Transfer Learning based Video Summarization in Wireless Capsule Endoscopy

Vrushali Raut
Electronics and Communication Engineering Department
MIT ADT University
Pune, India
vrushali.raut0210@gmail.com

Reena Gunjan
Electronics and Communication Engineering Department
MIT ADT University
Pune, India
reenagunjan@gmail.com

Abstract—**Wireless capsule endoscopy (WCE) is a noninvasive procedure to examine gastrointestinal tract. The main challenge in this medical procedure is the time required to examine the recorded video by the medical expert. The computer aided diagnosis of gastrointestinal disorders can prove great help in WCE examination procedure. In this paper a technique based on transfer learning is proposed for summarization of WCE video. In the methodology, Inception V3, a version of standard convolution neural network (CNN) architecture is used for transfer learning along with K-means clustering. The experimental results are evaluated by using F-measure and compressing ratio. The results show that the proposed method performs well in eliminating redundant frames and thus ultimately reduces the time of diagnosis in WCE video inspection.**

Keywords-K-means clustering; summarization; transfer learning; wireless capsule endoscopy

I. INTRODUCTION

Gastrointestinal Tract (GIT) can be examined by medical procedures such as colonoscopy, gastroscopy etc. Wireless Capsule Endoscopy (WCE) is a medical procedure that uses a capsule which is fitted with a camera. This capsule swallowed by the patient will take photographs of the internal parts of GIT. Capsule endoscope advances through GIT without any harm to the internal organs. This method of examining GIT is more convenient in approaching the interior parts of GIT which was limited in previous methods like endoscopy and colonoscopy [1]. The swallowed capsule takes thousands of pictures on the inside of the GIT and transmits them to a recorder which is generally placed on the belt around the patient's waist. The entire procedure will take approximately eight hours. The images taken are downloaded from the recorder into a computer, and those images are examined by the medical expert. The capsule move with the physical peristalsis and excreted by the patient.

The frame rate of capsule decides the number of images taken during the total procedure. For the frame rate of 0.5 Hz approximately 55000 to 60000 images are captured. If the frame rate increases the number of images can go up to 90000-100000. As physical peristalsis is responsible for capsule movement, the travel speed of capsule is very low. The slow speed of capsule movement leads to huge amount of redundant frames with fair amount of similarity in illumination, information and structure. A medical expert has to spend remarkable amount of time to examine the video. In some medical centers the assistant doctors examines the total video and summarize it for final diagnosis by senior medical expert. In manual summarization there are chances of dropping out images with abnormalities while inspecting such a huge load of images.

Image processing approaches such as segmentation, classification and detection are explored by many researchers in the field of WCE video analysis [2]-[5]. As compared to these areas of research in WCE video analysis, the area of WCE video summarization is less explored. Probable reason may be the preference for automated detection of abnormalities. Computer aided detection of abnormalities will surely help medical experts in analysis of video but in medical field the automated data analysis is very critical.

Summarized WCE video can be viewed by gastroenterologist to confirm the automated segmentation or detection results generated by the software thereby eliminating the chances of skipping the frames of critical importance in the WCE video inspection. In recent years, Internet of Things (IoT) environments are in great demand in health care sectors (physical activity monitoring and assessment, IoT personalized healthcare systems etc.). In IoT environment, it is inappropriate to forward entire WCE video to the medical expert or to the health center in view of inadequate provisions of smartphones and the prolonged video. Summarized video can serve the purpose in such conditions.

Informative frames of tissue and blood vessels are important for diagnosis. Physical peristalsis is responsible for random motion of capsule, which leads to several redundant and noninformative frames. These noninformative frames are blurred frames or of food particles, bubbles, intestinal cavity, fecal matter. Therefore, summarized WCE video without these redundant frames will remarkably decrease the examination time. The later section is organized in following way: Section 2 gives information about the background, Section 3 explains the proposed method with subsections including information about summarization, transfer learning, Inception V3, Section 4 is about implementation details. Experimental results are exhibited in Section 5 and finally, conclusion and set of references is mentioned.

