Lumen Detection in Endoscopic Images: A Boosting Classification Approach

Giovanni Gallo and Alessandro Torrisi
Department of Mathematics and Computer Science
***I**mage **P**rocessing **Lab**oratory*
University of Catania, Italy
{gallo,atorrisi}@dmi.unict.it

***Abstract*—Intestinal lumen detection in endoscopic images is clinically relevant to assist the medical expert in studying intestinal motility. Wireless Capsule Endoscopy (WCE) produces a high number of frames. Automatic classification, indexation and annotation of WCE videos is crucial to a more widespread use of this diagnostic tool. In this paper we propose a novel intestinal lumen detection method based on boosting. In particular, we propose a customized set of Haar-like features combined with a variant of AdaBoost to select discriminative features and to combine them into a cascade of strong classifiers. Experimental results show the efficacy of boosted classifiers to quickly recognize the presence of intestinal lumen frames in a video. To better assess the accuracy of the proposed boosted classifier, we present an experimental comparison with the results obtained with a Support Vector Machine using a linear kernel.**

***Keywords*-Classification; Pattern Recognition; Boosting; Wireless Capsule Endoscopy; Video Automatic Annotation; Support Vector Machine.**

I. INTRODUCTION

Wireless Capsule Endoscopy [2], [3] is a technique to explore small intestine regions that traditional endoscopy does not reach. A video-capsule, that integrates wireless transmission with image technology, is swallowed by the patient and it is propelled through the gut by intestinal peristalsis. Once activated, the capsule captures two frames per second and transmits images to an external receiver. The exam is concluded after about eight hours, that corresponds to the lifetime of the battery of the capsule. Images taken during the entire route of the capsule through the intestine are successively analyzed by an expert. He/She may spend up to one or more hours to gather the relevant information for a proper diagnosis. This greatly limits the use of the capsule as a diagnostic routine tool.

Such shortcoming may be overcome if the WCE video is automatically segmented into shorter videos, each one relative to a different trait of the bowels, and if reliable automatic annotation tools are available to the clinicians. Unfortunately, the goal of automatically producing a summary of the whole WCE video remains yet unaccomplished. Tools to extract semantic information from such videos are relevant research products for applied Pattern Recognition investigators.

In this paper we present a novel method to automatically discriminate a relevant subclass of frames. In particular, our classifier sorts the frames in two categories: "lumen frames" (images depicting the stages of an intestinal contraction where the shrinkage of lumen intestine is well visible) and "not lumen frames" (Figure 1). "Lumen frames" detection is clinically relevant because it announces the presence of a contraction and helps the physician to study the intestinal motility. Alteration of the physiological intestinal motility is an indicator of disorders in which the gut has lost its ability because of endogenous or exogenous causes. In particular, anomalies in contraction are a common symptom of irritable bowel syndrome, delayed gastric emptying, cyclic vomiting syndrome, and so on.

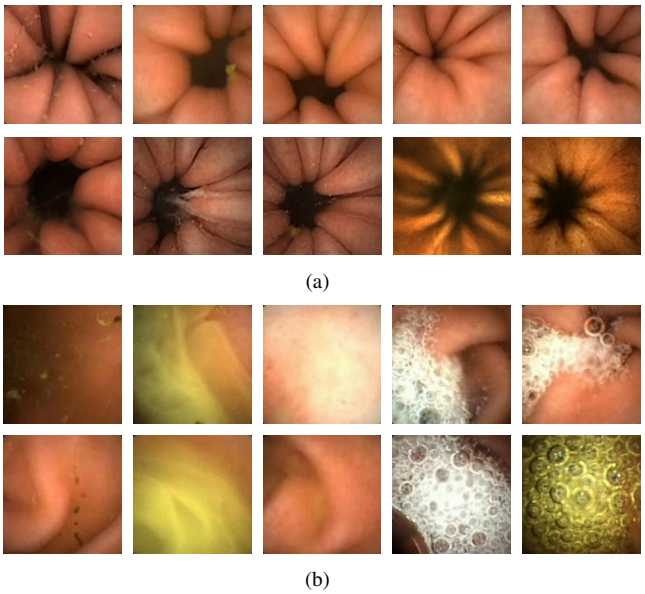

Figure 1. Examples of *lumen* (a) and *not lumen* (b) frames extracted from a WCE video.

Our summarization tool may be deployed in a diagnostic station providing real-time useful shortcuts to the middle phases of an intestinal contraction resulting in reduced time of analysis by the expert.

