Proceedings of the IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI 2012)
Hong Kong and Shenzhen, China, 2-7 Jan 2012

An Advanced WCE Video Summary Using Relation Matrix Rank

Jia Sen HUO, Yue Xian ZOU*, *Senior Member, IEEE*, Lei LI

Abstract—Wireless Capsule Endoscopy (WCE) is a promising new solution for gastrointestinal disease detection, which is able to view the entire gastrointestinal tract without pain. Since there are more than 50,000 frames in one WCE video from each test, the automatic computer-aided technique is highly demanded to remove the redundant frames while remaining the medical information. This paper presents a systematic method to achieve the effective WCE video summary using whole WCE video frames. A cascade color and texture feature based most representative frame extraction algorithm (CCTS-MRFE) has been developed. The color feature extraction method is developed in HSV space. Meanwhile, the texture feature extraction is determined by the proposed block edge directivity descriptor (BEDD) method in gray space. A relation matrix rank (RMR) based most representative frame (MRF) selection approach further improving the performance of the proposed CCTS-MRFE algorithm has been proposed. The preliminary experimental results show that the proposed methods are able to achieve the effective WCE video summary.

I. Introduction

Wireless Capsule Endoscopy (WCE) [1] was invented in year 2000, which made a significant breakthrough over the traditional endoscopy technique since it is a safe, noninvasive and reliable one. When a WCE is swallowed by a patient, it will be propelled by peristalsis to move forward through the entire gastrointestinal tract, and the corresponding images will be taken, sent and recorded during the process. The main advantage of this new technology is that the process of the physical examination does not require the sedation and is non-invasive which applies little pain to the patient.

However, a challenging problem is faced by the current WCE techniques. The WCE video contains more than 50,000 frames or images on average. It takes about two more hours for an experienced physician to review the whole video frames [2]. Meanwhile this examination process cannot guarantee all abnormal regions are detected due to possible human errors. Computer aided diagnosis method has been proved as an effective and reliable solution instead. The video summary method [3] has been considered as one of the solutions to achieve a satisfactory performance for the WCE techniques. Literature study showed that there are several WCE video summary methods have been proposed. Akovidis *et. al*. proposed an unsupervised methodology based on clustering and non-negative matrix factorization (NMF) to summarize the WCE video by keeping the MRFs from the whole examination. Their preliminary results showed that the proposed method is able to reduce up to 85% redundant frames without loss of abnormalities [4]. In [5], Meng *et. al*. developed a method based on the frame boundary detection and the linear discriminant analysis technique. The method reaches the reduction of 95% redundant frames with fidelity of 85% on average according to their experimental results.

In this paper, after carefully analyzing the WCE images, we proposed a novel method to summarize the WCE video sequences. Firstly, the color features of the images have been extracted and applied to segment the WCE video sequence into several clips. Secondly, a block edge directivity descriptor (BEDD) method has been proposed to extract the texture features of the WCE images. The texture features have been applied to further segment the clips into more accurate and smaller segmented clips, named as sub-clips. Thirdly, a novel Relation Matrix Rank (RMR) based MRF selection method is proposed for each sub-clip to further improve the performance of the WCE video summary.

The paper is organized as follows. Section II describes the proposed WCE video summary system and methodology. Section III presents a new method to select the MRFs from each sub-clip. Section IV shows the experimental results and performance analysis. Conclusions are drawn in Section V.

II. Proposed WCE Video Summary System And Methods

For an image, color and texture features are recognized as two major characteristics. Through careful evaluation of the WCE images, we have the following observations: 1) the frame color information is highly associated with the gastrointestinal organs. The transitional frames of different organs present different color patterns. It will be effective to use the color features of the WCE images to segment the whole WCE video into different clips in time order; 2) the frame texture features present more information of the structure of the digestive tract mucosa, which can be used to further segment the clips into sub-clips. The proposed WCE video summary system is illustrated in Fig. 1.

