

Received November 3, 2020, accepted December 3, 2020, date of publication December 14, 2020, date of current version January 25, 2021.

Digital Object Identifier 10.1109/ACCESS.2020.3044759

Summarization of Wireless Capsule Endoscopy Video Using Deep Feature Matching and Motion Analysis

B. SUSHMA  AND **P. APARNA**, (Senior Member, IEEE)

Department of Electronics and Communication Engineering, National Institute of Technology Karnataka, Surathkal 575025, India

Corresponding author: B. Sushma (sushmabg.177ec012@nitk.edu.in)


This work was supported by the Technical Education Quality Improvement Programme (TEQIP) of Government of India.

ABSTRACT Conventional Wireless capsule endoscopy (WCE) video summary generation techniques apprehend an image by extracting hand crafted features, which are not essentially sufficient to encapsulate the semantic similarity of endoscopic images. Use of supervised methods for extraction of deep features from an image need an enormous amount of accurate labelled data for training process. To solve this, we use an unsupervised learning method to extract features using convolutional auto encoder. Furthermore, WCE images are classified into similar and dissimilar pairs using fixed threshold derived through large number of experiments. Finally, keyframe extraction method based on motion analysis is used to derive a structured summary of WCE video. Proposed method achieves an average F-measure of 91.1% with compression ratio of 83.12%. The results indicate that the proposed method is more efficient compared to existing WCE video summarization techniques.

INDEX TERMS Autoencoder, convolutional neural network, deep learning, image similarity, keyframe extraction, video summarization, wireless capsule endoscopy.

I. INTRODUCTION

Wireless capsule endoscopy (WCE) is a non-invasive medical imaging procedure used to screen the entire gastrointestinal (GI) tract in order to detect various GI diseases [1]. WCE is considered as a robust diagnostic tool available for the analysis of GI diseases. When a patient swallows the capsule, it starts propelling through GI tract by peristalsis action and capture video frames of various parts of GI tract. The capsule capture images at the rate of 3 to 6 frames per second for over 8 hours and acquires around 90000-180000 frames [2]. The capsule travels at a very slow speed of about 0.16-1 mm/s and captures 2-12 frames for every 1mm of its travelling distance [3]. Slow movement results in huge number of redundant frames with high structural similarity. A physician has to invest a lot of time or appoint an assistant to inspect these huge number of frames and summarize the endoscopy video by eliminating redundant frames. The major disadvantage associated in manual summarizing is a chance of eliminating some of the frames

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad Zia Ur Rahman .

with lesion symptoms while inspecting thousands of images. Other methods work with the detection of lesions which includes tumours, ulcers, polyps and Crohn's disease. A few approaches are proposed for detection of lymphangiectasias, celiac disease and hookworms. All these methods deal with detection of only one or two type of abnormalities. Majority of the frames with other abnormalities still needs to be manually reviewed by the gastro-enterologist. To overcome all these drawbacks, developing an algorithm to generate video summary without missing frames with sensitive information is very significant. WCE video summarization tool allows the physician to get a quick glimpse of overall content in video and presence of possible abnormalities. Summarized video consisting of only keyframes is reviewed by the physician. Any frames with sensitive information is found, physician can always refer to the adjacent frames in the original video.

Problem of video summarization (VS) can be described as selecting a small batch of frames from the video stream consisting of large set of video frames that describe the whole content of original video. VS is a technique of parsing the sequence of video frames V into a shot set ψ and

extracting the set of keyframes ϕ [4]. Most of the state-of-the-art VS methods use three main common steps [5]:

- Feature extraction from each frame and latent space representation of extracted features.
- Temporal segmentation of video into shots. Each shot consists group of sequential frames with certain similar features.
- Finally, set of frames called as keyframes are extracted from each shot which describes the entire content of the shot.

Many VS techniques for WCE use fusion of extracted color, texture and shape features for shot segmentation [6]–[9]. These methods mainly use distance between information entropies of consecutive frame features with a threshold for segmenting video into several shots. Clustering techniques such as adaptive K-means and affinity propagation are adopted to extract keyframes. All these methods employ hand crafted features such as histogram oriented gradient (HOG) features, histogram based on Hue Saturation Value (HSV) colour space, Gray level co-occurrence matrix (GLCM) based texture features. All these feature extraction methods represent only low level attributes of a frame and fails to represent the high level semantic similarity between two consecutive frames [10]. In WCE video, colour and texture content varies very little from one frame to next consecutive frame. Therefore, colour and texture features are not adequate to detect significant changes between two successive frames [11]. Geodesic and Euclidean distance are used to remove the redundant frames from WCE video in [12]. Frames are considered as members of vector space to extract the set of keyframes called as orthogonal vectors using the computed distance. A sparse dictionary based method is used in [13] to select the more representative frames by fusing change in content information and gaze. Methods described above results in inappropriate summaries due to semantic gap. This poses a lot of practical challenges, when accuracy is an important criterion for medical diagnosis. All the constraints in the above methods lead the researchers to develop a supervised learning method using Siamese neural network (SNN) with linear SVM classifier [14]. SNN is capable of automatically learning semantic features in higher level required for discriminating the endoscopic images into similar or dissimilar pairs. But the modelling of network weights of the SNN depends on the robust labelled training data and needs physician's assistance for labelling. In another work, [15] singular value decomposition (SVD) method is used on both hand crafted and deep features to select keyframes. SVD is applied on a frame cluster and keyframes pertaining to a particular cluster are extracted. This will not capture more similarity in frames resulting due to slow movement of the capsule in some parts of GI.

To address all the above issues, unsupervised learning approach using convolutional autoencoders is proposed to extract features of endoscopic images. Many research outcomes have shown that unsupervised feature extraction of medical images lead to significant improvement compared

to conventional convolutional neural network [16]–[18]. Motivated by SNN, in the proposed work deep unsupervised image feature extraction network consisting of convolutional autoencoder neural network (CANN) is used. Euclidean distance computed between features extracted from a pair of GI tract images is used to classify the images as similar or dissimilar pairs. WCE video stream is segmented into different shots based on similarity measure. Finally, keyframes are extracted from each shot to remove redundant frames based on motion profile obtained by inter-frame motion energy and direction. In WCE the change in each frame is due to movement of the capsule. Therefore, it is possible to extract keyframes which covers the entire WCE video space of a shot with the help of motion analysis. Also, recent endoscopic VS techniques have shown efficiency in employing motion characteristics between consecutive frames for key frame extraction [11], [19].

The remaining sections of the paper are outlined as follows. In Section II proposed WCE VS technique is described. Section III discusses implementation considerations and performance of the proposed method for WCE video sequences. A conclusion comments of the paper are provided in Section IV.

II. PROPOSED METHOD

WCE video consists of an ordered set of consecutive frames represented as $V = I_1, I_2, I_3, \dots, I_m$, where I_i denotes i^{th} frame of the video and m is the number of frames in video. During WCE, capsule moves at a slow pace along the GI tract due to peristalsis action and captures a video sequence V which consists images of various parts of GI tract. Slow movement of the capsule in some parts of the GI tract such as small intestine results in small or no changes from frame to frame. In oesophagus, WCE video exhibits large changes from frame to frame due to fast movement of capsule. Therefore, a set of frames ϕ called as keyframes can be found which summarizes V by eliminating redundant frames. The task of finding ϕ given V involves the following function.

$$\left\{ I_V^{\phi_1}, I_V^{\phi_2}, \dots, I_V^{\phi_J} \right\} = \arg \min_{\phi_J} \{ D(\phi, V) | 1 \leq \phi_J \leq \kappa \}. \quad (1)$$

where D is measure of dissimilarity representing the criterion of VS [20]. Proposed WCE VS method for constructing ϕ from V is shown in Fig. 1.

A. CONVOLUTIONAL AUTOENCODER NEURAL NETWORK (CANN) FOR FEATURE EXTRACTION

CANN consists of an encoder and decoder networks. Encoder of the CANN generates high level feature map of the input by using several convolution and max-pooling layers. Decoder reconstructs the input from the feature map by using strided transposed convolution layers. The proposed CANN is designed based on the autoencoder network described in [21] and its architecture is shown in Fig. 2. Proposed feature extraction approach utilizes the power of convolutional filtering to train CANN in unsupervised way.

