Sparse coded handcrafted and deep features for colon capsule video summarization

Ahmed Mohammed[1], Sule Yildirim[2], Marius Pedersen[1], Øistein Hovde[3], Faouzi Cheikh[1]

[1] Department of Computer Science, Norwegian University of Science and Technology, 2815 Gjøvik, Norway.

[2] Department of Information Security and Communication Technology Norwegian University of Science and Technology, 2815 Gjøvik, Norway.

[3] Department of Gastroenterology, Innlandet Hospital Trust, Gjøvik, Norway, and Institute of Clinical Medicine, University of Oslo, Oslo, Norway

Abstract—**Capsule endoscopy, which uses a wireless camera to take images of the digestive track, is emerging as an alternative to traditional wired colonoscopy. A single examination produces a sequence of approximately 50,000 frames. These sequences are manually reviewed, which is time consuming and typically takes about 45–90 minutes and requires the undivided concentration of the reviewer. In this paper, we propose a novel capsule video summarization framework using sparse coding and dictionary learning in feature space. Video frames are clustered into superframes based on power spectral density, and cluster representative frames are used for video summarization. Handcrafted and deep features that are extracted for representative frames are sparse coded using a learned dictionary. Sparse coded features are later used for training SVM classifier. The proposed method was compared with state-of-the-art methods based on sensitivity and specificity on the KID dataset. The achieved results show that our proposed framework provides robust capsule video summarization without losing informative segments.**

Index Terms—**capsule endoscopy, deep features, KSVD, Dictionary learning, Random forest, informative frame.**

I. INTRODUCTION

CAPSULE VIDEO ENDOSCOPY has revolutionized the diagnostic work-up in the field of esophagus, small bowel and colon imaging. The colon traditionally has been examined via optical colonoscopy, a procedure perceived by many to be uncomfortable and embarrassing. Colon capsule endoscopy (CCE) is an alternative for visualizing the colon. Some of the commercially available CCE devices include PillCam COLON I and II from Given Imaging [1]. CCE devices are equipped with miniaturized camera, LED light source, radio transmitter and battery contained in an easy-to-swallow capsule of dimension 31x11 mm. PillCam COLON II has adaptive frame capture rate of 4 to 35 frames per second depending on its location and movement speed, and has capability of recording images for approximately 10 hours producing ~50,000 frames per procedure [2]. Visual inspection of video sequences is done offline by downloading the video from a receiver, which is worn by the patient during the procedure, to a workstation. Ideal informative (useful) frames depict tissue surface and blood vessel structures, which are crucial for diagnosis. However, due to capsule's zigzag or spinning, motion and purgative procedure used, a significant number of frames contains no useful information for the diagnosis. These frames contain mainly

Fig 1. Sample non-informative frames from the KID dataset: Frames with fecal matter, turbid fluids, motion blurred and food items.

fluids, bubbles, fecal materials, foods, turbid fluids or are blurred frames (Fig 1). Hence, automatic summarization of informative video segments will significantly reduce the reviewing time. For natural video summarization[3, 4, 5] most state-of-the-art methods mainly focus on the summarization of structured videos, such as sports, cartoons or surveillance videos. In comparison, the automatic summarization of unstructured data, e.g., endoscopic videos, is much more challenging. First, capsule videos contain deformable and low-texture context, which makes semantic information extraction a rather challenging task. Second, due to power and volume limitations, the images are taken under low illumination condition, highly compressed, contain noise from CMOS (complementary metal-oxide semiconductor) camera sensor and arbitrary movement of the camera, thus many of the CCE images are of poor quality, which makes accurate video summarization difficult. Finally, the objective of CCE video summarization is to assist doctors in diagnosis, so the video summary should highlight the suspected regions [6].

Many works [7-10, 18, 19] have been proposed to detect frames that are informative by anomaly detection. Evaluation is difficult as these methods are suitable for removing specific types of frames such as out-of-focus [7], bubble frames [8] and redundant frames [9]. These methods usually rely on specific image features, which are known to vary between patients due to different lighting conditions and uncontrolled capsule motion by peristalsis. In addition, non-informative frames contain variety of structures that may also be present in informative frames. Moreover, there is no publicly available dataset for comparing informative and non-informative sequences detection [10].

