33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Compressed Domain Topographic Classification for Capsule Endoscopy

N. Marques, E. Dias, J.P.S. Cunha, *Senior Member, IEEE*, M. Coimbra, *Member*, *IEEE*

Abstract— **In this paper we compare the classification accuracy of using compressed domain color (CDC) descriptors versus traditional full decoded images, for the purposes of topographic classification of wireless capsule endoscopy images. Results using a dataset of 26469 images, divided into stomach, small intestine and large intestine show a difference in classification accuracy below 1%. We also show that errors are mostly located near zone transitions (the pylorus and the ileo-cecal valve) and motivate the need for other visual descriptors (e.g. shape, motion) for addressing these specific areas. We conclude we can use the advantages of CDC in this type of classification with minor accuracy sacrifice.**

I. Introduction

CAPSULE endoscopy is a revolutionary pill shaped micro-device that explores the gastro intestinal (GI) tract reaching areas where the conventional endoscopy can't (Fig.1). It is ingested by the patient and captures 2 to 4 images per second for about 8 hours, typically reaching the colon. The signal transmitted by the capsule is captured by an external antenna array and stored on a portable hard drive carried in the patient's belt. According to its main manufacturer (Given Imaging, Israel), more than a million patients worldwide have undergone capsule endoscopy exams and more than 1,200 peer-reviewed clinical papers have demonstrated its clinical usefulness for detecting abnormalities within the GI tract [1].

Possibly the main setback of this technology is the time it takes to analyze an average of 60000 frames, which typically takes between 30-45 minutes by an experienced clinician [2]. Automation or at least semi-automation is thus necessary for reducing the costs of this procedure, paving the way for simple and effective GI screening mechanisms. Although the final goal of such systems is event detection, these computer vision algorithms can clearly benefit from an estimation of the organ being visualized (algorithms for the stomach need to be quite different from algorithms for the colon), motivating topographic segmentation which is defined as the segmentation of the full video exam into a set of smaller videos corresponding to different GI organs (e.g. stomach, colon, etc), thus mapping the different places inside the GI tract. Some interesting side effects of this task are that we can not only automatically estimate gastric and intestinal transit times, which is relevant information for a clinical diagnostic, but also save up to 15 minutes from the manual annotation procedure consisting of the marking of the locations of the pylorus and the ileo-cecal valve.

Fig. 1. The wireless endoscopic capsule (1 - Optical dome; 2 - Lens holder; 3 – Lens; 4 – Illuminating LEDs; 5 – CMOS imager; 6 – Battery; 7 – ASIC transmitter; 8 – Antenna).

Previous research on this topic has shown that it is possible to accomplish this objective using computer vision methodologies including MPEG-7 visual descriptors [3,4] and adapted Hue-Saturation histograms [5]. Most published literature, however, disregards the computational cost of these algorithms. If we intend to have autonomous screening systems in the future using wireless endoscopic capsules, it is highly desirable that these algorithms can run on simple portable hardware that might be incorporated, for example, next to the portable hard-drive carried in the patient's belt during the procedure. A research trend that the scientific community has explored for other fields is compressed domain processing, which means that we can exploit the information used to compress the video data for transmission and storage purposes, and extract relevant computer vision information from it. A good example is Coimbra's work on optical flow estimation [6].

In this paper we will show how this compressed domain information, namely DC coefficients of a MJPEG compressed video stream, can be used to perform

Manuscript received 24th March 2011. This work was partly supported by Fundação para a Ciência e Tecnologia under Grant PTDC/EIA-CCO/109982/2009.

M. Coimbra and Nuno Marques are with the Instituto de Telecomunicações, Department of Computer Science, Faculdade de Ciências da Universidade do Porto, Portugal (e-mail: {mcoimbra, nunomarques}@dcc.fc.up.pt).