A. Segmentation based on Color Features

As shown in Fig. 1, we firstly need to extract the color features of the WCE images. A lot of color spaces can be considered, such as RGB, Lab, CIE and et. al. In this paper, we use the HSV color space since it provides richer hue and saturation information compared to other color spaces.

For extracting the color features, we firstly transform the WCE images from RGB space to HSV space, where the S-, H-, V-images can be determined as follows, respectively:

$$s = \begin{cases} 0 & \text{if } \max(r,g,b) = 0 \\ 1 - \dfrac{\min(r,g,b)}{\max(r,g,b)} & \text{otherwise} \end{cases} \quad (1)$$

The authors are with Advanced Digital Signal Processing Laboratory, Peking University Shenzhen Graduate School. (* Email: zouyx@szpku.edu.cn)
This work was supported by the research project funded by Shenzhen ZiFu Technology Ltd.

$$h=\begin{cases}\text{undefined} & \text{if } \max(r,g,b)=\min(r,g,b)\\ \dfrac{60^\bullet\times(g-b)}{\max(r,g,b)-\min(r,g,b)}+0^\bullet & \text{if } \max(r,g,b)=r,\ g>b\\ \dfrac{60^\bullet\times(g-b)}{\max(r,g,b)-\min(r,g,b)}+360^\bullet & \text{if } \max(r,g,b)=r,\ g<b\\ \dfrac{60^\bullet\times(b-r)}{\max(r,g,b)-\min(r,g,b)}+120^9 & \text{if } \max(r,g,b)=g\\ \dfrac{60^\bullet\times(r-g)}{\max(r,g,b)-\min(r,g,b)}+240^\bullet & \text{if } \max(r,g,b)=b\end{cases} \quad (2)$$

$$v=\max(r,g,b) \quad (3)$$

The hue and saturation information will be extracted since we noted that the hue information of the image is more representative in differentiating the color differences in WCE images. Therefore, in our system, seven hues and three saturation levels are considered, and the H-histogram and S-histogram can be computed accordingly. From the H-histogram, the primary and secondary color chromaticity can be extracted by locating the highest and second highest bins, respectively. Similarly, from the S-histogram, we can determine the main saturation distribution feature by locating the highest bin. As a result, two hue features and one saturation feature have been used to form the color feature vector (CFV) for each frame.

Using the extracted frame CFV, the whole WCE video is segmented into different clips automatically, where the color segmentation condition is that if two frames have the same CFV, then these two frames can be segmented into one clip. It is noted that the number of the segmented clips by using CFV depends on the WCE video color features.

