Sudden Changes Detection in WCE Video

Giovanni Gallo¹, Eliana Granata¹, and Giuseppe Scarpulla²

¹ Dipartimento di Matematica e Informatica,
Viale A. Doria, 6, 95125 Catania, Italy

² Ospedale Maddalena Raimondi,
Via Forlanini, 5, 93017 San Cataldo (CL), Italy
`{gallo,granata}@dmi.unict.it`

Abstract. The direct visual inspection of WCE video by an expert is a tiring and cost activity and it is a true bottleneck to the widespread application of this diagnostic technique. In this paper we apply the texton approach to characterize with a numeric indicator the sub-sequences of a WCE that show sharp change and that are likely to represent relevant medical details. Experiments show that the proposed fully automatic technique may safely reduce the amount of frames that need further examination of up to 70%.

Keywords: Wireless Capsule Endoscopy (WCE), Video Segmentation, Textons, Video Classification, Gabor Filters.

1 Introduction

Wireless Capsule Endoscopy (WCE) is a recent technique which allows the visualization of the entire intestine without the need of any surgery. WCE has been proposed in 2000 [1] and it integrates wireless transmission with image and video technology. After FDA approved WCE in 2002 for the use in the US, this technique has been used routinely to examine the small intestine non invasively. WCE uses a small capsule of 11 mm diameter and 25 mm length. The front end of the capsule has an optical dome where white light illuminates the luminal surface of the intestine, and a micro colour camera sends images to a receiver worn by the patient. The camera takes 2 pictures per second. The stored data consists of about 50000 images, and is viewed as a video sequence using ad hoc software provided by the manufacturers [2]. The capsule contains a small battery that works for about 8 hours. The capsule is propelled through the intestine by normal peristalsis. This method has proved to be an important tool in the diagnosis of Crohn's disease [3], Celiac disease [4] and occult bleedings [5].

Medical specialists look for significative events in the WCE video stream by direct visual inspection, manually labelling clinical relevant frames. The time taken by this process, even for an experienced viewer, can be up to one or more hours. This is a bottleneck of this method, since it limits its general applicability and it incurs considerable costs. Labelling the video frames to automatically discriminate digestive organs such as esophagus, stomach, small intestine (duodenum, jejunum, ileum) and colon could hence be of great advantage. It would

make possible to find event boundaries that indicate either entrance to the next organ or to find unusual events within the same organ, such as bleedings, intestinal juices or obstructions, etc.

Manufacturers provide, in a Windows application named *Rapid Reader*, some functionalities to automatically analyze single images. This is done using a suspected blood indicator, that reveals the position in the video of areas with active bleedings. This tool has been reported to produce many false positives while at the same time it misses a large percentage of relevant events [6]. *Rapid Reader* includes also the localization function, which displays the route and the relative position of the capsule on a graphical torso model. To enable this functionality the viewer must find and mark two important locations in the video: first, the pylorus (the valve between stomach and intestine), and the ileo-caecal valve (between intestine and colon). Finding these two loci is difficult even with direct visual analysis because the tissues before and after them are very similar in both cases. Indeed the method proposed in this paper has been proved helpful to this task.

Computer Vision research has devoted a lot of effort to classify frames in a video or to discriminate still images from a database within some semantic classes. One of the state of art techniques to solve this problem is the use of textons [7]. In this paper we advocate the use of textons for the automatic discrimination of abrupt changes within a WCE video. In particular, we consider for each frame the high-frequency energy content and the features related to the texture by means of the response of the images to a bank of Gabor filters. Using textons, it is possible to calculate distances between intervals of frames to find points with abrupt change in pattern.

The experiments to test the method have been conducted on ten video segments extracted from WCE videos of different patients, in which the significant events have been previously labelled by experts. Results have shown that the proposed method may eliminate up to 70% of the frames from further processing while retaining all the clinically relevant frames.

This paper is organized as follows: Section 2 discusses related works. Section 3 describes the methodology that we use to discriminate meaningful events in a WCE video. Section 4 reports experimental results. Finally Section 5 draws conclusions and future works.

2 Related Works

In [8], Lee et al. propose an algorithm to detect event boundaries based on the energy function in the frequency domain. The segmentation of the video is performed taking into account the high frequency content function (HFC) of each frame. To find exact boundaries of digestive organs a bottom-up approach is used to determinate the correlation between consecutive frames. Their experimental results show a recall of 76% and a precision of 51%: the energy is hence a good indicator for relevant frame detection.

