Automatic Frame Reduction of Wireless Capsule Endoscopy Video

S. Tsevas, *Student Member, IEEE*, D. K. Iakovidis, *Member, IEEE*, D. Maroulis, *Member, IEEE*, and E. Pavlakis

***Abstract*—Wireless Capsule Endoscopy (WCE) is a non-invasive colour imaging technique that has been introduced for the screening of the gastrointestinal tract and especially the small intestine. WCE is performed by a wireless swallowable endoscopic capsule that transmits more than 50,000 video frames per examination. The visual inspection of the resulting video is a highly time-consuming task even for the experienced gastroenterologist. In this paper we propose a novel WCE video summarization approach which is subsequently evaluated using real world patient data. The proposed approach aims to the reduction of the number of the video frames to be visually inspected so as to enable significant reduction in the video assessment time. It is based on clustering using symmetric non-negative matrix factorization initialized by the fuzzy c-means algorithm and supported by non-negative Lagrangian relaxation to extract a subset of video scenes containing the most representative frames from an entire examination. Real world patient data that display abnormal findings at several sites in the small intestine were annotated by expert gastroenterologists in order to experimentally evaluate the proposed approach. The results demonstrate that the suggested approach leads to significant reduction of the total number of frames in the input video without losing critical information related to the abnormal regions of the small intestine.**

I. Introduction

STANDARD endoscopy enables the expert to view both ends of a patient's gastrointestinal tract including the foodpipe, stomach, duodenum, colon and terminal ileum. However, the examination of the small intestine, remains a difficult task which is still limited by the conventional endoscopic techniques. As a solution to the problem Wireless Capsule Endoscopy (WCE) was introduced [1]. This method represents a major departure from conventional endoscopy which is usually uncomfortable for the patient. It is performed by a swallowable capsule with the size of a large vitamin that includes a miniature colour video camera, a light, a battery and a video stream transmitter. By using this capsule, the expert can efficiently diagnose a range of gastrointestinal disorders, including ulcer, unexplained bleeding, and polyps.

Although WCE exhibits significant advantages over traditional examination techniques, there are challenging issues to deal with. A major problem is that the typical size for a WCE video for a patient is approximately 55,000 frames, and it usually takes more than an hour of intense labour for the expert in order to examine the whole frame sequence [2]. However, this manual examination process does not guarantee that some abnormal regions are missed. For example, it is quite common that an abnormal lesion is visible in only a few frames, or it can be so small or flat that it may escape from the examiners notice.

Computational approaches coping with the analysis of the WCE video include the development of special annotation tools to auto-bookmark abnormalities [3]; classification approaches that perform tissue discrimination either between normal and abnormal regions [4] or between different organs [5-7]; synergistic methodologies such as image registration techniques and L-G graphs for the detection of abnormal patterns in WCE images [8]; clustering techniques for blood detection [9]; neural network techniques for classification or detection of abnormal patterns [10,16]; intestinal motility assessment methodologies [13,15]; and other approaches that aim either to image enhancement [12] or to the rejection of invalid parts of the WCE video by performing intestinal juice detection [14].

However, no significant effort has been made to the direction of reducing the time required for the visual inspection of the WCE video. To cope with this issue, we further investigate an effective computational approach that drastically reduces the video frames to be inspected enabling this way faster inspection of the video sequence [16]. The proposed approach applies an unsupervised methodology based on clustering and non-negative matrix factorization (NMF) [18] to summarize the WCE video by keeping the most representative frames from the whole examination.

NMF was proposed in [19] in an effort to preserve much of the structure of the input data and, at the same time, to guarantee that both the resulting basis and its accompanying weights are non-negative. NMF's notion, lies in the process of using a low-dimensional subspace to approximate a much larger one. Lee and Seung [18-20] demonstrated that NMF is able to offer parts-based representations in contrast to other methods such as principal component analysis and vector quantization, leading to a more intuitive approach towards real world data.

The proposed summarization approach has been previously tested on a controlled dataset giving promising results [16]. However, no experimentation has taken place regarding the performance of the approach on real patient data.

