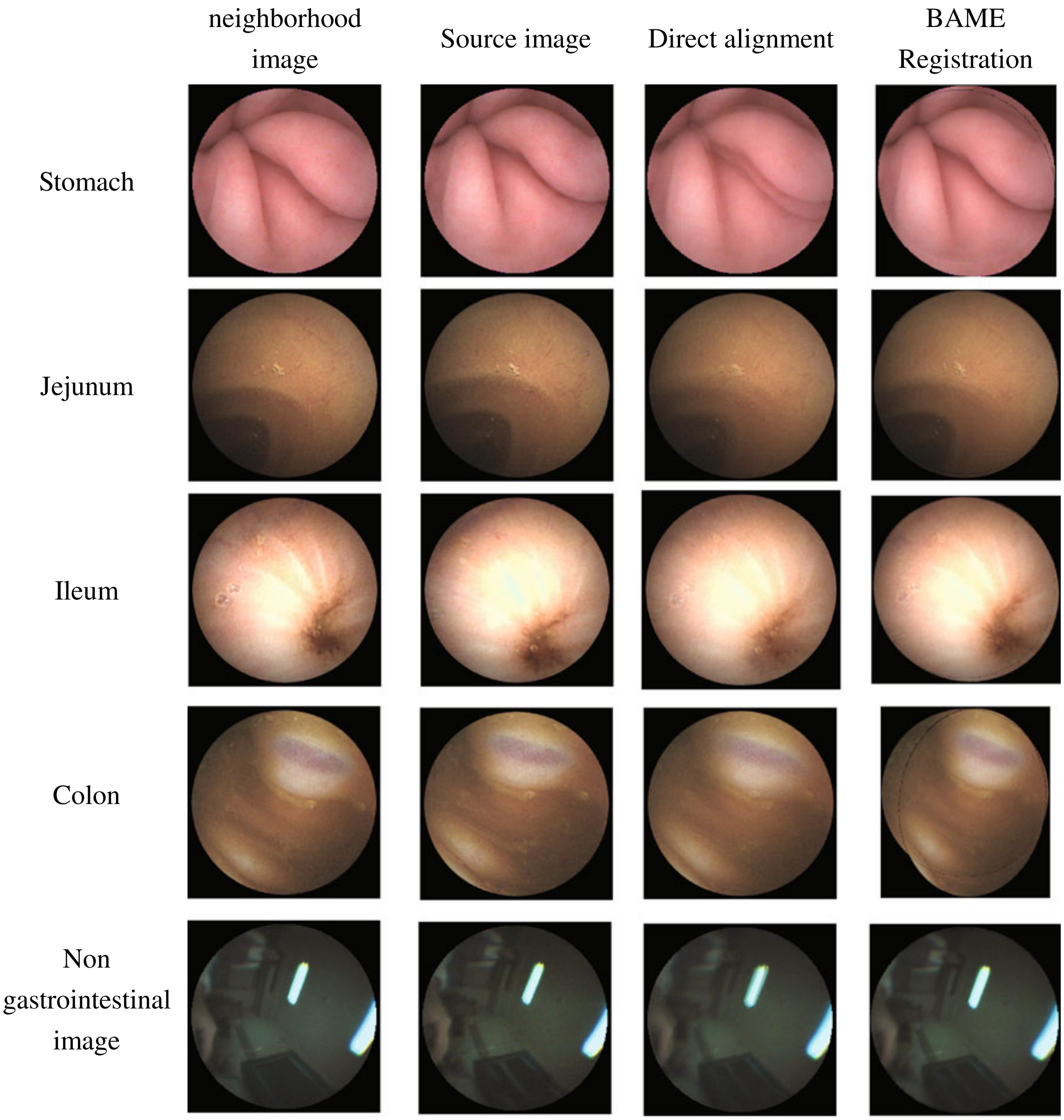


Fig. 8 The experiments of WCE camera motion estimation in gastrointestinal and nongastrointestinal image with BAME. Neighborhood and source images as a pair of images were chosen from different parts of WCE video: stomach, jejunum, ileum, colon and nongastrointestinal tract. We compare our BAME performance with image registration. The *third column* is the registration of image direct alignment without any transformation

