Design and Development of an Automatic Blood Detection System for Capsule Endoscopy Images

Pedro Pons[1], Reinier Noorda[2(✉)], Andrea Nevárez[3], Adrián Colomer[1], Vicente Pons Beltrán[3], and Valery Naranjo[1]

[1] Instituto de Investigación e Innovación en Bioingeniería (I3B), Universitat Politècnica de València, Valencia, Spain
adcogra@i3b.upv.es
[2] iTEAM Research Institute, Universitat Politècnica de València, Spain
reinoo@upv.es
[3] Unidad de Endoscopia Digestiva, Hospital Universitari i Politcnic La Fe, Digestive Endoscopy Research Group, IIS La FE, Valencia, Spain

Abstract. Wireless Capsule Endoscopy is a technique that allows for observation of the entire gastrointestinal tract in an easy and non-invasive way. However, its greatest limitation lies in the time required to analyze the large number of images generated in each examination for diagnosis, which is about 2 hours. This causes not only a high cost, but also a high probability of a wrong diagnosis due to the physician's fatigue, while the variable appearance of abnormalities requires continuous concentration. In this work, we designed and developed a system capable of automatically detecting blood based on classification of extracted regions, following two different classification approaches. The first method consisted in extraction of hand-crafted features that were used to train machine learning algorithms, specifically Support Vector Machines and Random Forest, to create models for classifying images as healthy tissue or blood. The second method consisted in applying deep learning techniques, concretely convolutional neural networks, capable of extracting the relevant features of the image by themselves. The best results (95.7% sensitivity and 92.3% specificity) were obtained for a Random Forest model trained with features extracted from the histograms of the three HSV color space channels. For both methods we extracted square patches of several sizes using a sliding window, while for the first approach we also implemented the waterpixels technique in order to improve the classification results.

Keywords: wireless capsule endoscopy · blood detection · machine learning · hand-crafted features · deep learning · convolutional neural networks

1 Introduction

1.1 Motivation

Currently, physicians have multiple techniques and instruments at their disposal to diagnose the many diseases that affect the human gastrointestinal tract (GI

tract). Traditional endoscopy techniques enables access to some of the GI tract areas with both diagnostic and therapeutic purposes. However, they share the same limitation: they do not allow observation of the complete small intestine. This greater section of the GI tract is only accessible through different, more invasive techniques such as push enteroscopy or intraoperative endoscopy.

Wireless Capsule Endoscopy (WCE), first introduced in 2001, does allow for minimally invasively observation of the entire GI tract. The patient only has to swallow a pill-sized camera (Fig. 1), which goes through the GI tract driven by peristaltic movements. The camera, installed on one side of the capsule, captures 2 to 6 images per second, while a LED light source illuminates the scene.

One of the main problems when using WCE is the large number of video frames generated per exam (over 150.000). Physicians spend up to two hours reviewing these, resulting in not only a high cost, but also a high probability of a wrong diagnosis due to fatigue. This situation is aggravated by the variable appearance of abnormalities and that they sometimes only appear in a single or few frames, requiring high concentration.

This paper focuses on blood detection in WCE images recorded by the PillCam^{TM} SB 3 capsule, as presence of blood in the bowel is a symptom of many diseases such as polyps, tumors, ulcers or Crohns disease. Therefore, blood detection is often a priority in analysis of WCE procedures. Even though RAPID^{TM} Reader, a software provided by the PillCam^{TM} manufacturer, contains an automatic blood detection tool, it cannot be used as a reliable tool for diagnosis, since several studies claim it has both a low specificity and a low recall [8, 3].

1.2 Literature review

Many researchers have taken up task of automatically detecting visible abnormalities in the GI tract, following different approaches. Regarding the type of features that are used, some authors try combining color features and texture features, although color features have been proven to be much more discriminative [5]. Most authors explore different color spaces aiming to find the features that best differentiate between blood and healthy tissue. While many researchers use the RGB color space because of its simplicity, it can be problematic as WCE images usually have an uneven illumination, which drastically affects the three RGB channels [1]. In contrast, in the HSV color space the inhomogeneous lightning only affects the V channel, which makes it a popular alternative.

Fig. 1. Endoscopic capsule PillCam SB3.

There are also different approaches regarding the strategies for processing the image. Pixel-level methods are the simplest and fastest ones, but they ignore the existence of spatial information, usually returning poor and incoherent results. Image-level methods benefit from spatial information, but fail to detect the smallest anomalies, whose features are masked by those from the healthy tissue. Patch-level methods are an intermediate step between the previous ones. While bigger patches are able to grasp more spatial information, smaller patches tend to report better sensitivity values [7].