II. BACKGROUND

Image representation is mostly done by extraction of different hand-crafted features in conventional WCE video

summarization techniques. These features are reflection of low-level characteristics of an image and hence cannot be considered as primary parameter to determine semantic similarity between images. Low level features are insufficient to perform the task, as they are not enough to depict high level semantic resemblance between two successive images.

Ahmed et al. explored different features extraction techniques in [6]. Those are (1) Color histogram (2) Color moment HSV (3) Local Binary Pattern (LBP) (4) Color moment RGB. Extracted features also include statistical features. Concept used for summarization uses cosine similarity which is computed between neighboring images with a threshold of 90%. In the same paper another method for summarization is proposed employing class means cosine similarity distance of identical frames.

Adaptive method for WCE video summarization is presented by Chen et al. in [7]. This method examines temporal correlation and similarity in features between neighboring WCE frames. Color features are extracted by using Color histogram (HSV) and texture features are extracted by Gray Level Co-occurrence (GLC). The mentioned feature similarity of adjacent frames is determined and similar images are grouped in a clip. Then by applying adaptive K-means clustering algorithm redundant frames are eliminated while retaining key frames. Performance metrics mentioned in the paper attains vales of 81.94% and 80.31% for F-measure and compression ratio respectively.

Zhan et al. concentrated on salient areas of the WCE images. Multiple features and extraction method relied on mutational and gradient identification is the way proposed in the paper for generating video abstract [8]. Chen et al. mapped similar frames together while different frames apart in a feature space [9]. Features used are high level semantic features. Supervised classification by using linear SVM is preferred for setting proper threshold which is generally performed manually in other proposed techniques. The performance metrics used are recall, precision, F-measure, compression ratio, computation time

Ahmed et al. proposed a technique in which learned dictionary is used to sparse code the handcrafted and deep features [10]. SVM classifier is trained by using these features. Feature set includes power spectral density, mean color, histogram (hue and opponent), pyramidal LBP, deep features extracted by GoogleNet. SVM classifier classifies all the images in to two categories, those are informative and non-informative. The performance metrics used are accuracy, sensitivity, specificity.

State-of-the-art summarized techniques generally observe following steps.

- Feature extraction from each image
- Latent space representation of extracted features.
- Shot Segmentation of entire video.
- Extraction of key frames from each shot

[11]–[14] uses combination of extracted features such as color, texture and shape for segmentation of shots. Shot segmentation is generally done with some threshold and distance between information entropies of adjacent frame. Critical frames are extracted by using clustering techniques like K-means, affinity propagation etc.

III. PROPOSED METHOD

In this paper we propose a technique for summarization of WCE video. The methodology includes following steps.

- Extract frames from video
- Employing Inception V3 for generating embeddings of each frame (Transfer learning)
- Applying K-Means clustering for grouping similar frames together
- Matrix operations for generating best sequences
- Stitching best sequences together for final output.

The later section gives detail explanation of the methodology and related topics.

A. Summarization

In summarization, epitomized tape of original prolonged video is generated that can be watched in well reduced amount of time as compared to the actual video.

B. Transfer Learning

Architecture construction from scratch is very complex and advanced task that needs a lot of research as well as time. Transfer learning is a machine learning technique used for deep neural networks in which a model is trained on one task and then it can be re-purposed on another related task. A CNN uses filters on the image to learn detail pattern compare to global pattern with a traditional neural network. Few standard CNN architectures are ResNet, AlexNet, VGGNet, GoogleNet, Inception etc. A network which is already trained (Pre-trained) is used to start the new process which may involve different type of datasets. The pre-trained network which has already learned features will learn the new data features in less amount of time than a new network. In transfer learning process, tuning a network is certainly much simpler and quicker. The learned features can be quickly transferred to a new job and the improved results can be obtained with small amount of data as compared to technique employing architecture designed from scratch.