In this paper, instead of finding frames containing specific pathology or defining a given image content such as bubbles or out-of-focus, we generalize non-informative frames through
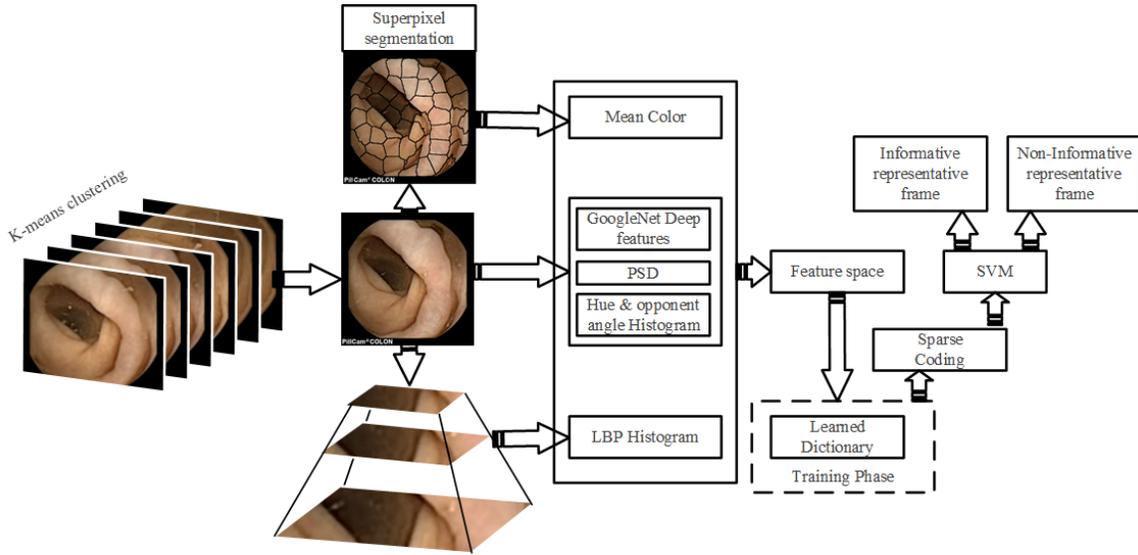
Fig. 2. Sparse coded handcrafted and deep features for colon capsule video summarization block diagram. The dictionary for sparse representation of image features is learned based on selected representative frames.

learning discriminative handcrafted and deep features. This is different from other works in that, the proposed method does not rely on predefined image content. The main contribution of the proposed method can be summarized in two parts. First, a ground truth dataset is developed with gastroenterologists that can be used for further comparison. The database developed contains a ground truth for informative and non-informative frames. Second, after analyzing CCE frames from different patients and different image features, we propose a framework for learning discriminative handcrafted and deep features for CCE video summarization framework, which improves state-of-the-art results in terms of sensitivity and specificity.

The outline of the article is as follows: in Section 2, we introduce previous works done to reduce the reviewing time of CCE. In Section 3 and 4, we present our approach and a detailed outline of the framework along the implementation of the proposed method. Evaluation and comparisons are presented in Section 5 and finally, in Section 6 we present discussions and conclusions respectively.

II. BACKGROUND

In the literature, there are two main research approaches to reduce the amount of time required by an expert for examining CCE video. These are anomaly detection and enabling better visualization. Anomaly detection has been identified as an indirect approach to reduce the review time. These anomalies include detection of causes of bleeding, polyps and other pathologies. For bleeding detection, color histograms with region growing [11], bag-of-visual-words [12], color wavelet features [13], and chrominance moments [14] have been applied in experimental settings. These methods rely on a single or a couple of image feature descriptors and their performance varies greatly when tested on full video sequences and open access datasets [15] . For polyps and lesions, similar types of feature descriptors such as color and texture [16] with second and higher order statistical measures [17], and other feature descriptors such as

salient pixels and image transformation were applied.

