Scan to know paper details and
author's profile

Machine Vision based Unsupervised Summarization of Wireless Capsule Endoscopy Video

Madhura Prakash M & Krishnamurthy G N

ABSTRACT

Wireless capsule endoscopy is a medical procedure that uses a small, pill-sized capsule with a tiny camera inside to examine the inside of the digestive system. The capsule contains a camera that takes pictures as it moves through the digestive tract. The images are transmitted to a small recording device that the patient wears on a belt or shoulder strap. The recording device stores the images, which are later downloaded and examined by a doctor. The pill-sized camera records a video for eight to twelve hours. The purpose of this work is to investigate and put into practise a technique for summarising the capsule endoscopy footage. The method clusters the video frames in order to accomplish this by looking for local maxima in the neighbouring distance graph that is generated from the colour histograms of each video frame. The method then determines the centroid frame for each cluster and produces a key-frame for each cluster. Visualising the conclusion of this investigation is its second objective. This is accomplished by placing each cluster on a timeline according to the size and calibre of the keyframe. The summary of the capsule endoscopy video is generated as a sequence of keyframes. These keyframes are representative frames of the video. The work also proposes a technique to assess the quality and the quantity of the keyframes extracted.

Keywords: wireless capsule endoscopy, video summarization, clustering, key-frames.

Classification: NLM: WN 208.5.C1

Language: English

Great Britain Journals Press

LJP Copyright ID: 392881

London Journal of Medical and Health Research

Volume 23 | Issue 4 | Compilation 1.0

Machine Vision based Unsupervised Summarization of Wireless Capsule Endoscopy Video

Madhura Prakash M[α] & Krishnamurthy G N[σ]

ABSTRACT

Wireless capsule endoscopy is a medical procedure that uses a small, pill-sized capsule with a tiny camera inside to examine the inside of the digestive system. The capsule contains a camera that takes pictures as it moves through the digestive tract. The images are transmitted to a small recording device that the patient wears on a belt or shoulder strap. The recording device stores the images, which are later downloaded and examined by a doctor. The pill-sized camera records a video for eight to twelve hours. The purpose of this work is to investigate and put into practise a technique for summarising the capsule endoscopy footage. The method clusters the video frames in order to accomplish this by looking for local maxima in the neighbouring distance graph that is generated from the colour histograms of each video frame. The method then determines the centroid frame for each cluster and produces a key-frame for each cluster. Visualising the conclusion of this investigation is its second objective. This is accomplished by placing each cluster on a timeline according to the size and calibre of the keyframe. The summary of the capsule endoscopy video is generated as a sequence of keyframes. These keyframes are representative frames of the video. The work also proposes a technique to assess the quality and the quantity of the keyframes extracted.

Keywords: wireless capsule endoscopy, video summarization, clustering, key-frames.

Author α: Research Scholar, Dept. of CSE, BNM Institute of Technology, Bangalore.
σ: Principal, BNM Institute of Technology, Bangalore.

I. INTRODUCTION

Medical endoscopy has developed into an important technique for minimally invasive surgery in recent years. Examination on multiple body areas and for minimally invasive abdominal surgery, joints and other body areas. The term "endoscopy" is derived from the Greek and describes a minimally invasive method of "looking inside" the human body. This is accomplished by inserting a medical device called an endoscope into a hollow organ or body cavity. Depending on the part of the body, it is inserted through a natural body opening or through a small incision that acts as an artificial access. Additional incisions are required during surgery to insert various surgical instruments.

Endoscopy is a catch-all phrase for a wide range of quite varied medical procedures. [1] Endoscopy has numerous varieties, each of which has unique qualities. They can be categorised using a variety of standards, including

- Body part (e.g., abdomen, joints, gastroin-testinal tract, lungs, chest, nose)
- Specialty in medicine (e.g., general surgery, gastroenterology, orthopaedic surgery)
- Therapeutic vs. diagnostic focus.