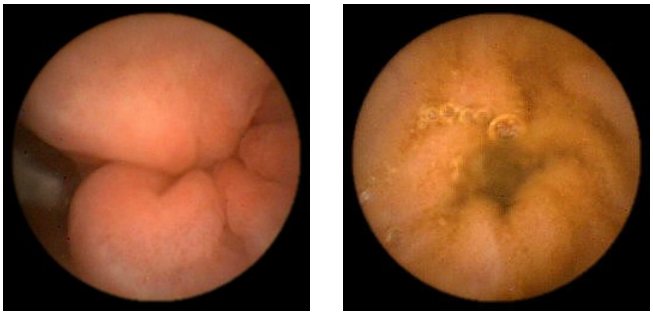
J.P.S. Cunha and E. Dias are with IEETA, Department of Electronics, Telecommunications and Informatics, University of Aveiro, Portugal (e-mail: {jcunha, edu}@ua.pt).

topographic classification of individual exam images as accurately as algorithms using fully decoded images. Details regarding the images used in this study can be found in Section II. Methods are detailed in Section III and results presented in Section IV. Observations and conclusions are drawn in Section V.

II. Materials

The capsule endoscope (Fig.1) is a disposable plastic capsule (M2A Capsule) which weighs 3.7 g and measures 11 mm in diameter × 26 mm in length. The contents include complementary metal oxide silicon (CMOS) chip camera, a short focal length lens, 4 white light emitting diode (LED) illumination sources, two silver oxide batteries, and a UHF band radio telemetry transmitter. Image features include a 140° field of view, 1:8 magnification, 1 to 30 mm depth of view, and, given the image sample resolution of the optical system, a minimum size of detection of about 0.1 mm. The activated capsule, after removal from the magnetic holder, provides image accrual and transmission at a frequency of 2 frames per second until the battery expires after 7 ± 1 hours. The capsule is passively propelled through the intestine by peristalsis.

The dataset used in this study includes 53 full endoscopic exams. All exams were annotated by a senior clinical specialist denoting the frame which represents the separation of the organs (Stomach / Small Intestine; Small Intestine / Large Intestine). Out of the 60000 usual frames in a capsule endoscopic exam, we picked 500 in order to discard the redundant frames, i.e. frames with a lot of resemblance, and to ensure that the classifier doesn't train and classify two images that are alike. There were 2257, 15165 and 9047 images of the stomach, small and large intestine respectively. All this data was obtained from the Capview.org database [7].



a) Stomach image b) Small intestine image

Fig. 2. Examples of endoscopic capsule images

III. Compressed Domain Topographic Classification

Although the final motivation for the experiments and results presented in this paper is topographic segmentation (dividing the full exam into its constituent organs), we will focus our attention simply on the topographic classification stage, meaning that we will classify each individual image into a specific class (organ). This implies ignoring the subsequent segmentation stage, which combines these individual classification results using methodologies such as global model fitting [3] or Hidden-Markov Models [5]. Since our contribution is to show that the extracted compressed domain feature vectors produce similar results to visual descriptors extracted using fully decoded images, we argue that we should inspect the direct impact of this change in the classification process, thus ignoring the error-correction ability of the segmentation stage.

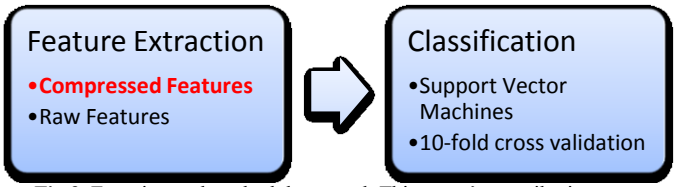


Fig 3. Experimental methodology used. This paper's contributions are highlighted in red.

Given that there are a large variety of video compression standards, we will focus on the ones using the DCT transform such as MJPEG. Our proposed methodology would need adaptation to standards that exploit temporal redundancy such as MPEG-2 [8], but current systems (Given Imaging, Olympus) tend to use JPEG [9] or MJPEG since it simplifies video browsing immensely and avoids additional compression artifacts.