As an alternative approach to pathology detection, [18] proposed a method that can detect frames with content that deviates from that of most of the frames in a video segment. Another method that is relevant to the current work was introduced by [8] to detect the informative frames. The authors used two steps in their method: the first step was to isolate highly contaminated non-bubbled (HCN) frames, and the second step was to signify bubbled frames. They used local color moments and the HSV color histogram, which characterized HCN frames. Then, a support vector machine (SVM) was applied to classify the frames. In the second step, a Gauss Laguerre transform (GLT) (based on texture feature) was used to isolate the bubble structures. The combinations of their proposed color and texture features showed average detection accuracies (86.42% and 84.45%).

Similar works proposed by [19] reveals a capacity for up to 85% frame reduction without loss of informative frames. However, evaluation of similar approaches on larger datasets indicates that the accuracy for detection of the most representative frames is rather low (66%)[15]. Most recently [6] proposed video summarization via similar-inhibition dictionary selection. The video summarization process was modeled as a problem of dictionary selection, i.e., to select an optimal subset of frames from the original video frames via dictionary learning under various constraints. The authors then defined similar-inhibition constraint and attention prior to build the dictionary model, which intends to reduce the redundancy between each selected element and to reinforce uniqueness.

III. PROPOSED METHOD

The proposed approach is based on the observation that CCE video frames exhibit high redundancy within and between frames. This can be exploited efficiently to find discriminative features for informative and non-informative frames. Given a video, the framework starts by temporally clustering the frames into superframes. Local features based on superpixel and global

features based on deep features as well as textural and color features are combined to form a feature vector. Dictionary learning is done in feature space, which are later sparse coded for compact representation as shown in Fig. 2. The detailed description of the proposed method is presented below.

A. Superframes

CCE procedure captures many uninformative and redundant frames based on the cleanliness of the bowel and the speed of the capsule. These uninformative frames usually depict food debris, turbid fluids, bubbles and other substances that block the view of the camera. In addition, the frames contain out-of-focus and blurred frames that are not helpful when visualizing the videos. Adopting ideas of over-segmentation in super pixels [20], in the proposed method, these frames are over segmented temporally using K-means clustering algorithm based on the power spectral density of the frames. For N by N image x, the Cartesian power spectral density is defined using Fourier transform as:

$$P(\omega) = \frac{1}{N^2}\left|Y(\omega)\right|^2 \tag{1}$$

where,

$$Y(\omega) = \sum_{n=0}^{N-1}\sum_{m=0}^{N-1} x_{nm} e^{-j\frac{2\pi}{N}(n+m)}, \tag{2}$$

and x_{nm} is pixel at position (n,m) in image x.

The power spectral density is used to cluster frames, as it is a good descriptor for representing blurriness and scene statistics. The Cartesian power spectral density is later converted to polar coordinates to account for capsule rotation as it moves through peristalsis. K-means algorithm is used to create temporal video segments based on the frames power spectral density. The frame, which is closest to the center of the cluster, is later used as the representative frame for further processing.

B. Superpixels

Starting with representative frames from initial over-segmented CCE videos, the representative frames are over-segmented using simple linear iterative clustering (SLIC) [20]. Superpixels provide convenient primitive regions from which to compute local image features. They capture redundancy within the image and greatly reduce the complexity of subsequent image processing tasks.

C. Handcrafted feature extraction

Most of the non-informative frames can be characterized by color and texture features. Food debris and fluids that hide colon walls and tissue surfaces are usually much colorful and have smooth textures. In the proposed method, local features based on superpixel and global features that represent color and texture are concatenated as an image descriptor. The details of the features are given below:

1) Color features

Color is an important feature for image representation, which is widely used for visual and automated CCE image analysis. Local and global image color features are extracted for image sequence. The average color of each superpixel segmented region is computed and aggregated to form local color features. The average color of each superpixel is computed in CIELAB color space, which is widely considered as a perceptually uniform color space. In addition, hue histogram and opponent histogram [21] are used for global image color descriptors as they are robust to photometric variations (i.e., shadow, shading, specular reflection, and light source changes) as well as geometrical variations (i.e., viewpoint, zoom, and object orientation), which are typical lighting conditions for the capsule. However, hue becomes unstable near the grey axis. To this end, [21] proposed a construction of hue histogram, where each sample of hue is weighted by its saturation. Given RGB image, hue H and saturation S can be computed from opponent colors O_1, O_2 as:

$$H = \arctan\left(\frac{O_1}{O_2}\right) = \arctan\left(\frac{\sqrt{3}(R-G)}{R+G-2B}\right), \tag{3}$$

$$S = \sqrt{O_1^2 + O_2^2}$$
$$= \sqrt{\frac{2}{3}\left(R^2 + G^2 + B^2 - RG - RB - GB\right)} \tag{4}$$

where O_1, O_2 are two components from opponent color space,

$$O_1 = \frac{1}{\sqrt{2}}(R-G)$$
$$O_2 = \frac{1}{\sqrt{6}}(R+G-2B) \tag{5}$$

Finally, the opponent angle θ^o, which is specular invariance, is defined as:

$$\theta^o = \arctan\left(\frac{O_1'}{O_2'}\right) \tag{6}$$

where O_1' and O_2' are the spatial derivatives of O_1 and O_2 respectively. The color histogram is built by taking 42 bins histogram of H and θ^o resulting in the final 84 bins histogram.

2) Blurred frames

Blurring occurs in CCE videos as the capsule progresses through the colon by peristalsis. Due to motion of the intestinal content as well as capsule, some of the frames of CCE video are blurred. Moreover, some of the frames are also out of focus resulting in obscured tissue and vessels structures. Using the fact that blurring an image is equivalent to low-pass filtering the frame, we used the power spectral density of the frame as in E.q. (1) to represent the degree of blur.

3) Texture features

Summarizing CCE videos or any medical data requires a robust representation to avoid any false negatives. In order to increase the performance we have included Local Binary Pattern (LBP), which is a simple, yet very efficient, texture operator. The basic LBP operator replaces pixel values with labels by binarizing 3×3 neighborhoods around each pixel with

the center pixel as a threshold. Pixel labels are then converted to decimal numbers. As the capsule moves through the colon, frame contents undergo scale and rotation transformation. Pyramidal representation is a type of multi-scale representation of the image by a set of image approximations from its different frequency-band images. This is obtained by subjecting the image to a repeated smoothing and subsampling. Image pyramidal representation is used to take into account size and scale variations. For the proposed method, LBP were applied with repeated smoothing and subsampling for three different resolutions.

D. Deep features

Convolutional Neural Networks (CNN) are biologically inspired variants of multilayer perceptron. CNNs are considered the state-of-the-art model in image recognition tasks. A pre-trained CNN, specifically; GoogleNet [22] is used as a feature extraction method. GoogleNet achieved a top-5 rank with an error rate of 6.67% on the 2014 ImageNet classification challenge. The basic building block of GoogleNet, the inception module, is a set of convolutions and poolings at different scales, done in parallel and then concatenated together. The deep features are extracted by removing the last fully connected layers. In particular we took the features after dropout in the cls3_pool layer of the GoogleNet model. The deep features were extracted using the MatConvNet [23] framework.

E. Dictionary learning and sparse coding

Visual inspection of many CCE videos shows that different segments of the video share similar local features including color and texture features. These features can be efficiently represented by using sparse coding techniques. Sparse signal representation has become very popular in the past decades and lead to state-of-the-art results in various applications such as face recognition [24], image denoising [25], and image classification [26]. The main goal of sparse modeling is to efficiently represent the images as a linear combination of a few typical patterns, called atoms, selected from a dictionary.

Here, we intend to use sparse representation of the handcrafted and deep features for classification of informative and non-informative frames.

1) Sparse coding

A dictionary is a collection of key feature patterns know as atoms. Sparse learning aims at finding a sparse representation of the input data in the form of a linear combination of the atoms. Given a dictionary matrix, $\mathbf{D} \in \mathbb{R}^{n \times k}$ that contains k atoms as column vectors $\mathbf{d}_j \in \mathbb{R}^n, j = 1,...,K$, the sparse coding problem of a signal $y \in \mathbb{R}^n$ can be stated as finding the most sparse vector $\mathbf{x} \in \mathbb{R}^k$ such that $y \simeq \sum_{j=1}^{k} x_j d_j$ or the representation error $\mathbf{R} = \mathbf{y} - D\mathbf{x}$ is minimized, therefore the optimization problem can be formulated as:

$$\min_{x} \|\mathbf{x}\|_0 \text{ subjected } \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2 \le \varepsilon, \tag{7}$$

where ε is the reconstruction error of the signal $\mathbf{y}$ using the dictionary $\mathbf{D}$ and sparse code $\mathbf{x}$.