Wireless Capsule Endoscopy (WCE) is a unique form of endoscopy procedure that falls under the diagnostic focus. The patient must swallow a tiny capsule with a built-in camera that transmits a huge number of frames over a prolonged period of time to an external receiver. The doctor then evaluates the footage once this process is complete. WCE is essential for small intestine exams in particular because neither gastroscopy or colonoscopy can access this part of the gastrointestinal tract.

Wireless capsule endoscopy is typically used to examine the small intestine, which is difficult to reach with traditional endoscopy procedures. It is often used to diagnose conditions such as Crohn's disease, celiac disease, and tumours or ulcers in the small intestine. The procedure is non-invasive and painless, and most people are able to go about their daily activities while wearing the recording device.

Capsule endoscopy is a medical diagnostic tool that records video of a patient's digestive tract. The method includes two devices. The first is a capsule with a camera and lights inside. The patient swallows this pill and sends its image to an external receiver. This external receiver is worn by the patient. Capsule endoscopy is used to capture images of the bowel that cannot be reached by conventional endoscopy methods. Capsule endoscopy is used to perform the following diagnostic procedures:

- Recognize inflammatory bowel illnesses; determine the cause of gastrointestinal bleeding.
- Identify cancer.
- Identify celiac illness.
- Look for polyps.
- After additional imaging examinations, perform follow-up testing.

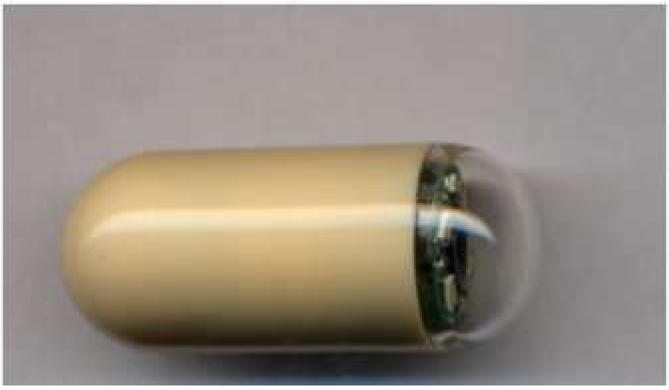

Figure 1: Image of an Endoscopy Capsule

Videos of endoscopic procedures are frequently several hours long. The issue of effective indexing and retrieval is brought on by the enormous amount of video data. It takes time to search through all of these films to find the one you need.

It is difficult to quickly extract the needed information from such a sizable video archive. Hence, methods are required to aid with the difficulty of managing video data. Video summarization is a fundamental method for handling video data.The sample frames from the capsule endoscopy procedure is depicted in Figure 2.