A. DC Images

The Discrete Cosine Transform (DCT) converts a signal, or in this case an image, from the spatial domain f(x,y) into a transform domain F(u,v), usually called the frequency domain. Equations (1,2) show its definition for an 8x8 block.

$$F(u,v)=\frac{C(u)C(v)}{4}\left[\sum_{x=0}^{7}\sum_{y=0}^{7}f(x,y)\cos\left(\frac{(2x+1)u\pi}{16}\right)\cos\left(\frac{(2y+1)v\pi}{16}\right)\right] \quad (1)$$

$$C(u)=\begin{cases}\frac{1}{\sqrt{2}} \Leftarrow u=0\\ 1, otherwise\end{cases} \quad C(v)=\begin{cases}\frac{1}{\sqrt{2}} \Leftarrow v=0\\ 1, otherwise\end{cases} \quad (2)$$

An important characteristic of the DCT transform is that it can be seen as a set of basis functions which given a known input array size (8x8) can be precomputed and stored. This involves simply computing values for a convolution mask (8x8 window) corresponding to each DCT coefficient giving us 64 DCT basis functions. We are especially interested in the DC coefficient of a luminance DCT block, given that it is proportional to the average grey level of the corresponding image block. Equation (1) simplifies greatly for this case (3):

$$F(0,0)=\frac{1}{8}\sum_{x=0}^{7}\sum_{y=0}^{7}f(x,y) \quad (3)$$

Extracting the average grey level of a block is therefore trivial as Equation (4) shows:

$$f(i,j)=\frac{1}{8}DC(i,j) \quad (4)$$

Using these DC coefficients, it is possible to obtain low-resolution versions of I-Frames with minimal decoding and processing. These images are called DC images.

B. Color DC Images

We can apply the same reasoning for the chromaticity components used in JPEG compression [9], extracting its DC component for both the red (*Cr*) and blue (*Cb*) chromaticity. Given the typical difference in resolutions between the luminosity and chromaticity components, there is the need for up-sampling the color components to produce the corresponding color DC images (Fig.4).

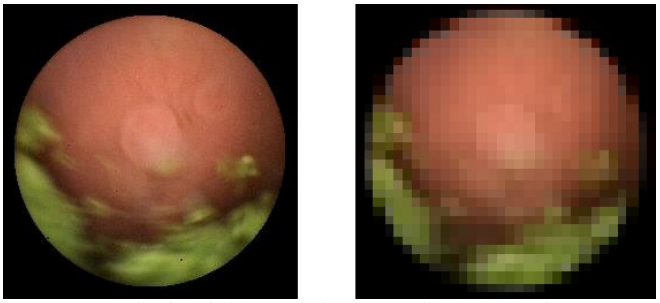


Fig 4. Fully decoded image and corresponding color DC image

C. Extracted Color Features

For the experiments in this study, we will use HSV (Hue, Saturation, Value) and HS (Hue, Saturation) histograms as color features, following the good results obtained in previous literature for topographic classification [3-5]. The number of bins used is 16 for Hue, 4 for Saturation and 4 for Value, giving us a HSV histogram with 256 coefficients, and a HS histogram with 64 coefficients. In order to test the performance of our proposed approach, we extract these same color features for both raw images (fully decoded, full resolution images), and color DC images, thus creating the following four distinct visual descriptors (Fig.5):

- **Raw HSV** – 256 coefficients, 16 bins for Hue, 4 for Saturation, 4 for Value, uses fully decoded image.
- **Raw HS** – 64 coefficients, 16 bins for Hue, 4 for Saturation, uses fully decoded image.
- **DC HSV** – 256 coefficients, 16 bins for Hue, 4 for Saturation, 4 for Value, uses color DC image.
- **DC HS** – 64 coefficients, 16 bins for Hue, 4 for Saturation, uses color DC image.

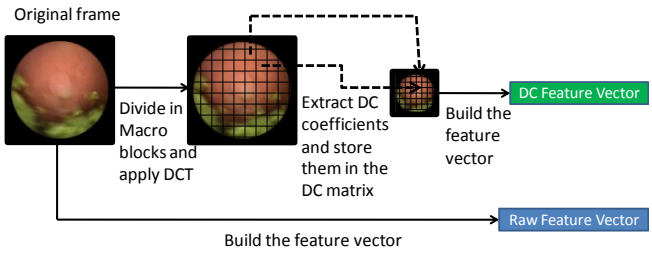


Fig 5. Feature extraction methodology

D. Image Classification

Classification was performed using the popular WEKA data mining platform (www.cs.waikato.ac.nz/ml/weka/). A total of 26469 instances were created, as detailed in Section II. 10-fold cross-validation was used. Images were classified into 3 distinct classes: stomach, small intestine, large intestine. Support vector machines (polynomial kernel) exhibited the best results from the various alternatives tested, and were used to produce the results here presented [10].