Alternatively, the optimization problem can be reformulated as

$$\underset{\mathbf{x}}{\arg\min} \sum_{j=1}^{k} \|x_j\|_0 + \lambda \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2 \tag{8}$$

The minimization problem above is not convex because of the ℓ_0 norm and solving this problem is NP-hard [27]. Hence, there are approximate solutions using greedy approach such as Orthogonal Matching Pursuit (OMP) [28]. For this work, we used the OMP method for sparse representation [29, 30].

2) Dictionary learning

A common setup for the dictionary learning problem starts with access to a training vector $\mathbf{Y} = [\mathbf{y_1}, \mathbf{y_2}..., \mathbf{y_m}]$, where each $y_i \in \mathbb{R}^m$. K-SVD (K-Singular Value Decomposition) is used to iteratively solve the optimization problem of Eq. 7, by alternatively computing the sparse approximation of $\mathbf{X}$ using OMP and then the algorithm proceeds to update the atoms of the dictionary $\mathbf{D}$. K-SVD [31] is an iterative method that alternates between sparse coding of the training set based on the current dictionary and a process of updating the dictionary atoms to better fit the data:

$$\underset{\mathbf{D},\mathbf{x}}{\arg\min} \sum_{j=1}^{k} \|x_j\|_0 + \lambda \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2 , \tag{9}$$

As noted earlier visual inspection of informative and non-informative CCE frames reveals that these images may contain similar texture and color features. Hence, the dictionary atoms are distinctive. We used the sparse representation of the feature space as a feature for classification of informative and non-informative frames as described in the next section.

IV. IMPLEMENTATION

The classification is performed using a Support Vector Machine (SVM) classifier [32]. SVM is a discriminative classifier formally defined by a separating hyperplane. A non-linear SVM is used with Radial Basis Kernel function (RBF). Superpixel segmentation with required number of regions 90 and compactness factor of 20. The feature vectors are aggregated into column vector of size 2429 for dictionary learning. These are summarized as follows, super-pixel segmented image region mean color (Fig. 2) (size 300), power spectral density (PSD) (size 253), hue histogram and opponent histogram (size 84), pyramidal LBP (3 Level, size 768), GoogleNet deep features (size 1024) are aggregated and normalized as follows:

$$F^k = \frac{F_i^k - \min(F^k)}{\max(F^k) - \min(F^k)} \tag{10}$$

where F^k is sub-feature vector k of the features for mean color, power spectral density, hue and opponent angle histogram , LBP features and F_i^k is the ith component of F^k. The dictionary was learned with KSVD algorithm for 100 atoms in the dictionary and 50 atoms for representation.

V. Dataset

500 sample images were chosen by a gastroenterologists from the KID dataset [33] with pathologies and normal images from different parts of the colon. The sample images are selected as representative frames for a complete CCE procedure. The images were taken by GivenImaging Pillcam COLON capsules with a resolution of 576x576 pixels. Multiple images from five different patients were included in the dataset. A gastroenterologists was asked to label the images as informative and non-informative frames. Non-Informative frames are defined as "Video frames that deliver no information for diagnosis or further analysis". After labeling, the dataset contains 339 non-informative frames and 161 informative frames.

VI. Results and Evaluation

In this section, we evaluate the proposed video summarization system on real CCE videos. A range of experiments was performed to assess the strengths and weaknesses of the proposed approach. Super-frame segmentation based on power spectral density provides acceptable clustering of the frames into similar temporal regions as it is shown on Fig. 3. Representative frames are chosen as described in section III.
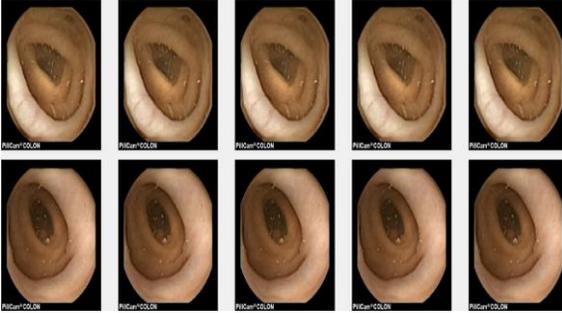

Fig. 3. Superframes segmentation result using K-means with temporal cluster size of 25.