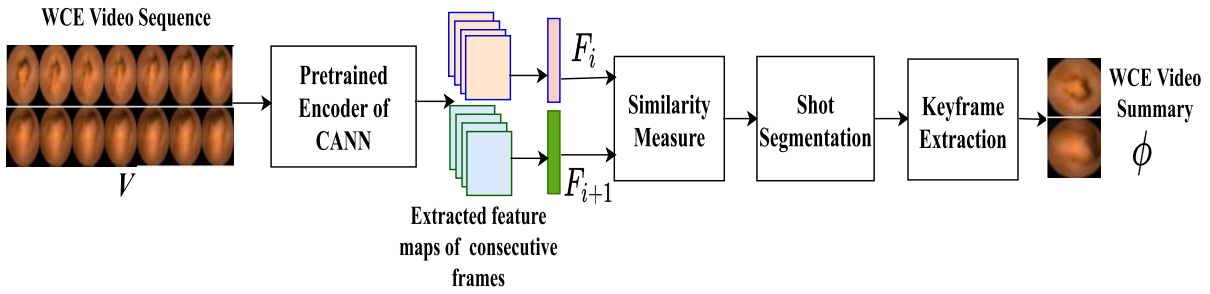


FIGURE 1. Proposed WCE video summarization method; F_i and F_{i+1} are the feature vectors of i^{th} and $(i + 1)^{th}$ sequential frames.

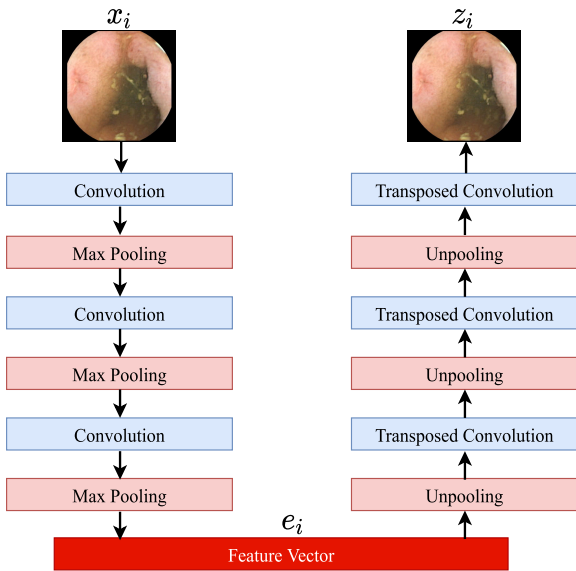


FIGURE 2. Convolutional autoencoder architecture showing encoder and decoder networks for extracting feature vector in endoscopic images.

The important characteristic of CANN is encoder-decoder neural network, which is trained for extracting high-level feature vector. Encoder consists three convolution layers and three max pooling layers. Decoder contains three transposed convolution layers followed by unpooling layers. Adding more layers will make the CANN more deeper and will improve the reconstruction of the input images at the decoder. But this increases the complexity of the model. CANN with three layers is capable of extracting the high level features from the image and decoder can reconstruct the image from the extracted features using three layers. The achieved reconstruction quality is sufficient for discriminating the pair of images into similar and dissimilar pairs at the encoder. The final goal of the CANN is to find feature vector for each input image by minimizing the mean squared error (MSE) between input and output over all image samples. The details of the encoder and decoder layer parameters are given in Table. 1.

Each convolution layer uses a non-linear activation function called Scaled exponential Linear Unit (SELU) instead of Rectified Linear Units (ReLU) used in other Convolutional networks. SELU activation function is close to zero mean and

unit variance. When propagated through multiple network layers, SELU automatically converge towards zero mean and unit variance. All these self normalizing parameters of SELU makes learning highly robust in network with many layers and utilizes strong regularization schemes [22]. For an input image matrix x_i , the encoder network computes encoder output e_i using

$$e_i = \sigma(x_i * f^n + b) \tag{2}$$

where σ denotes SELU activation function, $*$ represents 2D convolution operation, f^n is n^{th} convolutional filter kernel and b denotes encoder bias. The decoder reconstructs the encoded output using

$$z_i = \sigma(e_i * \tilde{f}^n + \tilde{b}) \tag{3}$$

where z_i is the reconstruction of the i^{th} input x_i , \tilde{f}^n is the n^{th} transposed convolutional filter and \tilde{b} is the bias of the decoder. Unsupervised training of the CANN aims to minimize the cost function given by:

$$J(\theta) = \sum_{i=1}^m (x_i - z_i)^2 \tag{4}$$

The gradients are calculated using the cost function given in (4) and the network parameters are optimized through stochastic gradient descent (SGD) to minimize the reconstruction loss. Autoencoder training is based on the work in [23]. Fig. 3 shows the training and testing performance of the CANN for 50 epochs. Similarity between the two consecutive frames is decided based on features extracted from the frames. To extract the features of an input image in an unsupervised method, both encoder and decoder networks are trained together. Input image is reconstructed by the decoder using encoder extracted features. Level of feature extraction is decided based on the reconstructed quality of images at the decoder. After the encoder is trained to extract the high level features, the decoder part of the CANN is removed and only the encoder is retained. Encoded features and reconstructed images of the CANN along with the input images are shown in Fig. 4.

B. SIMILARITY ESTIMATION

Two consecutive frames in WCE video sequence is considered as an image pair. For any input image I_i to the CANN,

TABLE 1. Layer parameters of convolutional autoencoder.

Layer	Type	Number of maps	Kernel Size	Output
1	Convolution	20	5x5	20x256x256
2	Max Pooling	20	2x2	20x128x128
3	Convolution	50	3x3	50x128x128
4	Max Pooling	50	2x2	50x64x64
5	Convolution	64	3x3	64x64x64
6	Max Pooling	64	2x2	64x32x32
7	Unpooling	64	2x2	64x64x64
8	Transposed Convolution	64	3x3	64x64x64
9	Unpooling	50	2x2	50x128x128
10	Transposed Convolution	50	3x3	50x128x128
11	Unpooling	20	2x2	20x256x256
12	Transposed Convolution	20	5x5	20x256x256

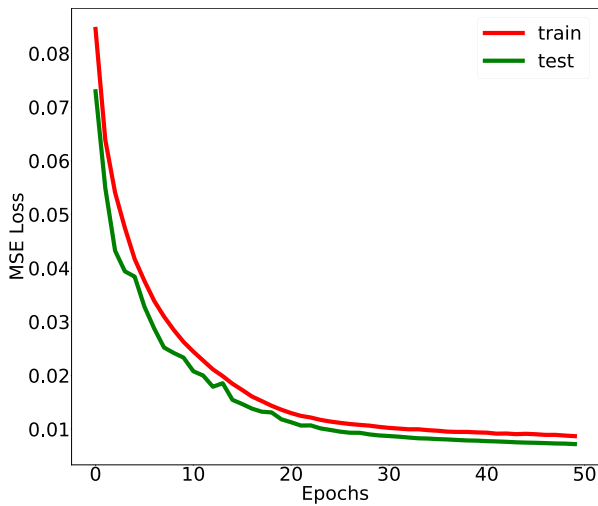


FIGURE 3. Training and testing losses for 50 epochs of the CANN.