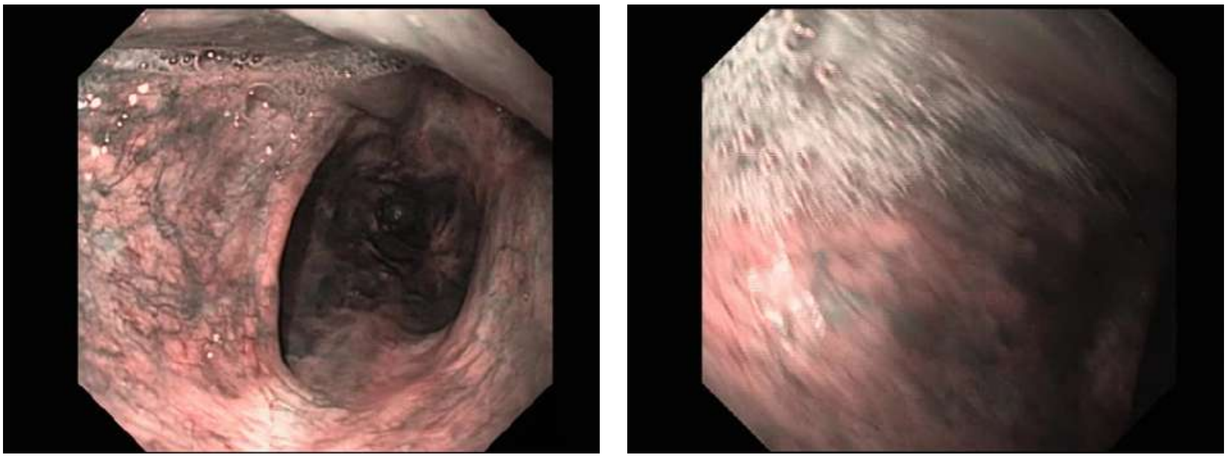

Figure 2: Sample Frames from Capsule Endoscopy Procedure

In order to extract the most important information, known as the keyframe of the video, video summarization attempts to limit the amount of duplicate data. It makes it possible for viewers to swiftly understand the key points of the video. It needs a thorough grasp of the video to produce a synopsis of it. Hence, it is hoped to create a framework that shows the viewer the useful components of video data by taking the information into account.

Video summary messages are typically run using two different approaches: static (keyframe-based) and dynamic (video hover-based) video analytics messages. The static video message contains a collection of a small but significant number of silent frames called keyframes, while the dynamic video message contains a collection of short important video clips. The video summary considers various characteristics such as representativeness, uniformity, static attention, temporal attention, and quality including hue, brightness, contrast, number of colours, edge distribution for keyframe selection. [2]

One of the major problems with capsule endoscopy procedures as mentioned above is that the gastro-intestinal video frames captured in the process are about 8 to 12 hours long. Reviewing the video frames is extremely time consuming for the physicians. The work's major objective is to shorten the time needed for reviewing videos of capsule endoscopy by pointing out the relevant portions of the video to the physician. To achieve this goal this, work aimed at generating a solution for extracting a strip of keyframes that can be presented as a summary of the original video.

With the intent of generating clusters that should be linked to the original video so that the user must be able to easily access the input video frames of every cluster, the two research questions addressed here are:

1. The process of summarising a lengthy capsule endoscopy video in a strip/timeline
2. The process of summarising groups of related frames by a single 'summary'

II. RELATED WORK

Due to the immense amount of video data being generated, there is an increased demand to analyse and summarise them. The work in the domain of video summarization has been carried out for a decade. For summarization, the videos will have to be sampled and divided into segments and shot boundaries. Since the medical videos can be short clippings or full-length videos with no shot-boundaries, extracting key frames is the most challenging task. The key frames can be extracted from the segments based on the several features of the video, either by supervised or unsupervised methods. The key frame extraction may consider low level global features of the video like colour, texture, edges, brightness and so on or the high-level semantics.

A technique for indexing and searching massive amounts of video data is called video summarization. To provide the user with a visual abstract of the video sequence, the approach outputs a brief summary of the video. A good video abstract aims to minimise noise while maximising the amount of information retained in the summary. Clustering techniques are frequently used for automatic video summarization, and they either extract a key-frame or a moving image for each cluster (video skims). In a clustering process, a cluster is a collection of objects that are more similar to one another than they are to the objects in the other clusters (groups). A good cluster will have minimal inter-cluster and intra-cluster variance.

The research study was focused on exploring the solutions available in literature for summarization of endoscopy procedures. The survey in this category consisted of perusing works from [3] to [12].

Ismail et al., [3] had proposed an unsupervised based approach for summarization of WCE videos. Here, the temporal descriptor as well as the colour and texture descriptors were used to represent each video frame. The possibilistic membership values and ideal feature weights inside each cluster were optimised by the authors using a probabilistic clustering and feature weighting technique with an objective function. A mean Jaccard coefficient of 15 had been attained using the technique. Chen et al., [4] had proposed a Siamese Neural Network (SNN) approach and Support Vector Machine (SVM) for summarization of WCE videos. In this approach SNN was used to map. Similar frame pairings were mapped closer using SNN, whereas dissimilar image pairs were mapped farther apart in the feature space.

Euclidean distance measure was computed in order to detect shot boundaries. AN F-measure of 84.75 % was achieved.

Emam et al., [5] had explored and investigated different feature extraction techniques, such as colour histogram, Local Binary Pattern (LBP), and statistical features, a pre-processing phase for WCE images. In this work a cosine similarity distance was computed to summarise the WCE videos and an average reduction rate of 92% was achieved. Lux et al., [6] had clustered similar frames based on low-level image features to extract key frames to summarise arthroscopy videos. Mehmood et al., [7] had proposed a cloud based binary classification model to classify the WCE frames as informative or non-informative.