C. Inception V3

In medical imaging, the informative frames have extensive variation in the locality of information. Capsule moves through the entire GIT which includes esophagus, stomach, small intestine and colon. The images taken in all these parts of GIT are different in multiple aspects. The challenging task in any computer aided diagnosis with CNN, is selecting the proper kernel size. If large numbers of convolutions are stacked then the networks are computationally expensive and prone to overfitting. Fig. 1 Shows simple inception module.

Inception V3 is a deep CNN architecture network that has already learned substantial feature representations for a wide range of images as it is trained by using millions of images

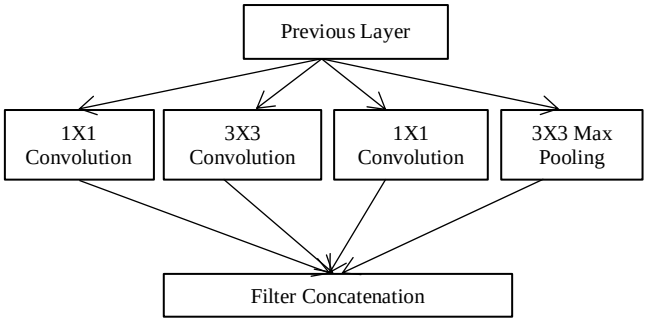

Figure 1. Basic Inception module

included in huge ImageNet dataset [15]. Advantages of Inception V3 are;

- Improved performance in terms of speed, accuracy and utilization of the computing resources.
- Increased depth and width of the network with optimum computational budget.

The network is 48 layers deep including parameter and polling layers with 9 inception modules. Table I gives detail information about the layers. Inception modules are made up of filters of varying sizes that work at the equal level. These modules are responsible for making the network broader than deeper. Convolution operations are performed on an input image. In inception versions generally the kernels used are of 1X1, 3X3 and 5X5. After completing all convolution operations max pooling is carried out and then the concatenated outputs are forwarded to the next inception module. The key metrics of Inception V3 are batch normalization in the auxiliary classifiers, RMSprop optimizer, factorized 7X7 convolutions and label smoothing (to avoid over fitting).

IV. IMPLEMENTATION

Fig. 2 shows block diagram of WCE video summarization process. Database includes KID dataset- video 1, two original WCE videos collected from medical center [16]. The KID dataset video received is with weak annotations.

TABLE I. INCEPTION V3 LAYERS

Layer	Filter Size	Stride	Depth
Convolution	7X7	2	1
Max Pooling	3X3	2	0
Convolution	3X3	1	2
Max Pooling	3X3	2	0
Inception (3a)			2
Inception (3b)			2
Max pooling	3X3	2	0
Inception (4a)			2
Inception (4b)			2
Inception (4c)			2
Inception (4d)			2
Inception (4e)			2
Max pooling	3X3	2	0
Inception (5a)			2
Inception (5b)			2
Average Pooling	7X7	1	0
Dropout (40%)			
Fully connected			
Softmax			

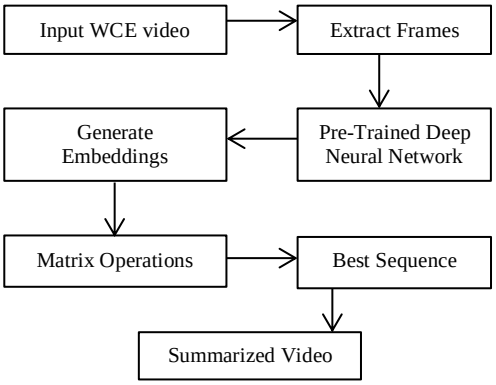

Figure 2. Video Summarization: Block diagram

The other WCE videos are annotated by medical expert i.e. all the key frames are labeled. By selecting optimum frame rate the frames are extracted from input WCE video.