The rest of this paper consists of three sections. Section II

S. Tsevas, D. K. Iakovidis and D. Maroulis are with the University of Athens, department of Informatics and Telecommunications, Panepistimiopolis, GR-15784, Athens, Greece (e-mail: s.tsevas@ieee.org, d.iakovidis@ieee.org, dmarou@di.uoa.gr).

E. Pavlakis, M.D., is with Aretaieion Hospital, Department of Surgery.

provides a description of the proposed methodology. Section III, presents the results of its experimental application on WCE video data with several amendments on the results provided in [16], and Section IV summarises the conclusions that can be derived from this study.

II. METHODOLOGY

The WCE video summarization approach we followed is based on the unsupervised data reduction methodology described in [17] and it is developed in three steps. In the first step dimensionality reduction of the initial dataset takes place that results in a square non-negative similarity matrix which is going to be used as an input to the steps that follow. In the second step fuzzy c-means (FCM) clustering takes place for the input video stream to group its frames into a predefined number of clusters, whereas in the final step two NMF algorithms are subsequently applied on the clustered frames in order to extract frames that are representative of the whole video. An overview of this methodology is illustrated in Fig.1.

Given a non-negative $m \times n$ matrix $\mathbf{V}$, the NMF algorithms seeks to find non-negative factors $\mathbf{W}$ and $\mathbf{H}$ of $\overline{\mathbf{V}}$ such that:

$$\mathbf{V} \approx \overline{\mathbf{V}} = \mathbf{W} \times \mathbf{H} \tag{1}$$

where $\mathbf{W} \in \Re^{m \times k}$ and $\mathbf{H} \in \Re^{k \times n}$.

Intuitively, we may think of $\mathbf{W}$ as the matrix containing the NMF basis and $\mathbf{H}$ as the matrix containing the non-negative coefficients. Consequently, NMF solves the following optimisation problem:

$$\min_{\mathbf{W},\mathbf{H}} \left\| \mathbf{V} - \mathbf{W}\mathbf{H} \right\|_F^2 , \ \mathbf{W}, \mathbf{H} \geq \mathbf{0} \tag{2}$$

where $\mathbf{H}$ actually is a reduced representation of $\mathbf{V}$ with lower dimensionality and F stands for the Frobenius norm of a matrix.

The dimensionality and the initial values of $\mathbf{W}$ and $\mathbf{H}$ (or just $\mathbf{H}$ in certain algorithms) are determined by means of the FCM algorithm. FCM performs soft clustering of the video frames so that they belong to more than a single cluster. After the application of the FCM, each frame holds a membership probability to the different clusters. These probabilities are stored in a $m \times k$ matrix $\mathbf{U_{FCM}}$. To prepare the input of the FCM, a $m \times m$ similarity matrix V is constructed according to the process described in [17].

The dimension k of the membership matrix of the converged FCM, $\mathbf{U_{FCM}}$, is set equal to the predefined number of clusters c and the values of its transpose are used to initialize $\mathbf{H}$. The neighbouring frames in the original m-dimensional vector space, are determined by calculating the $m \times m$ matrix $\mathbf{D}_E$ of the Euclidean distances. Using the $\mathbf{D}_E$, the calculation of the geodesic distance matrix takes place by finding shortest paths in a graph connecting neighbouring data points, resulting in a matrix $\mathbf{D}_G$ that contains the geodesic distances between the vectorial representations of the frames. Next, $\mathbf{D}_G$ is transformed into a pairwise similarity matrix according to Eq. (3),

$$V = e^{-\frac{D_G}{r}} \tag{3}$$

The symmetric NMF (SymNMF) which for a square matrix is:

$$\mathbf{V} \approx \mathbf{H} \times \mathbf{H}^{\mathrm{T}} \tag{4}$$

is applied on V so that it "unfolds" the clusters and make them more transparent. According to [22] the calculation of $\mathbf{H}$ is iterative according the following update rule:

$$H_{ik}^{j+1} = H_{ik}^{j} \left(1 - \beta + \beta \frac{(VH)_{ik}^{j}}{\left(HH^T H\right)_{ik}^{j}} \right) \tag{5}$$

