Automatic Discrimination of Duodenum in Wireless Capsule Video Endoscopy

L. Igual[1], J. Vitrià[2], F. Vilariño[1], S. Seguí[1], C. Malagelada[3], F. Azpiroz[3] and P. Radeva[2]

[1] Computer Vision Center, Universitat Autònoma de Barcelona, Bellaterra, Spain
[2] Departament de Matemàtica Aplicada i Anàlisis, Universitat de Barcelona, Barcelona, Spain
[3] Hospital de Vall d'Hebron, Barcelona, Spain

Abstract— **Wireless Capsule Video Endoscopy is a recent acquisition method providing an internal view of the gastrointestinal tract which is currently applied in a large quantity of methods for detecting different intestinal diseases. In some of these applications, the automatic identification of some regions of the small intestine is essential. However, the high amount of time needed for video visualization makes this task unfeasible. In this paper, we present a novel system for automatical labelling the transition from proximal to distal parts of the small bowel in the capsule endoscopy video based on textural descriptors. Results show an accuracy of the proximal-distal boundary detection of more than** 70%**.**

Keywords— **Wireless Capsule Video Endoscopy, automatic organ discrimination, small intestine, textural features**

I. Introduction

Wireless Capsule Video Endoscopy (WCVE) [1] provides an internal view of the gastrointestinal tract allowing a physician to examine the entire small intestine non-invasively.

Automatic characterization of the different regions of the intestinal tract in WCVE is a current open field of research. In [2] the authors proposed an automatic classification of digestive regions in WCVE using patterns of intestinal contractions. They characterize contractions by an energy function of intensity value of HSI color space in frequency domain from WCE images. Then, they define a condition for event detection by combining this energy and a High Frequency Content (HFC) function and imposing an empirical threshold. Finally, events corresponding to every digestive region are detected using a hierarchy structure. The method performance is tested on 10 different videos with good results for the determination of the entering stomach and entering duodenum. However, the lack of information of specialist annotation for entering cecum and, especially, for entering ileum, makes the evaluation of this segmentation difficult. In [3], the authors proposed a method to localize the end of the stomach using color and texture features of 28 sub-regions of each frame. Hue-Saturation histograms are chosen as color descriptors and a method based on singular value decomposition is used for texture measurements.

In this paper, we propose a new method for detecting the boundary between the proximal and distal regions of the small intestine (proximal-distal boundary). These two regions are visually distinguishable by their different textural appearances, corresponding to the particular shapes shown by the folds and wrinkles of the intestinal walls. The proximal part comprises the whole duodenum together with a small transition region of the jejunum-ileum complex, which share the same visual aspect. The distal part comprises the whole jejunum-ileum, except for the small transition part. The separation of these two regions arises as a relevant step in those methods which use the fold and wrinkle patterns for the characterization of intestinal events, such as intestinal contractions [4].

The limit from stomach to duodenum (gastric entrance or pylorus) is easily localized by specialists in the video, whereas the definition of the proximal-distal boundary is much more complicated. The tissue of proximal and distal regions can be characterized by the different texture of the intestinal wall; however the transition is smooth. For this reason, the specialists need several visual inspections of the whole video to have a global vision of it and correctly localize the boundary. Moreover, this could cause a considerable divergence in the annotation of the proximal-distal boundary of the same video by different specialists (inter-observer variability).

The structure of our method is shown in the scheme in Figure 1. First, in order to accelerate the process, we resample the video and filter the frames with intestinal contents which prevent the correct visualization of the intestine wall. Then, we process the rest of the frames by applying a bank of Gabor filters followed by a half-wave rectification. The obtained responses bring out the textural differences of the proximal intestinal region with respect to the rest of the intestine. This descriptor information is used to classify the video frames as belonging to the proximal or distal part. Finally, the result of the classification serves to approximate a step function and localize the proximal-distal boundary.

The paper is organized as follows: in Section II, we introduce the methodology for the automatic detection of the proximal-distal boundary, in Section III, we present the experimental results, and in Section IV, we expose the conclusions.

J. Vander Sloten, P. Verdonck, M. Nyssen, J. Haueisen (Eds.): ECIFMBE 2008, IFMBE Proceedings 22, pp. 1536–1539, 2008

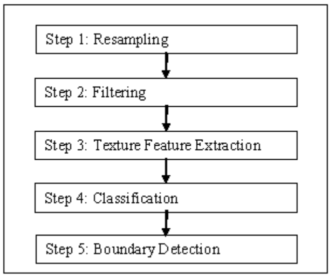
Step 1: Resampling
Step 2: Filtering
Step 3: Texture Feature Extraction
Step 4: Classification
Step 5: Boundary Detection

Fig. 1: Scheme of the method for automatic discrimination of small intestine proximal and distal regions.