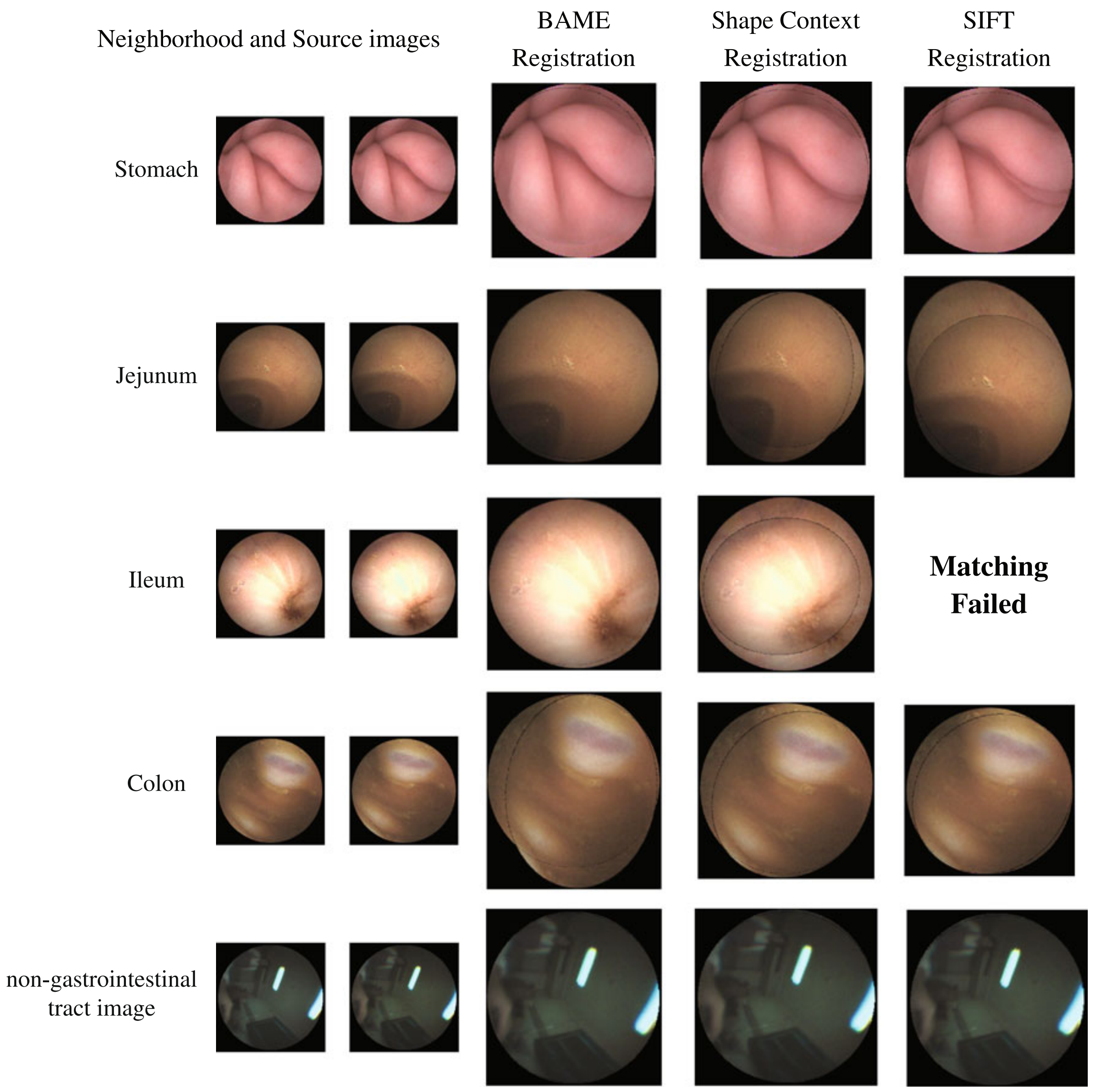

Neighborhood and Source images
BAME Registration
Shape Context Registration
SIFT Registration
Stomach
Jejunum
Ileum
Matching Failed
Colon
non-gastrointestinal tract image

Fig. 9 A comparison on image registration of BAME, SIFT, and Shape Context (SC). BAME is more effective on WCE motion estimation. The registration image of BAME has less visual artifacts than those of SIFT and SC

Then, invariant feature transform (SIFT) matching and shape context (SC) matching can be considered as a comparison with BAME in our experiments. SIFT and SC were set the same parameters as described in the papers [19, 20]. In the experiments of SIFT matching, we found that SIFT cannot extract dense matching points effectively between two successive images in most cases. This may be due to low resolution, poor structural information, and texture-less regions in WCE images. For SC Matching, we extract the edge (or contour) information of two successive WCE images first [13]. Then, we apply the SC algorithm to establish dense matching pair points between these images

Table 1 The comparison of BAME with SC and SIFT on PSNR and MSE

Methods	PSNR				MSE			
	Stomach	Jejunum	Ileum	Colon	Stomach	Jejunum	Ileum	Colon
Shape Context	25.56	27.28	19.35	28.36	182.03	122.60	760.31	95.55
SIFT	26.17	25.69	–	26.641	158.32	176.61	–	142.01
BAME	27.21	29.25	22.15	27.98	124.59	77.94	399.63	116.30

Table 2 The comparison of BAME with SC and SIFT on SSIM and MI

Methods	SSIM				MI			
	Stomach	Jejunum	Ileum	Colon	Stomach	Jejunum	Ileum	Colon
Shape context	0.83	0.78	0.67	0.77	1.05	1.07	1.05	1.05
SIFT	0.84	0.784	–	0.75	1.05	1.04	–	1.06
BAME	0.86	0.79	0.72	0.76	1.18	1.24	1.19	1.21

according to the edge information. Finally, the motion is estimated based on these matching points. For some cases, we could not ensure to extract edges (or contours) correctly from WCE images and also could not ensure that edges (or contours) extracted from WCE images is consistent or contains similar structural information between two WCE images. However, for the nongastrointestinal images, these images have obvious contours and structure; thus, both SIFT and SC can work well on the motion estimation of WCE camera. Experimental results of SIFT and SC matching are presented in Fig. 9. We use PSNR, MSE, SSIM, and MI to compare the registration performance of these methods. We notice that the performance of BAME is better than those of SIFT and SC in most cases. These results are shown in Tables 1 and 2.