A. *Training Details*

Total database is divided in to 3 groups
Training set: 80 %
Validation set: 20%
Batch size: 30
Epoch: 50
Input image size: 64X64X3
Final classes: 10

Trained model is saved as protobuf (.pb) file and imported in summarization process for generating embeddings of frames.

B. *Algorithms*

Algorithm 1

1) WCE video as an input to the process
2) Frames extraction from the input video
3) Use of Pre-trained model parameters for creating embedding for each frame (numpy array of each frame)
4) Apply K-Means clustering based on Euclidian distance with K=10
5) Generating best sequence in each matrix
   - Create matrix from each cluster embeddings
   - Transpose the Matrix
   - Normalize/flatten the matrix to get 1D matrix.
   - Find argmin values over the axis.
   - Get sequences from flattened matrix based on argmin values and index range.
   - To get a final summarized video stitch the best sequences from each matrix.

Clustering algorithm is used to cluster similar images. Results of multiple clustering algorithms were compared for finalizing a suitable clustering algorithm for WCE images. DBSCAN and spectral clustering are considered for comparison with K-means clustering. The experimental results of comparison helped in finalizing K-means clustering algorithm in the proposed technique. The algorithm proceeds as follows.

Algorithm 2

Input: Embeddings of all frames
Output: 10 clusters consisting similar frames
Assume: Set of data points be A = {a1, a2, a3,……..,an}
Set of centers be B = {b1, b2,…….,bx}

1) Random selection of 'x' Cluster Centers (CC).
2) Computation of the distance between each CC and every single data point
3) Allocate all the data points to the new CC. This step is performed based on minimum distance between data point and the CC.
4) Reevaluate the new cluster centers using:

$$Bi = \left(\frac{1}{X_i}\right)\sum_{j=1}^{x_i} A_i$$

X_i - Number of data points in i^{th} cluster.

5) Reevaluate the distance between latest allocated CC and each data point
6) If there is no change in assignment of data points then terminate the operation else repeat the procedure from point number 3.

V. EXPERIMENTAL RESULTS

The performance metrics include accuracy, sensitivity, specificity, and compression ratio. Following part explain meaning of basic terms in WCE video summarization analysis part. Table II and III shows comparison results for F-measures and compression ratio respectively.

- True positive (TP): The frame is rightly selected by the proposed method and the medical expert.
- False negative (FN): The frame which is incorrectly missed by the proposed method but selected by the medical expert.
- False positive (FP): The frame is incorrectly selected by the proposed method but not by the medical expert.

TABLE II MEAN F -MEASURE

Method	Video	F-measure
CCT-MRFE-RMR	Group I	83.26
	Group II	62.82
	Group III	70.19
	Mean	72.09
WCE-RIE	Group I	90.12
	Group II	73.40
	Group III	83.57
	Mean	82.36
Proposed Method	Group I	93.3
	Group II	86.92
	Group III	75.14
	Mean	85.12

TABLE III COMPRESSION RATIO COMPARISON

Method	Video	Redundant	Total	CR
CCT-MRFE-RMR	Video I	24888	30267	82.23
	Video II	31133	36949	84.26
	Video III	32092	35536	90.31
	Mean			85.60
WCE-RIE	Video I	23163	30267	76.53
	Video II	30407	36949	82.29
	Video III	29178	35536	82.11
	Mean			80.31
Proposed Method	Video I	2902	3261	88.99
	Video II	999	1087	91.90
	Video III	21161	28485	77.8
	Mean			86.23

- True negative (TN): The frame is not selected by both, proposed method and medical expert.
- N_R: Number of redundant frames
- N_T: Total number of frames

The results are compared with CCTMRFE-RMR [17] and redundant image elimination method WCE-RIE. Compression ratio is the ratio of N_R to N_T. F-Measure is calculated by using values of confusion matrix. Fig. 3 and Fig. 4 shows K-means clustering results. Fig. 4 shows ulcer images grouped together. Fig. 5 shows summarized video frames after elimination of redundant frames from a small part of original WCE video.