In [9] Vilarino et al. propose a method based on anisotropic image filtering and efficient statistical classification of the contraction features. They apply the

image gradient tensor to mine local directional information, and a sequence of descriptors to capture the star-pattern of the wrinkles that are the distinctive signature of contractions. Classification is performed with a SVM.

In [10] the previous authors observe that contractions represent only 1% of the entire video. They hence propose to use ROC curves together with simple and ensemble classifiers to identify the frames of contractions.

In [11] the same authors propose a method to reject part of the video with intestinal juices that are not meaningful for a diagnosis. They apply Gabor filters for the characterization of the frames containing intestinal juices.

In [12] Mackiewicz et al. describe a method that uses colour image analysis to discriminate the organs in WCE. The authors create a feature vector combining colour, texture and motion information of the entire image and valid sub-images. The video is segmented into meaningful parts using SVM and HMM.

Bleeding detection is investigated in [13], where multiple features are combined to analyze frames that present bleedings. A two steps process is proposed. The first step to discriminate images containing bleedings uses a block-based colour saturation method. The second step refines the classification using a pixel-based analysis of the luminance and colour saturation. The reported sensitivity in this case is of 88.3%.

3 Proposed Method

We consider the problem of classifying frames of a WCE video without imposing any constraints on the viewing or illumination conditions under which these images were specially obtained. The basic idea is that each digestive organ has a different visual pattern. Each pattern may be characterized by specific values of some observed features. Several candidate features may be considered: the texton method [7] allows to statistically combine all of them to produce a classifier. Our proposal integrates texture-based features (obtained as response to a bank of Gabor filters) with features like colour and luminance that are customarily of primary relevance to the clinician.

The general goal of automatic WCE segmentation is to split a video into shorter sequences each with the same semantic content. In the following we use the term “event” to indicate frame sequences of 6 consecutive frames that testify some abrupt and significative changes in the video. Our definition of “event” includes boundary transitions, intestinal juices, bile, bubbles, etc. We have make precise this term in accordance with medical experts that have manually labelled the test data. It should be pointed out that the term “event” has been used in the published research with different meanings than ours.

The features that we have chosen to build a classifier are: luminance and colour, energy and the responses to a bank of Gabor filters. These features are computed separately on frame sub-blocks. We apply k-means clustering to the set of feature vectors to build a texton dictionary. Frames are hence represented by mean of the histograms over the resulting dictionary. Comparison between histograms provides a way to assign a distance between frames.

3.1 Frame Pre-processing

The frames coming from the WCE videos have been pre-processed as described in this subsection in order to make further information extraction easier and more reliable. The original frames have a dimension of 576×576 pixels. This includes a large black area and textual information. We restrict the Region Of Interest within the circular area of the video, hence for each frame only a sub-image of dimension 352×352 pixels is considered. (Fig. 1)

Fig. 1. (a) original image; (b) extracted ROI

Each extracted ROI has been in turn transformed in the HSI colour space. It is indeed well known that this colour space grants a greater robustness for analysis. Frames, moreover, are partitioned into 484 squares each of 16×16 pixels. For each one of these squared sub-images we extract the features used for classification as described below. Hence each frame is eventually represented by the collection of the values of the features over all its 484 blocks.

3.2 Colour Features

Direct visual inspection by clinicians is largely based on the consideration of the chrominance value of the frames. Unfortunately colour features although informative do not provide sufficient clues to a classifier. Because of this we choose to include the average values of the hue, saturation and intensity of each of the 484 blocks of a frame among the representative features of the frame and to integrate them with more features as shown below.

3.3 Energy Features

Transitions, from an organ of the digestive tract to the next, are generally marked by frames that present a greater density of details. This fact has been exploited in [4] to characterize transitions. For this reason we include the high frequency content of blocks among the features used by the proposed classifier. Following [4] we consider the weighted sum of the energy function, linearly increased toward the high frequencies

$$HFC_i = \sum_{k=2}^{\frac{N}{2}+1} (|X_i(k)|^2 * k) \quad (1)$$

where the range $0 \cdots N$ is the index number range of the FFT frequencies in the frame; $|X_i(k)|^2$ is the squared module of the k^{th} component of the FFT and k is a weight to increase the relevance of higher frequencies.