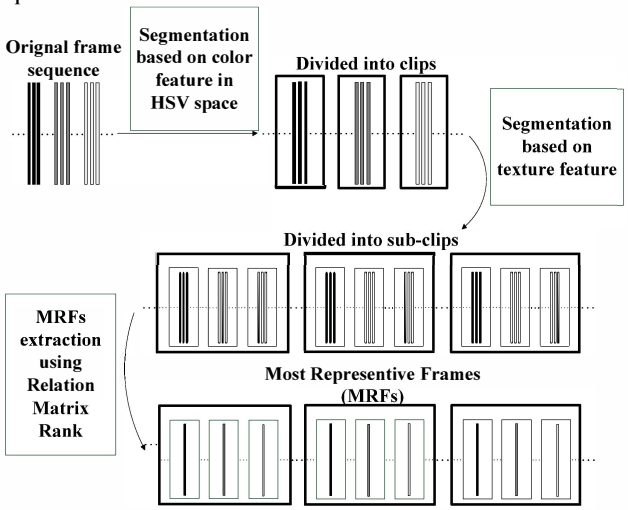

Fig 1 Conceptual illustration of the proposed WCE video summary system

B. BEDD method: Segmentation based on Texture Features

As discussed above, for the WCE images, their texture information associated with organs of the patients is able to be applied to further segment the clips. In this paper, we proposed a block edge directivity descriptor (BEDD) method to extract the texture features of the WCE images. The BEDD method can be described as follows: 1) each WCE frame is divided into 12 blocks and then for each block, the corresponding grayscale-block is computed; 2) the edge directivity descriptor (EDD) technique is employed to each grayscale-block to extract the block edge information at four directions. The EDD generates a 4-dimensional texture feature vector (TFV) for each grayscale-block; 3) the 48-dimensional texture feature vector, namely TFV_{48}, can be formed for each WCE frame by putting the texture feature vectors of 12 blocks into one vector. The main concerns for extracting TFV_{48} as the texture feature lies in: 1) blocks in one frame hold the texture location information, 2) EDD is an effective technique using the image edge information to represent the texture directional distribution [6], which has the ability to distinguish the images having small texture difference.

More specifically, in BEDD method proposed, the vertical and horizontal Sobel operators (V-Sobel and H-Sobel) are used. In principle, Sobel operator is a classic first order edge detection operator and is able to detect the edge information of the image. Therefore, the convolutions of the block with the Sobel operators generate the vertical and horizontal grayscale edge information matrixes, which can be used to provide the 4-dimensional texture feature vector accordingly **[7].** This proposed BEDD texture feature vector extraction process is shown in Fig. 2.

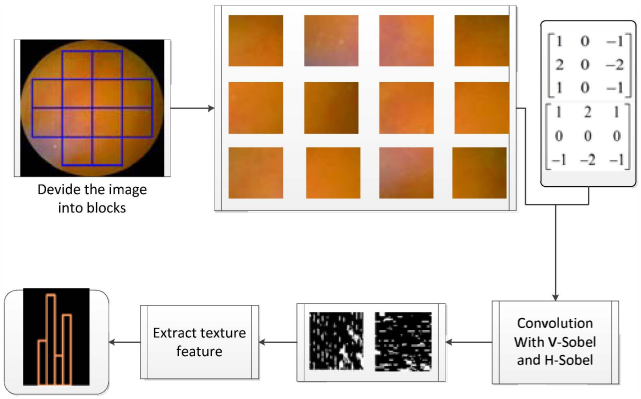

Fig 2 BEDD-based texture feature extraction process

Based on the clips segmented by the color feature vector described in II-A, we will segment each clip into several sub-clips making use of the TFV_{48}.

The frame difference (FD) between the frame f_i and f_j is defined as follows:

$$D(f_i,f_j)=\sum_{k=1}^{48}\left(1-\frac{|TFV_{48}^{i}(k)-TFV_{48}^{j}(k)|}{TFV_{48}^{i}(k)+TFV_{48}^{j}(k)}\right) \quad (4)$$

where TFV_{48} is the 48-dimensional texture feature vector extracted by the BEDD method presented above. The FD defined in (4) is used to segment the clips as follows: set f_1 as the first sub-clip; when the FD between f_j and f_{j+1} (j=2,3,4,…) is smaller than the threshold T_t (T_t is set experimentally), which means the similarity between f_j and f_{j+1} is poor, then one more sub-clip is added, and the f_{j+1} is put into the added new sub-clip. The process is repeated till the end of the clip. As a result, the frames segmented in one sub-clip should have similar color and texture properties. It is reasonable to randomly select any one frame in each sub-clip as the most

representative frame (MRF). From the derivation shown above, the proposed method has mainly involved in extracting color and texture features sequentially, which can be viewed as a cascade color and texture based MRF extraction algorithm, named as the CCT-MRFE algorithm.

For presentation clarity, the CCT-MRFE is summarized as:

1) For each frame in the WCE video, H-, S- images can be computed by (1) and (2), respectively;
2) The CFV can be computed from the H-histogram and S-histogram for each frame accordingly;
3) Segment the whole WCE video into different clips using CFV computed in 2);
4) Compute the 48-dimensional feature vector (TFV_{48}) of each frame using the BEDD method proposed in II-B.
5) Each clip obtained in 3) is segmented into sub-clips by using TFV_{48} computed in 4);
6) Randomly select one of the frames as the MRF in each sub-clip to form the final selected video frames.