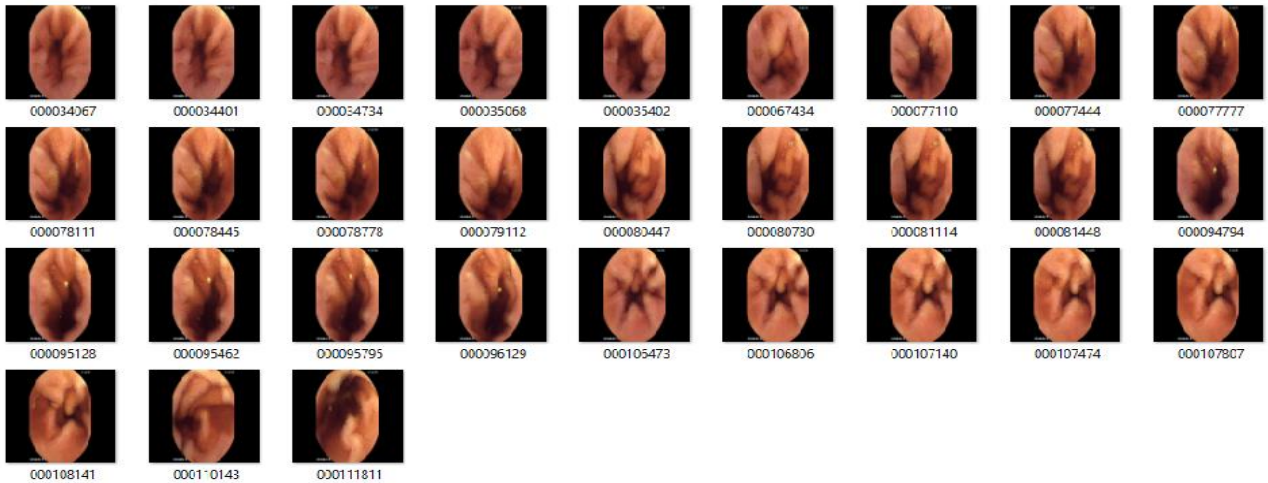

Figure 3. K –means clustering results

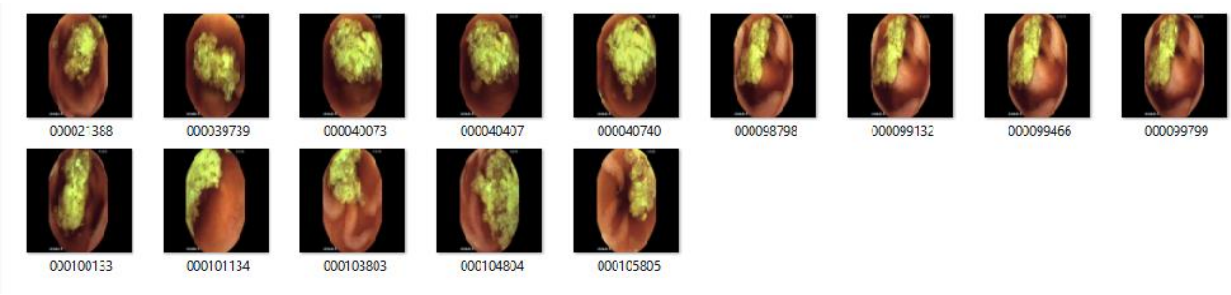

Figure 4. K –means clustering results showing ulcer images grouped together.