where j stands for the iteration index, H_{ik} is the (i,k) entry of $\mathbf{H}$ and $0 < \beta \leq 1$, with 0.5 advised as a good choice for beta value [23]. For j=0, $\mathbf{H}$ is set to $\mathbf{U_{FCM}}$. SymNMF iterates until

$$\left| L_{j+1} - L_j \right| < \varepsilon_1 \tag{6}$$

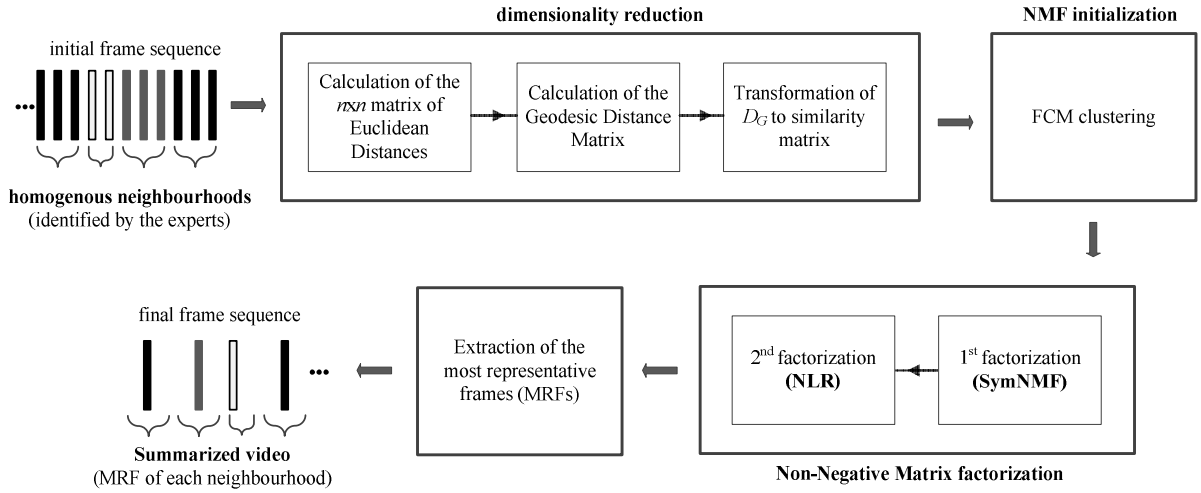

Fig. 1: Methodology for unsupervised summarization of WCE video.

where ε_1 is a small positive constant close to zero and L_j is the objective function of the SymNMF at the j-th iteration and is defined as:

$$L_j = \left\| \mathbf{V}_j - \mathbf{H}_j \mathbf{H}_j^T \right\|_F^2 \quad (7)$$

The final step of the methodology imposes orthogonality constraints on the output of the SymNMF so as to extract the most representative members of a given cluster. It is implemented by means of an NMF multiplicative update algorithm known as Non-negative Lagrangian Relaxation (NLR) [17, 23]. This algorithm iterates according to the following update rule:

$$X_{ik}^{j+1} = X_{ik}^{j} \sqrt{\frac{(VX)_{ik}^{j}}{(X\alpha)_{ik}^{j}}}, \ \alpha = \mathbf{X}^T \mathbf{V} \mathbf{X} \quad (8)$$

until:

$$\left| L_{j+1} - L_j \right| < \varepsilon_2 \quad (9)$$

where ε_2 is a very small positive constant close to zero and L_j is the objective function of the NLR at the j-th iteration and is defined as:

$$L_j = \mathrm{Tr}\left(\mathbf{X}^T \mathbf{V} \mathbf{T}\right) - \mathrm{Tr}\left(\alpha\left(\mathbf{X}^T \mathbf{X} - \mathbf{I}\right)\right) \quad (10)$$

Tr(·) stands for the trace of the matrix (the summary of the diagonal elements) and **I** is the identity matrix. For j=0, **X** is set to the result obtained by the SymNMF.

In NLR the entries of **X** are viewed as cluster indicators and as a result the interpretation of the results at convergence is straightforward allowing this way a relatively easy interpretation of the cluster structure.