In our approach "lumen frames" detection is obtained as a special case of object detection. To this aim we choose the Viola and Jones paradigm introduced in 2001 [4]. Although

other techniques, like neural networks, fuzzy rules systems, etc., could be deployed, the main motivation for our choice has been the following. Haar features based classification is readily customizable to recognize different kinds of objects; moreover, boosting allows fast learning even in presence of high dimensionality data. Indeed in the case of boosting as for all ensemble learning method, different classifiers are built using a tiny part of the available features. The classification obtained by combining the responses of different classifiers improves the performance achieved by a standard classification algorithm in a straightforward, efficient, principled way when adaptive boosting is adopted.

This paper is organized as follows: Section II reviews related works and reports examples of object detection based on approaches similar to the proposed one. Section III describes in detail how Viola-Jones technique is customized to address the present problem. Section IV reports the experiments conducted on real WCE videos. It also describes an interesting comparison between the results obtained using Boosting and Support Vector Machine. Finally, Section V draws conclusions and discusses some future works.

This paper is a revised and expanded version of the contribution presented by the same authors to "The Third International Conferences on Pervasive Patterns and Applications" [1].

II. Related Works

Most of the systems reported in literature to recognize intestinal lumen images refer to traditional probe-based endoscopy. The motivation behind these methods is to aid the physician to individuate lumen region to avoid or minimize the collision of the endoscope tip with the intestinal mucosa. In this context, Asari [5] proposes a Region Growing Segmentation to extract lumen from gray level endoscopic images.

Recently, the original WCE has been modified/updated to a novel configuration allowing the movements to be remotely monitored. In this context, the recognition of lumen could help the capsule to go through the intestine minimizing collisions and avoiding to record meaningless frames. To this aim, Zabulis et al. [6] propose a system based on a Mean Shift Segmentation algorithm variant to locate lumen regions in WCE frames.

The problem of the detection of frames with a clear narrowing of lumen in WCE videos to assist the diagnostic and clinical use of this imaging technique is not much investigated. Some works study the general problem of contraction finding to examine the intestinal motility [7]–[9]. If we associate a label to each "lumen frame" extending the selection to a certain number of adjacent images in the video, our task is roughly equivalent to the search of intestinal contractions.

The main idea exploited in this work is to customize the Viola-Jones method for object detection [4], [10]. Initially proposed for face detection, this technique is based on the use of simple features calculated in a new representation of the image. Based on the concept of integral image [11], a huge set of features is tested and the boosting algorithm AdaBoost is used to reduce this set [12]–[14]. The introduction of a tree of boosted classifiers provides a robust and fast detection and minimizes the false positive rate. This strategy has been proven effective to recognize various kinds of objects. Several systems have been proposed for different recognition problems, like face, hands and pedestrian [15]–[18]. The possibility to define a specific set of features and the more recent release of an open source implementation [19] have permitted to use extensively this method in many Computer Vision applications.

III. Proposed Method

In this section we describe an automatically trainable system to detect frames where the front shrinkage of intestinal lumen is well visible. The learning stage for the proposed system can be summarized in the following three steps:

- Evaluation of a customized set of Haar features to the integral images of the training samples.
- Selection of the best discriminative features through AdaBoost algorithm.
- Construction of a final boosted classifier based on a cascade of classifiers whose complexity is gradually increasing.

To obtain, through a reliable learning procedure, a good classifier we must guarantee two requirements: a comprehensive set of examples where the objects of interest may occur; a suitable selection of descriptors to describe each possible occurring pattern. In order to detect an object in an image we should in principle take into account the information provided by each single pixel. This search space may be reduced if we exploit the semantic information enclosed by "lumen frames". These images, indeed, show a strong geometrical coherence that may help in discriminating them from other kinds of frames. To this aim, Haar-like features, a set derived from Haar wavelets [20], recognize objects using intensity contrast between adjacent regions in an image.

Basic Haar features proposed by Viola-Jones and specialized for face detection do not have proper discriminative power for lumen investigation: it is necessary to define customized variations for the present case. In particular, the features needed in this work should provide a strong positive response on a rectangular region with low intensity called generically "lumen" and a brighter surrounding area corresponding to the gut wall. By combining a learned evaluation threshold to each feature, it is possible to assign an image to the appropriate category. Figure 2a shows an example of the first kind of our proposed features that we call "center-surround" feature.