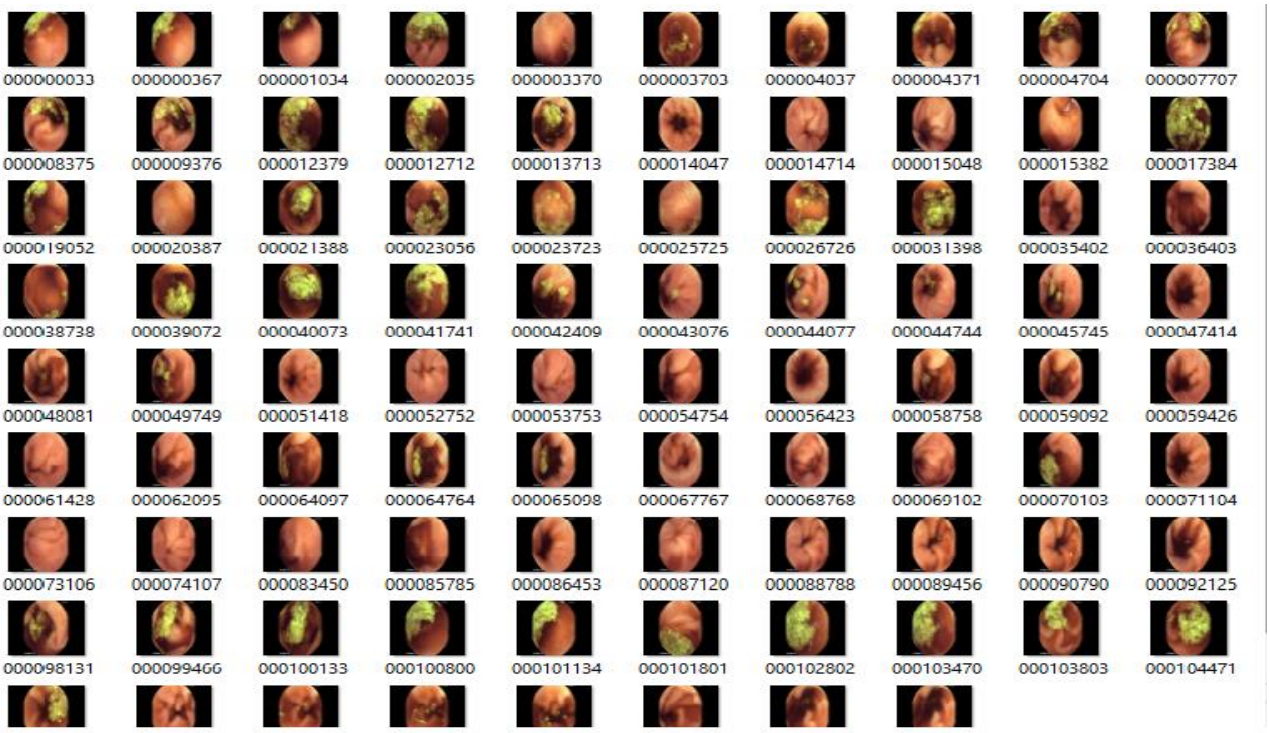

Figure 5. Summarized video frames after elimination of redundant frames.

CONCLUSION

Video summarization area is less explored by researchers in WCE examination. The proposed technique is based on transfer learning using Inception V3 and K-means clustering. The mean of F-measure values calculated for all groups is 85.12%. This value indicate that the proposed method effectively select the informative frames. The mean of all compression ratio values is 86.23% that indicate better capacity of proposed method to eliminate redundant frames. The video of 50000 frames can be summarized to approximately 1500-2000 frames which ultimately lead to decrease in WCE video inspection time. Summarized WCE video can be used by the medical expert to confirm the automated detection results of WCE abnormalities. The summarized video can serve the requirements of IoT environment while transferring the data to smartphones or medical centers. Future scope includes improvement in the experimental results by increasing the database. For actual application of methodology in clinical setting, validation of results can be done by comparing the results of different techniques on similar datasets.

ACKNOWLEDGMENT

We thank Dr. Dimitris Iakovidis, University of Thessaly, Greece for permitting access to the requested KID datasets.