IV. Results

The accuracy (5) of the classification results obtained using each of the four color descriptors can be seen in Table 1.

$$Accuracy = corr.\ class./\ total\ class. \quad (5)$$

Descriptor	Accuracy (%)
Raw HSV	85.2431
Raw HS	83.6110
DC HSV	84.5253
DC HS	80.9362

Table 1 – Accuracy results of topographic classification using the four proposed color descriptors.

We can observe that the compressed domain descriptors exhibit classification results that are quite close to the ones extracted from fully decoded images, especially for HSV histograms.

	Stomach	Small Intest.	Large Intest.
Stomach	1352	760	145
Small Intest.	279	13830	1056
Large Intest.	49	1617	7381

Table 2 – Confusion matrix for the *Raw HSV* descriptor

	Stomach	Small Intest.	Large Intest.
Stomach	1367	768	122
Small Intest.	323	13667	1175
Large Intest.	57	1651	7339

Table 3 – Confusion matrix for the *DC HSV* descriptor

A closer inspection of the confusion matrices associated with the HSV descriptors (Table 2 – Raw HSV, Table 3 – DC HSV) also shows a close similarity between both modalities, namely in the order of magnitude of the error entries, strengthening our belief that it is possible to perform topographic classification in the compressed domain with classification precisions comparable to the conventional fully decoded image domain.

Although full topographic segmentation is not performed in this paper, we have mapped the location of the classification errors inside the gastrointestinal tract, thus assessing if they have some logical explanation associated with the visual nature of the observed tissues. In order to produce this map we have normalized the location of each error to a percentage its corresponding organ. As an example, if an

error was found in image 585 of an exam where the stomach ranges from image 100 to 12000, its corresponding location inside the stomach would be: (585-100)/(12000-100) = 4,4%. By concatenating the three error histograms of the contemplated three classes, we have obtained the error plot displayed in Fig. 6 for the *DC HS* descriptor.

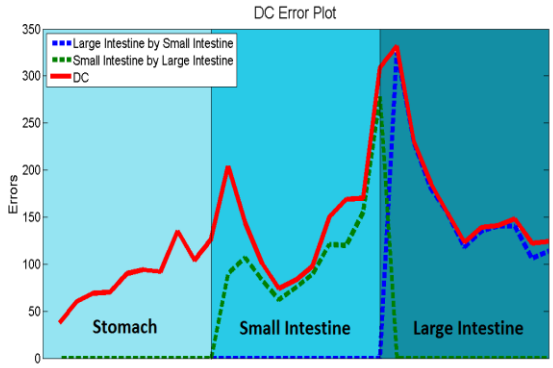


Fig 6. This error plot divides the wrongly classified images into 3 different zones, the Stomach, the Small Intestine and the Large Intestine. The red line is the total amount of errors done by the classifier. The green dotted line illustrates the amount of images the classifier did a mistake by classifying them as large intestine while they were from the small intestine. The blue dotted line shows the reverse classification of the green dotted line.

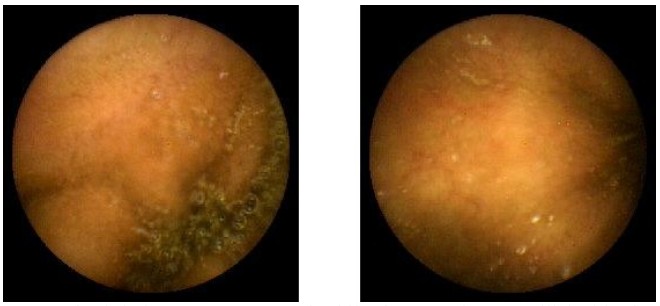


Fig 7. The left frame was captured while the capsule was still inside the ileum, while the right one was after crossing the ileo-cecal valve. Most exams exhibit this similarity, as explained in Section IV.