the corresponding extracted feature Fea_i is generated as,

$$Fea_i = G(I_i, W) \tag{5}$$

where $G(\cdot)$ is a non-linear mapping function of encoder and W is the network parameters of the encoder part of the CANN. Euclidean distance between features of the image pair is computed and classified as similar or dissimilar pair based on the fixed threshold. To learn the threshold, Euclidean distance between 20000 consecutive WCE image pairs in feature space is computed and the observations made are: i) Euclidean distance varies between around 0 to 270 ii) Similar pair of images have distance close to 0 iii) Dissimilar pair of images have larger distance close to 270 iv) Images with few dissimilar patches have distance approximately equal to 20. Based on all the above observations and suggestions from gastroenterologist, a threshold of 20 is fixed for classification. Losing frames with significant lesions can be avoided by selecting small threshold, despite small threshold results in few number of frames in each shot. Euclidean distance for similarity judgement is calculated as:

$$Dis_{i,i+1} = ||(Fea_i - Fea_{i+1})||_2 \tag{6}$$

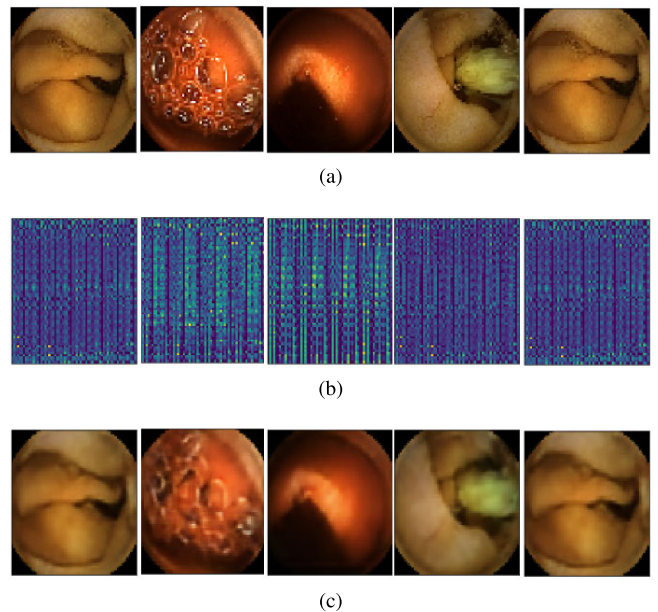


FIGURE 4. Visualization of extracted features and reconstructed images of CANN (a) Input images to CANN, (b) Features extracted by encoder network and (c) Decoded images by decoder network.

where $Dis_{i,i+1}$ is the Euclidean distance of the features Fea_i and Fea_{i+1} extracted from i^{th} and $i+1^{th}$ WCE frames respectively. Based on $Dis_{i,i+1}$, the image pair is considered as similar or dissimilar by (7). Similar and dissimilar pair of images are labelled as 1 and 0 respectively.

$$S_i = \begin{cases} 1, & Dis_{i,i+1} < 20 \\ 0, & Dis_{i,i+1} \geq 20 \end{cases} \tag{7}$$

C. SHOT SEGMENTATION

WCE video content contain similar content from frame to frame. Therefore, other shot detection methods which are proposed for multimedia cannot be used with WCE video [24]. In this article, the concept of video shot is defined as group of contiguous frames segmented based on similarity changes between two consecutive frames. Proposed shot segmentation method is shown in Fig. 5. Shot boundary is

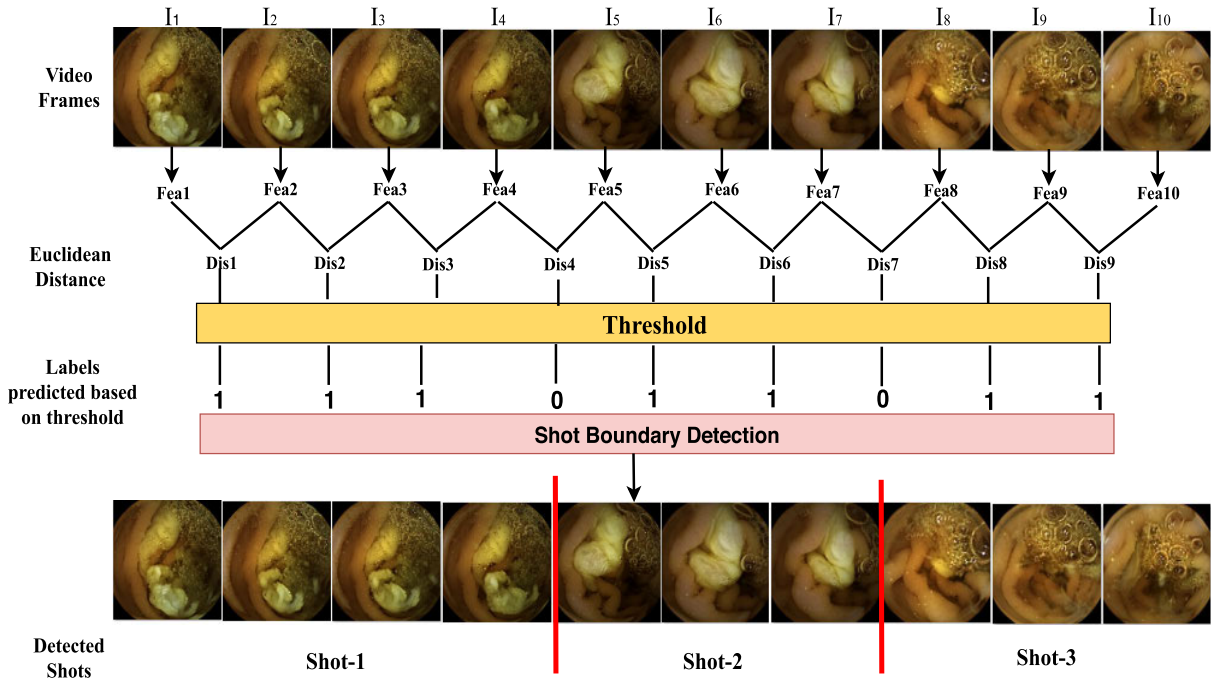


FIGURE 5. WCE video shot segmentation based on frame similarity.

detected when the similarity label is set to 0. WCE shot segmentation which incorporates frame matching, separates shots consisting of frames with high similarity.

D. KEYFRAME EXTRACTION

Video summarization of WCE video can be concluded with keyframe extraction in each shot. Frames in each shot has high similarity and consist a lot of redundancy. These redundant frames contributes very less or no information for covering the entire WCE video. Motion analysis between frames within a shot gives an idea of redundant frames [4]. If a pair of frames exhibit larger motion then the frames are likely to be considered as less redundant. Inherent intra-shot redundancy is reduced to retrieve keyframe representation by analysing capsule’s motion. Motion profile which constitutes motion score, motion direction and motion energy denoted as M_s, M_d, E_m respectively is derived from every shot before extracting keyframes.

First, the relative inter-frame motion score is estimated for a considered i^{th} shot S_{h_i} given as

$$M_{s_i} = \{M_{s_i}(n), n = 1, 2, \dots, n_i\}, \tag{8}$$

where n_i is the number of frames in S_{h_i} and $M_{s_i}(n)$ is intra-shot motion score between successive pair of frames. Motion score M_s for a frame pair (I, I') consisting of matched feature positions (X, X') , which is also the difference in average distance between (I, I') given by

$$M_s = \frac{1}{\alpha} \left\{ \sum_{m=1}^{\alpha} d(x_m, \hat{x}_m) - \sum_{m=1}^{\alpha} d(x'_m, \hat{x}'_m) \right\} \tag{9}$$

where \hat{x} is the X features center of mass computed by:

$$\hat{x} = \frac{1}{\alpha} \sum_{m=1}^{\alpha} x_m \tag{10}$$

where α is number of matched feature pairs detected [25] and each (x_m, x'_m) is a matched feature pair in (X, X') . Matched features in a pair of frames is as shown in Fig. 6. $d(x_m, \hat{x}_m)$ is the euclidean distance between x_m and \hat{x}_m .

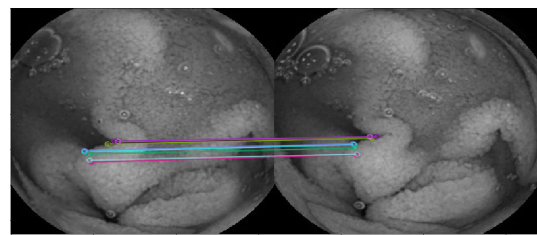


FIGURE 6. Matched features between pair of consecutive frames.