In order to evaluate the accuracy of the proposed method we used 10-fold cross validation with four metrics including Accuracy, Precision, Sensitivity and Specificity that are computed as follows:

$$Sensitivity = \frac{TP}{TP + FN}, \quad Specificity = \frac{TN}{TN + FP}$$

$$Accuracy = \frac{TN + TP}{TP + FP + FN + TN}, Precision = \frac{TP}{TP + FP}$$

Where TP, TN, FP, FN are true positive, true negative, false positive and false negative respectively. We compare our method with other methods without super-frame clustering algorithm. Namely, the proposed method was compared against [34-37]. The results are summarized in (Table I) as reported on respective works. Second and third column shows the image features used for informative frame detection and the metrics

used respectively. As it can be seen, exact comparison of the methods is difficult as the dataset and the source codes are not available. In addition, the dataset developed in this work is general as it contains all artifacts that reduce the informativeness of a frame, such as bubbles, food debris, blurred frames etc. Moreover, the validation method used in the previous works were not reported.
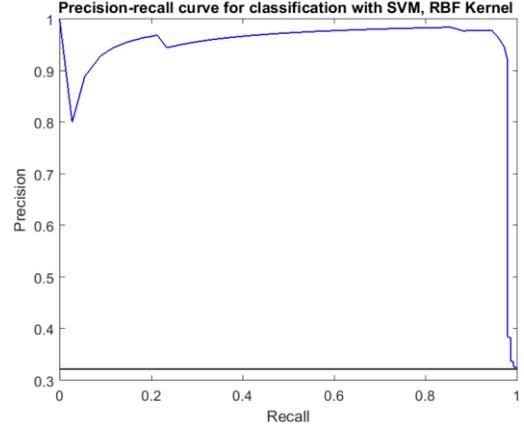

Fig.4. The precision-recall curve for SVM classifier.

Nonetheless, the proposed method's performance gives state of the art result for CCE video summarization with 10-fold average precision of 0.94 and 95.38 accuracy based on the reported scores. For completeness, we also provide the precision-recall graph in Fig. 4 and the Receiver operating characteristic (ROC) curve in Fig. 5. The area under the ROC curve (AUC) is 0.975.

TABLE I
COMPARISON OF METHODS FOR REDUCING CAPSULE VIDEO REVIEW TIME
AS REPORTED ON THE RESPECTIVE WORKS

Proposed	Features	Best average results	
		Metric	Value
[34]	Color and texture	Accuracy, Sensitivity/Specificity	91.6 80.1/93.1
[35]	Color	Accuracy, Sensitivity/Specificity	93.7 **95.1**/92.7
[36]	Colour	Sensitivity/Specificity	76.4/87.5
[37]	Color, texture and motion	Overall, frame reduction (%)	85.6
Proposed Method	Handcraf ted and deep features	10 Fold cross validation average Accuracy, Sensitivity/Specificity	**95.38** 91.29/**97.34**

Comparison of different methods for frame reduction. As it can be seen, the proposed method performance is robust for summarizing CCE videos as indicated by gastroenterologists

VII. Conclusions and discussion

In this paper, we proposed automated CCE video summarization framework, which is based on combination of handcrafted features and deep features. In addition, dataset is developed with gastroenterologists unlike the previous methods, which involve subjective decision by the authors in classifying the frames into informative and non-informative.
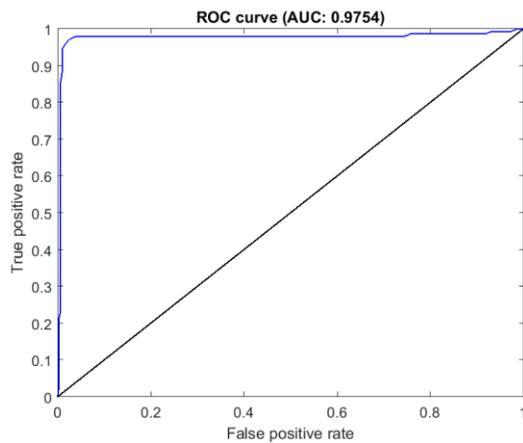
Fig. 5. The ROC curve for the final classifier. The area under the ROC curve is 0.975.