Analyzing Fig 6, one can observe two relevant error peaks immediately after the two zone transitions (pylorus and ileo-cecal valve). There are physiological reasons for this since in the beginning of the small intestine (duodenum) there is a strong color similarity with the stomach tissue. The visual distinction between these two areas includes shape (the typical 'tunnel'-like vision is expected in the duodenum, which does not happen in the stomach) and motion (free 'tumbling' inside the stomach as opposed to peristaltic burst of forward motion in the duodenum), both of which cannot be quantified using color descriptors alone. The most serious error peak happens near the ileo-cecal valve. We can argue that this is because of two different facts. First, there is a strong color similarity (as the examples on Fig. 7 show) between the ileum and the cecum tissues, and they are both often contaminated with faeces. Secondly, it is quite common that the capsule actually photographs the large intestine while still inside the ileum. This seemingly odd occurrence is easily explained by the fact that the capsule collides with the partially opened ileo-cecal valve, photographs the cecum throught this partial opening, and bounces back inside the ileum. This can go on for almost 30 minutes until there is a random synchronization between the capsule motion and the valve aperture, propelling the capsule through the junction.

V. Discussion and Conclusions

In this paper we have shown that compressed domain color descriptors exhibit near similar classification precision to the more conventional descriptors extracted from fully decoded images, for the task of topographic classification of capsule endoscopy images. This promising result reinforces the possibility of having in the future real-time, possibly even in-capsule, image processing and analysis of endoscopic capsule exams, by exploiting information that will be calculated anyway for the purposes of image transmission and storage. Although we do not explicitly calculate the computational cost savings of this approach, we can refer to previous literature where this has been addressed more carefully [11], stating that obtaining this DC image information accounts for less than 20% of the full decoding cost. Also, for creating the color histograms, DC images are 64 times smaller than raw images, also providing some interesting speed-ups. In the future, we expect to integrate these descriptors into a full topographic segmentation algorithm, and inspect if real-time performance can be successfully obtained.

References

[1] Given Imaging Home Page - www.givenimaging.com. [Online] 24 March 2011.

[2] A.F. Ravens, C.P. Swain, "The wireless capsule: new light in the darkness", in *Digestive Diseases*, vol. 20, 2002, pp. 127-133.

[3] J.P. Silva Cunha, M. Coimbra, P. Campos, J. Soares, "Automated Topographic Segmentation and Transit Time Estimation in Endoscopic Capsule Exams", in *IEEE Transactions in Medical Imaging*, vol. 27/1, Jan. 2008.

[4] M. Coimbra, P. Campos, and J.P. Silva Cunha, "Topographic Segmentation and Transit Time Estimation for Endoscopic Capsule Exams", in *Proc. of IEEE ICASSP 2006*, Toulouse, France, 2006.

[5] M. Mackiewicz, J. Berens and M. Fisher, "Wireless Capsule Endoscopy Color Video Segmentation," *IEEE Transaction on Medical Imaging,*, vol. 27, no. 12, December,2008

[6] M. Coimbra, and M. Davies, "Approximating optical flow within the MPEG-2 compressed domain" in *IEEE Transactions on Circuits and Systems for Video Technology*, Volume: 15 , Issue: 1 , Jan. 2005, pp. 103-107.

[7] M. Coimbra, M. Mackiewicz, M. Fisher, C. Jamieson, J. Soares and J. P. S. Cunha, "Computer vision tools for capsule endoscopy exam analysis", invited paper in *Eurasip NewsLetter*, vol. 18/1, March 2007, pp. 1-19.

[8] ISO/IEC JTC1 IS 13818 - 2 (MPEG-2) Information technology – generic coding of moving pictures and associated audio information, 1996.

[9] ISO International Standard 10918 – JPEG.

[10] C. J. Burges, "A tutorial on support vector machines for pattern recognition," Data Mining Knowledge Discovery, vol. 2, no. 2, pp. 121–167, 1998.

[11] Baskharan, and Konstantinides, *Image and video compression standards*, 2nd ed, Kluwer Academic Publishers, 1997.