In another work the authors Mehmood et al., [8] proposed a technique to manage the data generated by WCE procedure. A binary classification approach was proposed to either discard or keep the frame based on colour space conversion, contrast enhancement and curvature measurement.

Lakshmi Priya [9] had proposed a detection process that involved three steps: visual content representation for the feature extraction, construction of continuous signal for similarity assessment between the consecutive frames and the classification of continuity values for Transition identification. A central tendency-based shot boundary detection for video summarization was implemented here. Khan et al., [10] had proposed an ensemble saliency model, consisting of motion, contrast, texture, and curvature saliency for summarization of WCE videos. Sainui [11] had suggested a colour histogram feature and an optimisation method based on the quadratic mutual information statistical dependence measure for increasing coverage of the full video content and reducing redundancy among chosen key frames. For the purpose of summarising echocardiography films, domain-specific knowledge and automatic spatiotemporal structure analysis were integrated by Ebadollahi et al., [12]. Using the graph that was generated, the videos were time sampled.

The survey gave an understanding that the existing solutions focused on extracting statistical features from the individual frames in the video for the purpose of summarization and keyframe extraction. The existing works have not considered the global features of the video and the temporality of the video for summarization purposes. Further, these summarization techniques were mostly focused on WCE procedure videos.

III. IMPLEMENTATION

Implementation wise the algorithm designed is roughly divided into three parts. These three steps begin with the feature-distance calculation, clustering of the frames, and key-frame extraction from the clusters. The approach combines k-means and local maximum finding. The video frame distances are plotted on a graph. The chi-squared distances between the two-colour histograms of the frames are what these distances represent. The colour histograms have a 4x8x8 bin distribution and are in CIE lab colour space. Then, by looking for the local maxima, it establishes cluster boundaries. These local maxima are discovered with these characteristics. 1. The distance is the greatest distance found after comparing the distances of 2k symmetric neighbours. 2. The distance exceeds the second maximum by an amount n. The closest frame to the cluster mean is chosen by the algorithm to extract a key frame. The proposed algorithm consists of five steps that are depicted in a pipeline like shown in figure 3. The algorithm steps are given in Table 1.

Table 1: Algorithm Steps for Key Frame Extraction

Step 1: Creation of an array of colour histograms.
Step 2: Calculation of the distance between all consecutive frames and recording them in an array
Step 3: Cluster Creation
Step 4: Calculation of keyframe for each cluster
Step 5: Calculate fidelity of every keyframe

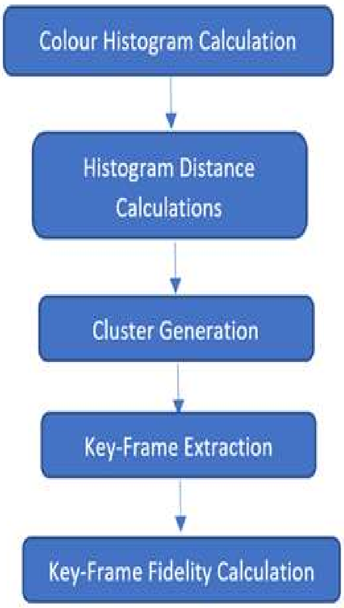
Colour Histogram Calculation
Histogram Distance Calculations
Cluster Generation
Key-Frame Extraction
Key-Frame Fidelity Calculation

Figure 3: Key-Frame Extraction Pipeline

The colour histograms from the capsule endoscopy video are extracted in the CIE lab colour space. This colour space is preferred, because it excludes the colours from the light intensity. It uses three values to represent colour: L* for perceived lightness and a* and b* for the four distinct hues seen by humans: red, green, blue, and yellow. The histogram has a 4x8x8 stack distribution, resulting in a 20-cell histogram. For each image, a histogram is created and stored in a table. The graph is cached because the graph computation is the slowest part of the algorithm.