Our preliminary results of the proposed CCT-MRFE algorithm show that the selection of the MRF affects the final result of the WCE video summary. Random selection of the MRF in each sub-clip possibly leads to get an undesired MRF, which only represents the minority in this sub-clip. In this case, the majority information will be lost. To solve this problem, in the following section, we will propose a new method to effectively extract the MRF for each sub-clip.

III. Most Representative Frame Extraction Based On Relation Matrix Rank

In this section, we are going to explore an effective method to extract the MRF for each sub-clip segmented in II-B.

It is noted that, in high-dimensional Non-Euclidean space, the space structure relationship between vectors is hard to be described [8]. The relation matrix is one of the effective solutions to describe the distribution information of vectors, which can be used to measure the space structure of the vectors. According to relation matrix theory [9], the larger rank of the relation matrix means the associated vector has higher confidence to represent the remaining vectors in the space. This is the rationale behind the following development of the relation matrix rank based MRF extraction method.

Let's assume that the whole WCE video ($\boldsymbol{V}$) is segmented into L non-overlapping sub-clips, which can be written as:

$$\boldsymbol{V} = \bigcup_{k=1\ldots L} \boldsymbol{S}_k \text{ where } \boldsymbol{S}_k = \{f_{ik}, i = 1,\ldots N\} \tag{5}$$

where $\boldsymbol{S}_k$ represents the k-th sub-clip and N is the number of frames in the $\boldsymbol{S}_k$. As discussed above, the texture feature vector TFV_{48} for each frame in $\boldsymbol{S}_k$ can be computed, and the difference between frame i and j can be calculated by (4) denoted as $D(f_{ik}, f_{jk})$. A relation matrix $\boldsymbol{M}_k$ of $\boldsymbol{S}_k$ is defined as:

$$\boldsymbol{M}_k = \begin{bmatrix} D(f_{1k}, f_{1k}) & D(f_{1k}, f_{2k}) & \cdots & D(f_{1k}, f_{Nk}) \\ D(f_{2k}, f_{1k}) & D(f_{2k}, f_{2k}) & \cdots & D(f_{2k}, f_{Nk}) \\ \vdots & \vdots & \vdots & \vdots \\ D(f_{Nk}, f_{1k}) & D(f_{Nk}, f_{2k}) & \cdots & D(f_{Nk}, f_{Nk}) \end{bmatrix} \tag{6}$$

where k is the sub-clip index (k=1,…,L), i, j is the frame index in the sub-clip (i,j=1,…,N). Obviously, $\boldsymbol{M}_k$ is in size of $N \times N$.

Moreover, we define a $N \times 1$ representative rank vector $\boldsymbol{R}_k$ as $\boldsymbol{R}_k = [r_1, \ldots, r_N]^T$ for the k-th sub-clip. So we can see that the value of r_i is the degree of accuracy that the frame f_i can be the representative in the sub-clip $\boldsymbol{S}_k$, which is in the range of (0, 1). The larger value of r_i gives higher confidence for the frame f_i to be the MRF in this sub-clip. Let j=0, the rank vector $\boldsymbol{R}_k(0)$ is the initial representative rank vector and its r_i can be set as $1/N$. The MRF is selected by the following iterative process:

$$\boldsymbol{R}_k(l+1) = \boldsymbol{M}_k \boldsymbol{R}_k(l), l = 0,1,2,\ldots \tag{7}$$

where $\boldsymbol{R}_k(l+1) = \left[r_1^{l+1}, \ldots, r_K^{l+1} \right]^T$ and $\boldsymbol{M}_k$ is defined in (6). The iteration loop in (7) will stop when the distance $\|\boldsymbol{R}_k(l_F+1) - \boldsymbol{R}_k(l_F)\| < \delta$. Then we get the $\boldsymbol{R}_k(l_F)$ as the result of the iteration and its maximum value can be determined, denoted as $r_i(l_F)$. Since the Relation Matrix Rank (RMR) is well suited to the MRF selection, the associated f_i of $r_i(l_F)$ corresponds to the MRF of the sub-clip accurately. Analysis shows that the proposed RMR-based MRF selection method has low computation complexity.