II. Methodology

WCVE frames visualize three essential parts: the intestinal wall, lumen and intestinal contents. The texture of the gut wall change from proximal to distal regions of the small bowel as it can be seen in the examples displayed in Figure 2 (first row images). The main difference between tissue in proximal and distal parts is that wrinkles and mucous folds are much more abundant in the first part. Therefore, we deal with the region discrimination by using textural analysis. Visual texture in intestinal images can be due to intestinal contents (turbid and bubbles) or wrinkles of the intestinal wall. We get rid of the textural component due to the intestinal contents and focus our textural analysis on the wrinkles.

Let us describe the five steps of our system for automatical labelling of the proximal-distal boundary of a video.

Step 1: Resampling. We perform a subsampling taking into account that the frame ratio is two frames per second. We keep 1 frame of every 5 efficiently reducing the computation time without a loss in accuracy.

Step 2: Filtering. We detect the frames with turbid intestinal contents and bubbles which can prevent the correct visualization of the intestinal wall and we remove them.

For turbid intestinal contents detection, we use a semi-supervised procedure using a Self-Organized Map Method (SOM) [5], where the distance measure is computed based on a color space. Some frames with intestinal juices are not filtered by this method due to a low presence of them or a different color characterization, as for instance could be the case of bubbles. Some frames with bubbles show a color that is slightly different from the general turbid paradigm. However, these bubbles have also impact in the textural analysis and hinder the correct classification of the frames, so they have to be detected. For that, we use the method in [6] based on Gabor filters for the characterization of the bubble-like shape of intestinal juices. This method returns the segmented areas with intestinal juice bubbles for the video frames. Then, we use the following criterion for the rejection decision: if more than 50% of the frame contains bubbles, then we filter this frame. Moreover, the bubble areas of a frame are used to define *non-valid* areas for textural analysis. The *non-valid* areas are avoided in the subsequent processing and only the pixels in *valid* areas are used (see Figure 3 for an example).

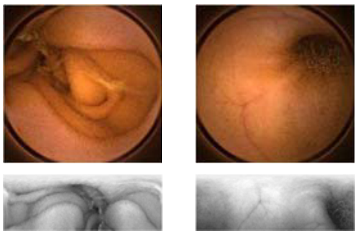

Fig. 2: Example of WVCE frames (top) and their polar representation (bottom). Proximal (left) vs. distal (right) frame.

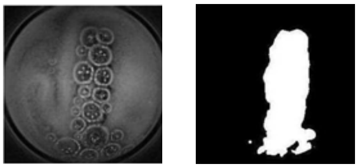

Fig. 3: Example of WVCE frame with bubbles (left) and the detected non-valid area (right).

Step 3: Textural Feature Extraction. The free movement of the camera and the intestine motion can make the identification of the proximal wrinkle paradigm difficult. Our main interest is to find frame descriptors invariant to translations and rotations. For these reasons, we compute the textural descriptors by applying a bank of Gabor filters on the polar representation of images (Figure 2).

A Gabor filter is a sinusoidal plane of particular frequency and orientation, modulated by a Gaussian envelope. These filters have been shown to possess good localization properties in both spatial and frequency domain and have been successfully applied in multiple tasks such as texture segmentation, edge detection, object detection, and image representation, among others [7, 4]. We denote $H(x,y,\sigma,\phi)$ the response of a Gabor filter, where σ is the standard deviation of the Gaussian kernel and ϕ represents the orientation.

For the construction of the bank of even-symmetric linear filters, we use two different scales and four different directions: $\sigma = [12.7205, 6.3602]$ and $\phi = [0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}]$, with an overall result of 8 filters in the bank. These parameters were obtained throughout an extensive empirical search.

We perform a convolution of the gray-scale version of the images with the bank of filters resulting in $R_i(x,y) = I * H_i(x,y,\sigma,\phi)$, where H_i denotes the *ith* Gabor filter and

$i \in \{1,...,8\}$. After the filter application, we perform a half-wave rectification [8] to avoid possible cancellations of positive and negative values. That is, we split the positive and negative parts of the filter response into $R_i^+(x,y)$ and $R_i^-(x,y)$.