Next, motion direction sequence $M_{d_i} = \{M_{d_i}(n), n = 1, 2, \dots, n_i\}$ is computed using (11), which classifies motion direction as forward, backward and no-motion depending on capsule’s movement inside GI tract.

$$M_{d_i}(n) = \begin{cases} forward = 1, & \text{if } M_{s_i}(n) \leq TH_f \\ backward = -1, & \text{if } M_{s_i}(n) \geq TH_b \\ nomotion = 0 & \text{otherwise} \end{cases} \tag{11}$$

Considering wide range of motion analysis through a large number of experiments, threshold values TH_f and TH_b

are selected as -0.12 and 0.12 . Finally, Motion energy $E_{m_i} = \{E_{m_i}(n), n = 1, 2, \dots, n_i\}$ is computed for the features associated with each frame pair by:

$$E_m = \sum_{m=1}^{\alpha} ||x_m - x'_m||^2 \quad (12)$$

An example motion profile of a 40 frames shot from video sequence captured in stomach in which the frame sequence exhibits forward, backward and no-motion is shown in Fig. 7. Based on the obtained motion direction signal M_d of the capsule endoscope, an example shot is segmented into 8 different continuous runs consisting forward, backward and no-motion indicated as F, B and NM respectively. Frame with minimum motion energy is selected as keyframe in each segment. Another motion profile of a short shot of 14 frames of video captured in colon which exhibits only forward and no-motion is shown in Fig. 8. Keyframe selection in a video shot of 70 frames captured in a small bowel with motion profile is shown in Fig. 9. In all the frame shots the keyframe indicators are marked by green circles.

III. RESULTS & DISCUSSIONS

A. DATASETS

The proposed WCE VS method is evaluated using two datasets. Both the datasets are having frame resolution of 320×320 . The dataset-1 is a publicly available dataset called as KID. It consists of 3 WCE videos and the complete description of this dataset is available in [26] and [27]. Another WCE dataset employed in this work is the dataset-2 collected from department of gastroenterology, Manipal hospital, Bangalore, India. Video sequences consists frames of complete WCE examination from 20 different patients. Both the datasets are captured by Intramedic Miro-Cam WCE capsule. Around 50000 WCE frames are resized to a resolution of 256×256 , captured at different location of the GI tract from different patients is used for training CANN. Batchwise training is performed using mini-batch size of 32 samples and the number of epochs used for each batch is 50. For evaluating the performance of the proposed technique, keyframes of around 20 video sequences including 3 video sequences of KID-dataset are located with the help of an expert gastroenterologist. Around 5000 similar and dissimilar pair of frames are identified to test the similarity judgement performance. These frame pairs and keyframes are used as ground-truth summary to compare the performance of the proposed technique with other WCE VS methods. The details on frame resolution, frame motion characteristics and GI organ at which the frames are acquired is given in Table. 2. Video sequences in KID dataset covers all the GI organs and exhibits organ dependent motion characteristics. The results of the proposed method and other methods are computed using an Intel core i5-7200 2.5GHz CPU, 8GB RAM and NVIDIA GeForce 940MX GPU.

B. EVALUATION PARAMETERS

The performance of the proposed method is evaluated by F-score and compression ratio (CR) on 3 video sequences of KID-dataset and 4 video sequences captured in different parts of GI tract. F-score is computed using (13) which is a function of precision (p) and recall (r) computed using (14) and (15) respectively.

$$F\text{-score} = \frac{2rp}{r+p} \quad (13)$$

$$p = \frac{TP}{TP + FP} \quad (14)$$

$$r = \frac{TP}{TP + FN} \quad (15)$$

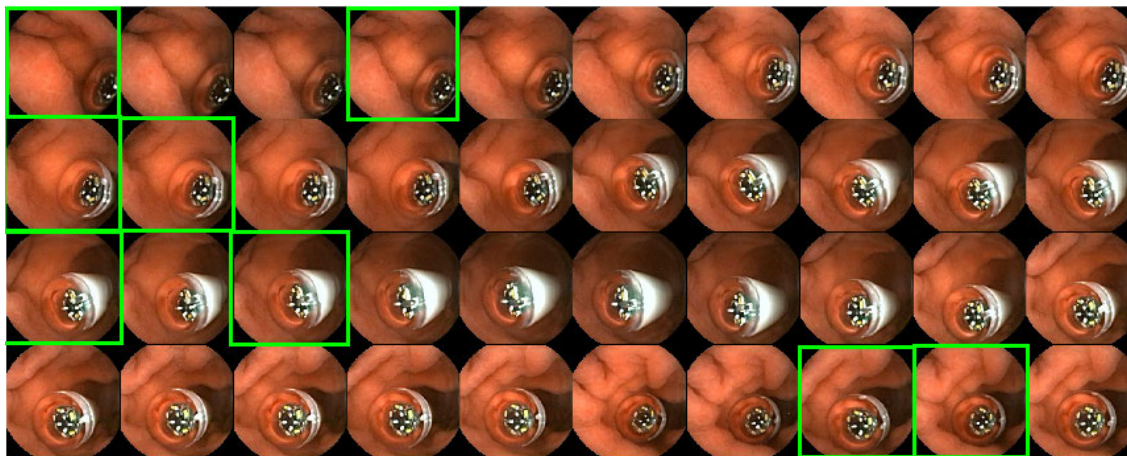
TP (True-positive) is the number of correct matches between keyframes extracted from proposed method and ground-truth summary. FN (False-negative) is number of frames which are in the result but not present in ground-truth summary. FP (False-positive) is number of frames in the ground-truth but not in result. Compression ratio (CR) is calculated using (16), where N_k is number of WCE keyframes extracted using proposed method and N_t is total number of keyframes in a video sequence. T_c is the time required to generate video summary of the considered video sequences.

$$CR = 1 - \frac{N_k}{N_t} \quad (16)$$

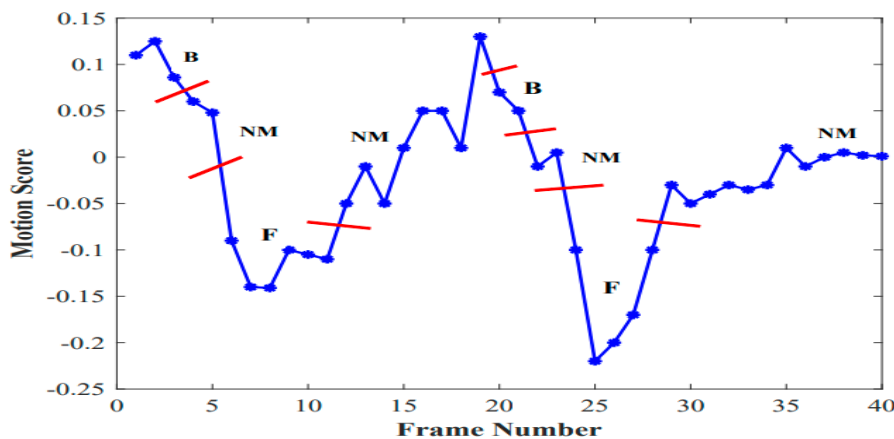
C. PERFORMANCE COMPARISON

Proposed method is compared with the other methods which involves different feature extraction, shot segmentation and key frame extraction methods. Methods with which the proposed method is compared are discussed below:

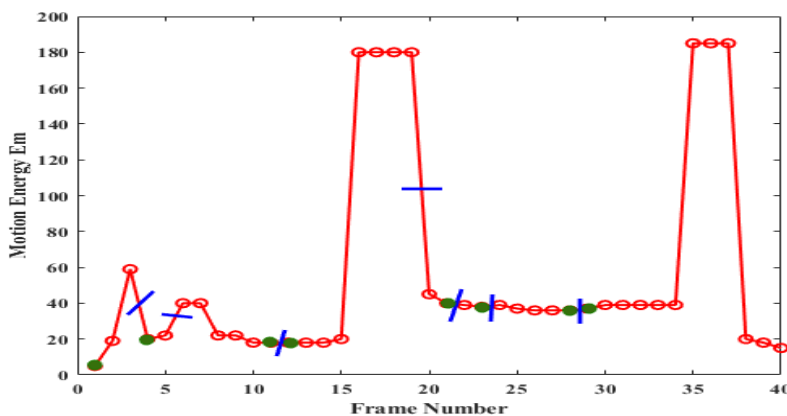
- HSV-KMC: WCE images exhibits different mucosal feature characteristics. Color is one of the significant feature. GI organ vary in color features for different organs. These color features are extracted in Hue Saturation Value (HSV) color space [28], since the information associated with H component is more indicative in representing the differences in WCE images. Color feature vector is extracted by using histogram of H and S. Shot is detected when the consecutive pair of frames are having different color feature vector [9]. In each shot the key frames are extracted by using K-means clustering (KMC) method [29], [30].
- CTS-KMC: In this method, fusion of color, texture and shape (CTS) features are considered for shot detection. Color feature vector is created in HSV color space. Local binary pattern (LBP) algorithm is used to extract texture features [31]. Shape features are represented using HoG [32]. Entropy of extracted features for each frame is used for segmenting the video into different shots [8]. KMC algorithm is used for key frames extraction.
- SIFT-KMC: Scale-invariant feature transform (SIFT) algorithm [33] is used to detect key feature points in an image using Difference of Gaussian (DoG) operator. Number of inlier matched features between pair