III. Results

In order to evaluate the performance of the proposed summarization approach experimentation took place on a real patient video. The video consists of 585 frames (m=585) and of about 5 min duration. Each frame of the video was visually inspected and annotated by two expert endoscopists, and two kinds of abnormal findings were identified in several sites of the small intestine; ulcers and phlebectasias. Following the approach we developed in [16] neighbourhoods of frames were extracted for each finding since each finding was visible in more than a single frame. Frame neighbourhoods exhibit very close similarity and were indicated by both experts without any interobserver variability, providing at the same time us with ground truth information. The composition of the video is the following: 150 ulcer frames, 35 phlebectasia frames and 400 normal frames, whereas phlebectasias were identified in 2 different sites in the small intestine while ulcers in one. Neighbourhood formation is presented in Tables I and II respectively.

In order to reduce the computational times the video frames were downscaled from 370×370 pixels to 101×101 pixels (n=10201). As it was noted in [16] Frames with dimensions smaller than 90×90 are not beneficial for the overall results. Finally images were converted to grayscale so as to form the initial m×n dataset matrix.

By following the process described in the previous section we calculated the similarity matrix V according Eq. 3, with r=100 [17], so as to proceed with the FCM calculations. Experiments have taken place for different number of clusters, thus FCM was executed for 2,3,4,5 and 6 clusters and application of SymNMF and NLR to V followed. Since the values of ε_1 and ε_2 do not contribute to better cluster separation according to the conclusions derived from [16] regarding the value of the computational times and the resulting cluster structure we used 1E-4 for both of them.

The membership of each frame to each of the three clusters as produced at the output of each algorithm in the 3 and 6 clusters case is illustrated in Figs. 2 and 3, respectively. It can be observed that on the one hand SymNMF application does not provide us with a rather clear cluster structure while on the other hand, application of the NLR results in clusters exhibits better separation than SymNMF but still the result is not satisfactory, though NLR enforces orthogonality. This is due to the fact that the number of iterations of both SymNMF and NLR are finite Actually, only a part of the examples are strictly 'orthogonal' to the members of other clusters. These members form the Most Representative Frames (MRFs) of the cluster.

TABLE I
NEIGHBOURHOODS PER FINDING AND NUMBER OF FRAMES PER NEIGHBOURHOOD

U	P
22	12
45	17
36	6
47	

U stands for Ulcer P stands for Phlebectasias. Each cell in the U and P columns of the table represents a neighbourhood of frames.

In order to extract the MRFs of each cluster, we apply the orthogonality condition with a mild deviation from the strict orthogonality according to [17]. Thus, we apply a threshold T to the entries of **X**. The general condition that the MRFs should meet in each cluster for a c cluster case is the following:

$$\text{Clust.1: } X_{i1} > \mathrm{T} \ \& \ X_{i2} < \mathrm{T} \ \& \ \ldots \ \& \ X_{ic} < \mathrm{T} \quad (11a)$$
$$\text{Clust.2: } X_{i1} < \mathrm{T} \ \& \ X_{i2} > \mathrm{T} \ \& \ \ldots \ \& \ X_{ic} < \mathrm{T} \quad (11b)$$
$$\vdots$$
$$\text{Clust.}c\text{: } X_{i1} < \mathrm{T} \ \& \ X_{i2} < \mathrm{T} \ \& \ \ldots \ \& \ X_{ic} > \mathrm{T} \quad (11c)$$

where i stands for the frame index within the initial WCE video and c represents the total number of clusters.