REFERENCES

[1] G. E N. Schoofs, J. Devière, and A. Van Gossum, “PillCam colon capsule endoscopy compared with colonoscopy for colorectal tumor diagnosis: a prospective pilot study.,” Endoscopy, vol. 38, no. 10, pp. 971–7, Oct. 2006.

[2] M. K. Bashar, T. Kitasaka, Y. Suenaga, Y. Mekada, and K. Mori, “Automatic detection of informative frames from wireless capsule endoscopy images,” Med. Image Anal., vol. 14, no. 3, pp. 449–470, 2010.

[3] Junzhou, C., et al. Contourlet based feature extraction and classification for Wireless Capsule Endoscopic images. in Biomedical Engineering and Informatics (BMEI), 2011 4th International Conference on. 2011. IEEE.

[4] Vladimir and A. Shvets, “TernausNet: U-net with VGG11 encoder pre-trained on imagenet for image segmentation,” arXiv preprint arXiv:1801.05746, 2018.

[5] D.K. Iakovidis, S.V. Georgakopoulos, M. Vasilakakis, A. Koulaouzidis, and V. Plagianakos, “Detecting and Locating Gastrointestinal Anomalies Using Deep Learning and Iterative Cluster Unification,” IEEE Transactions on Medical Imaging, 2018, doi:10.1109/TMI.2018.2837002

[6] 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.2015.7351879.

[7] 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," 2015 IEEE International Conference on Digital Signal Processing (DSP), Singapore, 2015, pp. 735-739, doi: 10.1109/ICDSP.2015.7251973

[8] C. Zhan, Y. Cai, N. Sheng, C. Qiu, Y. Cui and X. Gao, "Saliency based Wireless Capsule Endoscopy video abstract," 2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Datong, 2016, pp. 1423-1428, doi: 10.1109/CISP-BMEI.2016.7852940.

[9] J. Chen, Y. Zou and Y. Wang, "Wireless capsule endoscopy video summarization: A learning approach based on Siamese neural network and support vector machine," 2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, 2016, pp. 1303-1308, doi: 10.1109/ICPR.2016.7899817.

[10] A. Mohammed, S. Yildirim, M. Pedersen, Ø. Hovde and F. Cheikh, "Sparse Coded Handcrafted and Deep Features for Colon Capsule Video Summarization," 2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS), Thessaloniki, 2017, pp. 728-733, doi: 10.1109/CBMS.2017.13.

[11] Q. Zhao and M. Q.-H. Meng, “Wce video abstracting based on novel color and texture features,” in 2011 IEEE International Conference on Robotics and Biomimetics. IEEE, 2011, pp. 455–459.

[12] B. Li, M. Q.-H. Meng, and Q. Zhao, “Wireless capsule endoscopy video summary,” in 2010 IEEE International Conference on Robotics and Biomimetics. IEEE, 2010, pp. 454–459.

[13] Y. Yuan and M. Q.-H. Meng, “Hierarchical key frames extraction for WCE video,” in 2013 IEEE International Conference on Mechatronics and Automation. IEEE, 2013, pp. 225–229.

[14] J. S. Huo, Y. X. Zou, and L. Li, “An advanced WCE video summary using relation matrix rank,” in Proceedings of 2012 IEEE-EMBS International Conference on Biomedical and Health Informatics. IEEE, 2012, pp. 675–678.

[15] M. Berrimi and A. Moussaoui, "Deep learning for identifying and classifying retinal diseases," 2020 2nd International Conference on Computer and Information Sciences (ICCIS), Sakaka, Saudi Arabia, 2020, pp. 1-6, doi: 10.1109/ICCIS49240.2020.9257674.

[16] 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

[17] J. S. Huo, Y. X. Zou, and L. Li, "An advanced WCE video summary using relation matrix rank," in Biomedical and Health Informatics (BHI), 2012 IEEE-EMBS International Conference on, pp. 675-678, 2012.