(a)



(b)



(c)

FIGURE 7. Keyframe selection of a 40 frame shot in video sequence captured in stomach based on motion profile. (a) Keyframes. (b) Motion signal partitioned into segments. (c) Motion energy signal.

of frames is used to detect a shot. More feature matches between the corresponding pair of frames is considered as more similarity between the frames. A shot is detected when the feature matches are less than a set threshold. Key frames are extracted using KMC.

- SIFT-MA: This method uses SIFT for feature extraction and number of features matches to detect a shot. Key frames are extracted based on motion analysis (MA) as proposed in this method.
- SURF-KMC: Speeded Up Robust Features (SURF) method [34] is a common feature extraction method in

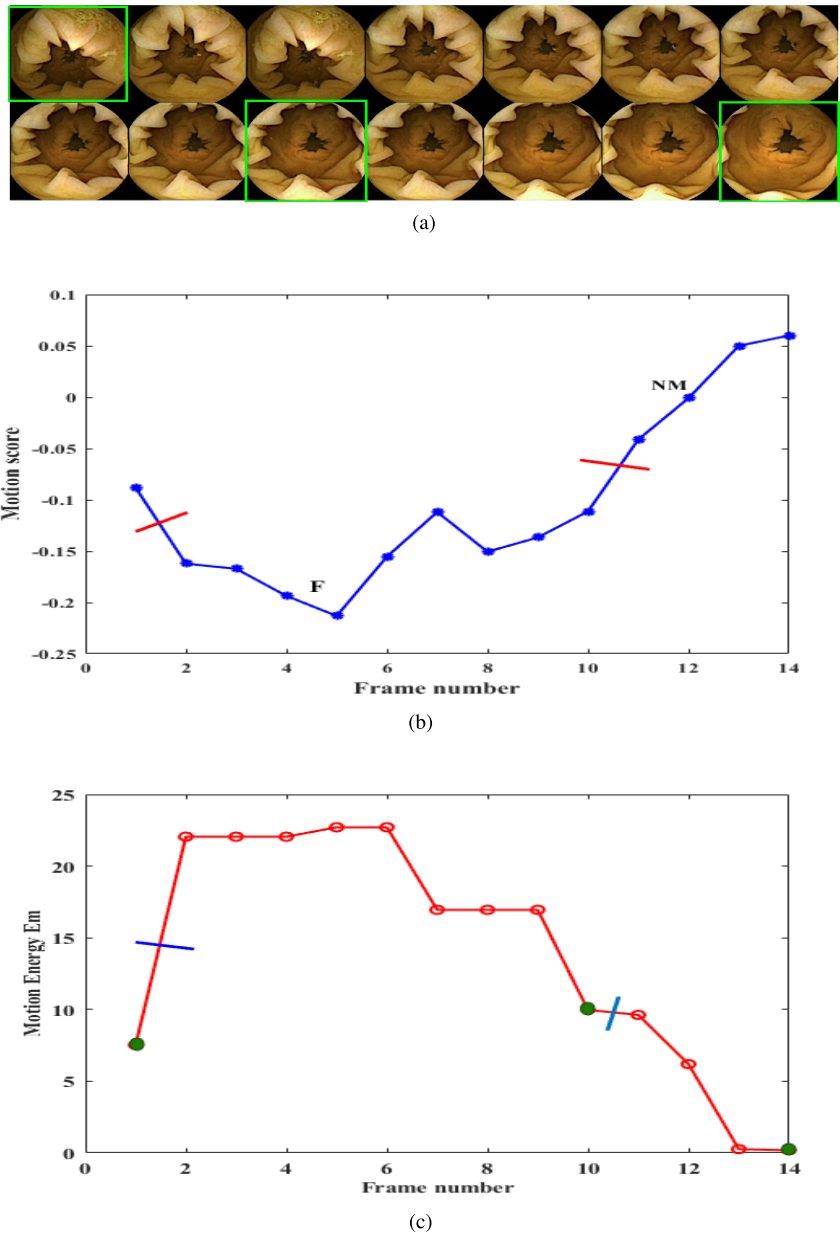


FIGURE 8. Keyframe selection of a 14 frame shot in colon video based on motion profile. (a) Keyframes. (b) Motion signal partitioned into segments. (c) Motion energy signal.

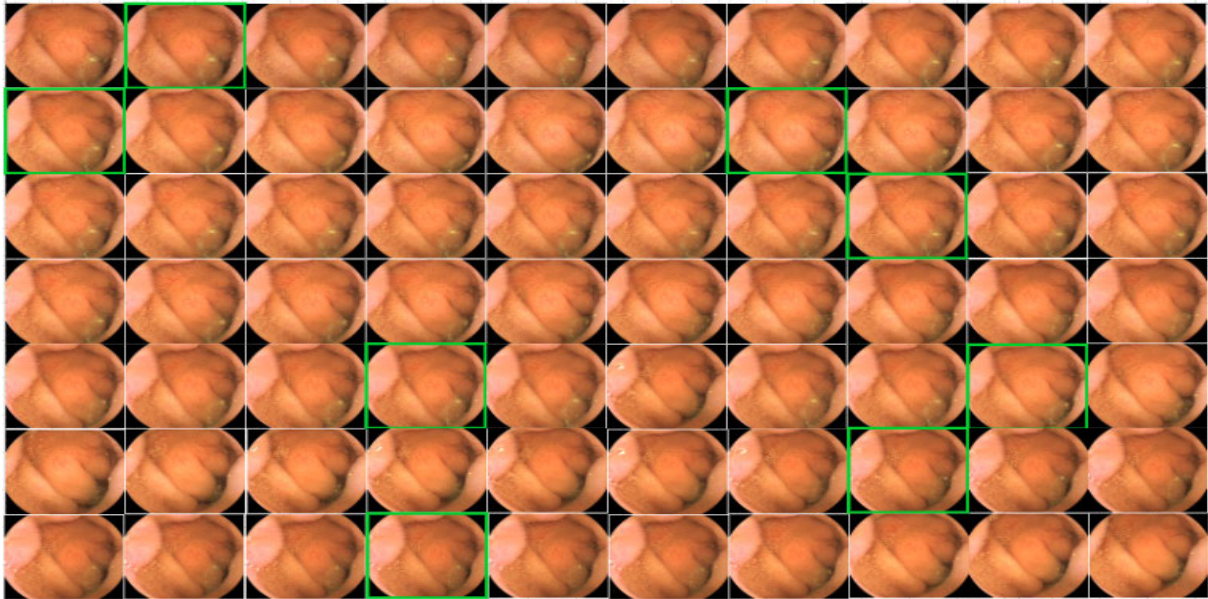
computer vision. Features are extracted using SURF and feature matches are used to detect a shot. Keyframes are extracted using KMC.

- SURF-MA: This method uses SURF for feature extraction and number of features matches to detect a shot. Key frames are extracted based on motion analysis. In the above SIFT and SURF based methods, matched feature points are retrieved between the pair of consecutive frames. The ratio of number of matched features to total number of features detected in both the frames is used to detect the shot. If the ratio is less than 0.15, it is considered as the two frames are in different shots.