Next, we use Eqs. (5) and (6) to estimate the local nonrigid motion of gastrointestinal tract based on the results from the motion estimation of WCE camera. In Fig. 10, we can notice that the alignment image with BAME-SIFTFlow is less blurred than the only motion estimation with BAME, which means SIFT flow can improve the accuracy of the motion

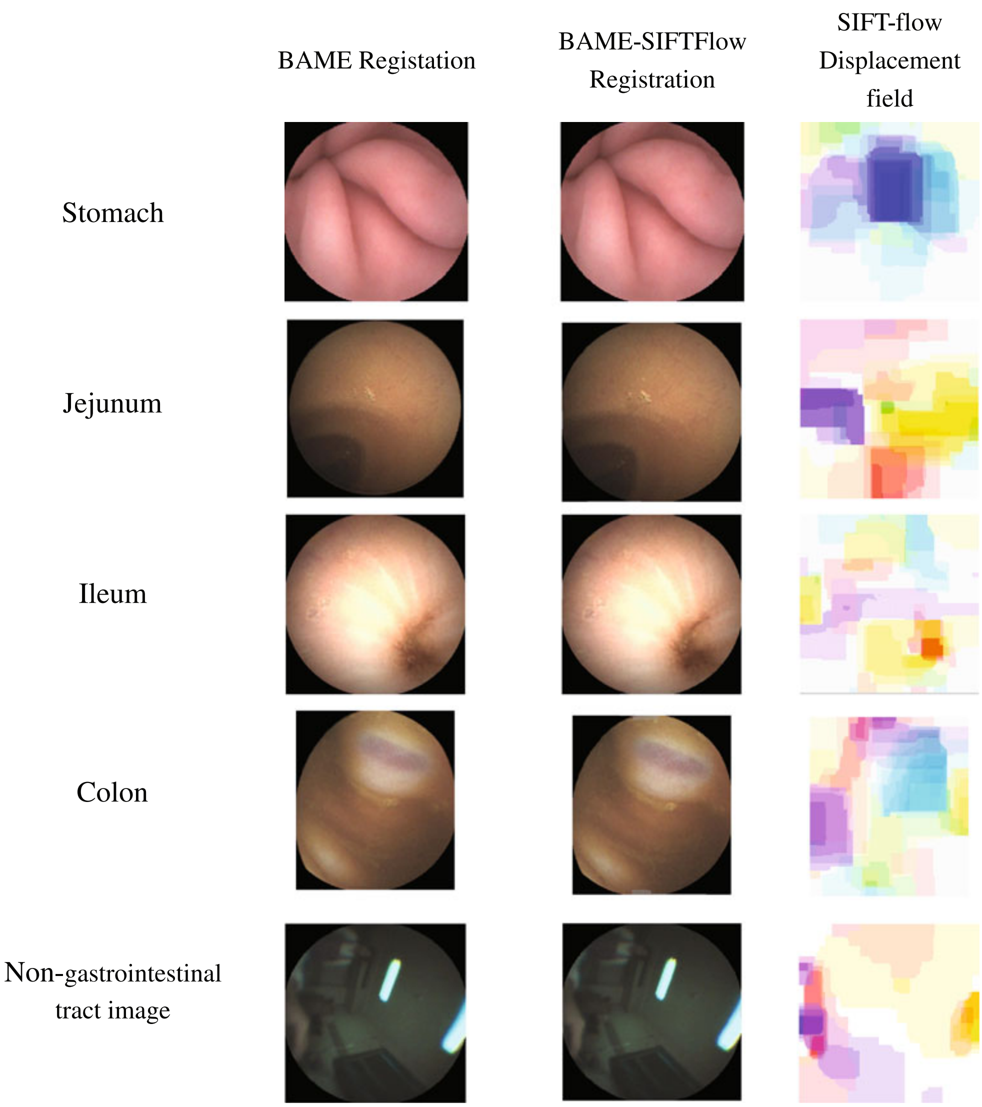

BAME Registation
BAME-SIFTFlow Registration
SIFT-flow Displacement field
Stomach
Jejunum
Ileum
Colon
Non-gastrointestinal tract image

Fig. 10 The local nonrigid motion estimation with SIFT flow. The registration image with BAME-SIFTFlow is less blurred than only motion estimation with BAME. SIFT-flow displacement field is a visualization of pixel displacements using the color-coding scheme of [23]