Once stored, they can be reused when calculating keyframe fidelity and when changing settings.

Distance is the similarity metric between two frames that are considered in a cluster. The distances between all consecutive frames in the cluster is computed and stored in the array.

TUsing the Chi-squared test specified in equation 1, the distance between two successive frames is computed using the histograms for the CIE lab colour space.

$$d\ (C1, C2) = \sum(C1 - C2)^2 / C1 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Where C1 and C2 are colour histograms in CIE lab colour space as of two consecutive frames.

Clustering is a method of grouping data points in a data set based on their similarity. The algorithm used here is based on finding local maxima.

Grouping is done according to the distance table. The algorithm covers all calculated distances. It will examine 2 k symmetrical neighbours at each distance point (k neighbours to the left and k neighbours to the right) and determine which neighbours have the highest value. A boundary is established between two points if the current distance is n times greater than the greatest distance between its neighbours (n1).

A key-frame acts as a representative frame that represents a cluster. The mean of the colour histogram is calculated and the frame closest in proximity to the mean histogram is chosen. The resulting key-frames from this technique were not of an optimal quality. The frame that was closest to the mean distance of the cluster's subsequent frames was chosen in the second iteration.

By contrasting a key-frame with the other frames in the cluster, the key-frame fidelity metric expresses the quality of a key-frame. The maximum 15 distances between the key frame and its cluster make up the key frame fidelity. The equation 2 given below is used to calculate the fidelity of the keyframe extracted

$$\text{fidelity}\ (KeyFr_k, Fr) = \max_i \text{distance}\ (Fr(i), KeyF_k)\ i = s_k, \ldots e_k \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Where $KeyFr_k$ is the key frame of cluster k, Fr is the set of frames in the video, s_k is the starting frame of the cluster k, e_k the final frame of the cluster k. The distance function is used to calculate distance between two frames. Keyframe fidelity is calculated so that the user can see the quality of the keyframe relative to other generated keyframes.

IV. RESULTS AND DISCUSSION

For the experiments four test videos with various symptoms are considered from the publicly available dataset Kvasir-Capsule Dataset [13]. The test videos have been resized to a resolution of 256x256. The frames are downsampled at 3 frames per second.

- *Angiodysplasie:* This video consists of 125608 frames.
- *Bleeding:* This video consists of 68939 frames.
- *Polyp:* This video consists of 124970 frames.
- *Ulcer:* This video consists of 121934 frames.

The symptoms in the test set videos considered are explained here. Angiodysplasie, also known as vascular malformation, refers to an abnormality in the blood vessels that can occur in various parts of the body, including the gastrointestinal tract, lungs, brain, and skin. In the gastrointestinal tract, angiodysplasia is a common cause of gastrointestinal bleeding, especially in older adults. It is characterised by the presence of small, dilated blood vessels in the mucosal lining of the intestines, which can rupture and cause bleeding.

A polyp is a growth that projects from the inner lining of a body organ. Polyps can occur in different parts of the body, including the colon, uterus, nasal passages, and stomach. An ulcer, also known as peptic ulcer, is an open sore that can develop in the lining of the stomach or small intestine. It is typically caused by the bacteria

Helicobacter pylori (H. pylori). The sequence of a keyframe strip is shown in Figure 4.