ACKNOWLEDGMENT

REFERENCES

[1] 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] C. E. Parker, C. Spada, M. McAlindon, C. Davison, and S. Panter, "Capsule endoscopy – not just for the small bowel: a review," *Expert Review of Gastroenterology & Hepatology*. Informa Healthcare, pp. 79–89, 16-Dec-2014.

[3] A. Ekin, A. M. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," *IEEE Trans. Image Process.*, vol. 12, no. 7, pp. 796–807, Jul. 2003.

[4] Yu-Fei Ma, Xian-Sheng Hua, Lie Lu, and Hong-Jiang Zhang, "A generic framework of user attention model and its application in video summarization," *IEEE Trans. Multimed.*, vol. 7, no. 5, pp. 907–919, Oct. 2005.

[5] Z. Cernekova, I. Pitas, and C. Nikou, "Information theory-based shot cut/fade detection and video summarization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 1, pp. 82–91, Jan. 2006.

[6] 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, 2016.

[7] J. Oh, S. Hwang, J. Lee, W. Tavanapong, J. Wong, and P. C. de Groen, "Informative frame classification for endoscopy video," *Med. Image Anal.*, vol. 11, no. 2, pp. 110–127, 2007.

[8] 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.

[9] Y. Chen, Y. Lan, and H. Ren, "Trimming the Wireless Capsule Endoscopic Video by Removing Redundant Frames," in *2012 8th I.Conf. on Wireless Comm, Netw. and Mob. Comp.* , pp. 1–4. 2012

[10] D. K. Iakovidis and A. Koulaouzidis, "Software for enhanced video capsule endoscopy: challenges for essential progress," *Nat Rev Gastroenterol Hepatol*, vol. 12, no. 3, pp. 172–186, 2015.

[11] S. Sainju, F. M. Bui, and K. A. Wahid, "Automated bleeding detection in capsule endoscopy videos using statistical features and region growing.," *J. Med. Syst.*, vol. 38, no. 4, p. 25, Apr. 2014.

[12] S. Hwang, "Bag-Of-Visual-Words Approach based on SURF Features to Polyp Detection in," no. i. S. Hwang, "Bag-Of-Visual-Words Approach based on SURF Features to Polyp Detection in wireless Capsule Endoscopy Videos" *Proc. Conf. on Adv. in Visual Computing* - vol 2, Pages 320-327,2012.

[13] C. S. Lima, D. Barbosa, J. Ramos, A. Tavares, L. Monteiro, and L. Carvalho, "Classification of endoscopic capsule images by using color wavelet features, higher order statistics and radial basis functions.," *Conf. Proc. ... Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng.* vol. 2008, pp. 1242–5, Jan. 2008.

[14] B. Li and M. Q.-H. Meng, "Computer-based detection of bleeding and ulcer in wireless capsule endoscopy images by chromaticity moments.," *Comp. Biol. Med.*, vol. 39, no. 2, pp. 141–7, Feb. 2009.

[15] M. Keuchel, N. Kurnianwan, P. Baltes, D. Bandorski, and A. Koulaouzidis, "Quantitative measurements in capsule endoscopy," *Comput. Biol. Med.*, vol. 65, pp. 333–347, 2015.

[16] A. V Mamonov, I. N. Figueiredo, P. N. Figueiredo, and Y.-H. R. Tsai, "Automated polyp detection in colon capsule endoscopy.," *IEEE Trans. Med. Imaging*, vol. 33, no. 7, pp. 1488–502, Jul. 2014.

[17] B. Li, M. Q. H. Meng, and L. Xu, "A comparative study of shape features for polyp detection in wireless capsule endoscopy images.," *Conf. Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Conf.*, vol. 2009, pp. 3731–4, Jan. 2009.