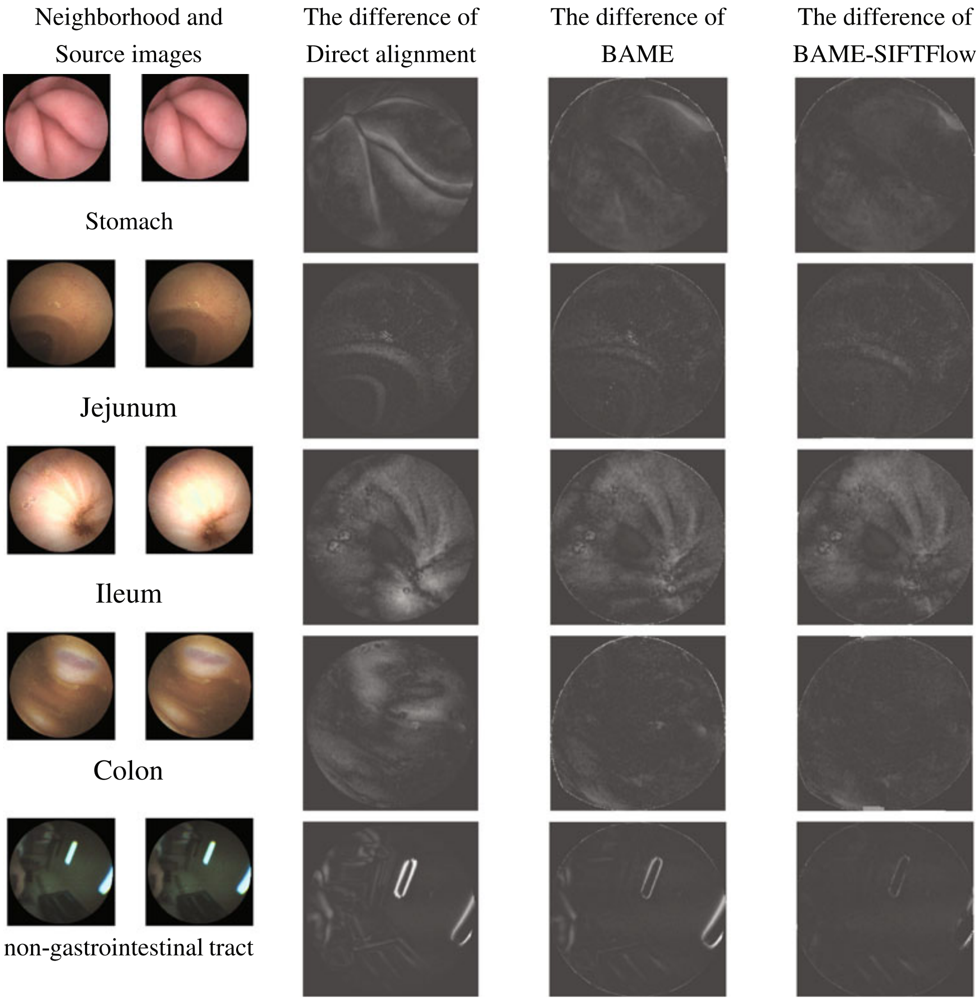

Neighborhood and Source images
The difference of Direct alignment
The difference of BAME
The difference of BAME-SIFTFlow
Stomach
Jejunum
Ileum
Colon
non-gastrointestinal tract

Fig. 11 The residual error is compared between the neighborhood image and transformed image with BAME and BAME-SIFTFlow. BAME-SIFTFlow does not present noticeable geometrical difference

estimation in the fine scale. This also verifies that the proposed the motion model is suitable for WCE video. From Fig. 11, we can find that the residual error is minimized between the neighborhood image and transformed image with BAME-SIFTFlow. Especially, the difference is smaller in nongastrointestinal images and stomach and colon images. However, for images of jejunum and ileum, we still found some obvious structure. The reason may be due to that the dense villi of small intestine cause an uncertain local motion.

To further test the performance of BAME-SIFTFlow, we compare the BAME-SIFTFlow with several popular nonrigid motion estimation algorithms, which include standard optical flow (HS), demon nonrigid registration method [21], and large displacement optical flow (LDOF) [22]. These methods are tested based on the results from the motion estimation of WCE camera with BAME and are named BAME-HS, BAME-Demon and BAME-LDOF. We again use PSNR, MSE, SSIM, and MI to evaluate the performance. From the comparison, we notice that the BAME-HS method is the worst of the local motion estimation among these methods. The BAME-LDOF has a poor performance compared with BAME-demon and BAME-SIFTFlow. However, our method has the better performance in most situations. The comparison is presented in Tables 3 and 4.

Table 3 The comparison of BAME-SIFTFlow with BAME-HS, BAME-Damon, and BAME-LDOF on PSNR and MSE

Methods	PSNR				MSE			
	Stomach	Jejunum	Ileum	Colon	Stomach	Jejunum	Ileum	Colon
BAME-HS	26.88	26.20	22.00	26.35	134.55	156.87	413.388	151.21
BAME-Damon	28.19	30.35	26.06	28.26	99.45	60.45	162.16	97.75
BAME-LDOF	26.79	26.97	21.90	26.49	137.24	131.56	423.30	147.16
BAME-SIFTFlow	28.44	28.17	22.16	27.98	93.94	99.80	398.47	104.23

Table 4 The comparison of BAME-SIFTFlow with BAME-HS, BAME-Damon, and BAME-LDOF on SSIM and MI