The typical appearance of a frame that shows an intestinal contraction consists in a dark area surrounded by the typical

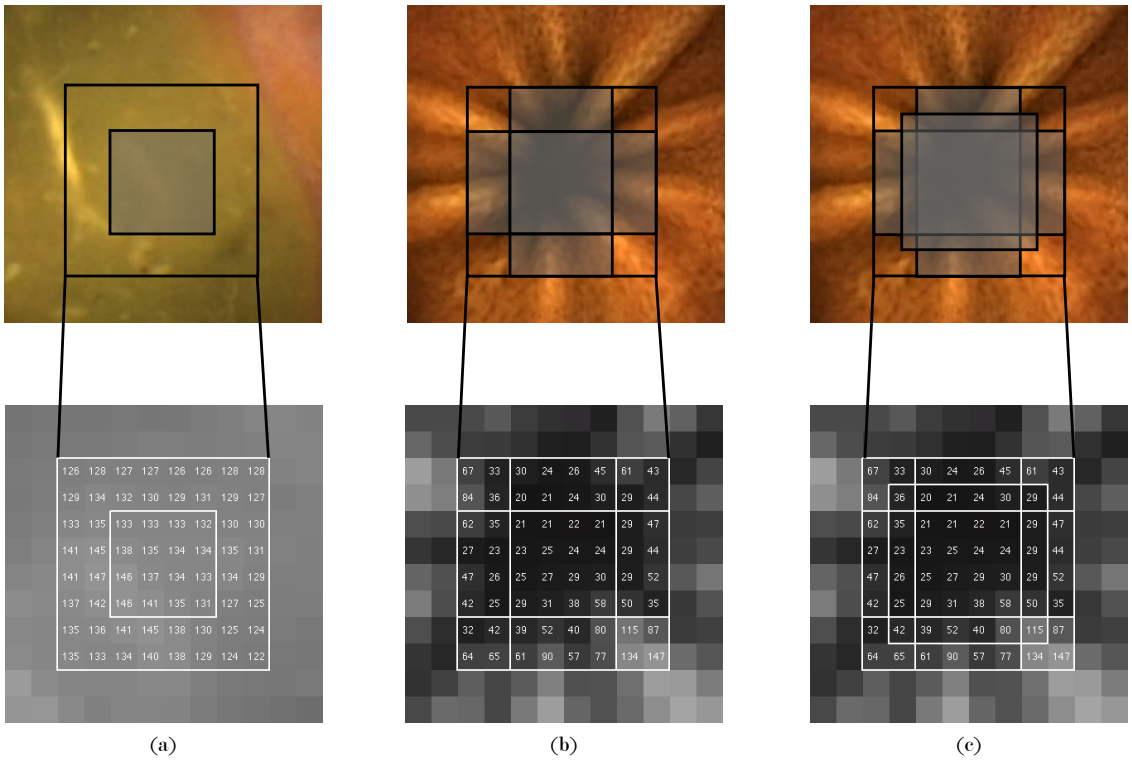

Figure 2. The three proposed kinds of features. For each feature we get a score S calculated as the difference of intensity between light and dark regions of the rectangle. In the first row are shown the images at the original resolution while in the second the images are rescaled to the base resolution 24×24 pixels zooming on the region of interest. (a) Evaluation of a "center-surround" feature in a "not lumen frame" ($S_a = 6348 - 2175 = 4173$). (b) Evaluation of the first cross feature in a "lumen frame" ($S_b = 1083 - 1766 = -683$). (c) Evaluation of the second cross feature in a "lumen frame" ($S_c = 861 - 1988 = -1127$).

rays that muscular tone produces due to the folding of the intestinal wall. We hence introduce two additional "cross-like" features that enhance the discriminative power produced by the simpler "center-surround" feature (Figure 2b - 2c). The computation of this second kind of features may be efficiently obtained as for the simpler "center-surround" feature from the integral image representation.

Using integral image representation, feature evaluation is accomplished by few memory accesses. It is straightforward to verify that to compute "center-surround" features, at any position or scale, only eight look-ups are needed. The remaining two kinds of features require more accesses due to greater number of rectangular areas. "Cross-features" require respectively 16 and 24 references from the integral image. The reader may easily convince himself that indeed this is the minimum number of look-ups needed from a direct analysis of this feature geometry.

Once a feature shape has been assigned, it is necessary to specify its position and scale within the region of interest. Actually, the features are scanned across the image top left to bottom right using a sliding offset of two pixels both in the horizontal and in the vertical directions. The process is iteratively repeated with different feature scales at each round. To keep the computation of the proposed features within the same number of look-ups into the integral image, we choose not to change the scale of the image but to variate instead the size of the features.

The exact representation for the three proposed types of features is as follows:

$$f = [x_w, y_w, s_{wx}, s_{wy}, x_b, y_b, s_{bx}, s_{by}, type, \theta, \rho] \qquad (1)$$

The first four elements x_w, y_w, s_{wx}, s_{wy}, refer to the larger square of the feature. Similarly, the following four elements x_b, y_b, s_{bx}, s_{by}, relate to the inner square. The $type$ parameter is an integer that indicates which type of feature is considered (1 for the "center-surround" feature, 2 and 3 for the two kinds of cross features respectively). The last two parameters are the optimal learned threshold and the polarity to register the category of images discriminated by that feature.