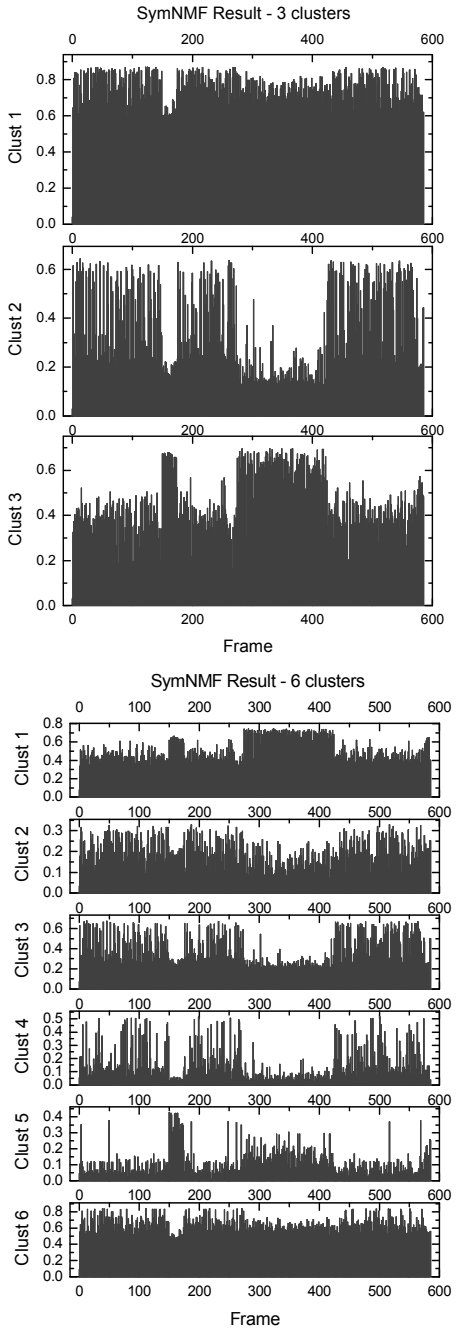

Fig. 2: Result of the SymNMF in the 3 (up) and 6 (down) cluster case.

The value of T controls the degree of summarization of the WCE video. Large values of T lead to more examples (frames) in the resulting set of MRFs. Figure 4 illustrates how the total number of frames in the resulting video varies with T for different clustering cases, whereas Fig. 5 shows the percentage reduction in the total number of frames of the initial video. It is apparent that for threshold values close to 1E-7 the total number of frames per cluster is substantially reduced. For a threshold value equal to 1E-2 summarization does not take place effectively for the 2-clusters case. Moreover, according to Fig. 5 the total number of frames may be reduced down to the 10% of the initial one (for T=1E-7 and 4 clusters), and since the number of frames is proportional to the visual inspection time a 90% reduction in this time is feasible.