Methods	SSIM				MI			
	Stomach	Jejunum	Ileum	Colon	Stomach	Jejunum	Ileum	Colon
BAME-HS	0.86	0.78	0.72	0.76	1.05	1.04	1.04	1.07
BAME-Damon	0.85	0.80	0.73	0.79	1.25	1.27	1.23	1.24
BAME-LDOF	0.85	0.79	0.72	0.77	1.04	1.04	1.04	1.06
BAME-SIFTFlow	0.89	0.81	0.74	0.79	1.21	1.24	1.19	1.24

Evaluation of the Reduction Scheme

Once the motion of a WCE video is estimated, we use it as salient features to reduce the number of images in a WCE video. In our experiments, we tested our reduction scheme on WCE video clips collected from different gastrointestinal tract scene, such as stomach, jejunum, ileum, and colon. Each WCE video clip contains 100 images, and each shot has 50 images. In each shot, we save the first image and last image as initial key frames. From the second image, we apply our reduction scheme as described in "Measuring Scene Changes in WCE Images." In experiments, DMC threshold is set to 5 pixels empirically. The number of key frames in each shot is shown in Table 5.

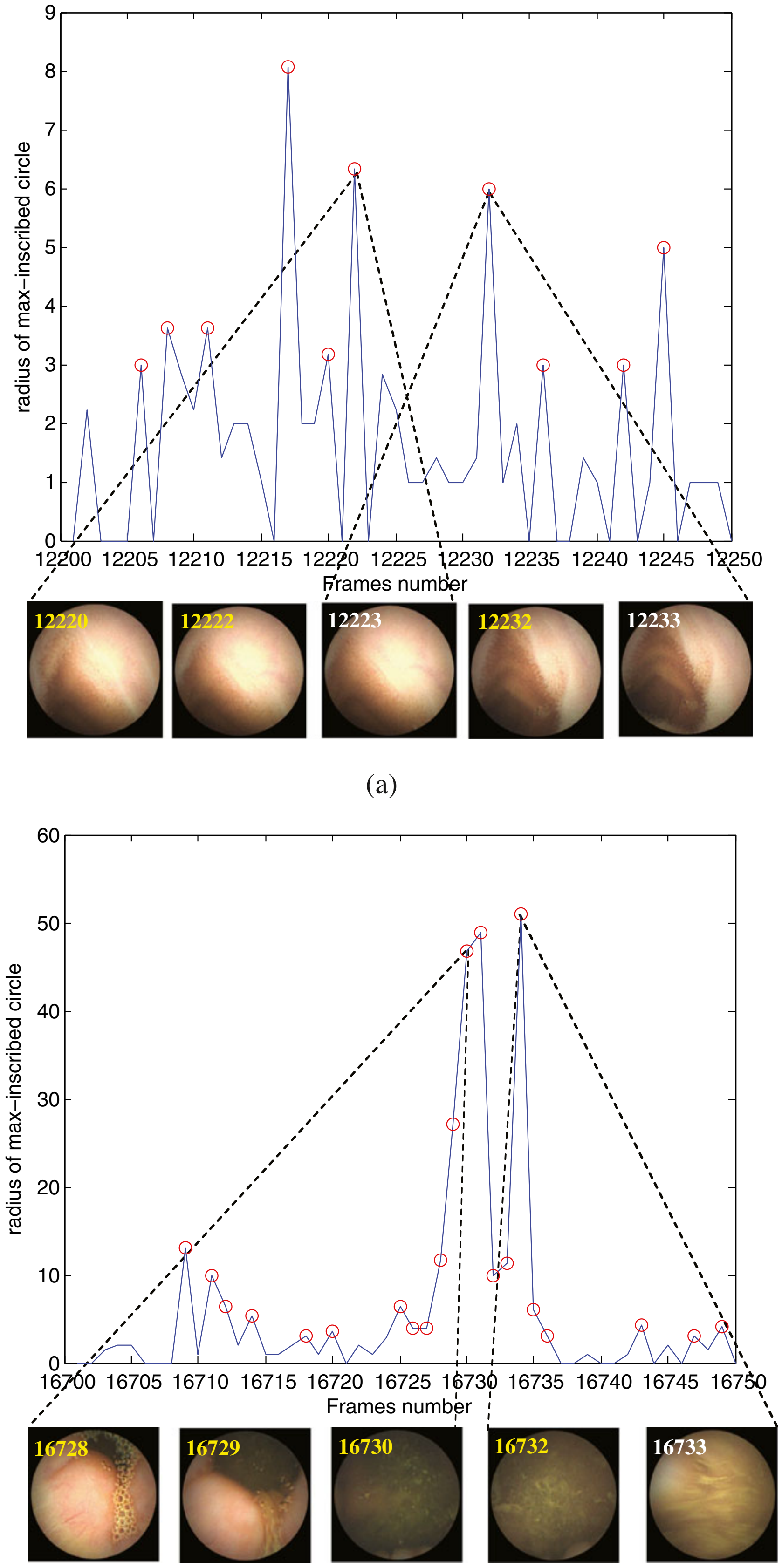
We plot the curves about the radius of max-inscribed circle of each WCE image invalid region in a shot and draw key frames and its neighborhood images (shown in Fig. 12). Red circle represents an image chosen by DMC threshold. In Fig. 12, key frames are marked with yellow number label, such as 1222, 12232, 16729, etc., and neighborhood images are marked with white number label. We notice that the key frame has an obviously scene changes. Figure 12b is an example of a WCE transition from ileum to colon, and key frame 16729 is ileocecal valve. Neighborhood images (16728 and 16730) have an obvious different scene; therefore, both of them are preserved as key frames. Figure 12b also explains a phenomenon that the movement of small intestine is larger than colon. However, even if the movement is less in colon, but food residue may still distort the motion estimation in colon images, which can explain why a shot of transition from ileum to colon preserves more images.