Finally, we obtain a 16-dimensional descriptor vector $d(t) = (d_1(t),...,d_8(t))$ for each frame at time t by computing the following averages of the filter responses:

$$d_i(t) = \Big(\frac{1}{N_{\mathbf{X}}}\sum_{\mathbf{x}}^{N_{\mathbf{X}}} R_i^+(\mathbf{x}), \frac{1}{N_{\mathbf{X}}}\sum_{\mathbf{x}}^{N_{\mathbf{X}}} R_i^-(\mathbf{x})\Big), \tag{1}$$

where $\mathbf{x} = (x,y)$ and $N_{\mathbf{X}}$ are the number of pixels of the valid areas of the frame. This descriptor vector is used as texture features and highlight the differences of the small intestine proximal frames with respect to the distal ones.

Step 4: Classification. The next step of the system uses the textural features for classifying each frame as belonging to the proximal or distal parts. We consider two different approaches, an *unsupervised classification* and a *supervised classification*.

In the unsupervised classification approach, the descriptor information is used to clusterize the video in four parts using a Normalized Cuts algorithm [9].

Regarding the choice of min-cut as a clustering method among others, we argue the following reasoning: A major drawback to clustering methods such k-means is that they cannot separate clusters that are non-linearly separable in input space. A recent approach has emerged for tackling such a problem: the spectral clustering algorithms, which use the eigenvectors of an affinity matrix to obtain a clustering of the data [9]. A popular objective function used in spectral clustering is to minimize the normalized cut, which is the approach taken in our work.

We associate one label l_i, $i \in \{1,...,4\}$ to each cluster, Then, we reduce the number of labels to only two in the following way: all the frames belonging to the two clusters with the highest cardinality (assume l_1 and l_2) will keep their letter. Those frames belonging to the clusters with the lower cardinality will adopt one of the letters of the other clusters as follows:

$$L(t) = \begin{cases} l_1 & \text{if } P_{l_2}(I) < P_{l_1}(I) \text{ and } I = [t-10, t+10], \\ l_2 & \text{otherwise} \end{cases}$$

where $P_{l_1}(I) = P(L(t) = l_1 | t \in I)$.

In this way, the video is dichotomized in two different classes. A further refinement is applied by means of a morphological closing in order to remove spurious frames.

In the supervised classification approach, the descriptor information is used in the training and testing a SVM classifier [10].

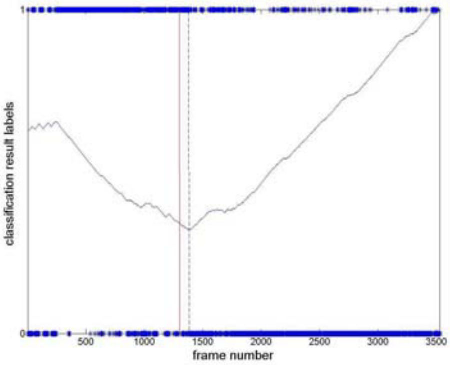

Fig. 4: Graphic of the classification results. The expert (solid line) and the system (dashed line) boundaries are displayed together with the error function E, and the frame label result are denoted by '*'.

Step 5: Step Function Approximation. We use the results of the classification for estimating the most probable position of the proximal-distal boundary by computing the best fit to a step function. Given the labels of the classification $L(t)$ for each frame at time t, we define the error function $E(t) = |L(t) - S(t)|$, where $S(t)$ is the step function defined as follows:

$$S(t) = \begin{cases} 0 & \text{if } x < t \\ 1 & \text{if } x \geq t \end{cases}$$

Then, we find the first distal frame at time t_0 such that $t_0 = \text{argmin}_t E(t)$.

III. Results

The experimental tests of the proposed method were performed on 13 different videos of healthy volunteers recorded using the endoscopy capsule developed by Given Imaging, Ltd., Israel [11] in the same conditions (fasting preparation). The total of frames to analyze were 349,100 and after resampling: 69,820. In average each video had 5,370 frames to be analyzed. Moreover, we had the following expert annotations: first post-gastric image, first distal image and first cecal image. We tested both approaches for classifying frames as proximal vs. distal frames: unsupervised classification and supervised classification.

Unsupervised classification. In Table 1, we show the results of the mean (μ) and median ($\mu_{1/2}$) of the classification of the 13 videos in terms of *Error*, *Sensitivity*, *Specificity*, *Precision* and *False Alarm Ratio*.