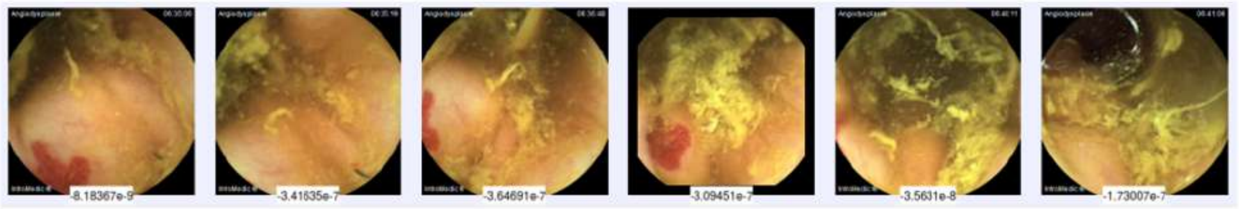

Figure 4: Extracted Key Frame Strip

Experiments have been conducted on all the test videos with the proposed algorithm and the different values for the parameters namely k and n are computed and analysed. The algorithm is tested on all the test videos for assessing the quantity and quality of the key frames. The quantity is measured as the number of frames divided by the total amount of frames in the video in order to compute the compression ratio. The assessment of the quality of the keyframes are done using the mean keyframe fidelities computation. The compression graph of an ulcer video is as shown in Figure 3. Here, the parameter n is represented by the x-axis, the compression rate by the y-axis, and the parameter k by the various lines.

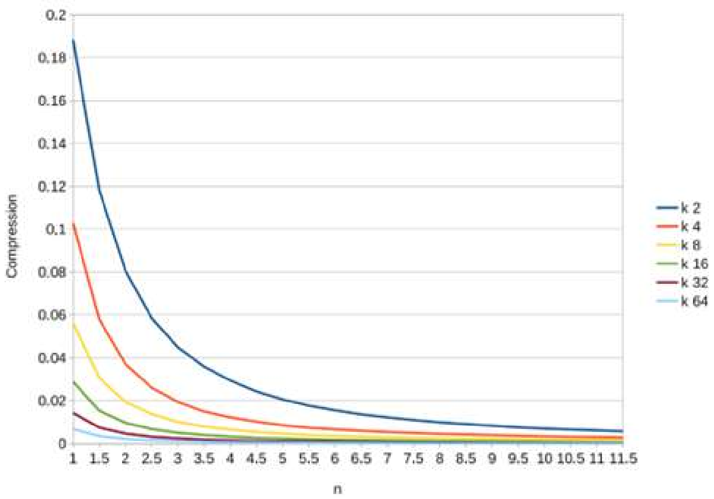

Figure 4: Compression Graph versus n of a test video

The Fidelity graph of the test ulcer video is shown in Figure 4. The distinct lines are the parameter k, the x-axis is the parameter n, and the y-axis is the average fidelity of the key-frames.

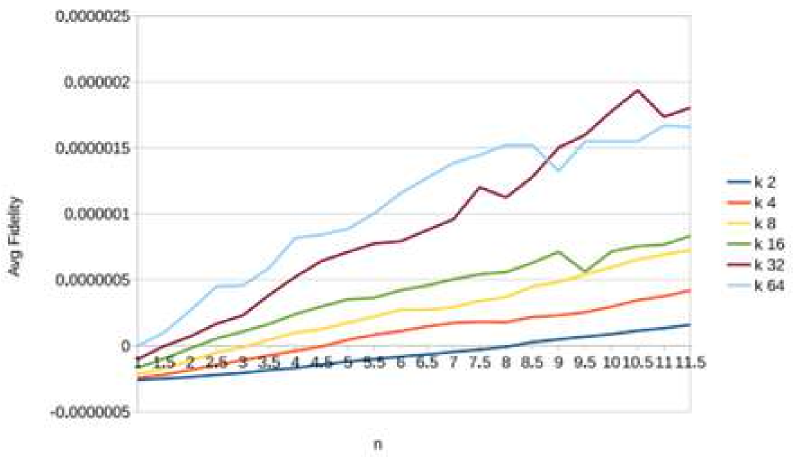

Figure 5: Average fidelity graph of a test video

The compression rate increases similarly to an exponential decay function as the parameters k and n are increased. The key-frame quality degrades roughly linearly as n and k increase. with a little bias in favour of the n-th parameter. This shows that the benefits of compression are diminishing in comparison to key-frame quality.

V. CONCLUSION

The endoscopy video is a special video domain having specific characteristics like specular light reflections, indistinct edges, occlusions, blurriness and artefacts like polyps, smoke, blood, or liquids.

The first two paragraphs can be merged into one Endoscopy videos are content that are unedited having highly similar information, in terms of colour and texture and no shot boundaries.