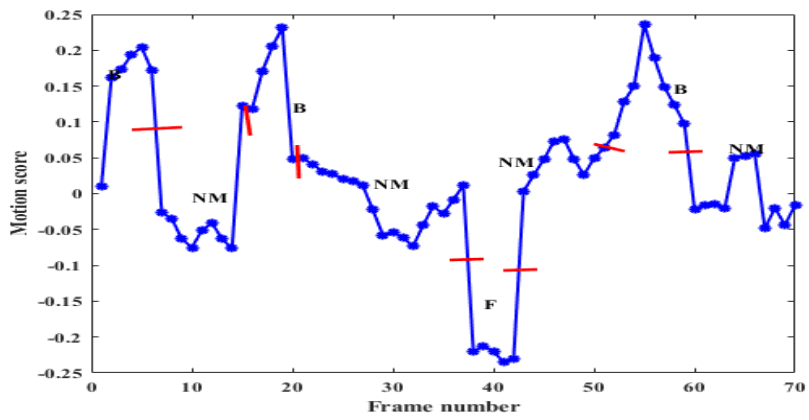
The comparison results for F-score on both the datasets is shown in Table. 3 and Table. 4. The CR comparison results are given in Table. 5.

It is very critical to achieve high accuracy in medical image analysis. The proposed method achieves high accuracy and high compression performance compared to other works. This indicates that it can eliminate redundant frames by extracting few keyframes which preserve informative frames.

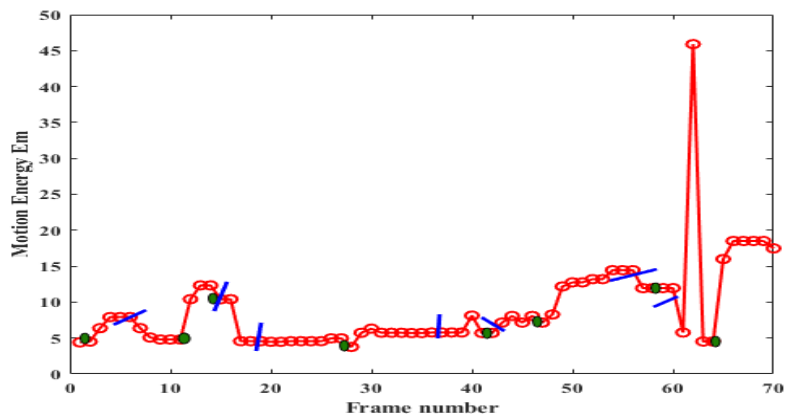
In the proposed method, WCE video is summarized based on two significant steps: shot identification and keyframe extraction. Shots are detected based on the frame feature matching. More similar features are required for



(a)



(b)

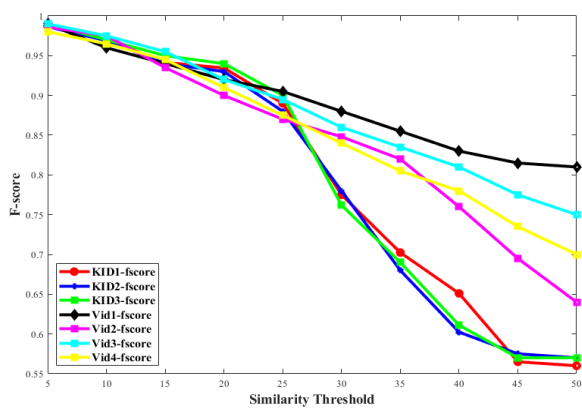


(c)

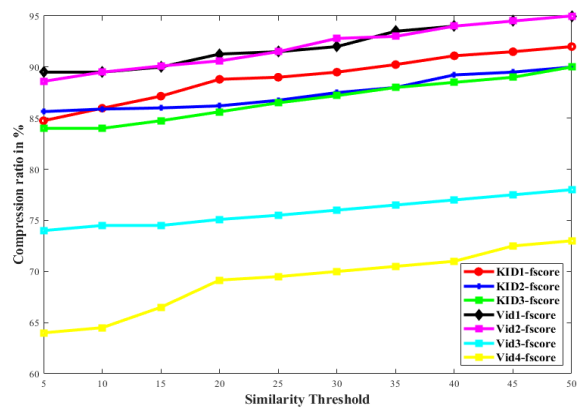
FIGURE 9. Visualization of a small bowel video shot with motion profile. (a) Keyframes (b) Partitioning of Motion signal. (b) Keyframe extraction based on motion energy.

TABLE 2. Test video sequences.

Test Video Sequence	Frame Resolution	Captured GI organ	Motion Type	Video length in frames	Capsule	Source
Video-1	320x320	Small Intestine	No or very less motion	5922	Intromedic Mirocam	Manipal Hospitals, Bangalore.
Video-2	320x320	Colon	Moderate to fast motion	3000		
Video-3	320x320	Esophagus	Fast motion	390		
Video-4	320x320	Stomach	irregular motion	450		
KID-1	320 x320	All GI organs	All types of motion	65000	MDSS Reserch Group [26]	
KID-2	320 x320	All GI organs	(irregular, fast, slow	62000		
KID-3	320 x320	All GI organs	and no motion)	62500		

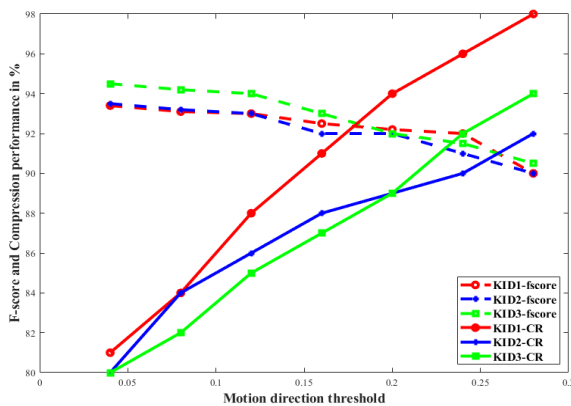


(a)

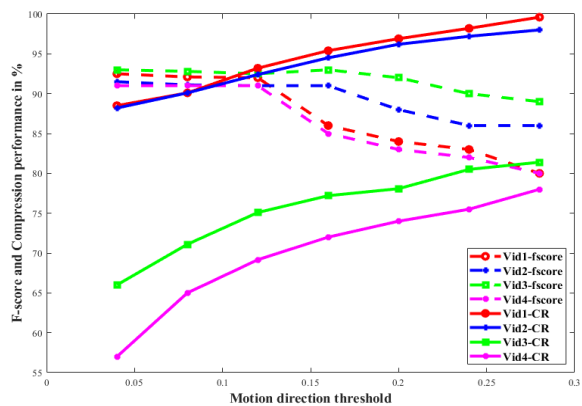


(b)

FIGURE 10. Performance test on similarity threshold. (a) F-score test on similarity threshold (b) Compression ratio test on similarity threshold.



(a)



(b)

FIGURE 11. F-score and compression performance for different motion direction thresholds. (a) KID-dataset (b) Dataset-2.

frame matching. The frame similarity threshold has direct influence on F-score as shown in Fig. 10a. Each plot for a specific test video sequence indicates the performance

measure variation according the change in similarity threshold. Also it can be observed from Fig. 10b that similarity threshold has less impact on compression performance.