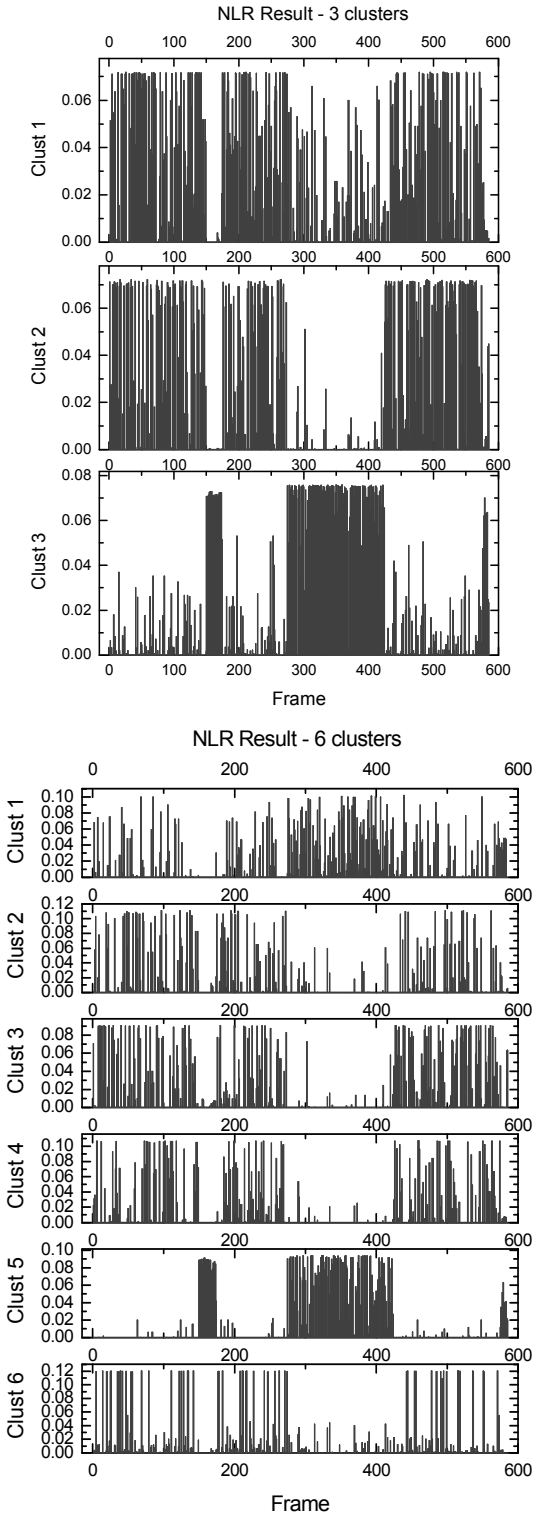

Fig. 3: Result of the NLR in the 3 (up) and 6 (down) cluster case.

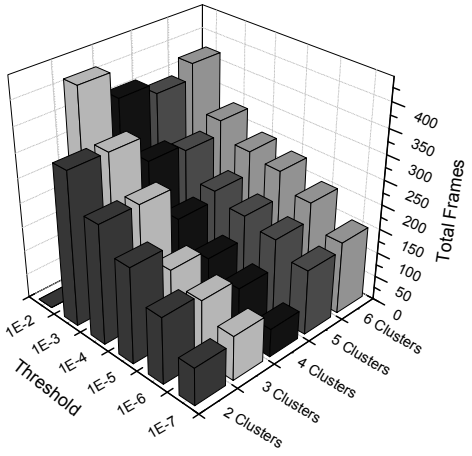

Fig. 4: Total number of frames per clustering-case for different thresholds T.

Although the summarized video is considerably shorter than the initial one, its content needs to be further examined. Since the video is of medical content any loss of frames containing abnormal findings may be critical for the patient. Thus, the summarized video should contain at least one representative frame from each of one of the abnormal findings neighbourhoods that experts indicated in the initial video. Any frame losses regarding the normal frame neighbourhoods are not of interest. The distribution of the representative frames per neighbourhood is presented in Tables II and III for the 3-clusters and 6-clusters cases respectively. It can be observed that for each neighbourhood in the summarized video there is at least one representative.

TABLE II
NUMBER OF REPRESENTATIVE FRAMES OF THE DIFFERENT NEIGHBOURHOODS IN THE SUMMARIZED VIDEO FOR THE DIFFERENT THRESHOLDS (3 CLUSTERS CASE)

T=1.0E-7	
U	P
2	3
9	5
4	1
1	

T=1.0E-6	
U	P
2	6
11	6
9	1
18	

T=1.0E-5	
U	P
4	6
13	7
10	1
23	

T=1.0E-4	
U	P
7	6
16	7
18	2
23	

T=1.0E-3	
U	P
8	6
23	9
21	2
30	

T=1.0E-2	
U	P
19	10
28	14
21	4
36	

U stands for Ulcer and P stands for Phlebectasias. Each cell in the U and P columns of the table represents a neighbourhood of frames.

By integrating a time stamp to each representative frame we can offer the expert the ability to return to the corresponding frame of initial video so as to further examine the area of interest.

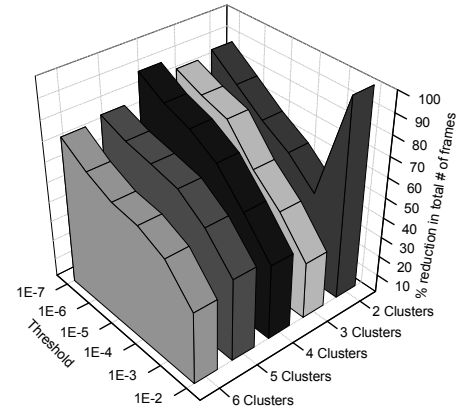

Fig. 5: Percentage reduction in the total number of frames for different thresholds T for different clustering-cases.