The "center-surround" features are evaluated considering difference between the sum of the pixels within two rectangular regions (Figure 3a). The second type of features considers a cross-shaped region to enhance lumen area. Location and size of this region are constrained by the size of correlated "center-surround" feature (Figure 3b). The third

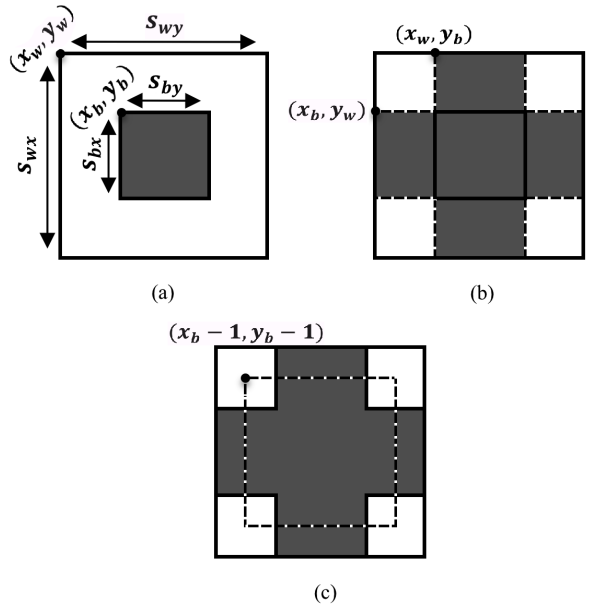

Figure 3. Schematic features representation. (a) Center-surround feature. (b) First cross feature obtained by center-surround feature considering the cross with width s_{by} and height s_{bx}. (c) Second cross feature obtained by the first taking into account a inner square of width and height greater than one pixel respect to the previous version.

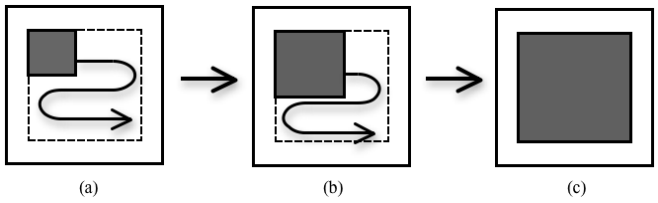

Figure 4. Given feature size, all regions of a fixed scale are considered in each location (a). This cycle is reiterated by increasing the size of the inner square (b) until maximum amplitude is achieved (c).

type of features is processed in a similar way. The central region of the cross is enlarged of one pixel both in the horizontal and in the vertical directions (Figure 3c). We consider the same total number of features for each type. Lumen area presents always a square aspect ratio, i.e., the bounding region of these areas is approximatively a square. This leads to a simplification of the feature definition (1) as follows:

$$f = [x_w, y_w, s_w, x_b, y_b, s_b, type, \theta, \rho] \quad (2)$$

We consider only squared features, i.e., those with equal horizontal and vertical even scale s_w. The internal region relative to lumen varies from a minimum size 2×2 up to $(s_w - 2) \times (s_w - 2)$ pixels. Once we have fixed the size of the external section, the descriptor associated with the lumen is shifted across the external descriptor with a resizing of two pixels at each step (Figure 4).

In this phase of processing the resolution of a WCE frame is reduced to 24×24 pixels. The total number of features per scale is hence equal to the total amount of different features in the image multiplied by the allowed variations of scale. For example, a 8×8 feature contains nine regions of size 2×2, four of size 4×4 and one of size 6×6 pixels. The total number of features of size 8×8 is 1134, equal to the number of windows in the image (assuming a horizontal and vertical offset of two pixels) for the total number of variations. Table I summarizes the feature counting for the chosen scales.

A. Training a cascade of strong classifiers

As it is stated above, during the training phase the dataset is rescaled to the base resolution 24×24 pixels. The integral image representation of gray tone training samples is used to compute feature scores. Application of AdaBoost provides a list of best discriminative features. In particular, we build a binary classifier for each feature (these are traditionally referred in the boosting community as weak classifiers). Initially all the examples have the same weight. For each boosting step, the determination of a new weak classifier involves the evaluation of the relevance of each feature on training data. The "best" feature is selected according to the weighted error that each feature shows on the training data. In the successive round, the samples are reweighted to emphasize the misclassified ones. Since this step has to be iterated several times, this is the most expensive section of the training module.

The result of the training module is a classifier (called "strong classifier" in the boosting jargon) computed as a weighted linear combination of the weak classifiers built during each round of boosting. The whole boosting process is, in turn, iterated, varying at each step the number of weak classifiers. The result is the realization of a cascade of strong classifiers with a gradually increasing number of features.