(In (1) we ignore the lowest two bins in order to avoid unwanted bias from low frequency components). We include the HFC of each 16x16 sub-image among the representative features of a frame.

3.4 Texture Features

Textures are powerful discriminators when one has to classify real world pictures. Indeed all the state of the art content based retrieval engines rely on texture analysis. It is natural to include texture descriptors among the features representing a WCE frame. Texture classification has a long history in Computer Vision, starting with Haralick [14] proposed features to the up today methods that use large sets of responses to family of linear filters. In this paper we choose a Gabor filter bank for texture representation. Full details about Gabor filter may be found in several signal processing textbooks, for example see the classical [15].

For the sake of the present application a Gabor filter is defined as follows:

$$H(u, v) = \frac{1}{2\pi\sigma_u\sigma_v} e^{-\frac{1}{2} \left[\frac{(u-u_0)^2}{\sigma_u^2} + \frac{(v-v_0)^2}{\sigma_v^2} \right]} \quad (2)$$

where $\sigma_u = \frac{1}{2\pi\sigma_x}$ and $\sigma_v = \frac{1}{2\pi\sigma_y}$ are the standard deviation of the Gaussian envelope along x and y directions. The set of parameters $(u_0, v_0, \sigma_x, \sigma_y)$ completely define a Gabor filter.

In particular in our usage we empirically found appropriate the following parameters set: $\sigma_x = \sigma_y = 2$; *phase* : 0, 2, 4, 8, 16, 32 and four directions: 0°, 45°, 90°, 135°. The rationale behind our choice has been to achieve a good compromise between the recall and precision of the resulting classifier. An optimal choice of the parameters set will be the subject of further investigations. Observe that also the response of a Local Binary Pattern operator, which can be seen as a special filter based texture operator, could be proved itself useful [16].

3.5 The Complete Features Vector

In our proposal a frame comes to be eventually represented as a vector of 28×484 components. The 28 features includes information about average colour and luminance (3 elements), HFC (1 element) and Gabor filter responses (24 elements) for each block.

3.6 The Texton Dictionary

In order to achieve a more abstract representation we pool together the vector of all of the 16x16 blocks of the frames in the video. In our experiments each of the videos is made of 500 frames. This leads to an ensemble of 242000 vectors.

The ensemble has been clusterized to get a small set of recurrent and typical “visual words”. Clusterization is performed with a standard K-clustering. The number of clusters is chosen to optimize the ratio of dispersion between cluster centers over the dispersion within clusters. We empirically found that a suitable value for the number of clusters in our experiments is 100.

3.7 From Frames to Histograms

In this phase the dictionary obtained in section 3.6 provides the buckets to compute, for each frame, the relative frequencies of “visual word”. In this way we come to a high level representation of a frame as a “bag of visual words”. The problem of evaluating similarity between frames is hence turned into the computation of histograms distance. Among the several available choices for histogram distance estimation we choose the Bhattacharyya distance [17]. In this way we define $d(f_i, f_j)$, as the distance between frame f_i and f_j , the Bhattacharyya distance between the corresponding histograms.

3.8 Finding Sudden Changes

Direct computation of the Bhattacharyya distance of a pair of consecutive frames is generally a weak indicator of changes in the video. This happens because occasionally a frame can be quite different from the previous one just because of casual disturbances and trasmission noise. To have a more robust indicator of sudden changes in the video we consider for each frame f_i the function $C(i)$ defined as follows:

$$C(i) = \frac{1}{9} \sum_{k=i}^{i+2} \sum_{j=3}^5 d(f_k, f_{i+j}) \quad (3)$$

$C(i)$ averages the distances between frames in a short sequence and it provides high values when a sudden change is happening or low values in more homogeneous segments (Fig. 2).

Thresholding the function $C(i)$ will naturally lead to select frames that are very likely to be loci of sudden changes (Fig. 4).