For notation clarity, the proposed WCE video summary algorithm using the RMR based MRF selection method is named as the CCT-MRFE-RMR algorithm.

IV. Experimental Results

To evaluate the performance of the proposed WCE video summary methods, several experiments have been carried out in this section. In our experiments, one recorded WCE video of a real patient has been used. The whole WCE video consists of 49084 frames in size of 480×480. The performance of the proposed methods is evaluated by the same standard fidelity and compression ratio (CR) used in [10], which are defined as follows:

$$\text{Fidelity} = N_{rep} / N_T \tag{8}$$

$$\text{CR} = 1 - N_{MRFs} / N_T \tag{9}$$

In (8), N_{rep} is the number of frames that can be represented by MRFs, N_T is the total frame number of the WCE video. In (9), N_{MRFs} is the number of the MRFs determined by the proposed algorithms. From the definitions given in (8) and (9), it is noted that a good WCE video summary algorithm should give both high fidelity and high compression ratio (CR).

Experiment 1: In this experiment, we select ten groups from the whole WCE video randomly, each group contains 1000 frames. The segmentation threshold T_t is set to be 5 for this experiment. The CCT-MRFE and CCT-MRFE-RMR algorithms have been evaluated. The experimental results are shown in TABLE 1, where N_{MRFs} is the number of MRFs. Fidelity-1 and Fidelity-2 are the fidelity results for CCT-MRFE and CCT-MRFE-RMR algorithms, respectively. From the results shown in TABLE 1, it is clear to see that the

proposed WCE video summary methods are able to achieve the average compression ratio of 77.9% with a high fidelity more than 95%. We also noted that the fidelity of the CCT-MRFE-RMR algorithm is higher than that of the CCT-MRFE algorithm. This verifies the effectiveness of the proposed RMR based MRF selection method.

TABLE 1 Performance comparison

group	N_{MRFs}	CR	Fidelity-1	Fidelity-2
1	304	69.6%	98.2%	99.5%
2	243	75.7%	96.1%	97.5%
3	246	75.4%	99.3%	99.7%
4	198	80.2%	97.6%	98.4%
5	102	89.8%	97.1%	99.5%
6	277	72.3%	97.8%	98.5%
7	175	82.5%	96.3%	97.4%
8	299	70.1%	98.9%	99.1%
9	267	73.3%	96.5%	97.7%
10	101	89.9%	96.7%	97.2%
Averag	221. 2	77.9%	97.6%	98.4%

Experiment 2: In this experiment, we will evaluate the influence of the segmentation threshold T_t on the performance of the CCT-MRFE-RMR algorithm. The WCE video and the experiment setup are the same as those in Experiment 1 except that T_t is set to be 3, 5 and 7. The compression ratio and the fidelity versus the group number are shown in Fig. 4 and Fig. 5, respectively. In Fig. 4., the red line with star, the green line with square and the blue line with circle indicates the results by using T_t = 3, 5, and 7, respectively.

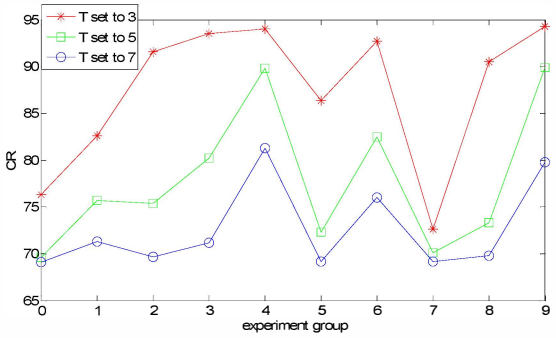

Fig 3 Compression ratio with different T_t

From Fig 3, we can see that the CR performance with T_t =3 is best among compared to other T_t values. However, from the red line with star, it is also clear to see that the dynamic range of the CR is largest, which indicates the weaker robustness of CCT-MRFE-RMR algorithm compared with CR results using bigger thresholds. From Fig. 4, we can see that the fidelity will decline rapidly when the threshold T_t gets smaller. The results in Fig. 3 and Fig. 4 suggest that the CR and the fidelity is conflict in selecting the threshold T_t, therefore, a tradeoff should be considered. From the simulation results, the threshold can be selected as 5 for this testing WCE video.