1.3 Objectives

In this work, we intend to develop an automatic blood detection system to reduce the time needed to review WCE videos, following two different approaches. The first consists in extracting "hand-crafted" features from the images and training machine learning algorithms capable of differentiating between blood and healthy tissue. The second consists in deep learning, concretely in training convolutional neural networks (CNNs), capable of extracting the needed features by themselves. Finally, we compared the results obtained with each procedure.

2 Materials and Methods

2.1 Data Collection

The available dataset consisted of 75 WCE images obtained with a PillCam SB3 capsule, all provided by the Digestive Endoscopy Center from Hospital Universitari i Politcnic La Fe (Valencia). A total of 40 of these contained blood, while the remaining 35 contained only healthy tissue.

2.2 Data Preparation

First, we manually created the ground truth of the images, i.e. the true target of each pixel in the images (Fig. 2). Next we extracted square patches of different sizes (32x32, 64x64 and 96x96 pixels) from each of the images in the data set, using sliding windows with 50% overlap in both the horizontal and vertical axis. The label of each patch was determined at the same time, labeling as blood only those of which the corresponding area in the ground truth was at least 10%.

Both to reduce the dependency of the results on the way the data were split and to improve the generalization of the models, we performed nested cross-validation [10]. This technique consists in training several models on different subsets of the data, through both an external and an internal cross validation loop. We used 5 folds in each loop. In each fold, a different partition of the the set of all areas was used for testing, while the remaining areas were used for training. Results of the models were averaged to obtain a single value for each performance metric. In the external loop, we ensured that each whole image participated only in a single partition and never in the corresponding training

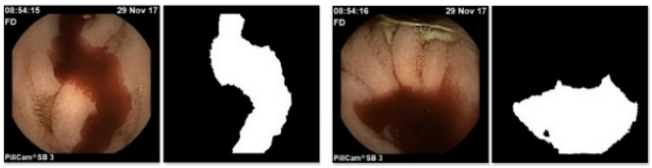

Fig. 2. Pixel-level ground truth of the images in the data set.

set. In the internal loop, we again partitioned the training data into different validation sets, using one such partition as validation and the remaining areas for training as before. This loop we used for optimizing the hyperparameters of our models. Additionally, each training set in the external loop was balanced by down-sampling the majority class.

2.3 Hand-crafted Feature Extraction

In our first approach, following the work done by other researchers, and after investigating whether either the LBP or the HOG histograms showed capacity to distinguish between the two classes, we chose to use only color features. Specifically, we used the RGB and HSV color spaces. We also tried to reduce the inhomogeneous illumination problem by converting the RGB images into the HSV space, applying a homomorphic filter to the V channel and converting the images back to RGB (Fig. 3). Concretely, the following features were tested to determine which of them behave better in our problem: histograms of all RGB channels, histograms of all HSV channels, histograms of the H and S channels of HSV and histograms of all RGB channels after our illumination correction.

After reviewing some of the most popular machine learning algorithms, we chose for Support Vector Machines (SVM) [4] and Random Forest [2] in this work because of their good performance in our early experiments, relative simplicity and resistance to overfitting. Regarding SVM models, we compared the performance of a simple lineal kernel and an RBF kernel, which usually outperforms the others and is relatively easy to implement. Regarding the Random Forest models, we estimated a convenient number of trees by observing that the out-of-bag error stayed practically constant from 80 trees onward.

The approach followed so far consisted in extracting features from square patches extracted from the images. However, the sliding window used for this purpose did not take the content of the image into account, thus sometimes capturing content of both classes in a single patch. In order to overcome this problem, we replaced this technique with a superpixel segmentation technique. Superpixels are image regions with similar features, of which the borders tend to adjust to those of the objects in the image. Specifically, we used a variant based on the watershed technique, called waterpixels [6], which has been reported to obtain better results than other popular methods [2]. One of the most interesting aspects of waterpixels is the freedom of choosing a criterion for selecting the points from which the waterpixels start growing. In this work, we used the

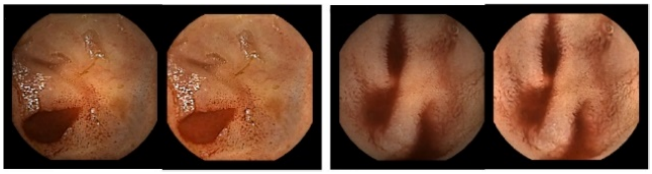

Fig. 3. WCE images after applying the homomorphic filtering.

minimum gradient points. In case several minima existed in the same cell, the one with the highest surface extinction coefficient prevailed.