An appropriate learning process requires that each strong classifier shows a prescribed detection rate, while main-

Table I
FEATURES NUMBER PER SCALE. THE FIRST COLUMN REFERS TO THE SIZE OF THE FEATURE WHILE THE SECOND IS RELATED TO MAXIMUM SCALE ALLOWED FOR THE LUMEN AREA.

Feature size	*Max Internal scale*	*#Features*	*#Variations*	*Total*
4×4	2×2	121	1	121
6×6	4×4	100	5	500
8×8	6×6	81	14	1134
10×10	8×8	64	30	1920
12×12	10×10	49	55	2695
14×14	12×12	36	91	3276
16×16	14×14	25	140	3500
18×18	16×16	16	204	3264
20×20	18×18	9	285	2565
22×22	20×20	4	385	1540
24×24	22×22	1	506	506
				21021

taining a definite rate of false positives. In particular, a minimum detection rate and a maximum false positive rate is required at every level of the cascade. For each strong classifier, a weak classifier is added until it reaches the required parameters for the current level of the cascade. Similarly, a new strong classifier is associated to the cascade until total false positive rate crosses a certain threshold.

One of the advantages of the proposed system is that the user only needs to define the feature set to be used and the false positives and detection rates for each level of the cascade. All the internal parameters are automatically selected during the training phase.

B. Testing a cascade of strong classifiers

In the proposed system, each test image is scaled to 24×24 pixels and it is labelled as "lumen frame" or "not lumen frame". This single scale procedure combined with selection of best features during training allows real time application of our system (up to 600 frames per second). Please notice that, differently than in the case where the object to recognize may appear at different scales, in the present case a "single-scale" choice has been shown adequate. Notice that in this simplifying choice of a single scale we differ from the original Viola and Jones approach. In the case of face detection the issue is to find faces that may appear at different scales within an image. These stringent requirements force Viola and Jones to include different scales in their detection procedure. In our case the problem is simpler: lumens are roughly all at the same scale and we do not require localization of them inside the frame but only to label the frame as a "lumen frame". This justifies our choice of a single scale.

IV. Experimental Results

A. Boosting based classification

In this section, we report the experiments carried out to verify the efficacy of the proposed method. To this aim, we have considered 10033 images extracted from real WCE videos of 12 patients of which 6 were healthy and 6 had suspected bowel disorders. Rather than considering only one training set as was done in an earlier version of this paper [1], we have extracted ten different training sets and control sets from the whole set at our disposal. This more extensive experiment has been aimed to verify if the behavior of the algorithm significantly changes according to the used learning set. To train each one of the cascades of strong classifiers, we take into account the integral images of 3000 images, 1000 positive and 2000 negative, rescaled to 24×24 pixels. The positive images have been previously manually selected from WCE videos labelled by an expert. The selected images represent a comprehensive set of scenes where the intestinal lumen can be present, including location and scale changes within the image. Differently, the negative examples have been randomly selected from videos that not contain any lumen. Both typical smooth images and images containing other judged negative events, like the presence of bubbles, bleedings, residuals, share this set.

During the learning module, we need to establish a maximum false positive rate and a minimum detection rate to satisfy for each layer of cascade. In particular, we require that 98% of positive images must be recognized at each level while maintaining a maximum amount of false positives equivalent to 80%. These values have been experimentally optimized. Notice, however, that higher positive images recognition rate are first of all rarely attainable and even when possible, they may introduce strong overfitting. At the next levels of the cascade these two values are computed relatively to the new dataset whose positives set is composed by every lumen recognized as such by the previous classifier; the negatives set includes the remaining false positives. A strong classifier will be added to the cascade until the total false positive rate drops to zero.

By iterating this process for each training set, we get ten different cascades of strong classifiers whose details are listed in Table II. It is straightforward to understand that the trained cascades are slightly different only in the total number of features, but the proportion of features is often the same: the cross-shaped features ($Cross\ 1$, $Cross\ 2$) are the most discriminative. The number of nodes in the cascade can not be deterministically calculated, but this also depends on the type of images used during learning. We do not impose any constraints on the number of features in each node. It is assured only that the node $i+1$ must have a greater or equal number of features than node i. To clarify this procedure, in Figure 5 is illustrated the cascade of strong classifiers relative to the 8-th dataset. The total detection rate of this cascade, D, and the final false positive rate F, are obtained as a combination of intermediate outcomes on the cascade:

$$D = \prod_{i=1}^{N} d_i = 97,98\% \qquad F = \prod_{i=1}^{N} f_i = 0\% \qquad (3)$$

where N is the total number of layers of the cascade. To test the effectiveness of trained cascades, we have considered

Table II
Details on trained cascades using ten different training sets.

