Video summarization approaches in wireless capsule endoscopy: A review

Vrushali Raut[1,*], *Reena Gunjan*[2]
[1]MIT School of Engineering, Pune, Maharashtra (India)
[2]MIT School of Engineering, Pune, Maharashtra (India).

Abstract. Wireless capsule endoscopy (WCE) is medical examination process for gastrointestinal tract (GIT). This noninvasive multi advantageous procedure can be made more popular by overcoming the problem of prolonged analysis time. Video summarization is a concise and meaningful representation of a video. Along with automated detection and segmentation methods, summarized video will serve as an additional source for confirming the analysis of WCE video before the final diagnosis without any dropouts. Nowadays, Internet of Things (IoT) environments are predominant in healthcare sectors. Considering limited resources of smart phones and long duration of WCE process, it is impractical to send all the WCE data to health-care centers or gastroenterologists. This paper reviews video summarization types and the techniques used for WCE video summarization by various researchers. Feature set selection, clustering methods and key frame selection techniques play important role in performance of summarization technique.

1Introduction

Initially capsule endoscopy was developed to see the inner area of small intestine, which is not easily reachable with the help of traditional endoscopy procedures. From 2001 onwards wireless capsule Endoscopy (WCE) was introduced for total gastrointestinal tract (GIT) inspection[1].

The WCE process involves swallowing of a camera loaded in a capsule which takes images of GIT during its movement through the entire tract. The images captured by the camera are then transmitted to the recording unit present on a belt worn by the patient[2]. These images are loaded in the work station (computer) in the form of video. The capsule is then expelled out from the body naturally after about seven to eight hours. During this total process, the number of images captured are significantly large (per patient 50,000 to 60,000) and with a diverse content, which makes the detection of gastro-intestinal (GI) abnormalities a challenging task. Inspecting a WCE video is laborious and time-consuming. An experienced doctor will generally take 60 to 90 minutes to evaluate the images of one patient. The workload of gastroenterologists can be reduced thereby increasing their efficiency by aiding the process with intelligent approaches in the diagnosis procedure. This way, the application of such approaches can contribute in speeding up the analysis of total WCE video consequently increasing clinical productivity and an overall cost reduction for healthcare systems is possible.
There is huge amount of redundancy in collected video. It is very difficult and time exhausting task for a medical practitioner to look for a peculiar abnormality from this large number of frames. Multiple image processing techniques like classification, detection and segmentation are proposed by researchers for analysis of WCE video. These techniques filter the video for specific images which may contains abnormality. However, in actual clinical practice, to avoid risk of missing something in the WCE examination, medical expert would always like to confirm the computer aided detection results. Acquiring summarized video can be a plausible solution for solving this problem. Video summarization process tries to abstract the vital occurrences, scenes, or items in a clip for providing easy interpretation of synopsis. In summarization process important steps are selection of features considering the application, different clustering methods and key frame selection. The later section gives information about different video summarization techniques based on multiple aspects and contribution by researchers in the field of WCE video summarization over past five years.

2Video Summarization

Video Summarization is the technique with which abstract view of entire video can be created and presented in very less amount of time as compared to original video.
Video summarization is analogous to summary creation of a video. Important aspects of video summarization are as follows:
- The created summary should be as short as possible
- Summarized video should contain the most important scenes and events from the video

*Corresponding author: vrushali.raut0210@gmail.com

- There should not be any redundancy

If an application needs processing of a video then summarized video will reduce processing time as well as computation complexity.

2.1 Static and dynamic

These are the two broad categories of summarization techniques. The static type summarized video contains set of key frames from the base video without consideration of time and sequence. The dynamic type summarized video contains the most significant, small portions of audio and video. They are also called as video skims and are similar to trailers of movies.

2.2 Summarization with low/high/multiple features

In the applications where quick response is needed and the scenes of video are not complex low level features does the work e.g. color, texture, motion etc. But use of only low level features may lead to loss of complex scenes having significant information.