Endoscopy videos contain a lot of unimportant content like small segments where nothing important happens. It is important and necessary to mine this video content to extract relevant portions that require attention from the physician. This would save the physician's time to a great extent. In this work the spatial features of the video frame namely the histogram distribution in CIE colour space has been considered for key frame extraction. The key frame extraction pipeline has been designed and implemented and the quantity and the quality of the frames extracted have been assessed. Since the video data consists of both spatial and temporal features it is important to consider both of these features in order to generate more meaningful summaries. Further work in summarization of endoscopy video should consider both spatial and temporal features.

REFERENCES

1. Münzer, B., Schoeffmann, K. & Böszörmenyi, L. "Content-based processing and analysis of endoscopic images and videos: A survey" Multimed Tools Appl (2018) 77: 1323. https://doi.org/10.1007/s11042-016-4219-z (Springer US)
2. Jung, Y., Cho, D., Woo, S., Kweon, I.S. (2020). Global-and-Local Relative Position Embedding for Unsupervised Video Summarization. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, JM. (eds) Computer Vision – ECCV 2020. ECCV 2020. Lecture Notes in Computer Science(), vol 12370. Springer, Cham. https://doi.org/10.1007/978-3-030-58595-2_11.
3. Ben Ismail, Mohamed maher & Bchir, Ouiem & Emam, Ahmed. (2014). "Endoscopy Video Summarization based on Multi-Modal Descriptors and Possibilistic Unsupervised Learning and Feature Subset Weighting". Intelligent Automation and Soft Computing. 10.1080/10798587.2014.890320.
4. Chen, Jin & Zou, Yuexian & Wang, Yi. (2016). "Wireless capsule endoscopy video summarization: A learning approach based on Siamese neural network and support vector machine". 1303-1308. 10.1109/ICPR. 2016. 7899817.

5. A. Z. Emam, Y. A. Ali and M. M. Ben Ismail, "Adaptive features extraction for Capsule Endoscopy (CE) video summarization," International Conference on Computer Vision and Image Analysis Applications, Sousse, 2015, pp.1-5. doi:10.1109/ICCVIA.2015.7351 879.
6. Lux, Mathias & Marques, Oge & Schoeffmann, Klaus & Böszörmenyi, Laszlo & Lajtai, Georg. (2010). "A novel tool for summarization of arthroscopic videos". Multimedia Tools Appl. 46. 521-544. 10.1007/s11042-009-0353-1.
7. Mehmood, Irfan et al. "Mobile-cloud assisted video summarization framework for efficient management of remote sensing data generated by wireless capsule sensors." Sensors (Basel, Switzerland) vol. 14,9 17112-45. 15 Sep. 2014, doi:10.3390/s140917112.
8. Mehmood, I., Sajjad, M. & Baik, S.W. J "Video summarization based tele-endoscopy: a service to efficiently manage visual data generated during wireless capsule endoscopy procedure" Med Syst (2014) 38: 109. https://doi.org/10.1007/s10916-014-0109-y, Springer.
9. G. G. Lakshmi Priya, "Medical Video Summarization using Central Tendency-Based Shot Boundary Detection", International Journal of Computer Vision and Image Processing, 3(1), 55-65, January-March 2013 55.
10. Khan Muhammad, Jamil Ahmad, Muhammad Sajjad and Sung Wook Baik, "Visual saliency models for summarization of diagnostic hysteroscopy videos in healthcare systems" Springer Plus (2016) 5:1495 DOI 10.1186/s40064-016-3171
11. Sainui, J. (2017). "Key Frame Based Video Summarization via Dependency Optimiza tion".
12. Shahram Ebadollahi, Shih-Fu Chang and HenryWu, "Echocardiogram Videos: Summarization, Temporal Segmentation and Browsing" IEEE International Conference on Image Processing, 22-25 Sept. 2002, ISBN: 0-7803-7622-6, ISSN: 1522-4880, DOI: 10.1109/ICIP.2002.1038098 (Conference).
13. Smedsrud, P.H., Thambawita, V., Hicks, S.A. et al. Kvasir-Capsule, a video capsule endoscopy dataset. Sci Data 8, 142 (2021). https://doi.org/10.1038/s41597-021-00920-z