$Train\ Data$	$Nodes$	$features$	$Center\ surround$	$Cross\ 1$	$Cross\ 2$
1	6	217	51	78	88
2	5	291	82	109	100
3	6	397	89	154	154
4	6	342	77	131	134
5	6	256	57	71	128
6	5	185	47	67	71
7	6	257	72	98	87
8	5	205	60	66	79
9	6	184	47	80	57
10	5	272	67	100	105

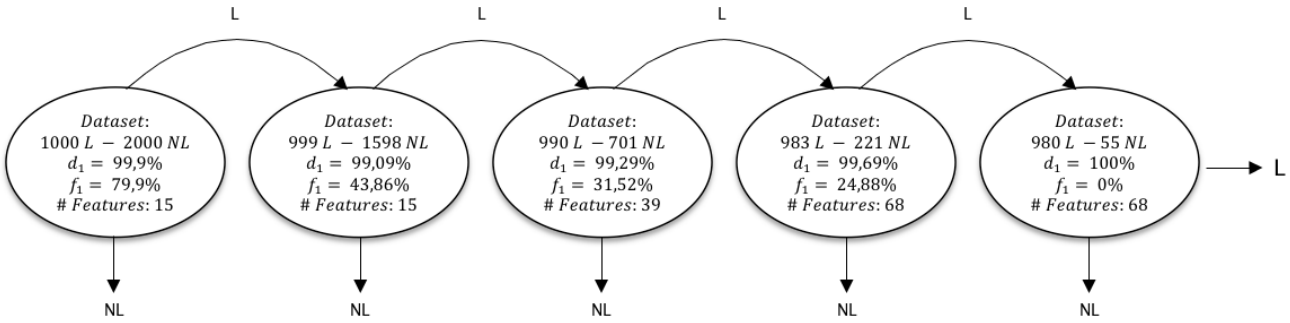

Figure 5. Cascade of strong classifiers. d_i and f_i represent detection and false positive rate at the i-th level of cascade. L and NL indicate *lumen* and *not lumen* frames, respectively.

ten different collections of 7033 images randomly extracted from a set of frames disjoined from each training set. During testing phase, we consider the integral images of test set rescaled to 24×24 pixels with the respective labels, the cascade of boosted classifiers as it has been obtained during training and, finally, a threshold that determines the rigorousness of the classifier. Each test sample gets through each single node of the cascade; a positive outcome is sent by the classifier i to the more complex classifier $i + 1$. An image is labeled as lumen if positively overcomes each node of the cascade. If at any point the test image is judged negative, it is rejected immediately without further test (Figure 5). The classification performance has been evaluated in terms of precision and recall by comparing our results with the annotations provided by the specialist. Table III shows the results. The labeling of images was previously made by a human expert. However, for certain images it is often difficult to understand, even to a skilled human observer, if what we hold as "lumen frame" is actually a particular fold of the intestinal tissue or vice versa.

Each strong classifier in the cascade is constrained by a rigidity threshold. Higher threshold values minimizes both detection and false positive rates. Similarly, a low threshold will lead to acceptance of a greater number of lumens images while increasing the probability of detecting false positives. The optimal value of threshold depends on the preferences of the physician. We expect that a higher amount of false positives than of false negatives is typically preferred. The presence of a high number of false positive results in more time spent by the expert to do a diagnosis. Losing a rightful lumen is a worse event because it means to miss a relevant event with the resulting inaccuracy in the final report. By varying the rigidity threshold from a minimum to a maximum value, we can construct a ROC curve comparing the detection rate versus the number of false positives. Figure 6 reveals that is possible to reach a detection rate above the 90%, keeping the amount of false positives at about 600 instances, i.e., 8% of the test dataset. All experiments have been conducted on a consumer level PC with Intel®Core™2 Duo processor and 4 GB of RAM. Calculations have been performed in MATLAB environment.

Table III
CLASSIFICATION RESULTS USING BOOSTING

Test Data	*Recall*	*Precision*	*Accuracy*
1	88,60%	72,06%	91,32%(6423/7033)
2	89,05%	71,64%	91,24%(6417/7033)
3	91,82%	69,11%	90,67%(6377/7033)
4	91,37%	67,86%	90,16%(6341/7033)
5	87,92%	70,73%	90,81%(6387/7033)
6	88,07%	71,76%	91,17%(6412/7033)
7	88,90%	69,06%	90,34%(6354/7033)
8	90,85%	70,78%	91,15%(6411/7033)
9	86,95%	73,40%	91,55%(6439/7033)
10	91,45%	68,33%	90,34%(6354/7033)

Figure 7 shows some examples of false positives obtained with the proposed method. In many circumstances, the intensity contrast between adjacent regions does not correspond to the presence of a lumen. This is maybe a consequence that Haar features are sensitive to illumination changes. Variations on the lighting conditions may cause the cascade to detect lumen that was not predicted during the training stage. Likewise, in some images, folds of the intestinal wall may produce contrasted regions that confuse the Haar features. If new kind of images are presented to the classifier, detection is difficult and the amount of false positives increases. To deal with this problem, training data must include as many examples as possible to predict only true lumen.