TABLE III
NUMBER OF REPRESENTATIVE FRAMES OF THE DIFFERENT NEIGHBOURHOODS IN THE SUMMARIZED VIDEO FOR THE DIFFERENT THRESHOLDS (6 CLUSTERS CASE)

T=1.0E-7	
U	P
9	1
14	5
10	2
17	

T=1.0E-6	
U	P
14	3
26	6
21	3
29	

T=1.0E-5	
U	P
16	3
28	7
21	5
31	

T=1.0E-4	
U	P
15	4
31	8
27	6
37	

T=1.0E-3	
U	P
18	4
33	8
29	6
40	

T=1.0E-2	
U	P
18	7
37	11
29	6
40	

U stands for Ulcer and P stands for Phlebectasias. Each cell in the U and P columns of the table represents a neighbourhood of frames.

The use of different number of clusters did not affect drastically the summarization result. However, the more the clusters are the smoother is the result of the summarization. This is due to the fact that the difference in total frames for different sequential thresholds becomes smaller as the increase in the number of clusters results in a better control of the summarization result. Nevertheless, increase in the number of clusters leads to an increase in the computation times needed. Figure 6 illustrates the normalized computation time necessary to perform FCM, SymNMF and NLR for different clusters.

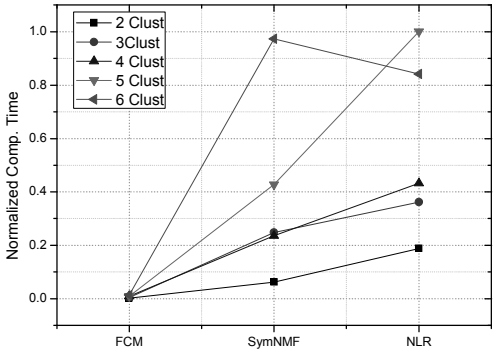

Fig. 6: Normalized computation times for different clusters.

IV. Conclusions

A novel approach to WCE video summarization is presented based on the methodology suggested in [17] that proposes the application of two subsequent NMF algorithms on a dataset formed by the frames of the video. The approach was evaluated experimentally by using a real patient video with multiple findings. The WCE video was annotated frame-by-frame by expert endoscopists who provided us with the ground truth information necessary to determine the effectiveness of the method. The results of the experimental evaluation of the aforementioned data demonstrated that the proposed approach leads to significant reduction of the total number of frames in the input video. Thus, its application may increase the productivity of the experts as it leads to smaller inspection times of a WCE video. Therefore more videos can be inspected in less time. Moreover, the produced summary contains representative frames from every frame neighbourhood in the input video that exhibit close similarity of the neighbouring frames.

Our effort is made towards the development of a robust intelligent system for WCE video summarization, and within this context fall the results presented here. Our future work includes further experimentation with the results of clustering since it seems that in many cases a rather good discrimination between normal and abnormal frames is achieved, utilization of various image features for the discrimination of other types of abnormal findings such as polyps and cancer as well as investigation of memory-efficient techniques to perform NMF on large WCE video streams.

Acknowledgment

We would like to give our special thanks to Dr. A. Polydorou, M.D. for sharing his surgical expertise so kindly with our research group. This research was partially funded by the special account of research grants of the University of Athens, Greece.

V. References

[1] G. Iddan, G. Meron, A. Glukhovsky, and P. Swain, "Wireless capsule endoscopy," *Nature*, 405 (6785), 2000, pp. 417-418.

[2] A. Maieron, D. Hubner,B. Blaha, C. Deutsch, T. Schickmair, A. Ziachehabi, E. Kerstan, P. Knoflach, R. Schoefl, "Multicenter retrospective evaluation of capsule endoscopy in clinical routine," *Endoscopy*, 36 (10), 2004, pp. 864-868.

[3] M. T. Coimbra, and J. P. S. Cunha, "MPEG-7 visual descriptors - contributions for automated feature extraction in capsule endoscopy," *IEEE Transactions on Circuits and Systems for Video Technology*, 16 (5), 2006, pp. 628-636.

[4] B. Li, M.Q.-H. Meng, "Analysis of the gastrointestinal status from wireless capsule endoscopy images using local color feature," *Information Acquisition*,. ICIA '07. International Conference, 8-11 July 2007, pp.553-557.

[5] M. Mackiewicz, J. Berens, M. Fisher, and D. Bell. "Colour and texture based gastrointestinal tissue discrimination," *ICASSP*, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings 2, 2006, pp. II597-II600.