In the applications where contents of the image are of critical importance, high level features contribute more in the task of summarization. Some of the high level features that plays major role in proper understanding of the contents of the video are object recognition, face detection, gesture detection, event and emotion detection etc. Summarization process with high level features is time consuming and computationally expensive.

Multiple features investigate the content from various aspects and utilize them for video summarization. The evaluation of summarization technique with multiple features is necessary considering the area of application, practicability and its cost efficiency.

2.3 Shot boundary based video summarization

A prolonged video can be segmented as per the shot or scenes.

- A scene in a video is a group of temporally and semantically associated class of elements of that video which communicates a clear meaning.
- A shot can be seen as a series of activities picked up by a single camera which does not encounter any major deviations in the visual data.

The shot boundary is located at sudden changes in the frames. Located shot boundaries can be used for summarization process. The process generally involves selection of key frame from each segmented shot. These key frames can be stitched later to form a summarized video. The algorithms used for such techniques need proper threshold value for execution of the method efficiently. If the threshold value chosen is very small, then there will be large number of segments that ultimately results in elongated summarized video. If the threshold value is a large number then there are chances of missing some important information from the video. So selection of threshold value is a critical procedure.

2.4 Clustering based Video Summarization (static)

In clustering procedure initially random frames are considered as cluster centres. Application of clustering algorithm gives considered number of clusters which contains frames with nearly similar contents. The cluster centers then can be revised and selected as representative key frame. There are many clustering algorithms like k means[3], guarded zone clustering, spectral clustering[4], MST (Minimum Spanning Tree) clustering[5] etc. In this procedure the representative key frames are not in the sequential order. The applications in which the contents are of much important than there occurrence sequence can prefer this methodology. Time constraint clustering algorithm to maintain the sequence of the video is proposed in [6].

2.5 Non- Clustering based Video Summarization

In this type of summarization, the representative frames are selected by using thresholding techniques. Sequential selection of key frames by using thresholding technique is presented in [7]. The procedure has following steps, 1) Calculate histogram of frames 2) Calculate sum of absolute differences of histograms 3) Selection of threshold 4) Selection of key frames: The frame which have the sum of absolute difference greater than the threshold value. The efficiency depends on the threshold value.

3Summarization in wireless capsule endoscopy

In literature WCE video summarization is done in two ways as shown in fig 1. The video can be summarized by eliminating the redundant frames and second way to summarize video is to select only those frames which have salient contents.

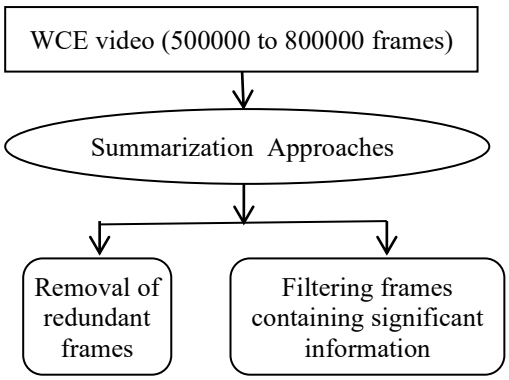

Fig. 1. Summarization approaches for WCE video.

Video summarization technique based on possibilistic clustering and feature weighting algorithm is proposed in [8] by Mohamed et al. The algorithm generates possibilistic membership which is used to detect the degree of similarity the WCE video frames. The frames are selected based on this membership function. The advantage of the algorithm is having an objective

function from each cluster which reduces optimization of possibilistic membership values and optimal feature weights. This optimization is achieved by updating the membership values, the feature weights and the centroids repetitively. The experimental result indicates that the possibilistic weights can identify noisy frames and provides superior description of the data. This method does not require determining number of clusters. The unique part in this methodology is cluster dependent feature subset weights.

Emam et al. proposed summarization based on adaptive feature extraction techniques [9]. As per the results LBP performance is superior as compared to statistical features. For similarity detection cosine similarity distance from Matlab is used. 90 % similarity threshold is applied for all feature extraction techniques.