Experiment 3: In this experiment, we will evaluate the performance of the proposed CCT-MRFE-RMR algorithm by using the whole WCE video with 49084 frames. The threshold T_t is set to be 5 according to the results from Experiment 2. The number of the extracted MRFs is 7533 frames, which gives the total compression ratio as 84.65% for the whole testing WCE video.

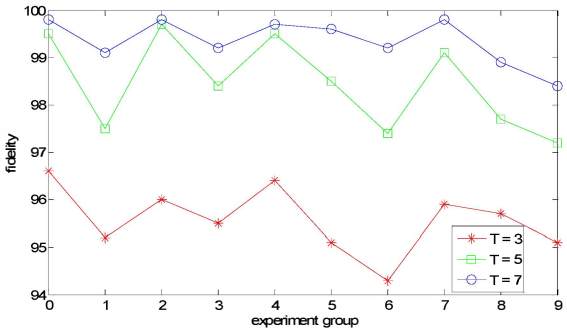

Fig 4 Fidelity result with different segment threshold T

V. CONCLUSION

This paper works on the WCE video summary technique. With the analysis of the WCE images, the color and texture features of the WCE images have been investigated to develop the cascade most representative frame extraction algorithms, named as CCT-MRFE algorithm. By proposing a new relation matrix rank based MRF selection method, an advanced CCT-MRFE-RMR algorithm has been proposed to further improve the video summary performance. Several experiments have been carried out. Testing results show that the proposed CCT-MRFE-RMR algorithm presents good WCE video summary ability. For this testing WCE video (with 49084 frames), it achieves 84.65% compression ratio at the high fidelity more than 95%.

ACKNOWLEDGMENT

This work was supported by the research project funded by Shenzhen ZiFu Technology Ltd. and the authors would thank Shenzhen ZiFu Technology Ltd. for providing us the WCE testing videos.

REFERENCES

[1] P. Swain, "Wireless Capsule Endoscopy," *Gut*, vol.52, ppiv48-iv50, 2003.

[2] D. G. Adeler, C J. Gostout, "Wirless Capsule Endoscopy," *Hospital Physician*, pp.14-22, May 2003.

[3] Ba T.Truong,Svetha Venkatesh. "Video abstract: A systematic review and classification,"*ACM Trans. Multimedia Comput. Commum. Appl.*, Vol.3, No.1.2007.

[4] D. K. Iakovidis, S. Tsevas, D. Maroulis, and A. Polydorou, "Unsupervised summarisation of capsule endoscopy video," *2008. the 4th International IEEE Conference on Intelligent Systems*,pp.3-15 - 3-20, 6-8 Sept. 2008.

[5] Max Q.-H Meng,Qian Zhao,"A Strategy to Abstract WCE Video Clips Based on LDA,"*International Conference on Robotics and Automation*,pp.454-459,May 2011.

[6] Davis L S, Mitiche A, "Edge Detection in Textures," *Computer Graphics and Image Processing*, 12:25-39,1980.

[7] Tomita F, Tsuji S. "Computer Analysis of Visual Textures", London: Kluwer, 1990.

[8] G. H. Golub, C. E. V. Loan, "Matrix Computations", The Johns Hopkins University Press, Baltimore, 1996.

[9] T. H. Haveliwala, S. D. Kamvar, "The Second Eigenvalue of Google Matrix", Stanford University Technical Report, 2003.

[10] C. Gianluigi, S. Raimondo, "An Innovative Algorithm for Key Frame Extraction in Video Summarization," *Journal of Real-Time Image Processing*, vol. 1,pp.69-88,2006.