4 Experimental Results

To assess the validity of the use of the indicator function $C(i)$ to discriminate meaningful sudden changes in WCE videos we have analized the performance of this index over 10 manually labelled sequences. The 10 randomly selected sequences have all been obtained from patients of the Ospedale “M. Raimondi” in the period between 2005 and 2008. The manual labelling protocol has been the following. Let $(f_i \dots f_N)$ be the sequence of frames in a video. We have formed the sequence of intervals $(I_1 \dots I_{\frac{N-3}{3}})$ where interval I_i is made of the six frames $(f_{3i-2} \dots f_{3i+3})$. For each interval the expert has judged if there is a significative change between the first 3 frames with respect to the last 3 frames. If this is

Fig. 2. The two rows represent two sequences of consecutive frames. The row (a) is relative to the pylorus: function $C(i)$ for it takes the value 0.41. The row (b) is relative to a smooth portion of stomach: the $C(i)$ takes the value 0.21.

Fig. 3. Plot of the function $C(i)$ for a WCE video sequence; peaks indicate loci of sudden change. The sequence in the upper row of Fig. 2 corresponds to the maximum (interval of frames: 355-360) while the sequence in the lower row of Fig. 2 corresponds to the minimum (interval of frames: 172-177).

the case the interval has been labelled as an “event”. In our setting hence an event is a relevant anatomical locus (esophagus, pylorus, etc.) or a pathological presence (bleedings, ulcerations, etc.) or a common non pathological disturbance (intestinal juices, bubbles, etc.) (Fig. 4). Observe that with this experimental framework we only need to compute function $C(i)$ (3) at the first frame of each interval. The typical graph of the $C(i)$ indicator is shown in (Fig. 3). The graph suggests that events may be found thresholding the indicator. However it is difficult at this stage of our research to automatically select a suitable threshold.

For this reason the performance of the index $C(i)$ has been tested as follows: intervals I_i of each video have been sorted according to the decreasing value of their $C(i)$ indicator. We have hence partitioned the sorted I_i ’s into ten groups of the same size. The first group contains the intervals with the top 10% $C(i)$ score, the last group contains the interval with the lowest 10% $C(i)$ score. For

Fig. 4. Examples of events among the WCE video frames; the row (a) corresponds to a pylorus; (b) is relative to the ileo-caecal valve; (c) represents frames with faecal residuals; (d) show the presence of bubbles

Table 1. Percentage of events in the ten groups of intervals sorted by $C(i)$ value

Video	10 th	9 th	8 th	7 th	6 th	5 th	4 th	3 rd	2 nd	1 st
1	93%	81%	31%	0%	0%	0%	0%	0%	0%	0%
2	81%	69%	0%	0%	0%	0%	0%	0%	0%	0%
3	100%	100%	38%	6%	13%	0%	0%	0%	0%	0%
4	88%	69%	0%	0%	0%	0%	0%	0%	0%	0%
5	94%	81%	31%	6%	0%	0%	6%	0%	0%	0%
6	94%	88%	19%	0%	0%	0%	0%	0%	0%	0%
7	100%	63%	0%	0%	0%	0%	0%	0%	0%	0%
8	100%	88%	0%	0%	0%	0%	0%	0%	0%	0%
9	100%	75%	19%	0%	0%	0%	0%	0%	0%	0%
10	100%	50%	6%	0%	0%	0%	0%	0%	0%	0%
mean	95%	76%	14%	1%	1%	0%	1%	0%	1%	0%

each group we have counted the number of intervals that the expert labelled as event. The statistics relative to the test data are reported in Table 1 and Table 2. It is evident from the experimental data that the proposed method may safely provide a filter to the clinician that could indeed concentrate the visual inspection on the intervals that score at the top 30% of the $C(i)$ index. Although this reduction is significative the number of intervals that have to

Table 2. Presence in % over the total number of events in the ten groups of interval sorted by $C(i)$ value

Video	10 th	9 th	8 th	7 th	6 th	5 th	4 th	3 rd	2 nd	1 st
1	45%	39%	15%	0%	0%	0%	0%	0%	0%	0%
2	54%	46%	0%	0%	0%	0%	0%	0%	0%	0%
3	39%	39%	15%	2%	5%	0%	0%	0%	0%	0%
4	56%	44%	0%	0%	0%	0%	0%	0%	0%	0%
5	43%	37%	14%	3%	0%	0%	3%	0%	0%	0%
6	47%	44%	9%	0%	0%	0%	0%	0%	0%	0%
7	62%	38%	0%	0%	0%	0%	0%	0%	0%	0%
8	53%	47%	0%	0%	0%	0%	0%	0%	0%	0%
9	50%	38%	9%	0%	0%	0%	0%	0%	3%	0%
10	64%	32%	4%	0%	0%	0%	0%	0%	0%	0%
mean	51%	40%	7%	1%	0%	0%	0%	0%	0%	0%

be examined is still high for a human observer but it may allow the realistic application of computationally more expensive pattern recognition algorithms to this restricted set of intervals.