Chen et al. addressed the problem of low level features that they cannot capture semantic similarity between WCE frames and used Siamese Neural Network (SNN)for the extracting high level semantic features[10]. This methodology requires pairs of similar and dissimilar WCE frames with labels. SNN training is done with contrastive loss function for extracting features with high level semantics. Linear SVM is used for classification based on similarity judgment. This process avoids the problem of setting correct threshold manually.

Adaptive redundant images elimination technique was proposed by Chen et al. in [11]. Similarity of features between consecutive WCE frames and temporal correlation is used for summary creation. HSV color histogram model and Gray Level Co-occurrence Matrix are used for generating color and texture features. W-parametric mean value threshold (WMVT) is developed which is data driven. The low level features similarity of consecutive WCE frames are compared with the threshold and then the matching images are grouped into the same clip. Adaptive K-means clustering algorithm is used to generate representative frames and to exclude redundant frames by using gradient dependent characteristic in each clip.

Hamza et al. proposed secure video summarization technique to aid diagnosis of WCE video in IoT (Internet of things) environment [12]. Recently the health care services started using IoT techniques for transmitting data to hospitals or medical practitioner. Paper proposes framework for transmission of WCE data considering summarization and encryption methods. In summarization part feature computation is done by using concept of integral image which is computed based on COC model. The features considered are multi-scale contrast, MI (moments of inertia) and curvature map. The saliency scores are calculated for all these features. The scores are then normalized in the range 0 to 1. Then the representative frames are uprooted using aggregated attention curve. This curve is generated by combining normalized scores.

Mohammed et al. in [13] used multiple features for colon video summarization. Methodology involves grouping of video frames based on power spectral density. Key frames are then selected for multiple feature extraction. Multiple features such as handcrafted (color and texture) and deep were extracted. Pre-trained GoogleNet model is used for extracting deep features. Next step is sparse coding of these features using a learned dictionary. SVM classifier is trained by using sparse coded features. The output is categorized as frames with salient content and frames without any information.

Figure 2 gives pictorial representation of key frame extraction process from redundant frames in WCE video. Process summary is - First WCE video is divided in to frames with proper frame rate then different feature extraction methods are used for optimal feature set selection. Key frame can be extracted by using clustering algorithm or thresholding techniques. The hurdles to overcome in this field are availability of dataset and standard evaluation parameters to validate the result for summarization. The dataset for training purpose should be developed with the help of guidance from gastroenterologist for annotating the images. Standard evaluation measures will ease comparison of summarization techniques based on their performances.

Table 1. Comparative analysis of related work

Year	Author	Features	Dataset	Performance metrics
2014	Mohamed et al. [8]	RGB Color Histogram HSV Color Moments Edge histogram	4 CE videos from 4 different patients. 75% for training and 25% for testing	Classification rate Relative cluster evaluation measures
2015	Emam et al. [9]	Color Histogram, Color Moment HSV, Color Moment RGB LBP Features, Statistical Features	10671 images	Reduction rate
2015	Chen et al.[10]	HSV color histogram Texture Features	One dataset = 500 continuous WCE frames (4000 data sets)	F-measure Recall Compression ratio (CR) Precision
2016	Chen et al.[11]	CNN as a feature extractor	Similar frames: 7875 pairs Dissimilar frames: 7986 pairs	Compressed ratio (CR) F-measure Computation time
2017	Mohammed et al.[13]	Mean color, power spectral density (PSD), hue and opponent histogram, LBP histogram, deep features	500 images	Average Accuracy Sensitivity Specificity

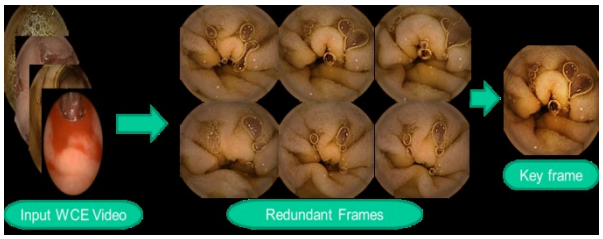

Fig. 2. Key frame extraction in WCE video

Source: Images adapted from [14]

4 Conclusion

This paper reviews types of video summarization techniques and particularly the methodologies used by various researchers to summarize WCE video in past five years. The aim of this paper is to survey various WCE video summarization techniques and provide a reference document which can be explored for further result improvement. Summarization techniques which are based on low level features are good for real time applications because of their speed and computationally inexpensiveness. The techniques which are based on high level features or user attention model are exclusively suitable for applications where the main constraint for summarization is accuracy rather than time. The various techniques have their own merits and demerits as per the application. The results are improved if semantic features are combined with visual descriptors. Video summarization field still demands more standard evaluation techniques to validate the results.