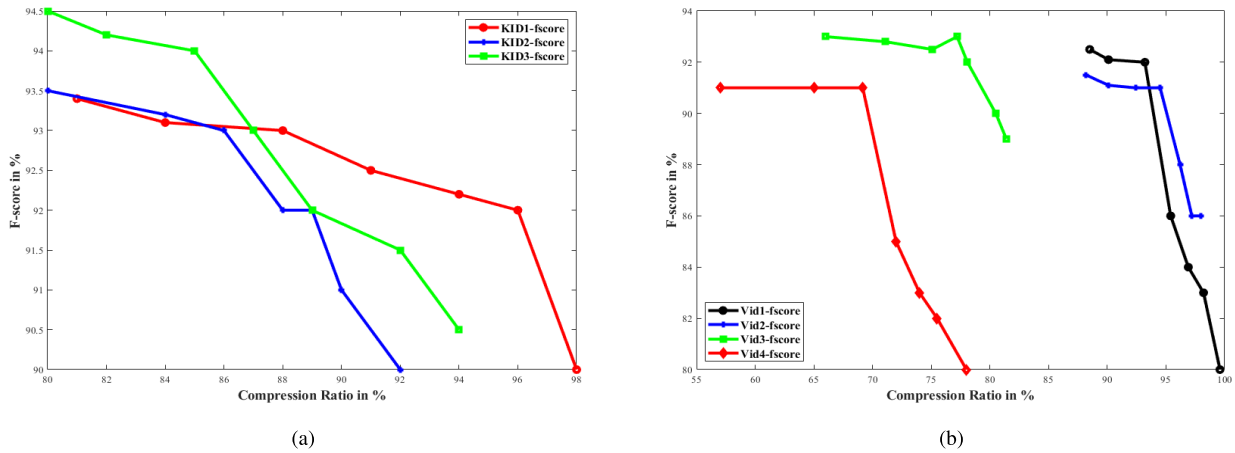


FIGURE 12. Comparison of summarization performance in-terms of F-score with compression ratio on (a) KID-dataset (b) Dataset-2.

TABLE 3. Comparison of Recall, Precision and F-score values of the proposed method with other methods on KID dataset.

Parameters	Test Video	HSV-KMC	CTS-KMC	SIFT-KMC	SIFT-MA	SURF-KMC	SURF-MA	Proposed Method
Recall	KID-1	0.57	0.79	0.82	0.80	0.78	0.78	0.94
	KID-2	0.54	0.74	0.80	0.79	0.73	0.73	0.92
	KID-3	0.51	0.72	0.81	0.83	0.69	0.76	0.93
	Average	0.54	0.75	0.81	0.81	0.73	0.75	0.93
Precision	KID-1	0.58	0.78	0.53	0.61	0.69	0.81	0.92
	KID-2	0.61	0.76	0.55	0.63	0.79	0.91	0.94
	KID-3	0.55	0.81	0.59	0.69	0.68	0.88	0.95
	Average	0.58	0.78	0.56	0.64	0.72	0.86	0.94
F-Score	KID-1	0.57	0.78	0.65	0.69	0.73	0.79	0.93
	KID-2	0.57	0.75	0.65	0.70	0.75	0.81	0.93
	KID-3	0.53	0.76	0.68	0.75	0.68	0.81	0.94
	Average	0.56	0.76	0.66	0.72	0.72	0.80	0.93

Setting high similarity threshold can tend to reject frames with significant lesions. Therefore, it is necessary to choose low threshold value. In shot detection a threshold of 20 is set for similarity estimation which detects even a small significant change between pair of frames and avoids loss of informative frames. Video shot is partitioned into different motion segments based on a threshold and a keyframe is extracted in each segment. Construction of motion profile proved to be strong over thresholds TH_f and TH_b . Low threshold values -0.12 and 0.12 are chosen to detect even a weak motion between the frames, because TH_f and TH_b has direct impact on F-score and CR. F-score and CR for datasets considered in this work for different motion direction thresholds is shown in Fig. 11. From the performance graph, it can be observed that F-score is maximum at 0.12. Video shot is partitioned into less number of segments at higher threshold and less keyframes are selected. Therefore, CR increases for the larger

thresholds. But accuracy is very important and threshold at which high accuracy is achieved is considered in the work. As shown in Fig. 12, summarization performance in-terms of F-score decreases as CR increases. Each plot for a particular test video sequence indicates how the F-score varies as CR varies. It can be observed from Fig. 12b that video with slow motion (Video-1 in dataset-2) has high F-score of around 93% at high CR of 95%. Video with fast motion (Video-4 in dataset-2) achieves high F-score of 91% with 70% CR. The accuracy drastically drops as the CR increases which is directly impacted by increase in TH_f and TH_b . Larger TH_f and TH_b gives high CR as more number of frame pairs are considered as no motion frames and this results in less motion segments. This will lead to an excessive rejection of significant frames and affects summarization performance interms of accuracy. The results clearly indicate that the proposed method is potential with consistent performance

TABLE 4. Comparison of Recall, Precision and F-score values of the proposed method with other methods on Dataset-2.

Parameters	Test Video	HSV-KMC	CTS-KMC	SIFT-KMC	SIFT-MA	SURF-KMC	SURF-MA	Proposed method
Recall	Video-1	0.76	0.89	0.72	0.70	0.78	0.82	0.91
	Video-2	0.72	0.85	0.74	0.71	0.70	0.75	0.89
	Video-3	0.69	0.78	0.76	0.72	0.72	0.78	0.90
	Video-4	0.64	0.72	0.73	0.69	0.73	0.77	0.92
	Average	0.70	0.81	0.73	0.70	0.73	0.78	0.90
Precision	Video-1	0.72	0.84	0.89	0.94	0.87	0.89	0.94
	Video-2	0.58	0.79	0.86	0.90	0.80	0.88	0.92
	Video-3	0.51	0.79	0.89	0.82	0.76	0.82	0.95
	Video-4	0.53	0.81	0.84	0.81	0.82	0.89	0.91
	Average	0.59	0.80	0.80	0.86	0.81	0.87	0.93
F-Score	Video-1	0.74	0.86	0.79	0.80	0.82	0.85	0.92
	Video-2	0.64	0.82	0.79	0.79	0.74	0.80	0.90
	Video-3	0.58	0.78	0.75	0.76	0.73	0.80	0.92
	Video-4	0.59	0.76	0.73	0.74	0.77	0.82	0.91
	Average	0.64	0.80	0.76	0.77	0.77	0.82	0.91

TABLE 5. Comparison of the proposed method with other methods interms of F-score (FS) and compression ratio (CR) results in %.

Test Video	HSV-KMC		CTS-KMC		SIFT-KMC		SIFT-MA		SURF-KMC		SURF-MA		Proposed Method	
	FS	CR	FS	CR	FS	CR	FS	CR	FS	CR	FS	CR	FS	CR
KID-1	57.24	74.60	78.37	71.5	65.34	62.30	68.87	63.4	73.30	64.30	79.21	63.80	92.78	88.80
KID-2	56.98	72.30	74.72	68.6	64.92	70.75	70.04	69.57	75.16	71.60	81.37	72.20	93.06	86.20
KID-3	53.43	70.96	76.25	66.3	68.45	70.36	74.86	68.78	68.12	72.43	80.91	73.62	94.21	85.62
Video-1	74.16	89.50	86.19	86.2	78.79	88.31	80.12	85.31	81.91	87.35	84.79	85.27	91.91	91.27
Video-2	64.38	77.90	81.88	74.2	78.88	83.57	78.76	81.5	74.19	84.20	79.79	82.60	89.92	90.60
Video-3	57.78	75.7	78.36	73.3	75.28	81.94	76.21	79.26	73.04	81.56	80.03	81.09	92.06	75.09
Video-4	58.84	82.5	76.49	80.5	73.14	81.59	74.13	79.94	76.81	79.32	81.89	75.16	91.14	69.16
Average	60.40	73.61	78.89	74.32	72.11	77.01	74.71	75.39	76.64	77.25	81.14	76.24	92.15	83.82

with greater than 90% accuracy achieved for video sequences of different motion characteristics.

IV. CONCLUSION

In this article a framework to obtain summary of WCE video content is proposed. Manual reviewing of the huge amount of frames captured during the WCE procedure is a challenging task for the physician interms of time and accurate diagnosis. In this work an efficient computer aided WCE video summarization method is presented. convolutional autoencoder is trained to extract high level deep features, which are suitable for segmenting video into shots. This method avoids laborious procedure of labelling large number of WCE image pairs. The change in two successive frames of WCE video is due to capsule motion, which varies in different parts of GI tract. Keyframes are extracted based on the motion profile constructed for each video shot. This method eliminates frames with very small temporal difference and retains candidate keyframes which covers sufficient WCE video.