B. Features analysis

One may reasonably ask if the proposed kind of features is optimal: may we obtain good classification results without one of these three kind of features? May we get away with only one kind? Adding some more elaborate Haar-like

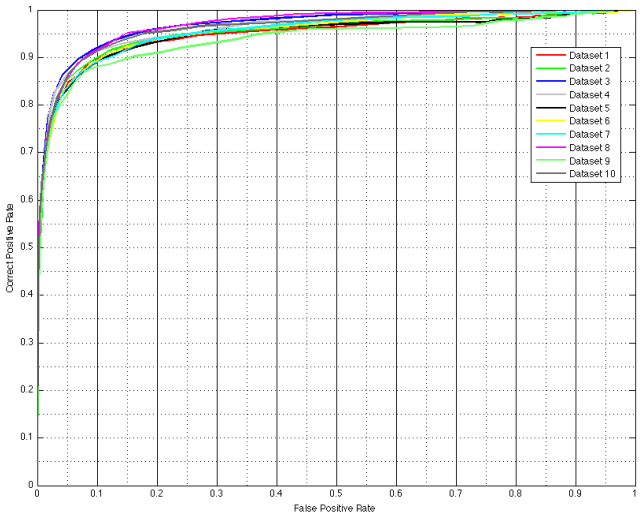

Figure 6. ROC curves for each dataset obtained by varying the stiffness threshold of each classifier from 0.1 to 1.

Figure 7. Example of some false positives detected by the system.

features is worth the gain in accuracy? The authors have tried to perform boosted classification using only one kind of feature among those proposed in this paper at each time. The results were only slightly different than those obtained using the whole set of features. This suggests that we might use only one kind of feature and achieve similar results. It is relevant to point out that the cross-shaped features have been introduced by the authors to improve not the results but the stability of the classifier. The availability of the whole set of features helps to keep down the number of classifiers in each node of the cascade. This happens because AdaBoost achieves more quickly the requirements fixed for the current classifier by the user. Also the number of nodes in the cascade is minimized. We can confirm that the use of additional features can only take effect on the structure of the classifier. The results would not be further significatively improved.

C. Comparing the boosted classifier with Support Vector Machine

The mean recall value we obtained using boosting is 89,5%. This result is efficiently attainable allowing a real-time performance. An interesting question is to compare the results provided by the boosting-based implementation with another "classic" classification method. The main problem in our data is the excessive dimensionality (63,063 features for each image to be classified). The high number of features suggests that comparison with other classification technique is fair only if these other techniques are adequate to handle these cases. For this reason, Support Vector Machine (SVM) is the ideal candidate for comparison. It is well know that SVM may easily deal with very high feature dimension; moreover, standard SVM implementation are available and this makes comparison easier and repeatable. SVM is a supervised learning algorithm used both for classification and regression. It indicates a binary classifier which projects the training samples in a multidimensional space looking for a separating hyperplane in this space. The hyperplane should maximize the margin, i.e., the distance from the closest training examples. SVM is well adapted to handle the curse of dimensionality and its performance has been tested in different application domains. We have considered the same data used in the previous experiments to train different SVMs using a linear kernel. We rely on a particular class of SVM called Least Squares SVM (LS-SVM). In this version it is possible to maximize the margin between support vectors by solving a linear equation with a least squares method. Classification results using this method are shown in Table IV. The superiority of the proposed boosting based technique is evident.

Table IV
CLASSIFICATION RESULTS USING SUPPORT VECTOR MACHINE.

Test Data	*Recall*	*Precision*	*Accuracy*
1	69, 92%	63, 84%	86, 79%(6104/7033)
2	71, 57%	63, 77%	86, 90%(6112/7033)
3	70, 82%	66, 39%	87, 67%(6166/7033)
4	69, 62%	64, 27%	86, 90%(6112/7033)
5	68, 79%	67, 58%	87, 83%(6177/7033)
6	70, 59%	65, 39%	87, 35%(6143/7033)
7	69, 17%	66, 14%	87, 44%(6150/7033)
8	70, 37%	65, 37%	87, 32%(6141/7033)
9	70, 37%	64, 11%	86, 92%(6113/7033)
10	72, 77%	63, 86%	87, 03%(6121/7033)

V. CONCLUSION

In this paper we introduced an automatic lumen detection algorithm for endoscopic images. Inspired by Viola-Jones object detection system, we show that using AdaBoost learning-based algorithm combined with a cascade of strong classifiers leads to a good rate of detection minimizing running time. Experimental results show that the proposed system detects positive images using exclusively Haar-like proposed features. Our detector is flexible and easily extensible to other semantic objects in endoscopic applications.