2.4 Deep Learning

Our second approach was based on CNNs, capable of learning good features to extract, thus avoiding potential loss of relevant information due to hand-crafted, human feature selection. The CNN models were trained on exactly the same aforementioned subsets of our data as the previous algorithms: we thus trained five models for each patch size for their performance metrics to be averaged at the end. The CNNs were trained parting from an VGG19 [9] model that was pre-trained on the ImageNet data set, through a technique known as "fine-tuning". This consists in freezing weights of the first layers of a pre-trained model, while optimizing those in the last layers to adapt the model to the new problem.

3 Results

Performance test results for all of our trained models are reported here for each of the different region types (patches of different sizes or water pixels). Regarding the SVM models, only results obtained with an RBF kernel are shown (Table 1), since we obtained significantly superior results compared to the linear kernel.

4 Discussion

Generally, we obtained worst results for histograms of the RGB channels directly extracted from the original image. However, if they were extracted after applying a homomorphic filter, the performance was significantly enhanced, with the results approaching those obtained for HSV. We also observed that the results obtained for HS were very similar to, or sometimes even better than, those for all HSV channels. A possible reason for this is that the V channel, despite containing additional information, also brings in illumination problems.

In Fig. 4, where the best results of each classifier are compared, we can see that CNN models appears best regarding accuracy and AUC, but when considering recall, which is considered the most important metric in this study due to false negatives having worse consequences for diagnosis than false positives,

Table 1. Average test results for SVM models with an RBF kernel.

Processed Areas	Features	Accuracy	Recall	Spec.	AUC
32x32 pixels	RGB	0,9221	0,9137	0,9279	0,9685
	HSV	0,9051	0,9090	0,9030	0,9589
	HS	0,9175	0,9260	0,9128	0,9718
	RGB-filtered	0,9236	0,9181	0,9274	0,9709
64x64 pixels	RGB	0,9045	0,8587	0,9353	0,9566
	HSV	0,8968	0,8931	0,8982	0,9571
	HS	0,9132	0,9162	0,9104	0,9694
	RGB-filtered	0,9148	0,8880	0,9326	0,9689
96x96 pixels	RGB	0,8497	0,7954	0,8906	0,9303
	HSV	0,8817	0,9026	0,8658	0,9549
	HS	0,8957	0,9034	0,8894	0,9637
	RGB-filtered	0,8991	0,8524	0,9339	0,9640
Waterpixels	RGB	0,9257	0,9485	0,9132	0,9752
	HSV	0,9086	0,9526	0,8851	0,9734
	HS	0,9471	0,9119	0,8919	0,9688
	RGB-filtered	0,9325	0,9500	0,9258	0,9769

Table 2. Average test results obtained for random forest models.

Processed Areas	Features	Accuracy	Recall	Spec.	AUC
32x32 pixels	RGB	0,9266	0,9206	0,9313	0,9756
	HSV	0,9252	0,9363	0,9198	0,9806
	HS	0,9238	0,9372	0,9169	0,9802
	RGB-filtered	0,9292	0,9342	0,9275	0,9787
64x64 pixels	RGB	0,9276	0,9189	0,9343	0,9769
	HSV	0,9318	0,9483	0,9210	0,9828
	HS	0,9316	0,9443	0,9231	0,9822
	RGB-filtered	0,9365	0,9357	0,9386	0,9823
96x96 pixels	RGB	0,9273	0,9225	0,9317	0,9760
	HSV	0,9378	0,9568	0,9233	0,9844
	HS	0,9357	0,9549	0,9209	0,9841
	RGB-filtered	0,9389	0,9364	0,9424	0,9849
Waterpixels	RGB	0,9058	0,9448	0,8835	0,9701
	HSV	0,9062	0,9614	0,8750	0,9791
	HS	0,8966	0,9612	0,8621	0,9777
	RGB-filtered	0,9073	0,9548	0,8873	0,9770

Table 3. Average test results obtained for the CNN models.

Patches	Accuracy	Recall	Spec.	AUC
32x32 pixels	0,9496	0,9390	0,9560	0,9901
64x64 pixels	0,9494	0,9208	0,9688	0,9917
96x96 pixels	0,8943	0,7545	0,9838	0,8920

Random Forest and SVM were significantly better. Using those recall values to compare the results obtained using different patch sizes, we observe that generally, greater patches resulted in lower recall. However, Random Forest models seem to be unaffected by changes in the patch size, since we can only observe differences in the order of hundreds and even appear slightly better for greater patches. For waterpixels we generally obtained higher recall than for patches, but while for SVM we also obtained similar or even slightly higher accuracy, both accuracy and specificity were significantly lower when using Random Forest.

Fig. 5 shows the results obtained on some of our test images. From each image we extracted patches (bottom row) or waterpixels (top row), which were classified and then combined to obtain a pixel-level classification of the image. Here we observe that the region labeled as blood appears to be more accurate pixel-wise when using waterpixels, as it does not produce block artifacts.