[6] J. Berens, M. Mackiewicz, and D. Bell. " Stomach, intestine and colon tissue discriminators for wireless capsule endoscopy images" *Progress in Biomedical Optics and Imaging*, Proceedings of SPIE 5747, 2005, pp. (I): 283-290.

[7] J. Lee, J. Oh, S. K. Shah, X. Yuan, and S. J. Tang, "Automatic classification of digestive organs in wireless capsule endoscopy videos," *Proceedings of the ACM*, Symposium on Applied Computing, 2007, pp. 1041-1045.

[8] N. Bourbakis, "Detecting abnormal patterns in WCE images," *Proceedings - BIBE 2005*, 5th IEEE Symposium on Bioinformatics and Bioengineering, 2005, pp. 232-238.

[9] S. Hwang, J. Oh, J. Cox, S. J. Tang, and H. F. Tibbals, "Blood detection in wireless capsule endoscopy using expectation maximization clustering," *Progress in Biomedical Optics and Imaging*, Proceedings of SPIE 6144 I, 2006.

[10] V.S. Kodogiannis, and M. Boulougoura, "Neural network-based approach for the classification of wireless-capsule endoscopic images," *Proceedings of the International Joint Conference on Neural Networks* 4, 2005, pp. 2423-2428.

[11] E. Wadge, M. Boulougoura, and V. Kodogiannis, "Computer-assisted diagnosis of wireless-capsule endoscopic images using neural network based techniques," *Proceedings of the 2005 IEEE International Conference on Computational Intelligence for Measurement Systems and Applications*, CIMSA 2005, pp. 328-333.

[12] B. Li, and M. Q. -H Meng, "Wireless capsule endoscopy images enhancement by tensor based diffusion," *Annual International Conference of the IEEE Engineering in Medicine and Biology Proceedings*, 2006, pp. 4861-4864

[13] F. Vilariño, L. I. Kuncheva, and P. Radeva., "ROC curves and video analysis optimization in intestinal capsule endoscopy," *Pattern Recognition Letters,* 27 (8), 2006, pp. 875-881.

[14] F. Vilariño, P. Spyridonos, O. Pujol, J. Vitrià, P. Radeva, and F. De Iorio, "Automatic detection of intestinal juices in wireless capsule video endoscopy," *Proceedings - International Conference on Pattern Recognition,* 4, 2006, pp. 719-722.

[15] F. Vilariño, P. Spyridonos, J. Vitrià, C. Malagelada, and P. Radeva, "A machine learning framework using SOMs: Applications in the intestinal motility assessment," *Lecture Notes in Computer Science* (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 4225 LNCS, 2006, pp. 188-197.

[16] D.K. Iakovidis, S. Tsevas, D. Maroulis and A. Polydorou, "Unsupervises Summarisation of Capsule Endoscopy Video," accepted for publication in *4th IEEE International Conference on Intelligent Systems IS'08*.

[17] O. Okun, and H. Priisalu, "Unsupervised data reduction," *Signal Processing* 87 (9), 2007, pp. 2260-2267.

[18] D.D. Lee and H.S. Seung, "Unsupervised learning by convex and conic coding," *Adv. Neural Inf. Process. Systems,* 9, 1997, pp. 515–521.

[19] D.D. Lee and H.S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature,* 401, 1999, pp. 788–791.

[20] D.D. Lee and H.S. Seung, "Algorithms for non-negative matrix factorization," *Adv. Neural Inf. Process. Systems,* 13, 2000, pp. 556–562.

[21] L.K. Saul and D.D. Lee, "Multiplicative updates for classification by mixture models," *Adv. Neural Inf. Process. Systems,* 14, 2002, pp. 897–904.

[22] C. Ding, X. He, H.D. Simon, "On the equivalence of nonnegative matrix factorization and spectral clustering," *Proceedings of the SIAM International Conference on Data Mining*, Newport Beach, CA, 21–23 April 2005, pp. 606–610.

[23] C. Ding, X. He, H.D. Simon, "Nonnegative Lagrangian relaxation of K-means and spectral clustering," *Proceedings of the Sixteenth European Conference on Machine Learning*, Porto, Portugal, 3–7 October 2005, pp. 530–538.