5 Conclusions and Future Works

In this paper we have shown that a trasformation of the visual data coming from the WCE video into the space of textons opens up the possibility to robustly and efficiently discriminate eventful intervals of a diagnostic video. The proposed technique computes a suitable indicator of the presence of meaningful events. Experimental results show that all the relevant intervals are found among the 30% intervals when they are sorted according to the decreasing value of the proposed indicator. This percentage may be safely furtherly reduced optimizing the ROI selection, the parameters of the Gabor filter banks and the size of the visual dictionary. Future works will concentrate on this issues. Inclusions of the responses of non linear filters, like Local Binary Pattern, among the features will also be considered. The proposed technique may, especially if coupled with fast and effective tools for interactive exploration and visualization of the data, improve the applicability of WCE to clinical pratice.

References

1. Iddan, G., Meron, G., Glukhovsky, A., Swain, P.: Wireless Capsule Endoscopy. Nature 405, 725–729 (2000)
2. Given Imaging, Ltd Israel, <http://www.givenimaging.com>
3. Gay, G., Delvaux, M., Key, J.: The role of video capsule endoscopy in the diagnosis of digestive diseases: A review of current possibilities. Endoscopy 36, 913–920 (2004)
4. Swain, P.: Wireless capsule endoscopy and Crohns disease. Gut. 54, 323–326 (2005)

5. Culliford, A., Daly, J., Diamond, B., Rubin, M., Green, P.H.R.: The value of wireless capsule endoscopy in patients with complicated celiac disease. *Gastrointestinal Endoscopy* 62(1), 55–61 (2005)
6. Signorelli, C., et al.: Sensitivity and specificity of the suspected blood identification system in video capsule enteroscopy. *Endoscopy* 37, 1170–1173 (2005)
7. Varma, M., Zisserman, A.: A statistical approach to texture classification from single images. *International Journal of Computer Vision* 62 (2005)
8. Lee, J., Oh, J., Shah, S.K., Yuan, X., Tang, S.J.: Automatic classification of digestive organs in Wireless Capsule Endoscopy videos. In: SAC 2007: Proceedings of the, ACM symposium on Applied computing (2007)
9. Spyridonos, P., Vilarino, F., Vitri, J., Azpiroz, F., Radeva, P.: Anisotropic feature extraction from endoluminal images for detection of intestinal contractions. In: Larsen, R., Nielsen, M., Sporring, J. (eds.) MICCAI 2006. LNCS, vol. 4191, pp. 161–168. Springer, Heidelberg (2006)
10. Vilarino, F., Kuncheva, L.I., Radeva, P.: ROC curves and video analysis optimization in intestinal capsule endoscopy. *Pattern Recognition Letters* 27(8), 875–881 (2006)
11. Vilarino, F., Spyridonos, P., Pujol, O., Vitri, J., Radeva, P.: Automatic detection of intestinal juices in Wireless Capsule Video Endoscopy. In: ICPR 2006: The 18th International Conference on Pattern Recognition (2006)
12. Mackiewicz, M., Berens, J., Fisher, M.: Wireless Capsule Endoscopy Color Video Segmentation. *IEEE Transaction on Medical Imaging* 27(12) (December 2008)
13. Lau, P.Y., Correia, P.L.: Detection of bleeding patterns in WCE video using multiple features. In: Proceedings of the 29th Annual International Conference of the IEEE EMBS, August 23–26 (2007)
14. Haralick, R., Shapiro, L.: Image segmentation techniques. *Comput. Vision Graphics Image Processing* 29, 100–132 (1985)
15. Gabor, C.T., Sanjit, K.M.: *Modern Filter Theory and Design*. John Wiley & Sons, Chichester (1973)
16. Ojala, M.P.T., Maenpaa, T.: Multiresolution grey-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* 24(7), 971–987 (2002)
17. Bhattacharyya, A.: On a measure of divergence between two statistical populations defined by probability distributions. *Bull. Calcutta Math. Soc.* 35, 99–109 (1943)