We verify the performance of our reduction scheme by recall (RC), precision (PC) and compression ratio (CR). The key frames of ground truth in WCE video shots are labeled by the clinician. We also compare our method with ARPS and FCM-NMF scheme. These schemes were set same parameters as described in the papers [9, 11]. These results of which are presented in Tables 6 and 7. For a good WCE video reduction scheme, we hope both recall and precision should be as high as possible. However, in practice, we are more concerned with the correct key frames detected in WCE video because the false positives may lead to a failure for medical diagnosis. Thus, recall is more important than precision in our application.

From experimental results, the average performance of our reduction scheme for recall is 74 %, precision is 58 %, and CR is 68 %. Although ARPS and FCM-NMF have higher compression radio (CR) than our method, the higher RC also causes these methods having lower RC and PC. This is due to ARPS and FCM-NMF cannot preserve the representative images in WCE image sequence. For example, in images sequence 16701 to 16750 (Ileum to Colon), both ARPS and FCM-NMF cannot preserve ileocecal valve (16728 and 16729) as key frames. In fact, the local nonrigid deformation is not considered in ARPS. ARPS is a template search scheme, which is not suited for the case of WCE movement because

Table 5 The number of key frames in each shot

The shot	Key frame number	Compression radio (CR) (%)
3001-3050 Stomach	3000, 3004, 3013, 3015, 3016, 3017, 3018, 3025, 3035, 3039, 3043, 3048, 3049, 3050	72
12201-12250 Jejunum	12201, 12206, 12208, 12211, 12217, 12220, 12222, 12232, 12236, 12242, 12245, 12250	76
12251-12300 Jejunum	12251, 12256, 12258, 12261, 12267, 12270, 12272, 12282, 12286, 12292, 12295, 12300	76
13001-13050 Ileum	13001, 13003, 13005, 13008, 13012, 13013, 13015, 13019, 13023, 13026, 13029, 13036, 13038, 13042, 13043, 13046, 13048, 13050	66
13051-13100 Ileum	13051, 13055, 13059, 13060, 13062, 13064, 13068, 13070, 13073, 13074, 13076, 13078, 13086, 13088, 13091, 13092, 13097, 13100	66
16701-16750 Ileum-Colon	16701, 16709, 16711, 16712, 16714, 16718, 16720, 16725, 16726, 16727, 16728, 16729, 16730, 16731, 16732, 16733, 16734, 16735, 16736, 16743, 16747, 16749, 16750	54



(a)

(b)

Fig. 12 The curve is the radius of max-inscribed circle of each WCE image invalid region in a shot. **a** The curve of a ilenum shot, and **b** the curve of a ilenum–colon shot. Key frames are marked with *yellow number label*, and neighborhood images are marked with white number label

Table 6 The comparison of our reduction scheme with ARPS k=2 and FCM-NMF c=5

Image Sequence	ARPS k=2			FCM-NMF c=5			BAME-SIFTFlow		
	RC (%)	PC (%)	CR (%)	RC (%)	PC (%)	CR (%)	RC (%)	PC (%)	CR (%)
3001-3050 Stomach	56	47	70	18	40	90	73	67	72
12201-12250 Jejunum	71	36	72	14	10	70	86	60	76
12251-12300 Jejunum	50	42	76	40	30	74	60	60	76
13001-13050 Ileum	56	71	72	38	50	72	56	63	66
13051-13100 Ileum	56	46	74	33	33	82	78	44	66
16701-16750 Ileum to Colon	33	31	54	17	17	76	92	52	54
Average	54	46	70	27	30	77	74	58	68

RC, PC and CR denote recall, precision, and compression ratio, respectively

the scheme assumes that the camera movements are just along horizontal and vertical directions. Actually, the movement of WCE can be in any directions. We also notice that it does not have any obvious improvement in experiments of FCM-NMF with increasing the number of clusters. Moreover, in the experiments of FCM-NMF, we cannot obtain a consistent result from same images sequence and parameters each time, which may be due to the drift of clustering centers in clustering procedure. In addition, for FCM-NMF scheme, this is also difficult to evaluate a right similarity and form the geodesics distance matrix between successive images just using Euclidean distances.

However, the average recall of our method is still 74 % (see Table 6). A major reason could be the objective criterion of reducing redundant data in a WCE video is very different from subjective criteria of the clinician. For clinicians, they are more interested in image content itself than in the WCE video. This reason also explains why the precision is not high because our scheme leads to more images taken as key frame in a shot than the ground truth marked by clinician. In addition, different clinicians may have different subjective criteria for the key frame extraction. Even for a clinician, it is still very difficult to judge if one image is better than another that should be preserved. Meanwhile, we find that different movement speed in different part of gastrointestinal tract can cause different reduction ratio. If we increase the CR, it may decrease the recall. Thus, we must make a tradeoff between the recall and CR.