In Figure 4, we display the result of the classification for one of the videos. The stars on the abcise axis indicate

	Error	Sens.	Spec.	Prec.	FAR
μ	32.61%	64.91%	67.63%	48.88%	118.40%
$\mu_{1/2}$	31.44%	65.00%	66.06%	55.27%	69.26%

Table 1: Results of the unsupervised classification.

	Error	Sens.	Spec.	Prec.	FAR
μ	17.72 %	82.65 %	82.31%	68.08%	89.50%
$\mu_{1/2}$	8.95%	100.00%	92.93%	78.68%	7.50%

Table 2: Results of the unsupervised classification after step function approximation (Step 5).

	Error	Sens.	Spec.	Prec.	FAR
μ	21.86%	40.88%	92.54%	73.24%	23.52%
$\mu_{1/2}$	20.60%	39.05%	94.51%	73.26%	17.00%

Table 3: Results of the supervised classification.

	Error	Sens.	Spec.	Prec.	FAR
μ	18.04%	42.58%	97.46%	95.78%	9.08%
$\mu_{1/2}$	16.65%	32.65%	100.00%	100.00%	0.00%

Table 4: Results of the supervised classification after step function approximation (Step 5).

the classification results at frame level. The minimum of the function E, depicted in the graphic, points the proximal-distal boundary that we have emphasized with the dashed line. The solid line indicates the position of the boundary defined by the specialist.

After the step function approximation (Step 5), we recomputed the error measures and we improved the mean and median of the results as shown in Table 2. We also computed the error between the estimated boundary and the boundary annotated by the specialist in minutes for all the videos. The mean and median are 26.09 and 18.04 minutes respectively.

Supervised classification. We performed a Leave-One-Video-Out Cross Validation with a SVM classifier for the same data and the mean (μ) and median ($\mu_{1/2}$) of the obtained results are displayed in Table 3. In Table 4, we show the mean and median of the results after computing the most probable position of the boundary (Step 5). The mean and median of the errors made in the boundary estimation are 28.28 and 21.13 minutes respectively.

IV. Conclusions

We have presented a system to automatically discriminate the frames of proximal and distal regions of the small intestine in WCVE using texture descriptors. Moreover, this method is able to localize the proximal-distal boundary with an accuracy of approximately 26 minutes in average. A potential application of the presented method is to define non-valid frames for intestinal contraction analysis, since the contraction paradigm changes due to the presence of proximal wrinkles. Moreover, clinical procedures based on the capsule position could potentially be improved with this new information.

Acknowledgements

This work was supported in part by a research grant from projects TIN2006-15308-C02, CSD2007-00018, FIS-PI061290 and FIS-PI071188.

References

1. Iddan G., Meron G., others . (2000) Wireless capsule endoscopy. Nature 405:417.
2. J. Lee, J. Oh, S. Shah, X. Yuan, S. Tang. (2007) Automatic classification of digestive organs in wireless capsule endoscopy videos. SAC '07: Proc. of the 2007 ACM symposium on Applied computing, :1041–1045.
3. Mackiewicz M., Berens J., Fisher M., Bell D.. (2006) Colour and Texture Based Gastrointestinal Tissue Discrimination. ICASSP'06 Proceedings. 2006 IEEE International Conference, IEEE.
4. Vilariño F.. A machine learning approach for intestinal motility assessment with capsule endoscopy. . PhD thesis (2006) . Universitat Autònoma de Barcelona.
5. Vilariño F., Spyridonos P., Vitrià J., Malagelada C., Radeva P.. (2006) A Machine Learning framework using SOMs: Applications in the Intestinal Motility Assessment. CIARP'06, :188-197.
6. Vilariño F., Spyridonos P., Pujol O., Vitrià J., Radeva P.. (2006) Automatic Detection of Intestinal Juices in Wireless Capsule Video Endoscopy. ICPR '06: Proc. of the 18th International Conference on Pattern Recognition, 30:719–722.
7. Jain A., Ratha N., Lakshmanan S.. (1997) Object detection using Gabor filters. Pattern Recognition, 30:295-309.
8. Malik J., Perona P.. (1990) Preattentive texture discrimination with early vision mechanisms. Journal of the Optical Society of America A. 7:923-932.
9. Shi J., Malik J.. (2000) Normalized cuts and image segmentation. IEEE PAMI. 22:888-905.
10. Vapnik V.. (1995) The Nature of Statistical Learning Theory. Springer-Verlag.
11. Given Imaging Ltd.. http://www.givenimaging.com