Similarity and motion detection thresholds have a key role in deciding the summarization performance interms of accuracy and compression ratio. Thresholds are set to get maximum accuracy in this work. With the set thresholds, proposed method achieves an average F-measure of 91.1% with compression ratio of 83.12%. Significant number of experiments prove the efficiency and potential of the proposed method. Improvement is required in several areas which accounts for future work. When a capsule travels in GI tract, the frames captured are completely or partially degraded due to poor illumination and obscured by secreted fluids. These frames are uninformative or partly informative. It is beneficial to detect and remove the uninformative frames to reduce the video content for the review. Restoration techniques for improving the quality of partially degraded frames warrant future work.

ACKNOWLEDGMENT

The authors are grateful to Dr. V. V. Raj (gastroenterologist) and his chief assistant for providing WCE videos.

REFERENCES

- [1] J.-D. Zeitoun, A. Chrysostalis, B. Terris, F. Prat, M. Gaudric, and S. Chaussade, "Portal hypertensive duodenal polyp: A case report," *World J. Gastroenterology*, vol. 13, no. 9, p. 1451, 2007.
- [2] P. Swain, "Wireless capsule endoscopy," *Gut*, vol. 52, no. 90004, pp. 48–50, Jun. 2003.
- [3] J. L. Toennies, G. Tortora, M. Simi, P. Valdastrì, and R. J. Webster, "Swallowable medical devices for diagnosis and surgery: The state of the art," *Proc. Inst. Mech. Eng., C, J. Mech. Eng. Sci.*, vol. 224, no. 7, pp. 1397–1414, Jul. 2010.
- [4] X. Zhu, A. K. Elmagarmid, X. Xue, L. Wu, and A. C. Catlin, "InsightVideo: Toward hierarchical video content organization for efficient browsing, summarization and retrieval," *IEEE Trans. Multimedia*, vol. 7, no. 4, pp. 648–666, Aug. 2005.
- [5] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, "Creating summaries from user videos," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, in Lecture Notes in Computer Science, vol. 8695. Cham, Switzerland: Springer, 2014, pp. 505–520.
- [6] Q. Zhao and M. Q.-H. Meng, "WCE video abstracting based on novel color and texture features," in *Proc. IEEE Int. Conf. Robot. Biomimetics*, Dec. 2011, pp. 455–459.
- [7] B. Li, M. Q.-H. Meng, and Q. Zhao, "Wireless capsule endoscopy video summary," in *Proc. IEEE Int. Conf. Robot. Biomimetics*, Dec. 2010, pp. 454–459.
- [8] Y. Yuan and M. Q.-H. Meng, "Hierarchical key frames extraction for WCE video," in *Proc. IEEE Int. Conf. Mechatronics Autom.*, Aug. 2013, pp. 225–229.
- [9] J. Sen Huo, Y. Xian Zou, and L. Li, "An advanced WCE video summary using relation matrix rank," in *Proc. IEEE-EMBS Int. Conf. Biomed. Health Informat.*, Jan. 2012, pp. 675–678.
- [10] G. Antipov, S.-A. Berrani, N. Ruchaud, and J.-L. Dugelay, "Learned vs. hand-crafted features for pedestrian gender recognition," in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 1263–1266.
- [11] M. J. Primus, K. Schoeffmann, and L. Boszormenyi, "Segmentation of recorded endoscopic videos by detecting significant motion changes," in *Proc. 11th Int. Workshop Content-Based Multimedia Indexing (CBMI)*, Jun. 2013, pp. 223–228.
- [12] D. K. Iakovidis, S. Tsevas, and A. Polydorou, "Reduction of capsule endoscopy reading times by unsupervised image mining," *Computerized Med. Imag. Graph.*, vol. 34, no. 6, pp. 471–478, Sep. 2010.
- [13] S. Wang, Y. Cong, J. Cao, Y. Yang, Y. Tang, H. Zhao, and H. Yu, "Scalable gastroscopic video summarization via similar-inhibition dictionary selection," *Artif. Intell. Med.*, vol. 66, pp. 1–13, Jan. 2016.
- [14] J. Chen, Y. Zou, and Y. Wang, "Wireless capsule endoscopy video summarization: A learning approach based on siamese neural network and support vector machine," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 1303–1308.
- [15] A. Biniaz, R. A. Zoroofi, and M. R. Sohrabi, "Automatic reduction of wireless capsule endoscopy reviewing time based on factorization analysis," *Biomed. Signal Process. Control*, vol. 59, May 2020, Art. no. 101897.
- [16] M. Kallenberg, K. Petersen, M. Nielsen, A. Y. Ng, P. Diao, C. Igel, C. M. Vachon, K. Holland, R. R. Winkel, N. Karssemeijer, and M. Lillholm, "Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1322–1331, May 2016.
- [17] D. Kumar, A. Wong, and D. A. Clausi, "Lung nodule classification using deep features in CT images," in *Proc. 12th Conf. Comput. Robot Vis.*, Jun. 2015, pp. 133–138.
- [18] M. Chen, X. Shi, Y. Zhang, D. Wu, and M. Guizani, "Deep features learning for medical image analysis with convolutional autoencoder neural network," *IEEE Trans. Big Data*, early access, Jun. 20, 2017, doi: 10.1109/TBDATA.2017.2717439.
- [19] U. von Öhsen, J. M. Marcinczak, A. F. M. Vélez, and R.-R. Grigat, "Keyframe selection for robust pose estimation in laparoscopic videos," *Proc. SPIE*, vol. 8316, Feb. 2012, Art. no. 83160Y.
- [20] B. T. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 3, no. 1, p. 3, Feb. 2007.
- [21] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *Proc. Int. Conf. Artif. Neural Netw. (ICANN)*, in Lecture Notes in Computer Science, vol. 6791. Berlin, Germany: Springer, 2011, pp. 52–59.
- [22] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 971–980.
- [23] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 153–160.
- [24] A. F. Smeaton, P. Over, and A. R. Doherty, "Video shot boundary detection: Seven years of TRECVID activity," *Comput. Vis. Image Understand.*, vol. 114, no. 4, pp. 411–418, Apr. 2010.
- [25] D. Sarghos, C.-I. Chen, C.-M. Tsai, Y.-F. Wang, and D. Koppel, "Feature detector and descriptor for medical images," *Proc. SPIE*, vol. 7259, Mar. 2009, Art. no. 72592Z.
- [26] (2017). *KID Dataset*. [Online]. Available: <https://mdss.uth.gr/datasets/endoscopy/kid/>
- [27] D. K. Iakovidis and A. Koulaouzidis, "Software for enhanced video capsule endoscopy: Challenges for essential progress," *Nature Rev. Gastroenterology Hepatology*, vol. 12, no. 3, p. 172, 2015.
- [28] G. Paschos, "Perceptually uniform color spaces for color texture analysis: An empirical evaluation," *IEEE Trans. Image Process.*, vol. 10, no. 6, pp. 932–937, Jun. 2001.
- [29] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 881–892, Jul. 2002.
- [30] Y. Zhuang, Y. Rui, T. S. Huang, and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," in *Proc. Int. Conf. Image ICIP*, Oct. 1998, pp. 866–870.
- [31] Z. Guo, L. Zhang, and D. Zhang, "A completed modeling of local binary pattern operator for texture classification," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1657–1663, Jun. 2010.
- [32] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 886–893.
- [33] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
- [34] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 346–359, Jun. 2008.



B. SUSHMA received the B.E. degree in electronics and communication engineering from Visvesvaraya Technological University, Belagavi, in 2002, and the M.Tech. degree in communications from the National Institute of Technology Karnataka, Surathkal, India, in 2006, where she is currently pursuing the Ph.D. degree. She has more than five years of industry experience and nine years of teaching experience. Her current research interests include multimedia signal processing, machine learning, and deep learning.



P. APARNA (Senior Member, IEEE) has been associated with NITK, Surathkal, since 2002, under various capacities, where she has been working as an Assistant Professor since 2008. Her research interests include biomedical signal processing, signal compression, computer architecture, and embedded systems. She has presented a number of research papers at various International conferences. She has published more than 25 research papers in various journals and conference proceedings. She is actively involved in research activities in the area of signal processing for ten years.

...