To verify whether our reduction scheme is more suitable than ARPS and FCM_NMF, we compare the sampling frequency of our scheme with ARPS and FCM-NMF on a WCE images sequence (500 frames) without ground truth labeled by clinician. The result is presented in Fig. 13. Black curve in Fig. 13 is intensity of an image sequence (500 frames). The changes of intensity can reflect the scene changes in image sequence. In the sampling curves (blue, red, and green), value one (1) represents a key frame extracted from images sequence. We found that our scheme is more consistent with the changes of intensity than ARPS and FCM-NMF. The more changes of scene, the more key frames need to be preserved, which can explain that our scheme can produce an acceptable reduction sequence for browsing and examination in WCE images sequence.

Conclusions

We presented a new reduction scheme for WCE video. This scheme makes use of the motion feature to preserve those

Table 7 The comparison of our reduction scheme with ARPS k=3 and FCM-NMF c=6

Image Sequence	ARPS k=3			FCM-NMF c=6			BAME-SIFTFlow		
	RC (%)	PC (%)	CR (%)	RC (%)	PC (%)	CR (%)	RC (%)	PC (%)	CR (%)
3001-3050 Stomach	27	43	86	18	40	90	73	67	72
12201-12250 Jejunum	57	49	80	43	17	64	86	60	76
12251-12300 Jejunum	20	29	86	50	28	64	60	60	76
13001-13050 Ileum	44	73	78	22	36	78	56	63	66
13051-13100 Ileum	44	50	84	33	33	82	78	44	66
16701-16750 Ileum-Colon	33	44	82	17	28	86	92	52	54
Average	38	48	83	31	30	77	74	58	68

RC, PC and CR denote recall, precision, and compression ratio, respectively

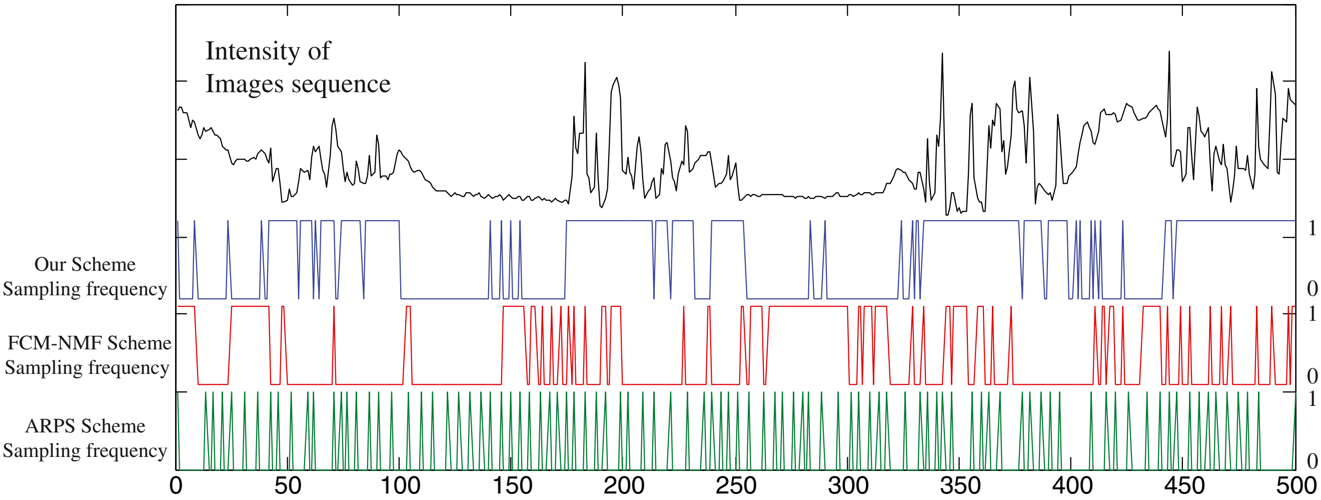


Fig. 13 A comparison on sampling frequency of ARPS, FCM-NMF, and our scheme. *Black curve at top row* is intensity of an image sequence (500 frames). The changes of intensity can reflect the sense changes in image sequence. The sampling frequency of our scheme (*blue curve*) is consistent with the changes of intensity

images that have obvious scene changes in the temporal neighborhood. To obtain the motion feature from successive WCE images, a new WCE video motion model is proposed. Based on this new motion model, the motion estimation of WCE video is divided into two levels, the coarse level and the fine level. In the coarse level, the WCE camera motion is estimated with BAME. In the fine level, the local gastrointestinal tract motion is estimated with SIFT flow.