References

1. A. G. & P. S. Gavriel Iddan, Gavriel Meron, "Wireless capsule endoscopy," Nature, vol. **405**, no. May, pp. 417–418, (2000)
2. A. Wang et al., "TECHNOLOGY STATUS EVALUATION REPORT: *Wireless capsule endoscopy," Gastrointest. Endosc.*, vol. **63**, no. 4, pp. 539–45, (2013)
3. S. E. F. De Avila, A. P. B. Lopes, A. Da Luz, and A. De Albuquerque Araújo, "VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method," *Pattern Recognit. Lett.*, vol. **32**, no. 1, pp. 56–68, (2011)
4. A. Ioannidis, V. Chasanis, and A. Likas, "Weighted multi-view key-frame extraction," *Pattern Recognit. Lett*., vol. **72**, pp. 52–61, (2016)
5. R. Panda, S. K. Kuanar, and A. S. Chowdhury, "Scalable video summarization using skeleton graph and random walk," *Proc. - Int. Conf. Pattern Recognit.*, pp. 3481–3486, (2014)
6. J. L. Lai and Y. Yi, "Key frame extraction based on visual attention model," *J. Vis. Commun. Image Represent*., vol. **23**, no. 1, pp. 114–125, (2012)
7. C. V. Sheena and N. K. Narayanan, "Key-frame Extraction by Analysis of Histograms of Video Frames Using Statistical Methods," *Procedia Comput. Sci*., vol. **70**, pp. 36–40, (2015)
8. M. M. Ben Ismail, O. Bchir, and A. Z. Emam, "Endoscopy Video Summarization based on Multi-Modal Descriptors and Possibilistic Unsupervised Learning and Feature Subset Weighting," *Intell. Autom. Soft Comput*., (2014)
9. A. Z. Emam, Y. A. Ali, and M. M. Ben Ismail, "Adaptive features extraction for Capsule Endoscopy (CE) video summarization," in *Proceedings - International Conference on Computer Vision and Image Analysis Applications, ICCVIA 2015*, (2015)
10. J. Chen, Y. Zou, and Y. Wang, "Wireless capsule endoscopy video summarization: A learning approach based on Siamese neural network and support vector machine," *Proc. - Int. Conf. Pattern Recognit.*, pp. 1303–1308, (2016)
11. J. Chen, Y. Wang, and Y. X. Zou, "An adaptive redundant image elimination for Wireless Capsule Endoscopy review based on temporal correlation and color-texture feature similarity," *in International Conference on Digital Signal Processing, DSP*, (2015)
12. R. Hamza, K. Muhammad, Z. Lv, and F. Titouna, "Secure video summarization framework for personalized wireless capsule endoscopy," *Pervasive Mob. Comput.*, vol. **41**, pp. 436–450, (2017)
13. A. Mohammed, S. Yildirim, M. Pedersen, O. Hovde, and F. Cheikh, "Sparse Coded Handcrafted and Deep Features for Colon Capsule Video Summarization," *Proc. - IEEE Symp. Comput. Med. Syst*., vol. **2017-June**, pp. 728–733, (2017)
14. Koulaouzidis, D. K. Iakovidis, D. E. Yung, E. Rondonotti, U. Kopylov, J. N. Plevris, E. Toth, A. Eliakim, G. W. Johansson, W. Marlicz, and others, "KID Project: an internet-based digital video atlas of capsule endoscopy for research purposes," *Endoscopy International Open*, vol. **5**, no. 06, pp. E477–E483, (2017)

* Corresponding author: vrushali.raut0210@gmail.com