REFERENCES

[1] G. Gallo and A. Torrisi, "Boosted wireless capsule endoscopy frames classification," in *Proc. of Third International Conferences on Pervasive Patterns and Applications, PATTERNS'11*, September 25-30, 2011, Rome, pp. 25–30.

[2] G. Imaging, "Expanding the scope of gi," Last accessed: May 2012. [Online]. Available: http://www.givenimaging.com

[3] G. Iddan, A. Glukhovsky, and P. Swain, "Wireless capsule endoscopy," *Nature*, vol. 405, pp. 725–729, 2000.

[4] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1, 2001, pp. 511–518.

[5] K. Asari, "A fast and accurate segmentation technique for the extraction of gastrointestinal lumen from endoscopic images," in *Medical Engineering and Physics*, vol. 22, 2000, pp. 89–96.

[6] X. Zabulis, A. Argyros, and D. Tsakiris, "Lumen detection for capsule endoscopy," in *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on*, September 22-26, Nice, 2008, pp. 3921–3926.

[7] P. Spyridonos, F. Vilarino, J. Vitria, F. Azpiroz, and P. Radeva, "Anisotropic feature extraction from endoluminal images for detection of intestinal contractions," in *Medical Image Computing and Computer-Assisted Intervention*, 2006, pp. 161–168.

[8] G. Gallo and E. Granata, "Lbp based detection of intestinal motility in wce images," vol. 7961, no. 1. SPIE, 2011, p. 79614T. [Online]. Available: http://link.aip.org/link/?PSI/7961/79614T/1

[9] J. Lee, J. Oh, S. K. Shah, X. Yuan, and S. J. Tang, "Automatic classification of digestive organs in wireless capsule endoscopy videos," in *Proceedings of the 2007 ACM symposium on Applied computing*, ser. SAC '07. New York, NY, USA: ACM, 2007, pp. 1041–1045. [Online]. Available: http://doi.acm.org/10.1145/1244002.1244230

[10] P. Viola and M. Jones, "Robust real-time object detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2002.

[11] F. C. Crow, "Summed-area tables for texture mapping," in *Proceedings of the 11th annual conference on Computer graphics and interactive techniques*, ser. SIGGRAPH '84. New York, NY, USA: ACM, 1984, pp. 207–212. [Online]. Available: http://doi.acm.org/10.1145/800031.808600

[12] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee, "Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods," *The Annals of Statistics*, vol. 26, no. 5, pp. 1651–1686, 1998. [Online]. Available: http://dx.doi.org/10.2307/120016

[13] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," in *Proceedings of the Second European Conference on Computational Learning Theory*. London, UK: Springer-Verlag, 1995, pp. 23–37. [Online]. Available: http://portal.acm.org/citation.cfm?id=646943.712093

[14] Y. Freund and R. Schapire, "A short introduction to boosting," *J. Japan. Soc. for Artif. Intel.*, vol. 14, no. 5, pp. 771–780, 1999. [Online]. Available: citeseer.ist.psu.edu/freund99short.html

[15] P. Viola, M. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, vol. 2, October 2003, pp. 734–741.

[16] L. Yun and Z. Peng, "An automatic hand gesture recognition system based on viola-jones method and svms," in *Computer Science and Engineering, 2009. WCSE '09. Second International Workshop on*, vol. 2, October 2009, pp. 72–76.

[17] M. Kolsch and M. Turk, "Analysis of rotational robustness of hand detection with a viola-jones detector," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 3, August 2004, pp. 107 – 110.

[18] M. C. Santana, O. Déniz-Suárez, L. Antón-Canalís, and J. Lorenzo-Navarro, "Face and facial feature detection evaluation - performance evaluation of public domain haar detectors for face and facial feature detection," in *VISAPP (2)*, 2008, pp. 167–172.

[19] OpenCV, "Open computer vision library," Last accessed: May 2012. [Online]. Available: http://opencv.willowgarage.com

[20] C. Papageorgiou, M. Oren, and T. Poggio, "A general framework for object detection," in *Computer Vision, 1998. Sixth International Conference on*, January 1998, pp. 555 – 562.