[18] D. K. Iakovidis, E. Spyrou, D. Diamantis, "Efficient homographybased video visualization for wireless capsule endoscopy", Proc. IEEE 13th Int. Conf.Bioeng.vol, pp. 1-4, 2013.

[19] M. M. Ben Ismail and O. Bchir, "CE Video Summarization Using Relational Motion Histogram Descriptor," *J. Image Graph.*, vol. 3, no. 1, pp. 34–39, 2015.

[20] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "SLIC Superpixels," *EPFL Tech. Rep. 149300*, no. June, p. 15, 2010.

[21] J. Van de Weijer and C. Schmid, "Coloring Local Feature Extraction," *Comput. Vis.*, vol. 3952, pp. 334–348, 2006.

[22] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D.Erhan, V. Vanhoucke, A. Rabinovich, C. Hill, and A. Arbor, "Going Deeper with Convolutions," *arXiv:1409.4842*, 2014.

[23] A. Vedaldi and K. Lenc, "MatConvNet," in Proceedings of the 23rd ACM inter. conf. on Multimedia - MM ' ,vol, pp. 689–692,2015.

[24] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, 2009.

[25] M. Elad and M. Aharon, "Image Denoising Via Sparse and Redundant Representations Over Learned Dictionaries," IEEE Trans. Image Process., vol. 15, no. 12, pp. 3736-3745, Dec. 2006.

[26] M. Rastgoo, G. Lemaitre, O. Morel, J. Massich, R. Garcia, F. et al.é, "Classification of Melanoma Lesions Using Sparse Coded Features and Random Forests," Proc. SPIE, Med. Imag.,vol 9785, 2016.

[27] M. Elad, "Sparse and redundant representation modeling-What next?," *IEEE Sig. Process. Lett.*, vol. 19, no. 12, pp. 922–928, 2012.

[28] S. Mallat, Z. Zhang, "Matching pursuits with time-frequency dictionaries", *IEEE Tran. Sig.Proc.*, vol. 41, pp. 3397-3415, 1993.

[29] G. Davis, S. Mallat, and M. Avellaneda, "Adaptive greedy approximations," *Constr. App.*, vol. 13, no. 1, pp. 57–98, Mar. 1997.

[30] Y. C. Pati, Y. C. Pati, R. Rezaiifar, Y. C. P. R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal Matching Pursuit: Recursive Function Approximation with Applications to Wavelet Decomposition," *Proc. 27, A.Conf. SIGNALS, Syst. Comput.*, vol. 1, pp. 40-44, 1993.

[31] R. Rubinstein, M. Zibulevsky, and M. Elad, "Efficient implementation of the K-SVD algorithm using batch orthogonal matching pursuit," *CS Tech.*, pp. 1–15, 2008.

[32] J.A.K. Suykens, J. Vandewalle, "Least squares support vector machine classifiers", Neu Proc. Let, vol. 9, no. 3, pp. 293-300, 1999.

[33] E. Spyrou, D. K. Iakovidis, et al. "Video-based measurements for wireless capsule endoscope tracking," *Meas. Sci. Technol.*, vol. 25, no. 1, p. 15002, Jan. 2014.

[34] S. Segui, M. Drozdzal, F. Vilarino, C. Malagelada, F. Azpiroz, P. Radeva, and J. Vitria, "Categorization and Segmentation of Intestinal Content Frames for Wireless Capsule Endoscopy," IEEE Trans. Inf. Tech. B.med., vol. 16, no. 6, pp. 1341–1352, Nov. 2012.

[35] O. HajiMaghsoudi, A. Talebpour, H. Soltanian-zadeh, and H. A. Soleimani, "Automatic informative tissue's discriminators in WCE," *IEEE Int. Conf. on Img. Sys. and Tech.* , pp. 18–23,2012.

[36] Yichen Fan, M. Q.-H. Meng, and Baopu Li, "A novel method for informative frame selection in wireless capsule endoscopy video," *IEEE Inter. Conf. Eng. in Med. and Bio. Soc.*, pp. 4864–67, 2011.

[37] D. Iakovidis, E. Spyrou, and D. Diamantis, "Efficient homography-based video visualization for wireless capsule endoscopy," *13th IEEE Conf. on B.Inf and B.Eng,* vol. pp. 1–4, 2013.