Through the empirical comparison, we find that the BAME-SIFTFlow method can estimate the motion between WCE images more accurate in most situations, especially when two successive WCE images have a large displacement. This is mainly due to the fact that the method takes two stages to estimate the motion. Therefore, the BAME-SIFTFlow is consistent with our gastrointestinal tract motion assumption and is robust for practical WCE video applications. Moreover, we notice that SIFT-flow has a better performance for local motion estimation than HS, Damon, and LODF in WCE video. We think that SIFT flow can be extended for other medical image applications. However, we also find that our reduction scheme has a lower recall in some situation. A major reason could be that WCE images preserved as key frames by a clinician may be more subjective and lack of an objective criteria. Therefore, it is still a challenging task to determine which images should be preserved as key frames of WCE video with unsupervised learning. Other reasons may be that the motion estimation is not enough as a feature to capture the scene changes in WCE images sequence. Therefore, it should be considered by combining the motion feature and other image features, such as color and texture, for future work.

References

1. Iddan G, Meron G, Glukhovsky A: Wireless capsule endoscopy. Nature 405(6785):417, 2000
2. Mackiewicz M, Berens J, Fisher M: Wireless capsule endoscopy color video segmentation. IEEE Trans Med Imaging 27:1769–1781, 2008
3. Karargyris A, Bourbakis N: Wireless capsule endoscopy and endoscopic imaging: a survey on various methodologies presented. IEEE Eng Med Biol Mag 29:72–83, 2010
4. Li B, Meng MQH: Computer-aided detection of bleeding regions for capsule endoscopy images. IEEE Trans Biomed Eng 56:1032–1039, 2009
5. Li B, Meng MQH: Computer-based detection of bleeding and ulcer in wireless capsule endoscopy images by chromaticity moments. Comput Biol Med 39:141–147, 2009
6. Girgis HZ, Mitchell BR, Dassopoulos T, Mullin G, Hager G: An intelligent system to detect Crohn's disease inflammation in wireless capsule endoscopy videos. Proceedings of the 7th IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI 2010). IEEE, Rotterdam, The Netherlands 1373–1376, 2010
7. Li B, Meng MQH: Texture analysis for ulcer detection in capsule endoscopy images. Image Vision Comput 27:1336–1342, 2009
8. Karargyris A, Bourbakis N: Detection of small bowel polyps and ulcers in wireless capsule endoscopy videos. IEEE Trans Biomed Eng 58:2777–2786, 2011
9. Tsevas S, Iakovidis DK, Maroulis D, Pavlakis E: Automatic frame reduction of wireless capsule endoscopy video. Proceedings of the

2008 8th IEEE International Conference on Bioinformatics and BioEngineering, IEEE, Athens, Greece, 1–6, 2008
10. Iakovidis DK, Tsevas S, Polydorou A: Reduction of capsule endoscopy reading times by unsupervised image mining. Comput Med Imaging Graph 34:471–478, 2010
11. Yao N, Kai-Kuang M: Adaptive rood pattern search for fast block-matching motion estimation. IEEE Trans Image Process 11:1442–1449, 2002
12. Improved Endo Capsule Software Enhances Diagnostic Experience. Available at http://www.olympusamerica.com/cpg_section/cpg_headlineDetails.asp?pressNo=713
13. Karargyris A, Bourbakis N: A video-frame based registration using segmentation and graph connectivity for wireless capsule endoscopy. Proceedings of the 2009 IEEE/NIH Life Science Systems and Applications Workshop (LiSSA 2009), IEEE/NIH, Bethesda, MD, USA, 2009, pp. 74–79
14. Karargyris A, Bourbakis N: Three-dimensional reconstruction of the digestive wall in capsule endoscopy videos using elastic video interpolation. IEEE Trans Med Imaging 30:957–971, 2011
15. Pham DT, Ghanbarzadeh A, Koc E, Otri S, Rahim S and Zaidi M: The Bees Algorithm-A Novel Tool for Complex Optimisation Problems. In Proceedings of the 2nd International Virtual Conference on Intelligent Production Machines and Systems (IPROMS 2006), Cardiff, UK, 2006, Elsevier, Oxford, pp. 454–459
16. Maes F, Collignon A, Vandermeulen D, Marchal G, Suetens P: Multimodality image registration by maximization of mutual information. IEEE Trans Med Imaging 16:187–198, 1997
17. Pluim JPW, Maintz JBAA, Viergever MA: Mutual-information-based registration of medical images: a survey. IEEE Trans Med Imaging 22:986–1004, 2003
18. Liu C, Yuen J, Torralba A, Sivic J, Freeman WT: SIFT flow: Dense correspondence across different scenes. Proceedings of the 10th European Conference on Computer Vision (ECCV 2008), Marseille, 2008, Springer, France, 3:28–42
19. Lowe DG: Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60:91–110, 2004
20. Belongie S, Malik J, Puzicha J: Shape matching and object recognition using shape contexts. IEEE Trans Pattern Anal Mach Intell 24:509–522, 2002
21. Thirion J-P: Image matching as a diffusion process: an analogy with Maxwell's demons. Med Image Anal 2:243–260, 1998
22. Brox T, Malik J: Large displacement optical flow: descriptor matching in variational motion estimation. IEEE Trans Pattern Anal Mach Intell 33:500–513, 2011
23. Baker S, Scharstein D, Lewis JP, Roth S, Black MJ, Szeliski R: A database and evaluation methodology for optical flow. Proceedings of the IEEE 11th International Conference on Computer Vision, IEEE, 1–8, 2007