One of the greatest limitations found during this study was the lack data. The absence of a public database forces researchers to search for new images and get them manually labelled, which is a slow and tedious task. The pixel-wise segmentation obtained by our algorithm could be used as a labelling tool for easier creation of ground truth images in the future.

5 Conclusion

We trained different models capable of automatically detecting blood in WCE images, following both classical machine learning and deep learning approaches. The color features we found to be most useful in this work were the histograms of the H and S channels from the HSV color space and the three RGB channels after applying a homomorphic filter to correct illumination variances.

Of the patch-based models, we obtained greatest recall (95,68%) for a Random Forest model based on HSV histograms, with 92,33% specificity. Additionally, we found that using waterpixels detected areas had visually better borders.

Acknowledgments. This work was funded by the European Unions H2020: MSCA: ITN program for the "Wireless In-body Environment Communication WiBEC" project under the grant agreement no. 675353. Additionally, we gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research.

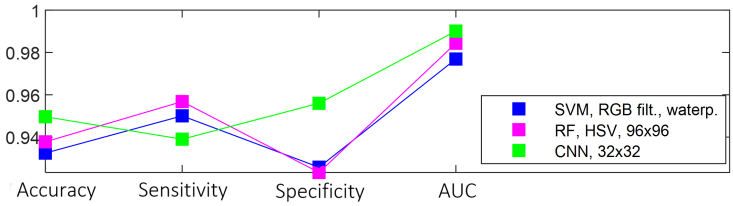

Fig. 4. Comparison of the best results from each type of model.

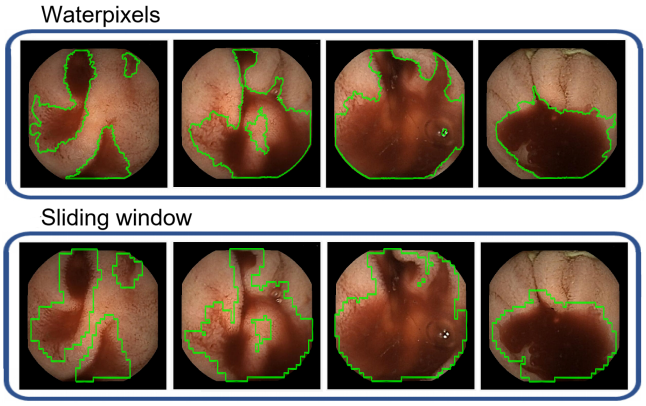

Fig. 5. Comparison of detected regions using waterpixels (top) and patches (bottom).

References

1. Berens, J., Finlayson, G.D., Qiu, G.: Image indexing using compressed colour histograms. IEE Proceedings-Vision, Image and Signal Processing **147**(4), 349–355 (2000). https://doi.org/10.1049/ip-vis:20000630
2. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001). https://doi.org/https://doi.org/10.1023/A:1010933404324
3. Buscaglia, J.M., Giday, S.A., Kantsevoy, S.V., Clarke, J.O., Magno, P., Yong, E., Mullin, G.E.: Performance characteristics of the suspected blood indicator feature in capsule endoscopy according to indication for study. Clinical gastroenterology and hepatology **6**(3), 298–301 (2008). https://doi.org/https://doi.org/10.1016/j.cgh.2007.12.029
4. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning **20**(3), 273–297 (1995). https://doi.org/https://doi.org/10.1007/BF00994018
5. Li, B., Meng, M.Q.H.: Computer-aided detection of bleeding regions for capsule endoscopy images. IEEE Transactions on biomedical engineering **56**(4), 1032–1039 (2009). https://doi.org/10.1109/TBME.2008.2010526
6. Machairas, V., Faessel, M., Cárdenas-Peña, D., Chabardes, T., Walter, T., Decencière, E.: Waterpixels. IEEE Transactions on Image Processing **24**(11), 3707–3716 (2015). https://doi.org/10.1109/TIP.2015.2451011
7. Novozámskỳ, A., Flusser, J., Tachecí, I., Sulík, L., Bureš, J., Krejcar, O.: Automatic blood detection in capsule endoscopy video. Journal of biomedical optics **21**(12), 126007 (2016). https://doi.org/https://doi.org/10.1117/1.JBO.21.12.126007
8. Signorelli, C., Villa, F., Rondonotti, E., Abbiati, C., Beccari, G., de Franchis, R.: Sensitivity and specificity of the suspected blood identification system in video capsule enteroscopy. Endoscopy **37**(12), 1170–1173 (2005). https://doi.org/10.1055/s-2005-870410
9. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
10. Varma, S., Simon, R.: Bias in error estimation when using cross-validation for model selection. BMC bioinformatics **7**(1), 91 (2006). https://doi.org/https://doi.org/10.1186/1471-2105-7-91